International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

A Multimodal Explainable Deep Learning Framework for Autism Severity and Behavioral Profile Prediction Using Clinical and Behavioral Data

[1]R.prathusha

[1]Assistant Professor, Department of Computer Science and Engineering, Methodist college of engineering and technology,Abids,Hyderabad

[1]Email: prathyusha88r@gmail.com

Abstract:

Autism Spectrum Disorder, or ASD, is a neurodevelopmental condition. People with ASD show differences in how they communicate socially. They may also have distinct behavior patterns. Development can look different from person to person. Getting checked early can help. It can also help teams act sooner. Still, standard screening tools might miss some details. Clinical and behavior signs can add extra clues. Conventional methods may not use all of that combined information. This study proposes a lightweight multimodal explainable deep learning framework for screening-based severity stratification and behavioral profile prediction using clinical and behavioral data. The framework processes clinical attributes through a multilayer perceptron (MLP) and ten behavioral screening responses through a bidirectional long short-term memory (BiLSTM) network. The resulting representations are combined using feature-level fusion and passed to two prediction heads for severity-category and behavioral-profile estimation. SHAP-based analysis is incorporated to examine the contribution of clinical features, while permutation-based analysis is used to assess the importance of individual behavioral screening items. Experiments were conducted on the UCI Adult Autism Screening dataset. After removal of duplicate records and basic data cleaning, 699 samples were retained. The multimodal model scored 99.05% accuracy. It reached 99.22% macro precision. It also got 98.81% macro recall. The macro F1 was 99.00% for the severity task that used screening scores. Its macro ROC-AUC came out to 0.9992. For the behavioral profile task, the results were lower. The micro F1 was 88.98%. The macro F1 was 88.06%. The results demonstrate the feasibility of combining clinical and behavioral information within an explainable deep learning framework. However, because the severity and behavioral-profile targets were derived from the screening responses available in the selected dataset, the findings should be interpreted as a proof-of-concept rather than as independent clinical severity diagnosis.

Key Words: Autism Spectrum Disorder, multimodal learning, deep learning, BiLSTM, clinical data, behavioral data, explainable artificial intelligence, SHAP, behavioral profile prediction, severity prediction.

1. Introduction

Autism Spectrum Disorder, or ASD, is a neurodevelopmental condition. It shows up with ongoing differences in social talk and social interaction. It also includes restricted or repetitive behaviors and interests. ASD

does not look the same in every person. Because of this, spotting it early and describing it well can be hard. Two people may have similar labels but show different signs. A lot of factors can shape what clinicians see. These can include how a person acts over time, their developmental past, age-related or other background traits, and notes from clinical observation. So the group of people who go through autism screening can be very mixed. For this reason, computer methods that combine different kinds of data can help. They may support the search for clear patterns in screening results.

Researchers have tried machine learning and deep learning for autism screening and prediction. The main idea is that these models can pick up patterns from age, clinical notes, and behavior data. Some older methods have also shown promise. Logistic regression can sort cases using measured screening items. Support vector machines, random forests, and gradient boosting are used too. These can still separate people when the input is in a clear, tabular form. Deep learning takes a different route. It can learn curved or non straight links between variables. It can also form new features from the behavioral and clinical inputs. Still, a lot of work has one big limitation. Some methods use only one type of data. Others merge mixed features into one input and do not keep clinical signals apart from behavioral ones. This can make it harder for the model to reflect what is truly different across those sources.

Behavioral screening answers can help with computational work on ASD. Each question on a form can reflect a different part of social life, how a person communicates, and other behavior traits. Clinical details and basic demographic facts also matter. They add context that can fit with what the answers show. A multimodal learning plan can handle these sources on their own first. Then it can merge what each part has learned. This can make the final model more stable. In healthcare tasks, the model also needs to be explainable. People using it want to know what parts of the input push the prediction. Methods from explainable AI can support this goal. For example, SHAP can show how features affect the model at a fine level. This can make the results easier to understand.

We present a small multimodal deep learning approach for two tasks. One task sorts severity levels from screening data. The other task predicts a behavioral profile. For the clinical data, we use a multilayer perceptron. For the screening responses, we apply a bidirectional long short term memory model. This part looks at the order of ten behavioral answers. After that, we join the outputs by concatenating the features. Then we send the merged signal to a shared layer. From there, the model splits into two prediction heads. The first head makes a three class severity decision. These classes come from screening score groups. The second head gives three indicators for the behavioral profile that comes from screening. To check what matters most, we use SHAP on the clinical features. We also run a permutation based feature importance test. That helps us see how each behavioral screening item affects the results.

2. Related Work

Recent Advancements in machine learning and deep learning has opened fresh paths for using tools to help screen autism and study behavior. Researchers have looked at many kinds of inputs. Some papers use clinic notes and basic demographic data. Others focus on answers from behavior checklists. There is also work using brain scans. More studies test additional types of data too.. Conventional machine learning algorithms can provide effective classification for structured datasets, while deep neural networks can learn nonlinear feature representations from complex behavioral and clinical information. However, many approaches remain focused on a single prediction task or a single type of input data. In addition, model interpretability remains an important consideration for healthcare-oriented applications because high predictive performance alone does not explain which characteristics influence a model's decision. The following subsections briefly review machine learning approaches for autism screening, deep learning approaches for behavioral analysis, explainable artificial intelligence in healthcare, and the research gap addressed by this study.

2.1 Machine Learning for Autism Screening

A lot of work has looked at machine learning for finding ASD in an automatic way. Many papers rely on inputs such as age and sex. They also include clinical notes and responses from questionnaires [1]. A few frequent model types are logistic regression, support vector machines, decision trees, random forests, and gradient boosting. These can pick up patterns in tabular screening data. In other words, they can help with links that are not purely linear, and they can separate groups based on the available features. These methods often fit better when the dataset is not huge [2]. They usually need less compute than large deep learning systems. Earlier studies have also found that questionnaire answers, when paired with demographic and developmental factors, can help tell apart people who have ASD from those who do not.

Most standard machine learning methods handle every variable as one big feature set. That setup does not clearly split clinical traits from behavior-related reactions. Yet those two types of data come from different parts of a person's screening record. Also, many studies stop at a simple yes or no ASD label [3]. Less effort goes into grading how severe the screening signals look or mapping out behavior patterns. Because of these issues, it makes sense to try multimodal features. The idea is to represent clinical information and behavioral information on their own first. Then they can be merged later for the prediction task.

2.2 Deep Learning for Autism for Behavirol Analysis

Deep learning methods provide an alternative to conventional machine learning by automatically learning hierarchical representations from input data [4]. Recurrent neural networks, including long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) architectures, are particularly useful when the input contains ordered observations because they can capture dependencies among successive features [5]. In autism-related applications, behavioral assessment

responses can be represented as a sequence of observations, allowing recurrent architectures to learn relationships among different screening indicators.

Even with these benefits, using deep learning for ASD screening can run into trouble. This happens when the data set is not large and when the samples differ from each other. If you train one model on all clinical and behavioral fields at once, it may not clearly learn what each part contributes [6]. A better option is a multimodal setup. In that design, each type of input gets its own feature encoder. Then the model merges the outputs from those encoders. This lets each modality keep its own cues, while the final system can still use links between the sources [7]. Here, a simple plan is used. An MLP handles the structured clinical inputs. A BiLSTM processes the behavioral screening responses. Together, they form a multimodal model that fits a small tabular data set.

2.3 Explainable AI in Healthcare

The increasing use of artificial intelligence in healthcare has highlighted the need for interpretable prediction systems [8]. Complex models may achieve strong predictive performance while providing limited insight into the factors responsible for individual predictions [9]. Explainable artificial intelligence (XAI) methods attempt to address this limitation by identifying the contribution of input variables to model outputs. Among these methods, SHAP is based on Shapley-value concepts and provides both global and local explanations of model predictions.

For structured healthcare data, SHAP can be used to rank features according to their average contribution and to examine whether individual observations increase or decrease a particular prediction [10]. This can be useful for understanding the behavior of machine learning systems and identifying potentially influential clinical or demographic attributes. For behavioral inputs, feature-importance techniques can similarly be used to investigate the relative contribution of individual screening responses [11]. In the proposed framework, SHAP is applied to the clinical feature representation, while permutation-based importance is used for the behavioral screening variables.

2.4 Research Gap

The reviewed approaches indicate several opportunities for improvement. First, many autism screening studies primarily focus on binary ASD classification rather than extending the analysis toward screening-based severity stratification and behavioral characterization. Second, clinical and behavioral attributes are often processed together without explicitly modeling their different characteristics. Third, predictive performance is frequently emphasized without sufficient analysis of feature contributions and model behavior. Finally, complex architectures may be impractical for small publicly available tabular datasets.

To address these gaps, this study develops a lightweight multimodal framework that separately processes clinical and behavioral information before feature-level fusion. The proposed approach combines an MLP-based clinical encoder with a BiLSTM-based behavioral encoder and uses a shared representation for screening-based severity and behavioral-profile prediction. An

explainability layer further analyzes the contribution of the input features. Rather than introducing computationally intensive architectures, the proposed framework emphasizes a simple, reproducible design that can be implemented on a relatively small ASD screening dataset.

3. Proposed Methodology

The proposed framework is designed as a lightweight multimodal deep learning system that combines structured clinical information with behavioral screening responses for autism-related prediction. The framework uses separate feature-processing branches for the two modalities, followed by feature-level fusion and multi-output prediction. Clinical attributes are processed using a multilayer perceptron (MLP), while the ten behavioral screening responses are modeled using a bidirectional long short-term memory (BiLSTM) network. The learned representations are subsequently concatenated to form a shared multimodal representation. This representation is used for screening-based severity classification and screening-derived behavioral profile prediction. Finally, explainability techniques are applied to examine the contribution of clinical and behavioral features.

3.1 Dataset Description

The experiments were conducted using the UCI Adult Autism Screening Dataset, a publicly available dataset designed for autism screening research. The original dataset contains 704 instances and 21 attributes, including ten autism screening responses, demographic variables, medical-history-related attributes, screening score information, and the ASD classification label.

The ten behavioral variables, denoted as A1–A10, represent individual screening responses. Additional variables include age, gender, ethnicity, history of jaundice, family history of autism-related conditions, country of residence, previous use of the screening application, respondent relationship, and the final ASD screening classification.

During preprocessing, categorical missing-value indicators represented by "?" were converted to missing values. Two missing age values were handled using median imputation, while an implausible age value of 383 was treated as an invalid observation and replaced through median imputation. Five duplicate records were removed, resulting in 699 observations for the experiments.

The result attribute corresponds exactly to the sum of the ten A1–A10 screening responses. Therefore, it was excluded from the model inputs to prevent information leakage. Similarly, the severity and behavioral-profile targets were constructed independently of the result column but are derived from the available screening responses. These targets are consequently interpreted as screening-derived research targets rather than clinically validated diagnostic labels.

3.2 Data Preprocessing

The cleaned dataset was divided into clinical and behavioral modalities. Clinical attributes consisted of age, gender, ethnicity, jaundice history, autism-related family history, country of residence, previous application usage, and respondent relationship. Behavioral information consisted of the ten binary screening responses from A1 to A10.

Numerical clinical variables were standardized using:

$$x' = \frac{x - \mu}{\sigma}$$

where x represents the original feature value, μ represents the training-set mean, and σ represents the training-set standard deviation. Categorical clinical variables were transformed using one-hot encoding. Missing categorical values were handled using the most frequent category.

The data were divided into training, validation, and testing subsets using a 70:15:15 ratio. Stratified splitting was applied according to the severity target to preserve its class distribution across the subsets. Importantly, preprocessing transformations were fitted only on the training data and subsequently applied to the validation and test sets.

The behavioral A1–A10 responses were retained as binary values and arranged into a ten-step sequence for the BiLSTM branch. Thus, each behavioral sample was represented using the shape:

$$10 \times 1$$

where ten represents the number of behavioral screening items and one represents the feature value at each step.

3.3 Clinical Feature Representation

The clinical modality is processed using a lightweight multilayer perceptron (MLP). The clinical input consists of demographic and clinical attributes including age, gender, ethnicity, jaundice history, autism-related family history, country of residence, previous application usage, and respondent relationship. After preprocessing and one-hot encoding, the resulting clinical features are provided to the MLP.

The clinical branch consists of a dense layer with 64 neurons and ReLU activation, followed by a dropout layer with a rate of 0.2. A second dense layer containing 32 neurons generates the final clinical representation. This compact representation captures nonlinear relationships among the structured clinical attributes and is subsequently provided to the multimodal fusion stage.

3.4 Behavioral Feature representation

The behavioral modality consists of the ten A1–A10 autism screening responses. These responses are arranged as a ten-step sequence and provided to a bidirectional long short-term memory (BiLSTM) network. The BiLSTM processes the behavioral sequence in both forward and backward directions, allowing relationships among the screening responses to be represented.

The behavioral branch uses a 64-unit BiLSTM layer followed by dropout with a rate of 0.2. A dense layer containing 32 neurons then produces the final behavioral representation. This representation captures patterns across the ten behavioral screening indicators and is subsequently combined with the clinical representation.

3.5 Multimodal Feature Fusion

The clinical and behavioral representations are combined using feature-level concatenation. Each modality produces a compact 32-dimensional representation, which are joined to form a unified multimodal representation.

The fused representation is passed through a dense layer containing 64 neurons with ReLU activation, followed by dropout. A final dense layer with 32 neurons produces

the shared representation used by both prediction tasks. This simple fusion strategy allows information from clinical and behavioral modalities to contribute jointly to the final predictions without introducing computationally complex fusion mechanisms.

3.6 Screening – Based Severity Prediction

The first prediction head performs three-class screening-based severity classification. The severity categories are derived from the total A1–A10 screening score. Scores from 0–3 are categorized as Low, scores from 4–6 as Moderate, and scores from 7–10 as High.

A softmax output layer is used to generate the probability of each severity category. The category with the highest predicted probability is selected as the model output. Because these categories are derived from the screening responses available in the dataset rather than independently validated clinical severity annotations, they are treated as screening-based severity proxies rather than clinical severity diagnoses.

3.7 Behavioral Profile Prediction

The second prediction head estimates three screening-derived behavioral profile indicators: Social/Communication, Behavioral/Repetitive, and Social/Adaptive. The A1–A10 screening responses are grouped into these dimensions to provide a simple representation of different behavioral characteristics.

A sigmoid-based output layer generates an independent prediction for each behavioral profile. This allows more than one behavioral characteristic to be predicted simultaneously for the same individual.

3.8 Model Training

The complete multimodal model is trained jointly using both prediction objectives. Categorical cross-entropy is used for the three-class severity task, while binary cross-entropy is used for the three behavioral-profile outputs.

The Adam optimizer is used with a learning rate of 0.001. A batch size of 32 and a maximum of 50 epochs are employed. Early stopping with a patience of 8 epochs is used to reduce overfitting, with the model weights from the best validation performance restored after training.

3.9 Explainability Analysis

Explainability is incorporated to examine the contribution of the input features. SHAP is applied to the clinical-only model to identify the relative contribution of encoded clinical attributes to the prediction of the high screening-severity category. Mean absolute SHAP values are used to rank the most influential clinical features.

For the behavioral modality, permutation-based feature importance is used. Each A1–A10 screening feature is individually shuffled, and the reduction in prediction accuracy is measured. Features producing a larger accuracy reduction are considered more influential to the model's behavioral prediction.

4. Experimental Results

This section evaluates the proposed multimodal framework using the held-out test set. The evaluation focuses on screening-based severity prediction and screening-derived behavioral profile prediction. To examine the contribution of each modality, the proposed model is also compared with clinical-only and behavioral-only baseline models. Explainability results

obtained using SHAP and permutation-based feature importance are presented separately.

4.1 Experimental Setup

After preprocessing and duplicate removal, the final dataset contained 699 samples. The data were divided into training, validation, and testing subsets using a 70:15:15 ratio. Clinical attributes were processed using imputation, one-hot encoding, and standardization, while the A1–A10 behavioral responses were provided as sequential inputs to the BiLSTM branch.

We used a model with four parts. First, there is a clinical MLP. Next, we add a behavioral BiLSTM. After that, a layer mixes the features. Finally, there are two prediction heads. For training, we ran Adam with a learning rate of 0.001. We chose a batch size of 32. We let training run up to 50 epochs. To stop early, we watched the validation loss. When it did not improve, we ended training.

Three models were considered for comparison:

- Clinical-only MLP: uses only clinical attributes.
- Behavioral-only BiLSTM: uses only A1–A10 behavioral screening responses.
- Proposed multimodal model: combines clinical and behavioral representations.

Accuracy, precision, recall, and macro-F1 were used to evaluate severity prediction. Behavioral-profile prediction was evaluated using precision, recall, and F1-score for each profile, together with overall micro-F1 and macro-F1.

4.2 Severity Prediction Performance

The multimodal model achieved a 99.05% score on the test set it had not encountered during training. For macro metrics, the outcomes were 99.22% precision, 98.81% recall, and 99.00% F1. The overall ROC-AUC was calculated to be 0.9992. That figure indicates the model effectively distinguished the three severity categories.

Table 1: Severity prediction performance of the proposed model

Class	Precision	Recall	F1-score	Support
High	1	0.96	0.98	28
Low	1	1	1	35
Moderate	0.98	1	0.99	42
Macro Average	0.99	0.99	0.99	105
Overall Accuracy			0.9905	105

The confusion matrix further demonstrates the strong classification performance, with predictions closely concentrated along the correct-class diagonal. The ROC analysis also indicates excellent discrimination across the three screening-based severity categories.

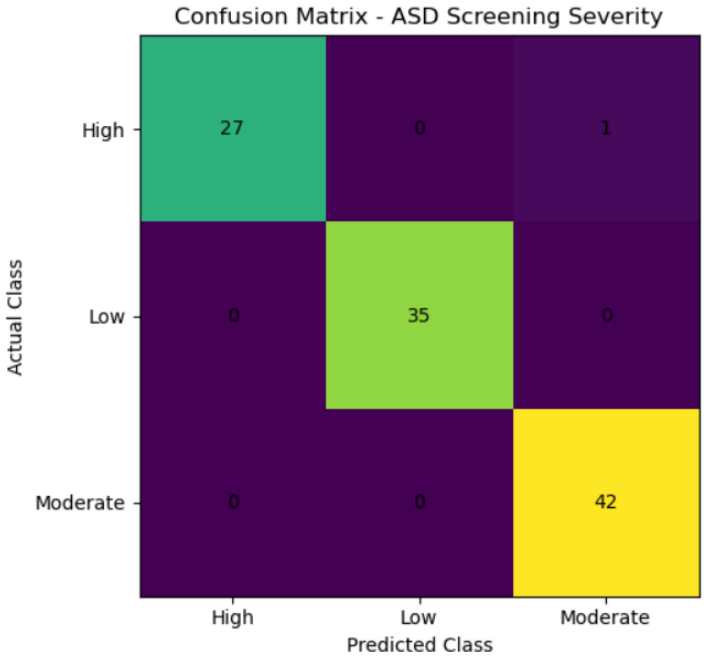


Figure 1: Confusion matrix of the proposed multimodal model for screening-based severity prediction

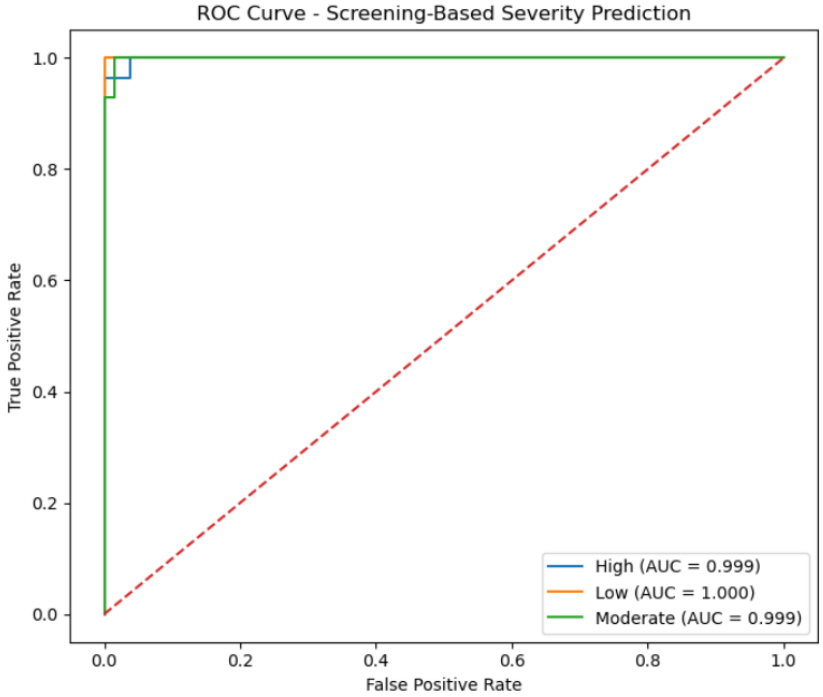


Figure 2: ROC curve of the proposed multimodal model for screening-based severity prediction

4.3 Behavioral Profile Prediction

The model outputs three behavior groups from the screening data. These groups are Social/Communication, Behavioral /Repetitive, and Social/Adaptive. For Social/Communication, the accuracy is

93.33%. The F1 score is 91.57%. For Behavioral/Repetitive, the accuracy is 86.67%. The F1 score is 79.41%. For Social/Adaptive, the accuracy is 93.33%. The F1 score comes to 93.20%. Across all three behavioral outputs, the model achieved a micro-F1 of 88.98% and a macro-F1 of 88.06%.

Table 2: Behavioral Profile Prediction Performance

Behavioral Profile	Accuracy	Precision	Recall	F1-score
Social/Communication	93.33%	90.48%	92.68%	91.57%
Behavioral/Repetitive	86.67%	79.41%	79.41%	79.41%
Social/Adaptive	93.33%	94.12%	92.31%	93.20%
Overall Micro-F1	—	88.98%	88.98%	88.98%
Overall Macro-F1	—	—	—	88.06%

The results indicate that the model provides relatively consistent predictions across the three screening-derived behavioral dimensions, although the Behavioral/Repetitive profile shows lower performance than the other two profiles.

4.4 Comparision with Baseline Models

To examine the contribution of different input modalities, the proposed framework was compared with a clinical-only MLP and a behavioral-only BiLSTM. The clinical-only model achieved an accuracy of 37.14% and a macro-F1 of 37.62%. In contrast, the behavioral-only BiLSTM achieved 100% accuracy and 100% macro-F1 for the screening-based severity task. The proposed multimodal model achieved 99.05% accuracy and 99.00% macro-F1.

Table 3: Comparision of severity prediction models

Model	Accuracy	Macro Precision	Macro Recall	Macro-F1
Clinical-only MLP	37.14%	37.20%	38.41%	37.62%
Behavioral-only BiLSTM	100.00%	100.00%	100.00%	100.00%
Proposed Multimodal MLP-BiLSTM	99.05%	99.22%	98.81%	99.00%

The comparison demonstrates that behavioral screening responses contain substantially more information for the constructed screening-based severity target than the available clinical variables. The multimodal model also achieves near-perfect performance while incorporating both information sources.

However, the 100% performance of the behavioral-only model must be interpreted carefully. The severity categories were constructed from the A1–A10 screening score, which is also the input to the behavioral model. Therefore, the behavioral-only result is expected to be very high and should not be interpreted as evidence of clinically validated severity classification.

4.5 Explainability Analysis

SHAP analysis was performed to investigate the contribution of clinical features to the high screening-severity prediction. Among the encoded clinical variables, the highest

mean absolute SHAP values were observed for ethnicity_White-European, country-of-residence indicators such as United Arab Emirates and United States, Jordan, ethnicity_Middle Eastern, and gender-related features.

Table 4: Top Clinical Features identified by SHAP

Rank	Clinical Feature	Mean Absolute SHAP
1	ethnicity_White-European	0.0422
2	contry_of_res_United Arab Emirates	0.0287
3	contry_of_res_United States	0.0282
4	contry_of_res_Jordan	0.0267
5	ethnicity_Middle Eastern	0.024
6	gender_f	0.0213
7	gender_m	0.0208
8	contry_of_res_United Kingdom	0.0172
9	austim_yes	0.0164
10	jundice_no	0.0129

The SHAP analysis provides an interpretable ranking of the clinical variables used by the model. However, these values represent model associations rather than causal clinical relationships. In particular, demographic variables such as ethnicity and country of residence should not be interpreted as biological determinants of autism severity because they may reflect sampling characteristics and demographic composition within the dataset.

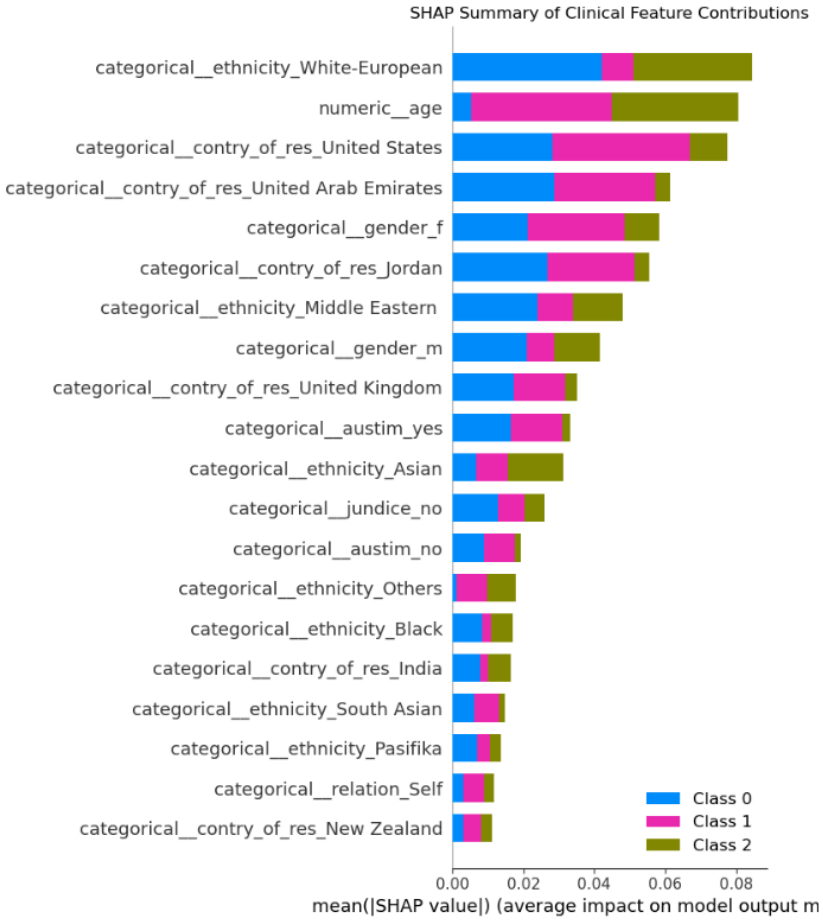


Figure 3: SHAP summary plot for clinical feature contributions

For the behavioral modality, permutation analysis identified A10, A4, A2, A3, and A5 as the five most influential screening items based on the reduction in prediction accuracy after feature permutation.

Table 5: Behavioral feature importance

Rank	Behavioral Feature	Accuracy Reduction
1	A10_Score	0.2286
2	A4_Score	0.219
3	A2_Score	0.1905
4	A3_Score	0.1714
5	A5_Score	0.1619
6	A1_Score	0.1524
7	A8_Score	0.1524
8	A6_Score	0.1143
9	A9_Score	0.1143
10	A7_Score	0.0952

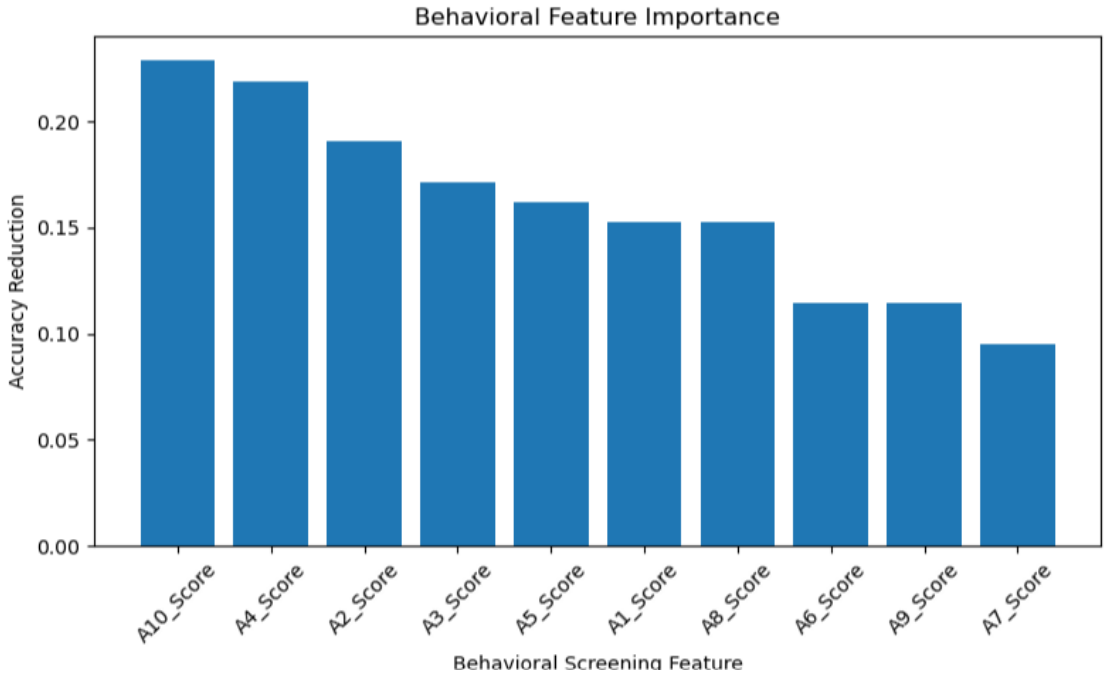


Figure 4: Permutation-based importance of behavioral screening features

4.6 Training Performance

The training and validation curves were examined to assess the learning behavior of the proposed model. The training process was controlled using early stopping to prevent unnecessary training after validation performance stopped improving. The resulting curves provide evidence of the model's convergence during training.

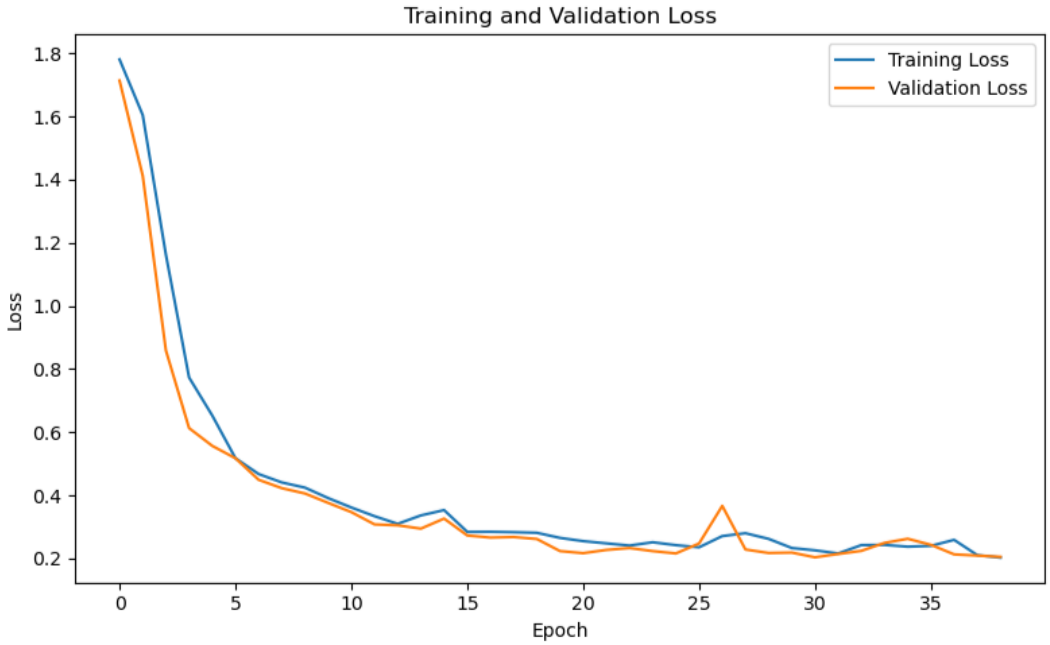


Figure 5: Training and validation loss of the proposed multimodal model

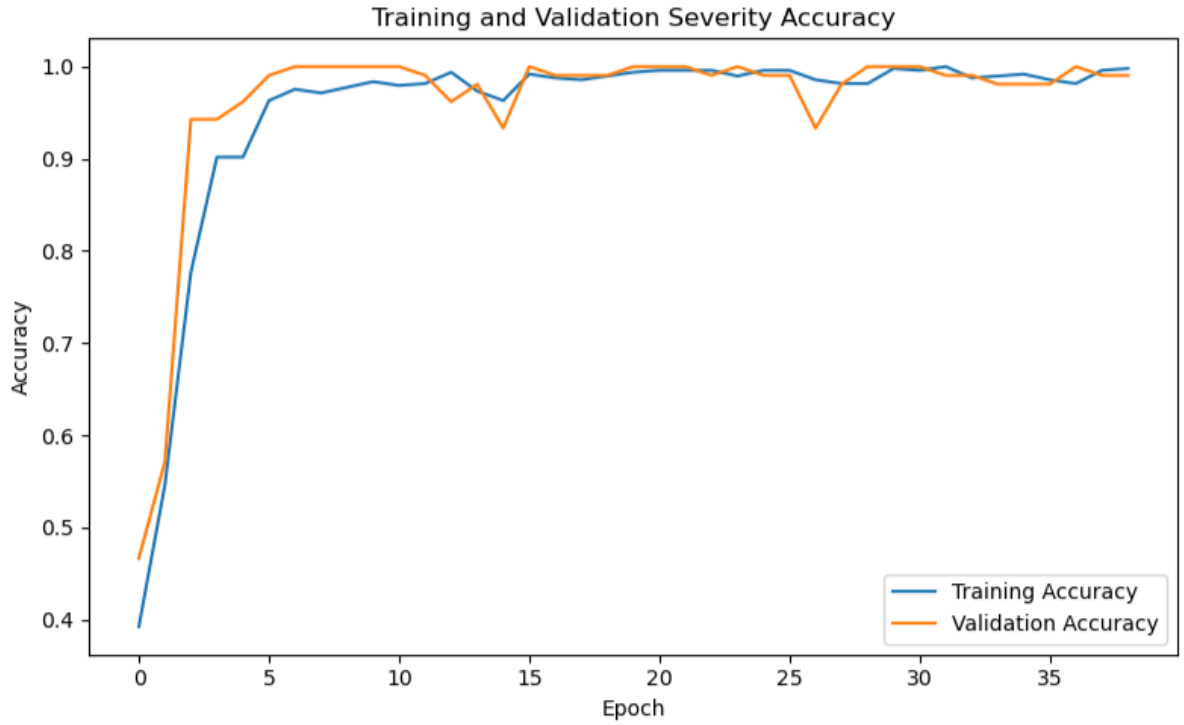


Figure 6: Training and validation severity accuracy of the proposed multimodal model

5. Discussion

The experimental results demonstrate the feasibility of combining clinical and behavioral information within a lightweight multimodal deep learning framework for autism-related screening analysis. The proposed architecture achieved a severity prediction accuracy of 99.05%, with a macro F1-score of 99.00% and a macro ROC-AUC of 0.9992. These results indicate that the fused representation was able to effectively learn patterns associated with the screening-derived severity categories. The behavioral profile prediction task also produced encouraging results, with a micro F1-score of 88.98% and a macro F1-score of 88.06% across the three derived behavioral profiles. The difference between the performance of the individual behavioral profiles suggests that some behavioral dimensions are easier to distinguish from the available screening responses than others.

The comparison with the baseline models provides an important interpretation of the multimodal design. The clinical-only MLP obtained a considerably lower macro F1-score of 37.62%, whereas the behavioral-only BiLSTM achieved perfect performance for the screening-derived severity target. However, the latter result should not be interpreted as evidence of clinically perfect autism severity prediction. The severity categories were constructed directly from the A1–A10 screening responses, which are also the inputs to the behavioral branch. Consequently, the behavioral-only model has direct access to the information from which the target was derived. The result therefore primarily demonstrates that the BiLSTM can reconstruct the predefined screening-score categories rather than independently predict clinically validated autism severity. Similarly, the lower clinical-only performance is expected

because the clinical variables do not directly contain the screening-score information used to construct the severity target.

The behavioral profile results provide a more informative assessment of the model's ability to distinguish different groups of screening responses. The Social/Adaptive profile achieved the highest F1-score of 93.20%, followed by the Social/Communication profile at 91.57%. The Behavioral/Repetitive profile produced a lower F1-score of 79.41%, indicating greater difficulty in identifying this profile from the available features. This variation may be related to differences in the distribution and discriminative characteristics of the individual A1–A10 screening responses. Therefore, future studies using independently annotated behavioral phenotypes would be valuable for determining whether the observed differences remain when the targets are not directly derived from the input variables.

The explainability analysis further supports the interpretability objective of the framework. SHAP analysis identified several clinical variables with relatively high contributions to the model's high-severity predictions, while permutation analysis indicated that A10, A4, A2, and A3 were among the most influential behavioral screening features. These findings provide a feature-level view of how the model uses the available information. However, the clinical SHAP results should be interpreted cautiously. In particular, demographic variables such as ethnicity and country of residence represent associations within this dataset and should not be interpreted as causal or biological determinants of autism severity. Their importance may instead reflect sampling characteristics, dataset composition, or correlations with other variables.

Overall, the study demonstrates that a relatively simple combination of an MLP and BiLSTM can integrate heterogeneous clinical and behavioral information without requiring computationally intensive architectures. The use of separate modality-specific representations followed by feature fusion also provides a straightforward structure for incorporating additional modalities in future work. Nevertheless, the present study should be considered a proof-of-concept based on screening-derived targets rather than a clinically validated autism severity prediction system. Independent clinical severity annotations, larger and more diverse cohorts, and external validation are required before such a framework could be considered for real-world clinical decision support.

6. Conclusion

This study presented a lightweight multimodal explainable deep learning framework for autism screening-based severity and behavioral profile prediction using clinical and behavioral information. The framework combines a clinical multilayer perceptron, a behavioral bidirectional long short-term memory network, and a feature-fusion mechanism with separate prediction heads. On the cleaned UCI Adult Autism Screening dataset containing 699 instances, the proposed framework achieved 99.05% accuracy and a macro F1-score of 99.00% for the screening-derived severity categories. For the three screening-derived

behavioral profiles, the framework achieved a micro F1-score of 88.98% and a macro F1-score of 88.06%. SHAP and permutation-based analyses were additionally incorporated to provide feature-level explanations of model behavior.

Despite these promising results, the findings should be interpreted within the limitations of the dataset and target construction. The severity categories and behavioral profiles were derived from the A1–A10 screening responses rather than obtained from independent clinical assessments. Therefore, the results demonstrate the ability of the proposed framework to model and reconstruct screening-derived patterns rather than establish clinically validated autism severity prediction. Future work should focus on larger and more diverse datasets containing independently assessed severity levels and clinically annotated behavioral phenotypes. External validation, longitudinal data, additional behavioral and clinical modalities, and prospective clinical evaluation could further determine the practical value of multimodal explainable learning for autism assessment and decision-support applications.

7. Limitations and Future Work

7.1 Limitations

The primary limitation of this study is the nature of the available dataset and target variables. The UCI Adult Autism Screening dataset was originally developed for autism screening and does not provide independently clinically validated autism severity labels or detailed behavioral phenotype annotations. Therefore, the severity categories and behavioral profiles used in this study were derived from the screening responses. Although this provides a practical basis for developing and evaluating the proposed framework, these derived targets should not be considered equivalent to clinical severity assessments.

Another limitation is the relatively small dataset size of 699 samples after preprocessing. The dataset also contains missing values and categorical variables with uneven representation across different groups. These characteristics can affect model generalization and explainability results. In addition, the behavioral-only model achieved perfect performance for the severity task because the target was directly derived from the same A1–A10 responses provided to the behavioral branch. This dependency limits the usefulness of that result as evidence of independent predictive performance. Finally, the study uses a single dataset without external validation, meaning that the reported performance cannot yet establish robustness across different populations, healthcare settings, or demographic distributions.

7.2 Future Work

Future research should evaluate the proposed framework using larger, multi-center datasets containing independently assessed autism severity levels and clinically validated behavioral profiles. External validation on independent cohorts would provide a stronger assessment of generalization. Additional modalities, such as longitudinal behavioral observations, speech characteristics, eye-tracking information, physiological signals, and electronic health record information, could also be incorporated to provide a more

comprehensive representation of autism-related characteristics.

The explainability component can also be extended through multimodal explanation techniques that jointly analyze clinical and behavioral contributions. Future studies could investigate uncertainty estimation, fairness assessment, calibration, and subgroup performance to improve the reliability of the framework. With appropriate clinical validation and prospective evaluation, the proposed lightweight architecture could serve as a foundation for developing interpretable screening and decision-support systems rather than being used as a standalone diagnostic tool.

References

[1] H. Abbas, F. Garberson, E. Glover, and D. P. Wall, "Machine learning approach for early detection of autism by combining questionnaire and home video screening," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 1000–1007, May 2018, doi: 10.1093/jamia/ocy039.

[2] F. Thabtah *et al.*, "Autism screening: an unsupervised machine learning approach," *Health Information Science and Systems*, vol. 10, no. 1, Sept. 2022, doi: 10.1007/s13755-022-00191-x.

[3] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, vol. 13, no. 1, p. 9605, June 2023, doi: 10.1038/s41598-023-35910-1.

[4] U. Jabbar *et al.*, "Deep learning based approach for Behavior classification in diagnoses of Autism Spectrum Disorder using naturalistic videos," *Frontiers in Computational Neuroscience*, vol. 20, Mar. 2026, doi: 10.3389/fncom.2026.1626315.

[5] Tal Barami, Liora Manelis-Baram, H. Kaiser, M. Ilan, and Ilan Dinstein, "Automated Analysis of Stereotypical Movements in Videos of Children With Autism Spectrum Disorder," *JAMA Network Open*, vol. 7, no. 9, p. e2432851, Sept. 2024, doi: 10.1001/jamanetworkopen.2024.32851.

[6] N. Alsharif, M. H. Al-Adhaileh, M. Al-Yaari, N. Farhah, and Z. I. Khan, "Utilizing deep learning models in an intelligent eye-tracking system for autism spectrum disorder diagnosis," *Frontiers in Medicine*, vol. 11, July 2024, doi: 10.3389/fmed.2024.1436646.

[7] Tamilvizhi Thanarajan, Y. Alotaibi, S. Rajendran, and Krishnaraj Nagappan, "Eye-Tracking Based Autism Spectrum Disorder Diagnosis Using Chaotic Butterfly Optimization with Deep Learning Model," *Computers, materials & continua/Computers, materials & continua (Print)*, vol. 76, no. 2, pp. 1995–2013, Jan. 2023, doi: 10.32604/cmc.2023.039644.

[8] C. C. Yang, "Explainable Artificial Intelligence for Predictive Modeling in Healthcare," *Journal of Healthcare Informatics*

Research, vol. 6, no. 2, Feb. 2022, doi: 10.1007/s41666-022-00114-1.

[9] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," *Sensors*, vol. 23, no. 2, p. 634, Jan. 2023, doi: 10.3390/s23020634.

[10] M. Shadan, A. Sai Chaithanya, C. Vaishali, S. Mohammed Musharraf, and M. Gulam Muqeeth, "Hybrid BiLSTM-CNN Model with Attention and XAI (SHAP & LIME) for ECG Arrhythmia Classification Using PTB-XL," *International Journal of Science and Research (IJSR)*, pp. 2242–2249, Dec. 2025, doi: 10.21275/sr251226130813.

[11] E. Kyrimi *et al.*, "Explainable AI: definition and attributes of a good explanation for health AI," *AI and Ethics*, vol. 5, no. 4, pp. 3883–3896, Mar. 2025, doi: 10.1007/s43681-025-00668-x.