SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

AUTISM SPECTRUM DISORDER DETECTION USING MACHINE LEARNING ON COMBINED SCREENING DATA

Faisal Shahzad[*1], Ayesha Maryam[2], Arbish Awan[3], Mishal Ijaz[4]

[*1,2,3,4]*Department of Computer Science, Bahauddin Zakariya University, Multan 60800, Pakistan*

[*1]faisal.shahzad.research@gmail.com, [2]ayeshamaryam2133@gmail.com, [3]arbishmscs@gmail.com, [4]mishalijaz868@gmail.com

DOI: https://doi.org/10.5281/zenodo.24156129

Keywords
Autism Spectrum Disorder; ASD screening; machine learning; questionnaire data; Logistic Regression; Support Vector Machine; feature interpretation; classification

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Faisal Shahzad

Abstract

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition associated with differences in social communication and interaction, together with restricted or repetitive patterns of behavior. Because characteristics may become apparent during early childhood while access to formal assessment can be limited, computational screening approaches have attracted increasing research interest [1] [2] [7] [16]. The present study develops a machine-learning framework for ASD screening using a combined age-group questionnaire dataset. The working dataset contained 1,100 records and 21 original columns, with 707 non-ASD and 393 ASD cases. The target variable was encoded as a binary outcome. Missing numerical values were replaced using the median, and missing categorical values were represented as "Unknown." To reduce target leakage, the rule-derived result field was removed before model development, and the age_desc field was excluded because it duplicated age-group information. Numerical variables were standardized and categorical variables were transformed using one-hot encoding. Seven supervised classifiers—Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Neural Network, Decision Tree, and Naive Bayes—were evaluated using a stratified 80:20 train-test split. Performance was assessed using accuracy, precision, sensitivity, specificity, F1-score, area under the ROC curve (AUC), and Matthews correlation coefficient (MCC). Logistic Regression obtained 100.00% on accuracy, precision, sensitivity, specificity, and F1-score, with AUC = 1.000 and MCC = 1.000 on the held-out test set. The corresponding confusion matrix contained 141 true negatives and 79 true positives, with no false positives or false negatives. SVM was the second-highest performing model, with 98.64% accuracy and AUC = 0.999. The results indicate that the combined questionnaire data contain strong predictive patterns; however, the perfect test-set performance should be treated as an experimental result rather than evidence of clinical readiness. Independent validation, repeated cross-validation, and evaluation across age and demographic subgroups are required before practical deployment.

INTRODUCTION

Autism Spectrum Disorder (ASD) comprises a diverse group of developmental conditions that can affect social communication, interaction, behavior, and everyday functioning. Autistic people's characteristics and support needs vary

considerably, and early identification can connect individuals and families with appropriate assessment and support [1] [7]. Behavioral research also reports differences in areas such as response inhibition and repetitive or stereotyped behavior, while genetic and biological studies demonstrate that ASD is associated with heterogeneous biological mechanisms rather than a single causal pathway [2] [3] [4]. Machine learning provides a complementary research approach because structured screening records can be represented as feature vectors and analyzed as classification problems. Earlier work has explored behavioral questionnaires, computational intelligence, neuroimaging, and deep learning for ASD-related classification. For example, Heinsfeld et al. demonstrated deep-learning classification using resting-state fMRI data from the ABIDE dataset, while Thabtah et al. investigated computational intelligence approaches for behavioral screening [5] [6]. Reviews have also emphasized the growing role of machine learning in ASD behavioral research and the need for careful interpretation of predictive models [8] [16]. Questionnaire-based screening is particularly attractive for tabular machine-learning studies because responses can be encoded directly as predictive variables. Thabtah's work framed ASD screening as a machine-learning classification problem, and the UCI Machine Learning Repository has served as an important public source for machine-learning datasets, including ASD screening data [13] [14]. Digital screening resources such as ASD Tests further illustrate the use of structured behavioral questions in preliminary screening contexts [15]. The present study builds on this line of research by combining age-group screening records into one dataset and evaluating a broader collection of supervised classifiers. This design relates to Erkan and Thanh's earlier study, which evaluated KNN, SVM, and Random Forest on separate child, adolescent, and adult ASD datasets [20]. Rather than reproducing that three-model, age-specific comparison, the present work evaluates seven classifiers after a unified preprocessing pipeline and selects the model according to a multi-metric evaluation.

Study Contributions

- A unified preprocessing strategy is applied to a combined ASD screening dataset, including missing-value treatment, categorical encoding, numerical scaling, and removal of target-related fields [6] [13] [14] [16] [20].
- Seven supervised learning algorithms are compared using multiple diagnostic metrics rather than accuracy alone [16] [17] [18] [19] [20].
- A Random Forest feature-importance analysis is used as a supplementary interpretation method to identify influential questionnaire variables [19].
- The manuscript explicitly discusses the limitations of a single held-out split and the need for independent validation before any clinical use [5] [8] [11] [12] [16] [20].

Literature Review

ASD Characteristics and the Need for Early Screening

The World Health Organization describes autism as a diverse group of conditions related to brain development and notes that characteristics may be detected in early childhood even though diagnosis may occur later [1]. Stakeholder research in South Africa has similarly highlighted the importance of early detection and intervention while describing practical barriers that can affect access to autism-related services [7]. The biological literature further indicates that ASD is heterogeneous. McDermott et al. described a variable clinical spectrum associated with TBR1-related ASD, illustrating the importance of genotype-phenotype variation, while Saresella et al. reported activation of multiple inflammasome complexes in ASD samples, providing evidence for biological complexity rather than a single disease mechanism [3] [4].

Machine Learning for ASD Screening

Machine-learning research has approached ASD identification through both behavioral and neuroimaging data. Heinsfeld et al. used deep learning with ABIDE resting-state fMRI data and reported that functional brain-activation patterns could be used to distinguish ASD participants

from controls [5]. Pagnozzi et al. reviewed structural MRI biomarkers and discussed the utility of machine learning for identifying neuroimaging patterns, while Thabtah reviewed machine-learning applications in ASD behavioral research and highlighted methods including SVM, decision trees, and logistic regression [8] [16]. Behavioural screening has also been studied as a direct classification problem. Thabtah et al. proposed a computational-intelligence approach for identifying autistic traits and reported feature-selection and classification experiments using behavioral data [6]. Thabtah's DSM-5-oriented screening work similarly framed ASD screening as a classification task based on historical cases and controls, supporting the use of structured questionnaire attributes as model inputs [14]. The present study therefore focuses on tabular behavioral and demographic information rather than neuroimaging. This choice is consistent with the accessibility of questionnaire-style screening data and with prior work using UCI-based ASD records [13] [20]. The goal is not to replace clinical assessment but to investigate whether machine-learning models can identify predictive patterns in structured screening information.

Classification Algorithms Used in the Study

KNN is a local, instance-based approach in which predictions are influenced by nearby observations in feature space. The nearest-neighbor literature provides the statistical foundation for such local methods [17]. SVM, introduced by Cortes and Vapnik for two-group classification, constructs a decision boundary in a transformed feature space and is widely used for binary classification [18]. Random Forest combines multiple decision-tree predictors and introduces randomization in feature selection and tree construction; Breiman's original work established the method and its generalization properties [19]. Logistic Regression models the relationship between predictors and a binary outcome through a logistic function and is particularly attractive when an interpretable linear decision structure is appropriate. Neural Networks can learn nonlinear relationships through layered transformations, while Decision Trees create rule-like partitions of the feature space. Naive Bayes provides a probabilistic baseline based on conditional independence assumptions. The seven-model comparison is motivated by the need to evaluate different modelling assumptions rather than relying on a single classifier [16] [20].

Materials and Methods

Dataset

The study used the Combined_Autism_Data.csv file associated with the updated Python experiment. The dataset contained 1,100 records and 21 original columns. The variables included ten AQ screening scores, age, gender, ethnicity, jaundice status, family-history information, country of residence, application-use information, the respondent category, and the Class/ASD outcome [13] [14] [20]. The target distribution was 707 records labelled NO and 393 labelled YES. The dataset structure follows the broader AQ-based screening approach used in publicly available ASD datasets.

Table 1. Dataset description and preprocessing design used in the proposed ASD screening framework

Property	Value
Dataset file	Combined_Autism_Data.csv
Total records	1,100
Original columns	21
Target column	Class/ASD
Class distribution	NO = 707; YES = 393
Removed score bias field	Result
Removed duplicate field	age_desc
Training/testing split	80% training; 20% testing
Training samples	880
Testing samples	220
Best model	Logistic Regression

Data Preprocessing

Preprocessing was designed to prevent incomplete or redundant fields from distorting the classifiers. Question-mark entries were treated as missing observations. Missing numerical values were imputed with the median, whereas missing categorical values were replaced with "Unknown." The Class/ASD field was converted into a binary target representation. The result column was removed because it is closely related to the final screening outcome and could allow a model to learn information that is effectively derived from the target [6] [13] [14] [20]. The age_desc variable was also removed because it duplicates age-group information already represented elsewhere. Numerical features were standardized, and categorical features were transformed using one-hot encoding.

Train-Test Strategy

After preprocessing, the data were divided into 80% training and 20% testing subsets using stratified sampling. The resulting split contained 880 training records and 220 test records while preserving the class distribution. A single held-out split was used for the reported experiment; therefore, the results should be interpreted as test-set performance for this dataset rather than as a population-level estimate [5] [8] [16] [20].

Table 2. Machine Learning Models

Model	Role in the experiment
Logistic Regression	Binary classification using a linear decision function expressed through logistic probabilities.
Support Vector Machine	Margin-based classification capable of representing complex class boundaries.
K-Nearest Neighbors	Instance-based classification determined from nearby observations.
Random Forest	Ensemble of randomized decision trees.
Neural Network	Layered nonlinear model capable of learning complex feature relationships.
Decision Tree	Rule-based recursive partitioning of the feature space.
Naive Bayes	Probabilistic classifier using conditional feature assumptions.

Algorithm references: [17] [18] [19].

Evaluation Metrics

The models were evaluated using accuracy, precision, sensitivity (recall), specificity, F1-score, AUC, and MCC. Accuracy represents the proportion of all correctly classified observations. Precision indicates the proportion of predicted positive cases that are actually positive, while sensitivity measures the proportion of actual positive cases correctly detected. Specificity measures the proportion of negative cases correctly identified. The F1 score summarizes precision and recall through their harmonic mean. AUC summarizes discrimination across classification thresholds, while MCC provides a correlation-based measure that incorporates all four confusion-matrix components. Using several metrics is important for screening studies because a model with high accuracy can still perform poorly for one class. Earlier ASD studies have also used combinations of accuracy, sensitivity, specificity, and related classification measures [6] [11] [16] [20].

Proposed Workflow

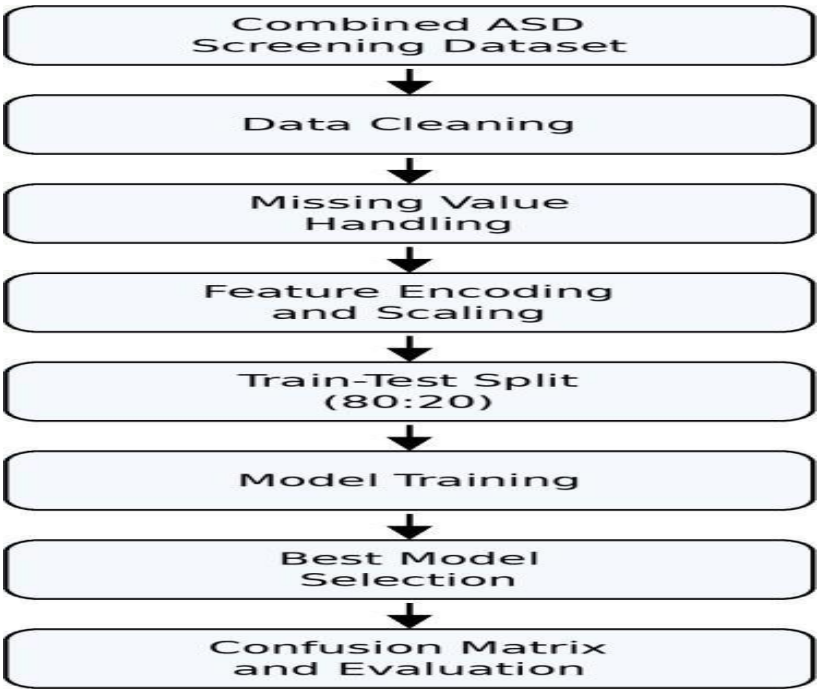


Fig. 1. Workflow diagram of the proposed ASD detection model.

Results and Analysis

Overall Model Performance

Table 2 reports the performance obtained on the 220-record held-out test set. Logistic Regression achieved the highest values across every reported metric. SVM also performed strongly, whereas Decision Tree and Naïve Bayes produced substantially lower overall results. The pattern illustrates why multiple classifiers and multiple metrics were considered instead of selecting a model from accuracy alone.

Table 2. Performance of all machine learning models on the combined ASD test set.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	MCC
Logistic Regression	100.00	100.00	100.00	100.00	100.00	1.000	1.000
SVM	98.64	98.72	97.47	99.29	98.09	0.999	0.970
KNN	95.00	91.46	94.94	95.04	93.17	0.979	0.893
Random Forest	94.09	95.83	87.34	97.87	91.39	0.986	0.871
Neural Network	93.18	93.24	87.34	96.45	90.20	0.988	0.851
Decision Tree	84.55	80.00	75.95	89.36	77.92	0.827	0.661
Naïve Bayes	38.64	36.27	93.67	7.80	52.30	0.511	0.027

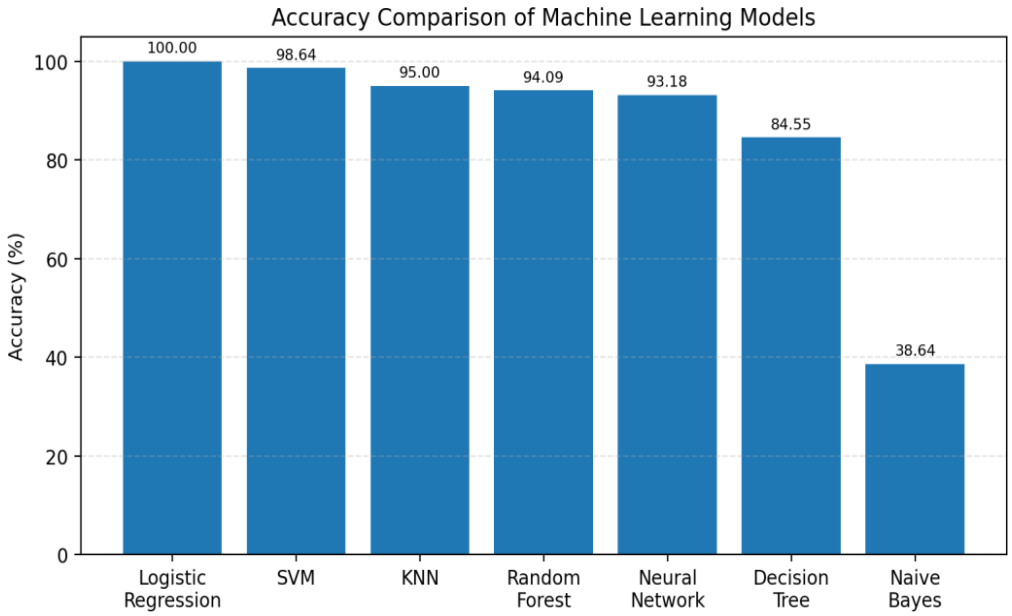


Fig. 2. Accuracy comparison of all models evaluated in the updated ASD experiment

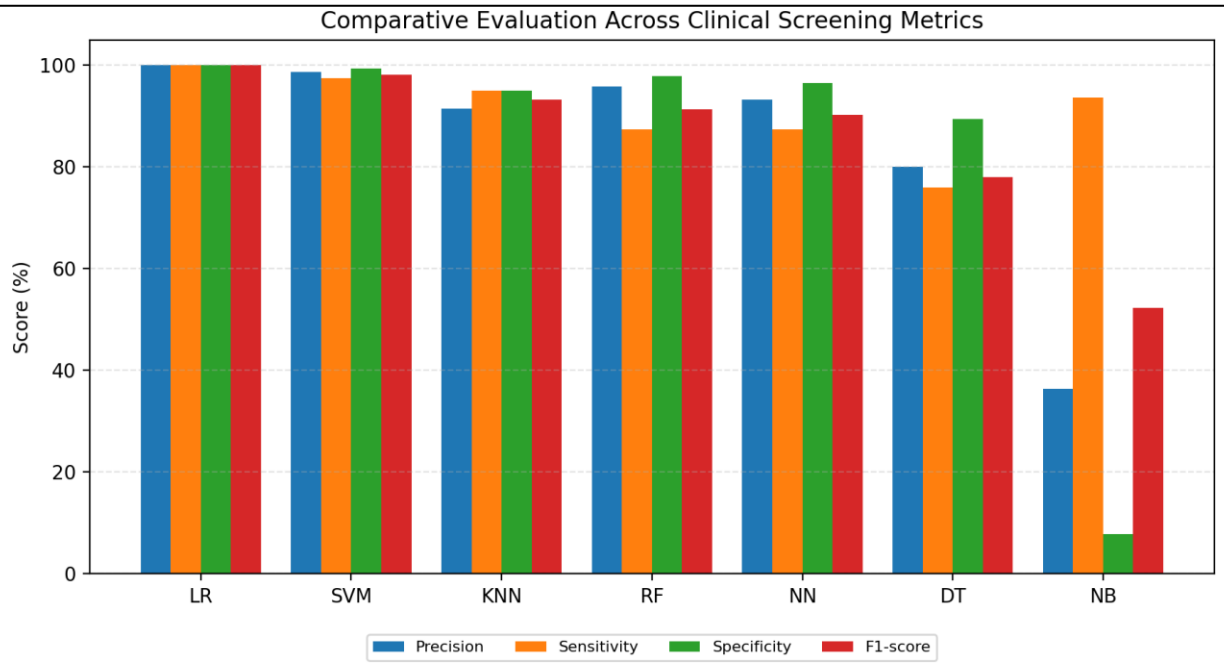


Fig. 3. Comparison of precision, sensitivity, specificity, and F1-score across the evaluated classifiers

Selected Model: Logistic Regression

Logistic Regression produced 100.00% accuracy, precision, sensitivity, specificity, and F1-score, together with AUC = 1.000 and MCC = 1.000. On the present test split, these values indicate complete separation of the two target classes. This result differs from the earlier age-specific comparison by Erkan and Thanh, in which Random Forest and SVM were the strongest approaches among the three tested algorithms [20]. The difference can reasonably be attributed to the different dataset composition, preprocessing decisions, model set, and experimental split used in the present study; it should not be interpreted as evidence that one algorithm is universally superior for ASD screening.

Table 3. Classification report of the selected best model on the test set.

Class	Precision	Recall	F1 score / Support
No ASD	1.00	1.00	1.00 / 141
ASD	1.00	1.00	1.00 / 79
Overall accuracy			1.00 / 220

Confusion Matrix

The Logistic Regression confusion matrix contained 141 true negatives and 79 true positives, with zero false positives and zero false negatives. Consequently, all 220 observations in this test split were classified correctly. Confusion-matrix analysis is particularly informative in screening research because it reveals whether strong aggregate performance is achieved at the expense of either positive-case detection or negative-case recognition [11] [16] [20].

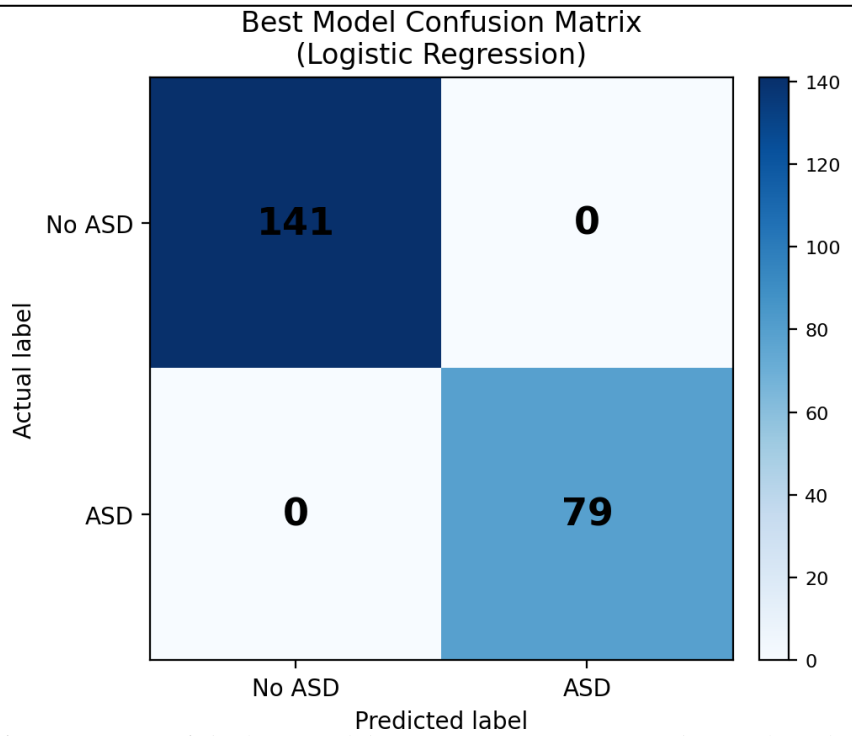


Fig. 4. Confusion matrix of the best model, Logistic Regression, on the combined ASD test set

Feature Importance and Interpretation

Although Logistic Regression was selected as the final classifier, Random Forest feature importance was used as a supplementary interpretation procedure. The most highly ranked variables were predominantly AQ questionnaire scores, including A9_Score, A5_Score, A6_Score, A4_Score, A3_Score, and A10_Score. This pattern indicates that behavioral questionnaire responses carried substantial predictive information in the present dataset. Such a finding is consistent with behavioral ASD screening research that treats questionnaire responses as central predictive attributes [6] [14] [16] [19] [20].

Discussion

The results demonstrate that a combined AQ-based screening dataset can be classified effectively when preprocessing is explicitly controlled. The removal of the result field is especially important because rule-derived or outcome-related fields can create an unrealistically easy prediction task. Prior ASD screening research has emphasized the use of behavioral variables and careful feature construction, making leakage control an important methodological consideration [6] [14] [16] [20]. In the present experiment, removing results before model training was intended to ensure that the classifiers learned from questionnaire and demographic information rather than from a field closely tied to the final label.

Logistic Regression achieved perfect performance on the selected test split. Its performance is noteworthy because the model is comparatively simple and interpretable relative to more complex nonlinear approaches. At the same time, a perfect result on a single held-out split can arise when the dataset contains highly discriminative questionnaire variables, strong class-separating structure, or residual dependencies between predictors and the target. The result therefore warrants external validation rather than immediate clinical interpretation.

SVM was the second strongest classifier, reaching 98.64% accuracy, 97.47% sensitivity, 99.29% specificity, and AUC = 0.999. This is consistent with the established use of SVMs for two-group

classification and with previous ASD screening research in which SVM was among the strongest models [18] [20]. KNN also performed well, with 95.00% accuracy and 94.94% sensitivity, which is compatible with the usefulness of local-neighbor methods for structured feature spaces [17].

Random Forest and Neural Network produced accuracy values of 94.09% and 93.18%, respectively. Random Forest also supplied the feature-importance analysis used for interpretation. Its ensemble structure makes it useful for identifying variables that repeatedly contribute to tree-based decisions, although feature importance should not be interpreted as proof of causal influence [19]. The Neural Network result suggests that nonlinear modelling was effective but was not necessary to obtain the highest test-set score in this experiment.

Naive Bayes produced 93.67% sensitivity but only 7.80% specificity. Thus, it identified many ASD-labelled cases but also classified a large number of non-ASD cases as positive. For a screening context, sensitivity is important, but extremely low specificity could lead to a high number of unnecessary follow-up assessments. The result reinforces the importance of reporting class-specific measures rather than accuracy alone [16] [20].

The broader biomedical machine-learning literature also illustrates why evaluation and validation must be separated from claims of clinical readiness. Studies in Alzheimer's disease have used machine learning to estimate disease-related risk from imaging and to integrate multimodal features, while work in Parkinson's disease has investigated feature-driven classification for early diagnosis [9] [10] [11] [12]. These studies provide methodological context for applying machine learning to health-related prediction, but they are not direct evidence about ASD and are therefore used here only to support the general importance of feature-based modelling, multimodal/high-dimensional analysis, and careful evaluation.

Neuroimaging-based ASD studies provide another important comparison. Deep-learning work with ABIDE data and systematic reviews of structural MRI machine learning demonstrate that ASD classification has been investigated using complex brain-imaging representations as well as behavioral questionnaires [5] [8]. The current study deliberately addresses the lower-dimensional, questionnaire-based setting, which is easier to reproduce with tabular public data but does not provide the same type of biological information as neuroimaging.

Finally, the perfect confusion matrix should be interpreted cautiously. The model was evaluated using one stratified 80:20 split, and the dataset itself is a screening dataset rather than an independent clinical cohort. Independent external datasets, repeated stratified cross-validation, calibration analysis, and subgroup evaluation
by age and demographic characteristics are
needed before considering practical decision-support applications [5]
[7] [8] [16] [20].

Limitations and Future Work

Single train-test split: The reported results come from one stratified 80:20 split. Repeated stratified cross-validation should be used to estimate variability [5] [8] [16] [20].

External validation: Testing on an independent dataset is necessary to determine whether the observed performance generalizes beyond the current combined dataset [5] [8] [16] [20].

Age and subgroup effects: The combined dataset contains multiple age-related records; future work should examine performance separately by age group and relevant demographic subgroups [5] [8] [16] [20].

Potentially strong questionnaire structure: AQ scores may contain highly discriminative information. Future studies should examine feature dependence and leakage risks in greater depth [5] [8] [16] [20].

Interpretability: Future work can add SHAP or LIME explanations and compare them with model-specific feature importance [5] [8] [16] [20].

Clinical use: The system should be considered a screening-support research model, not a replacement for professional clinical assessment [5] [8] [16] [20].

Conclusion

This study presents a machine-learning framework for ASD screening using a combined questionnaire-based dataset containing 1,100 records. The workflow included missing-value treatment, binary target encoding, removal of outcome-related and redundant fields, categorical encoding, numerical scaling, stratified train-test splitting, and multi-metric model evaluation. Seven supervised classifiers were compared: Logistic Regression, SVM, KNN, Random Forest, Neural Network, Decision Tree, and Naive Bayes. Logistic Regression produced the highest test-set results, reaching 100.00% across accuracy, precision, sensitivity, specificity, and F1-score, with AUC and MCC both equal to 1.000. Random Forest feature importance indicated that several AQ screening scores were among the most influential variables [6] [16] [19] [20]. The findings support the feasibility of machine learning for research on structured ASD screening data, but the perfect test-set result should not be interpreted as evidence of clinical deployment readiness. Future validation using repeated cross-validation, independent datasets, subgroup analysis, and clinically collected records is essential.

Data Availability

The experimental dataset was supplied as Combined_Autism_Data.csv with the updated Python notebook used for the reported experiment. The underlying ASD screening-data context is related to publicly available machine-learning resources and AQ-based screening datasets documented through the UCI Machine Learning Repository and prior ASD screening research [13] [14] [20].

REFERENCES

[1] World Health Organization. Autism spectrum disorders [Internet]. 2017 [cited 2018 Dec 4]. Available from: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[2] Hlavata P, Kasparek T, Linhartova P, et al. Autism, impulsivity and inhibition: a review of the literature. Basal Ganglia. 2018;14:44–53. https://doi.org/10.1016/j.baga.2018.10.002

[3] McDermott JH, Clayton-Smith J, Briggs TA. The TBR1-related autistic-spectrum-disorder phenotype and its clinical spectrum. European Journal of Medical Genetics. 2018;61(5):253–256. https://doi.org/10.1016/j.ejmg.2017.12.009

[4] Saresella M, Piancone F, Marventano I, et al. Multiple inflammasome complexes are activated in autistic spectrum disorders. Brain, Behavior, and Immunity.2016;57:125–133. https://doi.org/10.1016/j.bbi.2016.03.009

[5] Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical. 2017/2018;17:16–23. https://doi.org/10.1016/j.nicl.2017.08.017

[6] Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Informatics. 2018;117:112–124. https://doi.org/10.1016/j.ijmedinf.2018.06.009

[7] Franz L, Adewumi K, Chambers N, Viljoen M, Baumgartner JN, de Vries PJ. Providing early detection and early intervention for autism spectrum disorder in South Africa: stakeholder perspectives from the Western Cape province. Journal of Child and Adolescent Mental Health. 2018;30(3):149–165. https://doi.org/10.2989/17280583.2018.1525386

[8] Pagnozzi AM, Conti E, Calderoni S, Fripp J, Rose SE. A systematic review of structural MRI biomarkers in autism spectrum disorder: a machine learning perspective.

International Journal of Developmental Neuroscience. 2018;71:68–82. https://doi.org/10.1016/j.ijdevneu.2018.08.010

[9] Casanova R, Barnard RT, Gaussoin SA, et al. Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. NeuroImage. 2018;183:401–411. https://doi.org/10.1016/j.neuroimage.2018.08.040

[10] Parisi L, RaviChandran N, Manaog ML. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Systems with Applications. 2018;110:182–190. https://doi.org/10.1016/j.eswa.2018.06.003

[11] Escudero JH, Ifeachor E, Zajicek JP, Green C, Shearer J, Pearson S. Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease. IEEE Transactions on Biomedical Engineering. 2013;60(1):164–168. https://doi.org/10.1109/TBME.2012.2212278

[12] Liu S, Liu S, Cai W, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Transactions on Biomedical Engineering. 2015;62(4):1132–1140. https://doi.org/10.1109/TBME.2014.2372011

[13] Dua D, Graff C. UCI Machine Learning Repository. University of California, Irvine. Available from: https://archive.ics.uci.edu/

[14] Thabtah F. Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics. Taichung City, Taiwan. 2017:1–6. https://doi.org/10.1145/3107514.3107515

[15] Thabtah F. ASDTests: a mobile app for ASD screening [Internet]. 2017. Available from: https://www.asdtests.com/

[16] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. Informatics for Health and Social Care. 2019;44(3):278–297. https://doi.org/10.1080/17538157.2017.1399132

[17] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992;46(3):175–185. https://doi.org/10.1080/00031305.1992.10475879

[18] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. https://doi.org/10.1007/BF00994018

[19] Breiman L. Random forests. Machine Learning. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324

[20] Erkan U, Thanh DNH. Autism spectrum disorder detection with machine learning methods. Current Psychiatry Research And Reviews. 2019;15(4):297–308. https://doi.org/10.2174/2666082215666191111121115