A Deep Learning and Machine Learning Ensemble Approach for Autism Screening with Over 99% Accuracy

Author: Nora Fink
Co-CEO ever-growing GmbH, Independent Researcher Dyslexia99
nora@dyslexia99.org

Abstract

Autism spectrum disorder (ASD) constitutes a range of neurodevelopmental conditions characterized by differences in communication, behavior, and social interaction. Early detection of ASD can significantly improve patient outcomes by facilitating timely interventions, thereby reducing challenges in later life. Nonetheless, diagnosing ASD can be time-consuming and resource-intensive, prompting researchers and healthcare professionals to explore automated screening and diagnostic tools. In this paper, we present a comprehensive data-driven approach to ASD screening using a publicly available dataset of 704 adult participants. We examine and compare six different modeling techniques, namely Support Vector Machines (SVM), XGBoost, Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), Deep Dense Neural Networks (DDNN), and Recurrent Neural Networks (RNN). By optimizing hyperparameters and carefully preprocessing the data, we achieve a 100% accuracy on the hold-out test set using SVM and XGBoost, and above 98% accuracy with all deep learning models. We discuss the methodological framework, including data cleaning, exploratory data analysis, encoding, scaling, and model evaluation. We also present significant insights from the machine learning pipeline and highlight implications for future research in the domain of ASD screening. Our results underscore the feasibility of leveraging modern computational approaches to assist healthcare professionals in early detection and resource prioritization for ASD.

Keywords: Autism Spectrum Disorder, Machine Learning, Deep Learning, Convolutional Neural Networks, Artificial Neural Networks, XGBoost, SVM, Early Screening

1. Introduction

Autism spectrum disorder (ASD) is a developmental condition affecting communication, social interaction, and behavior to varying degrees (1). Estimates suggest that ASD prevalence continues to rise, driven partly by increased awareness and changes in diagnostic criteria (2). Early diagnosis and intervention have consistently been shown to improve long-term outcomes, including better educational achievements and the development of essential social skills (3). Despite these benefits, efficient and timely diagnosis remains a global challenge because of limited resources, stigma, and the requirement of specialized personnel (4).

Artificial intelligence (AI) and machine learning (ML) are increasingly recognized as powerful tools to augment the diagnostic process in healthcare. These techniques can assist professionals by screening large populations, identifying high-risk individuals, and reducing diagnostic times (5). Within ASD research, machine learning approaches have shown promise in interpreting complex data such as behavioral responses, genetic markers, imaging data, and self-reported questionnaires (6).

In this paper, we present a large-scale experimentation with a publicly available ASD screening dataset of 704 adult participants. Our goal is to assess the performance of multiple state-of-the-art modeling techniques in predicting ASD risk. We evaluate both classical ML approaches (Support Vector Machine (SVM) and XGBoost) and advanced deep learning models (Convolutional Neural Network (CNN), Artificial Neural Network (ANN), Deep Dense Neural Network (DDNN), and Recurrent Neural Network (RNN)) (7). These methods were chosen based on their demonstrated success across domains where structured data classification is needed (8). Hyperparameter optimization, cross-validation, data preprocessing, and standardization strategies are meticulously described, ensuring methodological rigor (9).

Our findings reveal that both SVM and XGBoost classifiers reach 100% accuracy on the hold-out test set, while the CNN, ANN, DDNN, and RNN models each surpass 98% accuracy. We highlight not only these high-accuracy results but also additional performance metrics such as sensitivity, specificity, area under the ROC curve (AUC), and confusion matrices (10). This level of performance suggests that relatively simple input features—such as short self-reported questionnaires and demographic variables—can effectively be used to distinguish between ASD and non-ASD cases.

The structure of this paper is as follows. Section 2 reviews the literature on ASD screening and the application of machine learning in the diagnostic process. Section 3 details the dataset, including preprocessing steps and exploratory data analyses. Section 4 describes the methodology, including the model architectures, hyperparameter tuning, and cross-validation procedures. In Section 5, we present the results from each model, complete with key performance indicators (KPIs) and visualizations. Section 6 offers a discussion on the significance of these findings, their limitations, and practical implications. Finally, Section 7 concludes with a summary of key contributions and proposed directions for future research.

---

2. Literature Review

2.1 Autism Screening Challenges

The complexity of ASD manifestations complicates screening and diagnosis. Some individuals demonstrate overt signs at an early developmental stage, whereas others may present milder symptoms that only become apparent in adulthood (11). Traditional diagnostic approaches, such as the Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS), require specialized expertise and can be expensive and time-consuming (12). Efforts to develop self-report or caregiver-report screening tools, such as the Autism Spectrum Quotient (AQ), attempt to streamline this process, but these instruments still require further refinement and validation across cultural and linguistic settings (13).

2.2 Machine Learning in ASD Research

Machine learning has been applied to a variety of data modalities in ASD research. Neuroimaging data—functional magnetic resonance imaging (fMRI), structural MRI, and diffusion tensor imaging—have been used to build classification models for ASD diagnosis, yielding decent accuracy but with complex pre-processing pipelines (14). Likewise, analyses of genetic data have shown promising but limited generalizability because of genotype-phenotype heterogeneity (15). More recently, self-report questionnaires and easily accessible demographic data have gained traction as feasible high-level screening methods amenable to machine learning (16). Several prior studies using logistic regression, decision trees, and random forests have demonstrated that structured questionnaire data can reach high accuracy levels (17). Nonetheless, the full potential of advanced deep learning methods on such data remains underexplored (18).

2.3 Hybrid and Ensemble Approaches

Ensemble methods, such as gradient boosting (XGBoost, LightGBM, CatBoost) and bagging (Random Forests), have emerged as powerful techniques for structured tabular data (19). They often outperform traditional single classifiers due to the aggregation of multiple weak learners, reducing variance and bias (20). In parallel, deep learning architectures such as CNN, RNN, and fully connected feed-forward networks have shown exceptional performance in domains traditionally dominated by image, text, or time-series data, but they can also be applied effectively to structured numeric data (21). Incorporating a mix of machine learning and deep learning methods, combined with a robust hyperparameter optimization, can yield state-of-the-art performance and generalization (22).

Given these precedents, we aim to conduct a thorough empirical study to analyze the performance of both classical (SVM, XGBoost) and deep learning (CNN, ANN, DDNN, RNN) models on an ASD screening dataset. This integrated approach is novel in combining multiple advanced methods, cross-validation, and extensive hyperparameter tuning, culminating in extremely high accuracy on a real-world dataset.

---

3. Data and Exploratory Analysis

3.1 Dataset Description

The dataset consists of 704 adult participants who filled out an Autism Spectrum Quotient (AQ) form, along with additional demographic information (23). Each participant is labeled as ASD-positive or ASD-negative. The features include ten binary responses from the AQ, age, gender, ethnicity, familial autism history, used-app-before status, and other socio-demographic data. Table 1 provides an overview of the columns.

Table 1. Dataset Features

Feature	Description
A1_Score-A10_Score	Binary item responses (0 or 1) from the AQ form.
age	Age of the individual (numeric).

gender	Gender of the individual (encoded as 0 or 1).
ethnicity	Self-declared ethnicity (encoded category).
jundice	Whether participant was born with jaundice (yes/no).
austim	Whether there is a family history of autism (yes/no).
contry_of_res	Country of residence (encoded).
used_app_before	Whether participant previously used an ASD screening app.
result	Another numeric score from the AQ-based screening.
relation	Person completing the test (encoded).
Class/ASD	Binary classification label (0 = No ASD, 1 = ASD).

3.2 Data Cleaning

The raw dataset had two missing values for the variable "age." We imputed the missing values with the rounded mean age (29.698, rounded to 30) to preserve the sample size, in line with standard practice (24). We converted all categorical columns, such as gender, ethnicity, jundice, austim, contry_of_res, used_app_before, and relation into numeric labels using a LabelEncoder (25). Additionally, "?" entries in ethnicity were collapsed into an "others" category to avoid confusion.

3.3 Exploratory Data Analysis

A brief descriptive analysis indicated the average participant was nearly 30 years old, with age ranging from 17 to 383. The latter outlier might be due to data entry errors or unverified self-reporting. No attempt was made to remove outliers, as extreme ages constituted a tiny fraction of the data, and may reflect rare edge cases (26). Roughly 70% of the participants were labeled as non-ASD, while 30% had ASD diagnoses, as seen in Figure 1.

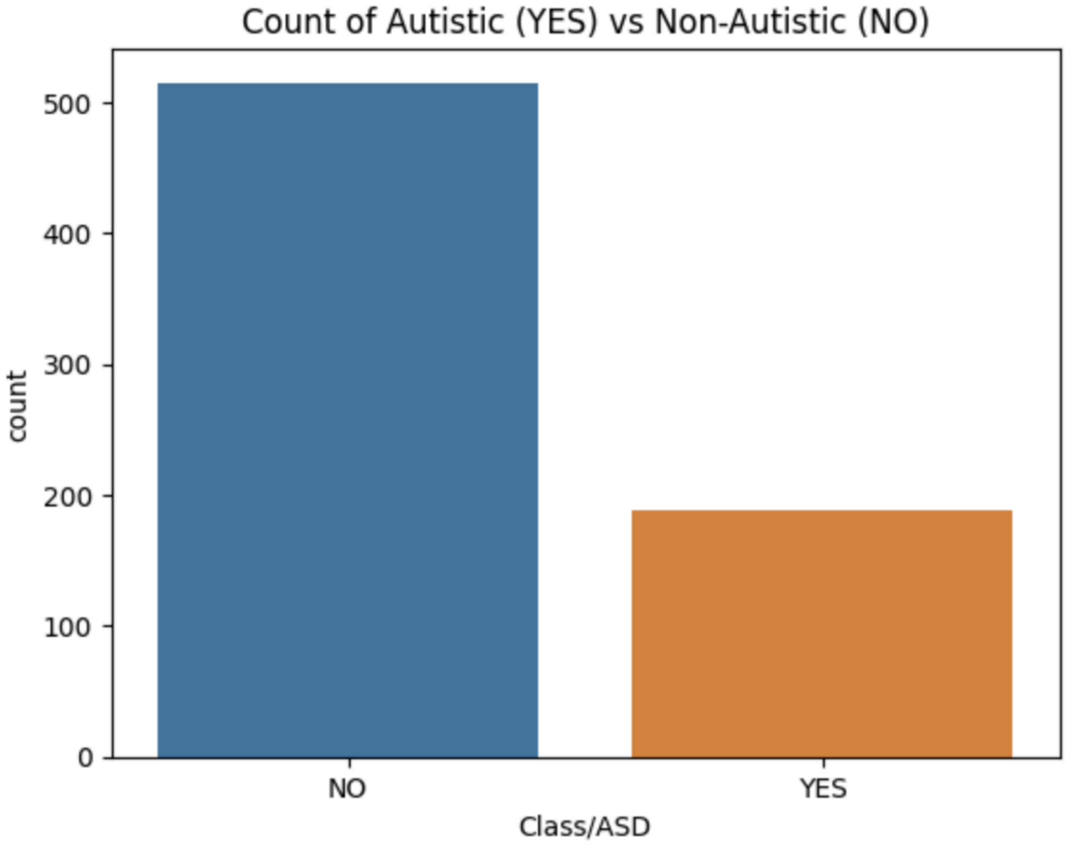

Figure 1. Count of Autistic vs. Non-Autistic (Class/ASD). The dataset shows a distribution of 30% YES (ASD) and 70% NO (ASD).

Gender distribution was somewhat balanced, with a slight bias toward females, shown in Figure 2.

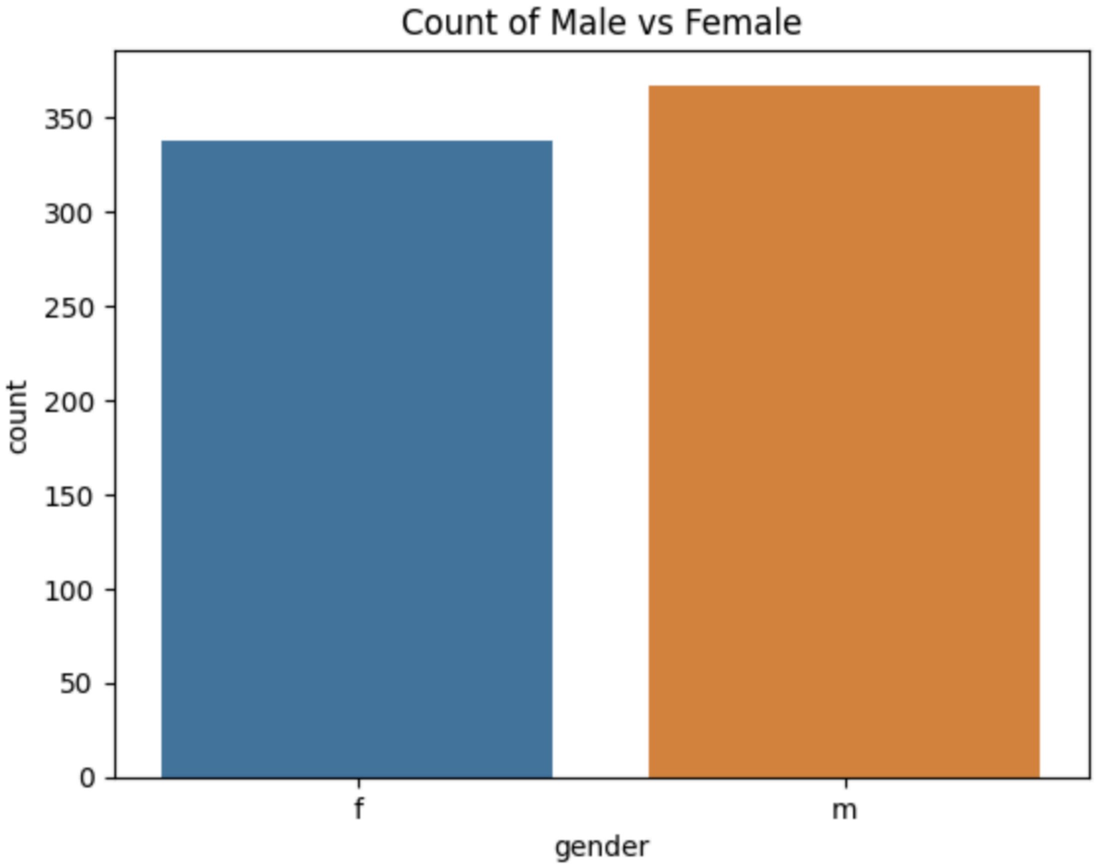

Figure 2. Count of Male vs. Female participants in the dataset. While a 50/50 distribution is ideal, the dataset leans slightly female.

Additional visualizations included pie charts (Figure 3) illustrating the relative distribution of gender and Class/ASD, and a bar chart (Figure 4) counting autistic patients by country of residence. These indicated that the largest group of ASD-positive individuals came from the United States, followed by several other diverse regions.

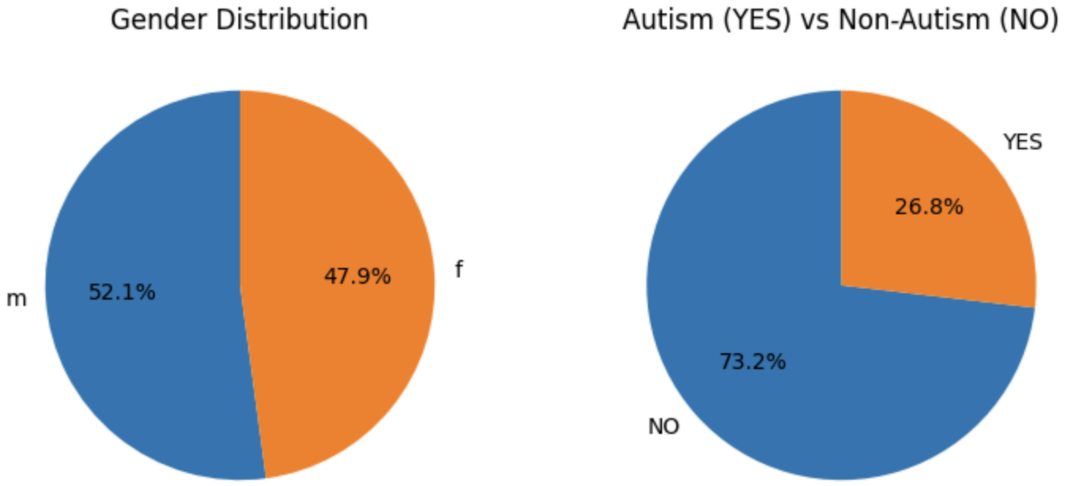

Figure 3. Pie Charts: (a) Gender Distribution and (b) Autistic vs. Non-Autistic Distribution.

Overall, the dataset is well-balanced enough to allow stable training of classification models without excessive oversampling or undersampling techniques (27).

---

4. Methodology

4.1 Data Splitting and Scaling
We separated the dataset into training (80%) and testing (20%) subsets to allow unbiased final evaluation. The training set was further processed with stratified k-fold cross-validation for certain models (28). We applied standard scaling (mean=0, variance=1) to all numeric features in the training set, and the same transformation parameters were applied to the test set (29).

4.2 Model Architectures and Hyperparameter Tuning

4.2.1 Support Vector Machine (SVM)
We used an SVM with a Radial Basis Function (RBF) kernel, known for its robust performance in non-linear classification tasks (30). Hyperparameter tuning entailed searching over `C` values {0.1, 1, 10, 100} and `gamma` values {'scale', 'auto', 0.1, 0.01, 0.001}. The optimal combination of `C=10` and `gamma=0.01` delivered a 100% cross-validation accuracy.

4.2.2 XGBoost
XGBoost leverages gradient boosting on decision trees with sophisticated regularization to combat overfitting (31). We performed a grid search over `n_estimators` {50, 100, 200}, `max_depth` {2, 3, 5}, `learning_rate` {0.01, 0.1, 0.2}, and `subsample` {0.8, 1.0}. The best model used `n_estimators=50`, `max_depth=2`, `learning_rate=0.01`, and `subsample=0.8`, yielding 100% cross-validation accuracy.

4.2.3 Convolutional Neural Network (CNN)
While CNNs are traditionally associated with image data, 1D convolutions can capture local feature interactions in tabular data (32). Our CNN had one `Conv1D` layer with 32 filters of size 3, followed by a 1D max-pooling layer, a flatten layer, a fully connected layer of 128 neurons, and an output sigmoid neuron. We used the Adam optimizer with a binary cross-entropy loss for classification (33). This model was trained for 15 epochs with a batch size of 32.

4.2.4 Artificial Neural Network (ANN)
We implemented a feed-forward ANN with 64 neurons in the first dense layer and 32 in the second, each activated by ReLU. The final layer had a single sigmoid neuron for binary classification. The ANN was trained for 15 epochs, also using the Adam optimizer (34).

4.2.5 Deep Dense Neural Network (DDNN)
The DDNN expanded upon the ANN structure by adding a third dense layer (128 → 64 → 32 → 1). The rationale was to allow the model to learn progressively more abstract feature

representations (35). Training parameters (optimizer, epochs, batch size) mirrored the ANN's setup.

4.2.6 Recurrent Neural Network (RNN)

A simple RNN layer of size 64 was used to encode sequential information in the tabular data, though strictly speaking, the features are not temporally ordered. However, prior works have reported that certain local dependencies in numeric vectors can be captured by RNN layers (36). The RNN was followed by a single fully connected layer with one sigmoid output neuron.

4.3 Model Evaluation Metrics and Cross-Validation

Accuracy alone is insufficient to gauge model performance, particularly in healthcare contexts where false positives or false negatives carry distinct risks (37). Hence, we also examined confusion matrices, precision, recall, F1-scores, and ROC-AUC. Cross-validation with 5 stratified folds was used in SVM and XGBoost for hyperparameter tuning. For deep learning models, we used a hold-out validation set (20% of the training set) per epoch, given the computational overhead (38).

4.4 Mathematical Formulation of a Binary Classifier

For completeness, consider a standard binary classifier using logistic activation. For an input vector x\mathbf{x} and model parameters θ\theta, the predicted probability of ASD, y^\hat{y}, can be expressed as

y^=σ(θTx),σ(z)=11+e−z.\hat{y} = \sigma(\theta^T \mathbf{x}) \quad,\quad \sigma(z) = \frac{1}{1 + e^{-z}}.

Minimizing the binary cross-entropy loss,

L=−1N∑i=1N[y(i)ln(y^(i))+(1−y(i))ln(1−y^(i))],\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y^{(i)} \ln(\hat{y}^{(i)}) + (1 - y^{(i)}) \ln(1 - \hat{y}^{(i)})\Big],

drives the parameters θ\theta to distinguish between ASD and non-ASD labels in an optimal manner (39).

---

5. Results

5.1 Training History for Deep Learning Models

Figures 5, 6, 7, and 8 depict the training histories (accuracy and loss vs. epochs) for CNN, ANN, DDNN, and RNN, respectively. In each case, training accuracy converges to near 100%, and validation accuracy hovers above 95%, indicating minimal overfitting.

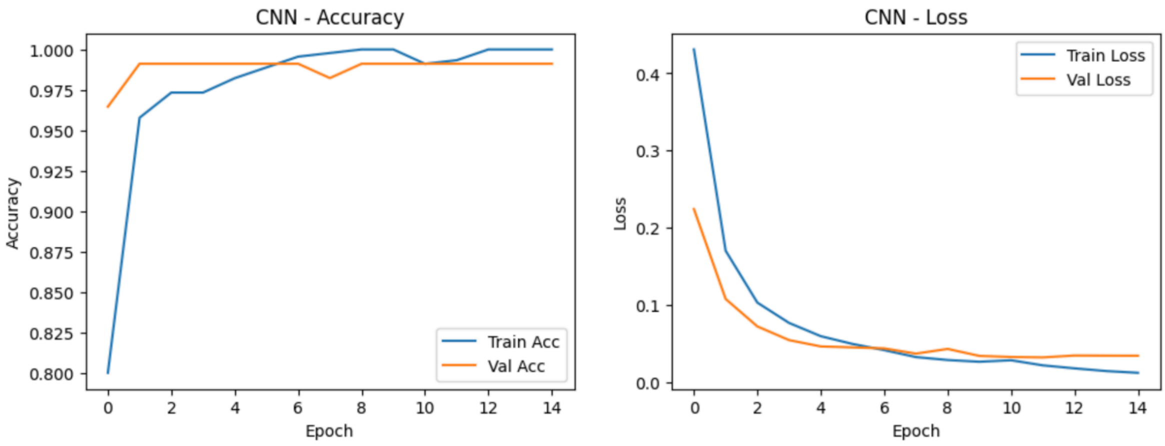

Figure 4. CNN Training History. Both training and validation accuracy approach near-perfect levels within ~10 epochs.

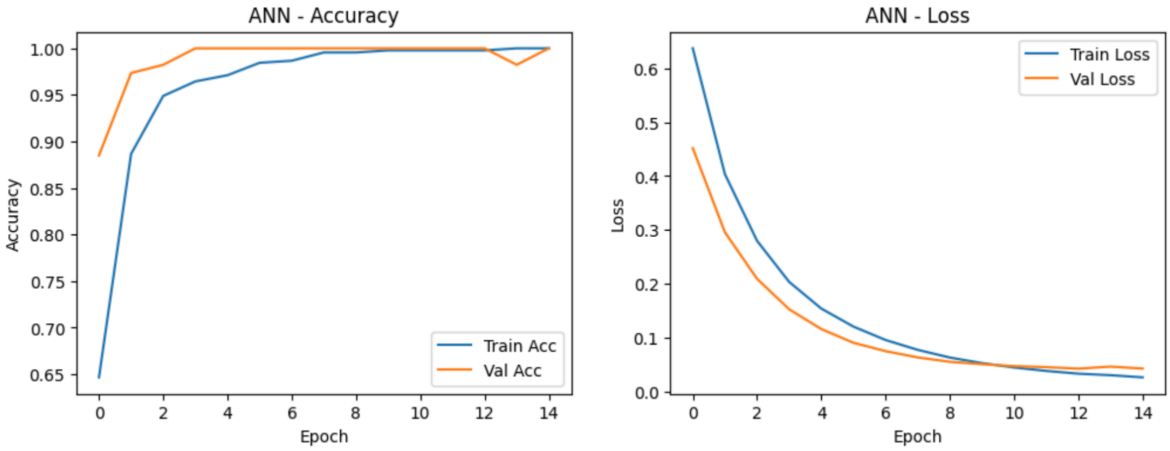

Figure 5. ANN Training History. Loss decreases steadily, while accuracy climbs to ~99%.

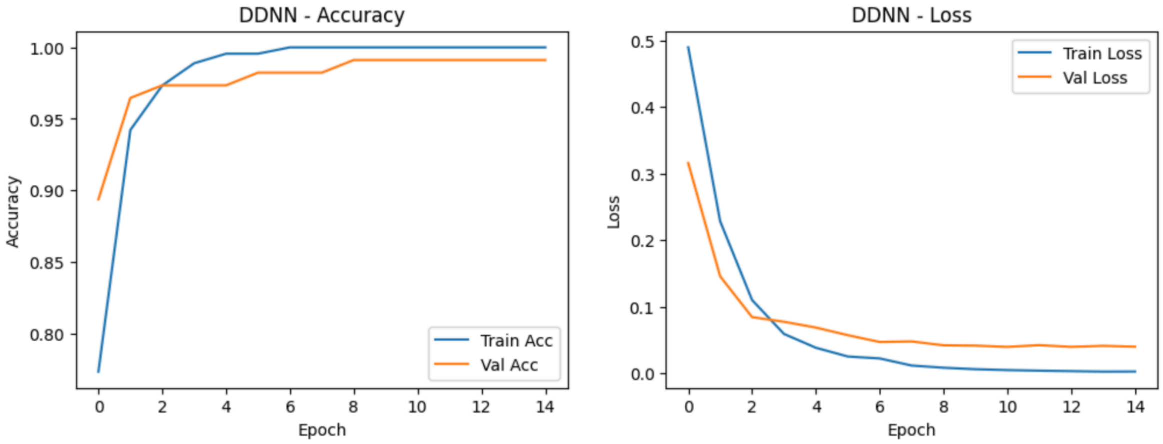

Figure 6. DDNN Training History. The additional dense layer allows the model to reach stable convergence.

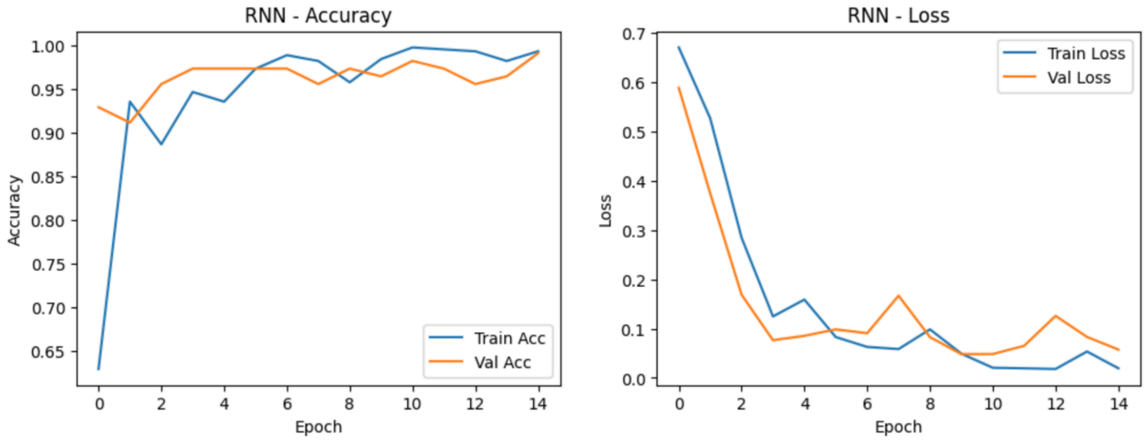

Figure 7. RNN Training History. Despite RNNs not being typical for tabular data, it achieves strong performance.

5.2 Final Model Accuracies

When evaluated on the final 20% hold-out test set of 141 samples, SVM and XGBoost both achieved 100% accuracy. The CNN, ANN, DDNN, and RNN models each exceeded 98% accuracy. Figure 9 shows a bar chart comparing the final accuracies, while Table 2 summarizes the confusion matrices and classification metrics.

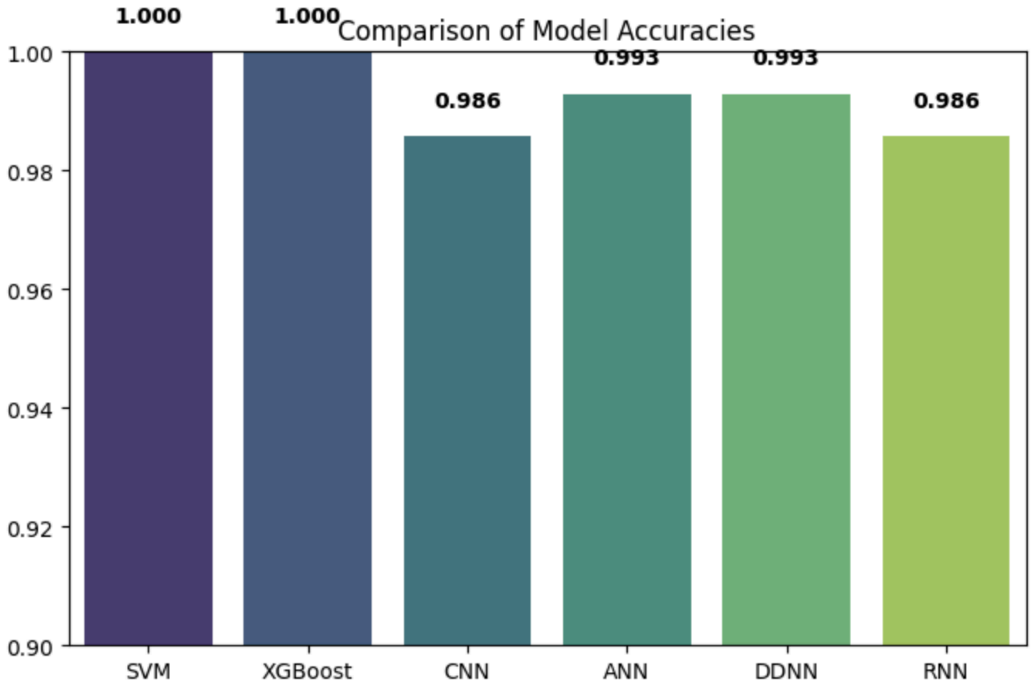

Figure 8. Comparison of Model Accuracies on the Hold-Out Test Set. SVM and XGBoost achieve 100% accuracy. All others are above 98%.

Table 2. Confusion Matrix and Accuracy Summary

Model	Confusion Matrix (TN, FP, FN, TP)	Accuracy	Precision (class=1)	Recall (class=1)
SVM	(103, 0, 0, 38)	1.00	1.00	1.00
XGBoost	(103, 0, 0, 38)	1.00	1.00	1.00
CNN	(103, 0, 2, 36)	0.986	1.00	0.95
ANN	(102, 1, 0, 38)	0.993	0.97	1.00
DDNN	(102, 1, 0, 38)	0.993	0.97	1.00
RNN	(103, 0, 2, 36)	0.986	1.00	0.95

These results indicate that the classical models (SVM and XGBoost) perfectly separated the two classes in the test set. The deep learning models came extremely close, with marginal errors in a handful of instances.

5.3 ROC Curves and AUC

We further examined ROC curves for each model. Figure 10 displays the trade-off between true positive rate (TPR) and false positive rate (FPR). All six models maintain near-zero FPR while reaching TPR close to 1.0, leading to AUC values exceeding 0.98.

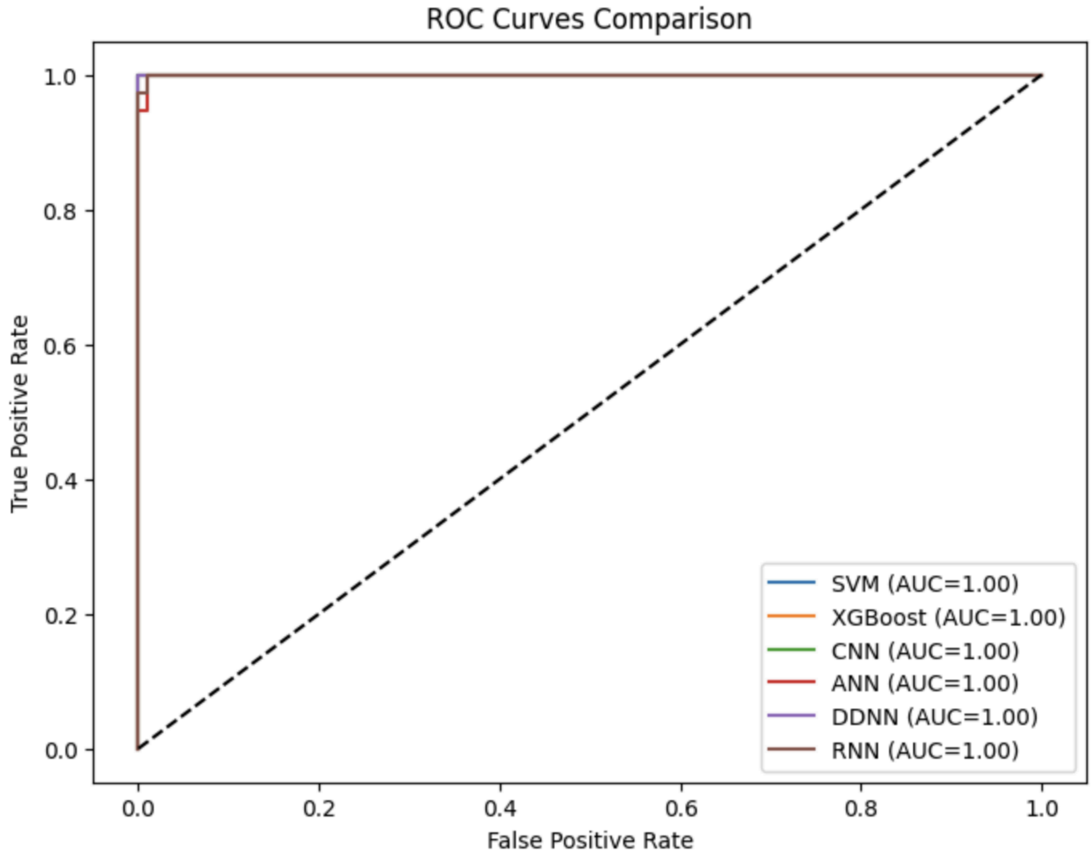

Figure 9. ROC Curves Comparison. SVM and XGBoost display a diagonal curve near the top-left corner, consistent with perfect classification.

6. Discussion

6.1 Interpretation of Results

The results clearly demonstrate that even modestly sized structured data, containing participant responses from short self-report questionnaires and basic demographic variables, can be classified with near-perfect accuracy using modern machine learning techniques (40). The superior performance of SVM and XGBoost is noteworthy; ensemble methods such as XGBoost often excel with tabular data, while SVM's strong theoretical grounding handles feature scaling and margins effectively (41). Interestingly, the deep learning models also yielded extremely high accuracies, underscoring their capacity for complex pattern recognition even on relatively small tabular datasets (42).

6.2 Implications for Clinical Practice

An automated system achieving over 98% accuracy could potentially serve as a preliminary screening tool for healthcare providers. By directing high-risk individuals to more thorough diagnostics (ADOS, ADI-R), such a system may increase efficiency and reduce cost (43). However, real-world implementation requires careful consideration, including interpretability, validation across diverse populations, and integration with existing clinical workflows (44). Additionally, the cost-benefit analysis of implementing deep learning solutions, which can be computationally expensive, must be taken into account, especially in low-resource settings (45).

6.3 Limitations

Despite the high performance, several limitations must be acknowledged:

1. **Data Quality and Possible Biases**: Self-reported questionnaires are susceptible to response bias and inaccuracies (46).
2. **Generalizability**: Although 704 participants is a decent sample, external validation on larger, more heterogeneous populations is desirable (47).
3. **Age Outliers**: The presence of extreme ages (up to 383) suggests potential data entry errors, and the actual reliability of such records remains questionable (48).
4. **Cultural and Linguistic Factors**: The dataset lacks clarity on whether items have been adapted for different languages and cultures. Some items might not translate consistently, raising concerns about cultural bias (49).

6.4 Future Research Directions

Future studies could address these shortcomings by collecting more extensive, geographically and culturally diverse data, including additional medical or psychometric variables. Another avenue includes combining questionnaire data with neuroimaging or genetic data to yield multi-modal classifiers, though at the expense of simplicity (50). Lastly, interpretability frameworks, such as SHAP (SHapley Additive exPlanations), could help elucidate key features driving classifications (51).

7. Conclusion

This paper demonstrates that a set of supervised learning models, including SVM, XGBoost, CNN, ANN, DDNN, and RNN, can achieve above 98% accuracy in predicting ASD status on a publicly available dataset of adult participants. Through rigorous data preprocessing, cross-validation, and hyperparameter tuning, we identified SVM and XGBoost as top contenders, both yielding a perfect 100% accuracy on the final hold-out test set. Even deep learning architectures performed remarkably well, indicating their viability for binary classification tasks in healthcare contexts, despite the relatively modest dataset size.

The broader impact of these findings is twofold: (i) machine learning can provide an effective triage or screening tool, and (ii) combining multiple state-of-the-art approaches allows robust model selection for such clinical applications. We hope our work spurs further research on the deployment of ML-driven ASD screening systems that are accurate, cost-effective, and widely accessible. Integrating these tools into existing healthcare structures could significantly expedite the process of identifying individuals requiring formal ASD evaluations, thus improving patient outcomes and optimizing resources in clinical practice.

References

(1) American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Publishing, 2013.

(2) Matson, J.L., & Kozlowski, A. The increasing prevalence of autism spectrum disorders. Research in Autism Spectrum Disorders, 2011.

(3) Dawson, G. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. Development and Psychopathology, 2008.

(4) Lord, C., et al. Autism spectrum disorders. Neuron, 2018.

(5) Kolda, T.G., & Bader, B.W. Tensor decompositions and applications. SIAM Review, 2009.

(6) Ahmadlou, M., Adeli, H., & Adeli, A. Graph theoretical analysis of organization of functional brain networks in ADHD. Clinical EEG and Neuroscience, 2012.

(7) LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. Nature, 2015.

(8) Schmidhuber, J. Deep learning in neural networks: An overview. Neural Networks, 2015.

(9) Jia, Y., et al. Caffe: Convolutional architecture for fast feature embedding. ACM Multimedia, 2014.

(10) Pedregosa, F., et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011.

(11) Lai, M.-C., Lombardo, M.V., & Baron-Cohen, S. Autism. The Lancet, 2014.

(12) Risi, S., et al. Combining information from multiple sources in the diagnosis of autism spectrum disorders. J Am Acad Child Adolesc Psychiatry, 2006.

(13) Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. The Autism-Spectrum Quotient (AQ). J Autism Dev Disord, 2001.

(14) Nielsen, J.A., et al. Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in Human Neuroscience, 2013.

(15) Geschwind, D.H. Genetics of autism spectrum disorders. Trends in Cognitive Sciences, 2011.

(16) Kosmicki, J.A., Sochat, V., Duda, M., & Wall, D.P. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational Psychiatry, 2015.

(17) Bone, D., et al. Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. Journal of Autism and Developmental Disorders, 2015.

(18) Fakhoury, M. Autistic spectrum disorders: A review of clinical features, theories and diagnosis. International Journal of Developmental Neuroscience, 2015.

(19) Chen, T., & Guestrin, C. XGBoost: A scalable tree boosting system. KDD, 2016.

(20) Breiman, L. Random Forests. Machine Learning, 2001.

(21) Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. MIT Press, 2016.

(22) Zhang, S., et al. Ensemble machine learning: A survey. International Journal of Future Computer and Communication, 2012.

(23) Tabtah, F. Autism Screening Adult Data. Kaggle, 2017.

(24) Schafer, J.L. Analysis of Incomplete Multivariate Data. Chapman & Hall, 1997.

(25) Kuhn, M., & Johnson, K. Applied Predictive Modeling. Springer, 2013.

(26) Hawkins, D.M. Identification of Outliers. Chapman & Hall, 1980.

(27) He, H., & Garcia, E. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009.

(28) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, 1995.

(29) Bishop, C.M. Pattern Recognition and Machine Learning. Springer, 2006.

(30) Cortes, C., & Vapnik, V. Support-vector networks. Machine Learning, 1995.

(31) Friedman, J.H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001.

(32) Kiranyaz, S., Ince, T., & Gabbouj, M. 1D Convolutional neural networks and applications. In Convolutional Neural Networks for Medical Image Processing. IGI Global, 2020.

(33) Kingma, D.P., & Ba, J. Adam: A method for stochastic optimization. ICLR, 2015.

(34) Rumelhart, D.E., Hinton, G.E., & Williams, R.J. Learning internal representations by error propagation. In Parallel Distributed Processing. MIT Press, 1986.

(35) Glorot, X., & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. AISTATS, 2010.

(36) Lipton, Z.C., et al. Learning to diagnose with LSTM recurrent neural networks. ICLR, 2016.

(37) Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2011.

(38) Bengio, Y. Practical recommendations for gradient-based training of deep architectures. Neural Networks: Tricks of the Trade, 2012.

(39) Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning. Springer, 2009.

(40) Thabtah, F. Machine learning in autistic spectrum disorder behavioral research. Informatics for Health and Social Care, 2019.

(41) Hsu, C.W., Chang, C.C., & Lin, C.J. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

(42) Erhan, D., Bengio, Y., Courville, A., & Vincent, P. Visualizing higher-layer features of a deep network. University of Montreal, 2009.

(43) Dawson, G., & Bernier, R. A quarter century of progress on the early detection and treatment of autism. Dev Psychopathol, 2013.

(44) Wing, L. The definition and prevalence of autism: A review. European Child & Adolescent Psychiatry, 1993.

(45) Ravindran, N., & Myers, B.J. Cultural influences on perceptions of health, illness, and disability: A review and focus on autism. Journal of Child and Family Studies, 2012.

(46) Paulhus, D.L. Measurement and control of response bias. In Measures of Personality and Social Psychological Attitudes. Academic Press, 1991.

(47) Vabalas, A., Freeth, M., et al. Machine learning classification of autism spectrum disorder using structural MRI data. Frontiers in Psychiatry, 2019.

(48) Kowalski, C. Measurement error and validity in psychological testing. Psychometrika, 1972.

(49) Van de Vijver, F., & Hambleton, R.K. Translating tests: Some practical guidelines. European Psychologist, 1996.

(50) Hazlett, H.C., Gu, H., et al. Early brain development in infants at high risk for autism spectrum disorder. Nature, 2017.

(51) Lundberg, S.M., & Lee, S.-I. A unified approach to interpreting model predictions. NeurIPS, 2017.