

Feature-engineered stacking and AutoGluon framework for deep learning-based autism detection

Savita Rani¹, Chandra Shakher Tyagi¹, Devendra Kumar²

¹Department of Computer Science and Engineering, SRM Institute of Science and Technology Delhi NCR, Modinagar, India

²Department of Computer Applications, ABES Engineering College, Ghaziabad, India

Article Info

Article history:

Received Sep 8, 2025

Revised Jul 18, 2026

Accepted Aug 12, 2026

Keywords:

Autism spectrum disorder

AutoGluon

Ensemble model

Shapley additive explanations

Stacking

ABSTRACT

Autism spectrum disorder (ASD) is a brain development disorder that distracts a person's perceptions and social interactions with others. Current studies emphasize conventional machine learning and typical ensemble methods and little focus have been placed on merging automated machine learning (AutoML) systems with stacking ensemble structures. To bridge this gap, this study evaluates stacking ensembles with an artificial neural network (ANN)-based meta-classifier and the AutoGluon framework. For enhancing model stability and generalizability across various ASD datasets, systematic RandomizedSearchCV along with 10-fold stratified cross-validation is utilized. The stacking model outperformed the AutoGluon framework and all individual classifiers on dataset-I, achieving 99.50% accuracy, 99.53% precision, 99.53% recall, and a 99.97% receiver operating characteristic-area under the curve (ROC-AUC). With 97.28% accuracy, precision 96.33%, 95.38 F1-score, and a 99.76% ROC-AUC on dataset-II, the model demonstrated great generalization and continuously outperformed traditional ensemble and boosting techniques. Stacking model perform well on both datasets and maintain consistence, whereas AutoGluon give competitive performance on dataset-I and it exhibited a notable decline on dataset-II. K-fold internal and external cross-validation were used to access model's generalization and consistency. The outcomes of stacking ensemble model indicate that model offers higher accuracy, greater resilience, better consistency across difference datasets and offers a robust approach to early ASD prediction.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Savita Rani

Department of Computer Science and Engineering, SRM Institute of Science and Technology Delhi NCR

Modinagar, Uttar Pradesh, India

Email: ss0004@srmist.edu.in

1. INTRODUCTION

The neurological disorder known as autism spectrum disorder (ASD) can be defined by recurrent trends of behavior, limited interests, and ongoing difficulties with social communication [1], [2]. The rising prevalence of ASD across the world has increased the need to find the way to have an accurate, scalable, and early diagnostic solution that can provide timely intervention and enhance the long-term developmental outcomes [3]. Early identification of ASD is especially important because early developmental therapies have been demonstrated to greatly improve behavioral, social, and cognitive functioning [4]. Although traditional methods are commonly used for diagnosing ASD, these methods are not suitable for widespread screening and implementation in low-resource environments because they are costly, time-consuming, and highly reliant on clinical competence [4], [5]. Various screening tools including childhood autism rating scale

(CARS), autism spectrum quotient (AQ), and the screening tool for autism in toddlers and young children (STAT) have been used to diagnose ASD. These assessments facilitate early intervention and help detect symptoms while measuring the severity of the condition. This study needs advanced procedures to overcome the deficiencies of conventional procedures, especially those involving artificial intelligence (AI). Recent developments in AI and have shown encouraging results in the automated screening of ASD through the identification of discriminatory trends in behavioral, demographic, and medical data sets [6], [7]. One of essential processes in machine learning are feature engineering and feature selection as removing of non-informative attribute eliminates unnecessary details, reduces dimensionality, and enhances model interpretability so indirectly affects the performance of the model [8], [9]. Feature selection and explainable artificial intelligence (XAI) approaches based deep learning frameworks have identified most significant features for ASD prediction [8]–[10].

Standalone deep learning models have excellent prediction capabilities; however, they are not applicable to small-scale or tabular ASD datasets. Models of this type typically need large amounts of labeled data, great computing power, and heavy hyperparameter adjustment [11]. To overcome these challenges, ensemble learning techniques have gained increasing attention. Single models are usually less robust and less accurate than ensemble methods for the ASD detection problems. Ensemble methods like bagging, boosting, and stacking integrate several classifiers to achieve a more robust and better predictive capability and significantly surpass single models in ASD detection problems [12]. Among ensemble techniques, stacking has appeared as a prevailing strategy due to its ability to optimally combine heterogeneous base learners through a meta-classifier [13]. Latest research suggests that stacking-based and hybrid ensemble frameworks can enhance the accuracy and generalization capability of ASD detection on different datasets [14], [15]. Table 1 describes various studied which are using stacking-based models for ASD detection.

Table 1. Comparative analysis of ASD detection model

Study / year	Dataset type	Method/ model	Feature engineering	AutoML used	Explainability / XAI	Key findings	Limitation/gap
Eslami <i>et al.</i> [1], 2019	fMRI	Hybrid CNN + RNN	None	No	Limited	High accuracy in neuroimaging data Good performance on early screening	Requires large fMRI datasets; limited interpretability
Ayub <i>et al.</i> [2], 2024	Behavioral/ tabular	Classical ML (SVM, RF, and gradient boosting)	None	No	No		No ensemble; lacks interpretability
Mumenin <i>et al.</i> [3], 2025	Behavioral/ tabular	Ensemble learning	Basic feature selection	No	Moderate	Improved accuracy and robustness Superior accuracy vs single classifiers	No AutoML; limited comparative analysis
Nishat <i>et al.</i> [4], 2024	Behavioral/ tabular	Stacking ensemble (tree + meta-classifier)	Principal component analysis (PCA) + SMOTE	No	Moderate		No AutoML; limited dataset generalization
Nafisah <i>et al.</i> [5], 2025	Behavioral/ tabular	CNN + feature selection	Deep feature selection	No	Moderate	Identified key behavioral indicators Improved prediction with reduced dimensionality	Computationally expensive; limited interpretability
Kumar <i>et al.</i> [6], 2023	Behavioral/ tabular	Modified CNN	PCA	No	No	Improved prediction with reduced dimensionality	Deep learning requires tuning; small datasets
Dick <i>et al.</i> [7], 2025	Administrative and birth registry	Transformer-based ensemble	Minimal	No	No	Captures temporal and hierarchical patterns	Complex model; requires large administrative data
Temiz <i>et al.</i> [8], 2025	Behavioral/ tabular	Tree ensembles	PCA + SMOTE	No	Shapley additive explanations (SHAP)	High accuracy; interpretable features Good interpretability; cross-dataset generalization	No AutoML; lacks comparison with AutoML ensembles
Alsubai [9], 2026	Behavioral/ tabular	Deep transfer learning	Feature selection	Partial	SHAP		Computationally intensive; limited stacking comparison
Yin <i>et al.</i> [10], 2023	Behavioral/ tabular	Unsupervised feature learning + Path signature	PCA	No	SHAP	Early feature detection; interpretable	Limited model comparison; no AutoML
Proposed study	Behavioral/ tabular	Hybrid stacking () and AutoML ensembles	PCA + SMOTE	Yes (AutoGluon)	SHAP	High accuracy, robust performance, clinically interpretable	Demonstrated comparative advantage over prior studies

However, it can sometimes be difficult and time-consuming to design the best stacking architectures and optimize various models through manual means. Concurrently, automated machine learning (AutoML) is seen as an efficient framework that can automate the model selection, hyperparameter optimization, and ensemble building processes. In particular, AutoML framework [16] employs multi-layer stacking, bagging, and boosting strategies to automatically generate high-performing ensemble models for tabular data and minimize human engineering but, manual stacking provides greater transparency and control over model selection compared to AutoGluon. Additionally, manual stacking reduces computational overhead and improves reproducibility, making it more suitable for small and tabular ASD datasets. Recent studies have shown how scalable, robust, and auto-tuning capable ensemble models based on AutoML can be beneficial in healthcare and diagnosis of neurodevelopmental disorders [17], [18]. However, there were not many studies that report a comparative analysis between AutoML-generated ensembles and manual stacking framework design for ASD when used in conjunction with feature engineering like PCA [19] and synthetic minority over-sampling technique (SMOTE) [20], [21]. Interpretability is another key requirement for clinical decision support systems. XAI techniques, particularly SHAP [22], have been broadly accepted to interpret ensemble and deep learning models by quantifying feature contributions to predictions. Incorporating explainability in stacking-based frameworks and AutoML to enhance clinicians' trust and improve transparency in ASD. In response to these issues, the study proposes an explainable stacking-based hybrid ASD prediction framework. This framework utilizes AutoGluon-based AutoML ensembles and manually stacks models through light gradient boosting machine (LightGBM) [23], random forest (RF) [24], and logistic regression (LR) [25]. The method proposed is assessed using a behavioral screening dataset obtained from questionnaires and developmental characteristics. Also, this study carries out comparison analysis on AutoML ensemble models' performance with stacking based on PCA and SMOTE under similar experimental settings.

2. PROPOSED MODEL

Figure 1 illustrates the proposed machine learning framework developed to enhance classification performance while ensuring robustness and interpretability. Data preprocessing, dimensionality reduction, class balancing, ensemble learning and model explainability are all integrated into the framework's complete workflow. The process begins with the raw dataset, which may have different feature types, missing values, and unnecessary identifiers and to deal with these missing values, appropriate imputation techniques are applied. This stage enhances the quality of the data and guarantees consistency among features. After that categorical attribute are encoded into numeric values using some encoding techniques, enabling compatibility with machine learning algorithms while conserving feature information. To ensure uniform feature contribution, feature scaling is applied prior to dimensionality reduction using PCA. To overcome the issue of class imbalance, the SMOTE is being used. By creating artificial samples for the minority class, SMOTE helps learning models capture decision boundaries more accurately, increasing recall and F1-score. Subsequently, SHAP-based XAI is used for feature selection using LightGBM model. It chooses the most discriminative features of both datasets. After that hyperparameter optimization was performed using RandomizedSearchCV combined with 10-fold stratified cross-validation. This strategy enabled efficient exploration of the hyperparameter space while providing robust and unbiased performance estimates.

The cleaned set is then used for two parallel learning strategies using an AutoGluon-based AutoML (WeightedEnsemble_L2) technique as well as a stacking model. AutoGluon, an AutoML framework [26], was utilized to create predictive model for ASD detection. It applies techniques of weighted ensembling like WeightedEnsemble_L2 to create a strong final prediction from models like gradient boosted trees, LightGBM, and RF. The strength of AutoGluon is its capacity to find many great models with minimal effort required on the user's part while also achieving strong performance across variety of datasets. The model's performance is listed using standard evaluation measures, and its robustness and resilience are confirmed using cross-validation accuracy and overfitting analysis. Compared to existing studies that rely on single classifiers or limited preprocessing, the approach integrates dimensionality reduction, class balancing, and advanced ensemble learning into an integrated pipeline. Additionally, the black-box constraint noted in earlier research is addressed by the incorporation of SHAP-based explainability, which makes the suggested method more appropriate for practical medical applications. Proposed research also performs a systematic comparative analysis between an AutoML-based ensemble and a manually constructed stacking ensemble. This comparison demonstrates the trade-offs between expert-driven ensemble creation and automated model optimization.

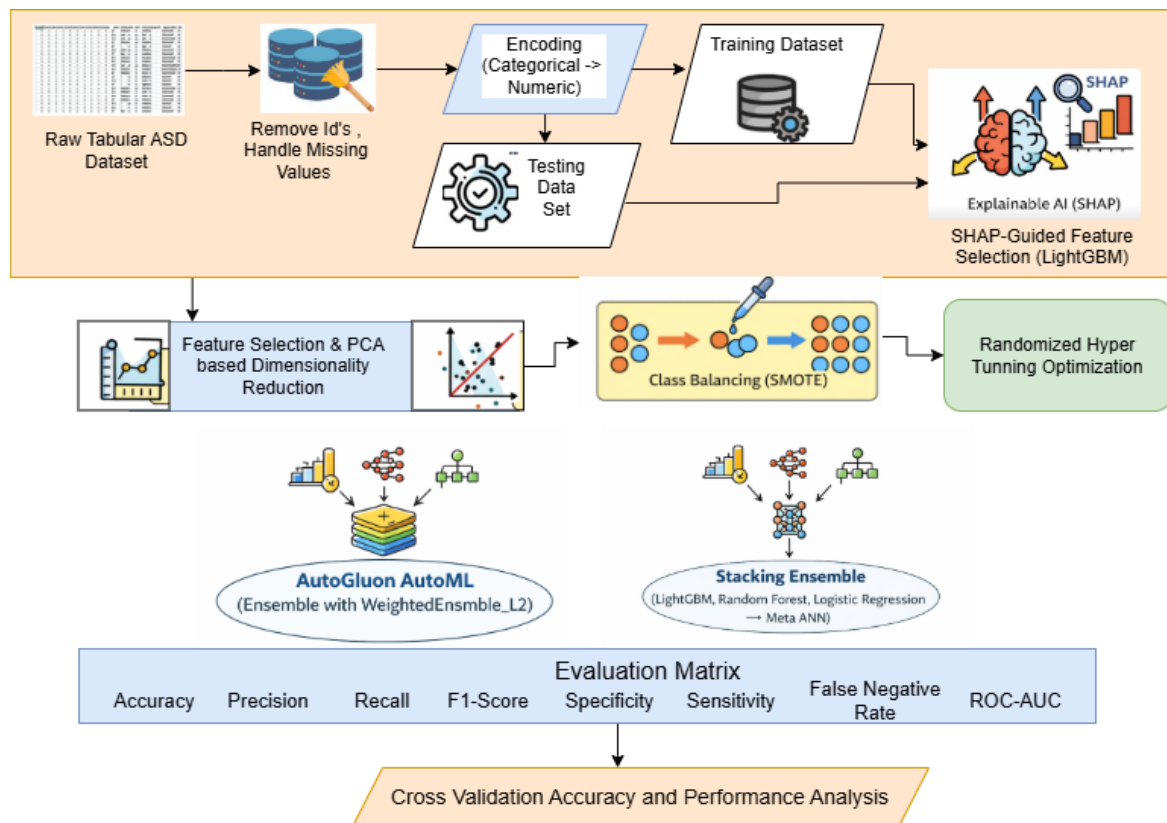


Figure 1. Proposed framework for ASD detection using stacking ensemble model and AutoGluon

3. METHOD

3.1. Dataset description

To ensure robust performance evaluation, two separate datasets are employed for assessing the proposed model. Dataset-I is from the University of Arkansas's Department of Computer Science's Autism Research at <https://www.kaggle.com/datasets/uppurimadhuri/dataset> having 1985 record and 28 features. Another dataset was developed by Dr. Fadi Fayeze using a mobile app called ASDTests (ASDtests.com) to screen autism in toddlers and is available on Kaggle at <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>. The dataset contains 6075 record and 15 features. Both data set includes several factors that are associated with children's ASD and final column indicates whether the individual would eventually exhibit ASD symptoms. For external cross validation 3 more data set including custom ASD children traits dataset: i) children at <https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children>, ii) adolescent at <https://archive.ics.uci.edu/dataset/420/autistic+spectrum+disorder+screening+data+for+adolescent>, and iii) adult at <https://archive.ics.uci.edu/dataset/426/autism+screening+adult>, were also used.

3.2. Data preprocessing

This study utilized two independent datasets that support the generalizability and resilience of the proposed ASD detection framework. The first dataset contains attributes related to ASD traits whose target variable is → "ASD_traits" and the second dataset contain a combined autism screening dataset with class label → "Class". Both datasets underwent the same preprocessing pipeline. This enables fair comparison and reproducibility of this work. Initially, the raw datasets were loaded and separated into feature matrices and their corresponding class labels. This study retained the non-target attributes as predictive features; the non-target labels were isolated for supervision. Before preprocessing, the dataset content was checked by studying the total number of features and instances. With both data sets made up of categorical and numerical attributes, the categorical features were turned into numerical codes using label encoding. This conversion allowed for the effective application of machine learning techniques while retaining their categorical nature. Non-numeric class labels were encoded in a similar manner. In order not lose any samples, mean imputation was done on all numerical features using a simple imputer. After imputation and encoding, the datasets were split into training and test subsets in an 80:20 stratified manner. In both subsets, stratification was applied to

ensure that the original class distribution is preserved, which is important in the case of ASD detection, as class imbalance is a common occurrence in such tasks. Standardization was then used in feature scaling to ensure that all feature distribution were treated equally in model training.

3.3. Dataset bias analysis

The bias analysis of dataset defined in Table 2 revealed noticeable variations across the all ASD datasets in terms of class distribution, demographic composition, and dataset size. Since datasets were questionnaire-based, a demographic bias assessment was conducted to evaluate the representativeness of the study population. Dataset-II contained the largest number of instances (6,075) and exhibited moderate class imbalance, whereas dataset-I showed the most balanced ASD class distribution and the highest entropy score, indicating greater class diversity. The children dataset showed a greater percentage of male participants, indicating possible demographic bias. Gender representation differed significantly amongst datasets. To address class imbalance, stratified sampling and SMOTE [20], [21] were used, and PCA decreased noise and redundant features to enhance model generalization. SHAP [22] was used to identify key features influencing ASD predictions, ensuring transparency, interpretability, and clinically meaningful decision-making. Together, these techniques improve the robustness, fairness, and interpretability of the proposed ASD detection framework.

Table 2. Dataset biasness analysis

Dataset	Instances	Attributes	ASD+ %	ASD- %	Male %	Female %	Mean age	Missing %	Imbalance ratio	Entropy
Dataset-II	6075	14	29.70	70.30	57.74	42.26	19.84	0.00	2.37	0.8775
Dataset-I	1985	27	54.11	45.89	72.90	27.10	9.62	0.11	1.18	0.9951
Adult	704	20	26.85	73.15	52.13	47.87	29.70	0.01	2.72	0.8393
Children	292	20	48.29	51.71	71.23	28.77	6.35	0.07	1.07	0.9992
Adolescent	104	20	60.58	39.42	48.08	51.92	14.13	0.00	1.54	0.9675

3.4. Feature selection

A cutting-edge technique in the field of XAI, SHAP [27] offers insightful information on the output of ML models. Although SHAP mainly concentrates on local explanations, one can gain global insights into the model by combining various explanations. This makes it easier to understand how the model behaves generally, including the characteristics that have the biggest impact and how they interact. To ascertain the influence of input features on the corresponding predictions, SHAP was applied to LightGBM, one of the base classifiers. In data.csv of 28 columns 15 features that exert the greatest influence were selected for developing our proposed model. The choice of features has an effect on how well different machine learning models perform when trained on datasets. Figure 2 represent the explanations generated using SHAP on dataset-I. The SHAP feature-importance results show that the model's predictions are mainly determined by a small set of key variables. In the first analysis, Qchat_10_Score is the most significant feature, followed by ethnicity and sex, while individual assessment items (A1–A10) contribute at a moderate level. Factors related to family history, developmental conditions, and genetic disorders have relatively low impact. Figure 3 represent the explanations generated using SHAP on dataset-II. In this dataset top 12 features exert the greatest influence were selected for developing our proposed model. In the second analysis, age is the most important predictor, while a few assessment attributes (A3-A10) have significant influence, and sex, family ASD history, and jaundice only have subtle effect. The results show that overall, the assessment scores and the age contribute primarily to the prediction process, and the demographic and clinical variables have a comparatively smaller prediction influence.

3.5. Hyperparameter optimization

Machine learning algorithms are controlled by configuration variables known as hyperparameters. When there are many hyperparameters, manual hyperparameter search becomes unfeasible and frequently takes a long time. Therefore, to improve predictive performance and model robustness, RandomizedSearchCV integrated with 10-fold stratified cross-validation was employed for hyperparameter optimization [28]. In contrast to exhaustive grid search, randomized search is a hyperparameter tuning strategy that investigates random combinations of hyperparameters within predetermined distributions or ranges with less processing complexity. Moreover, the proposed framework is also made more practical by using the technique of randomized hyperparameter optimization and stratified validation, which allows for a compromise among predictive accuracy and computational efficiency. The optimized hyperparameters

obtained through this procedure were subsequently used to construct the proposed stacking ensemble, contributing to its strong generalization performance across unseen test data.

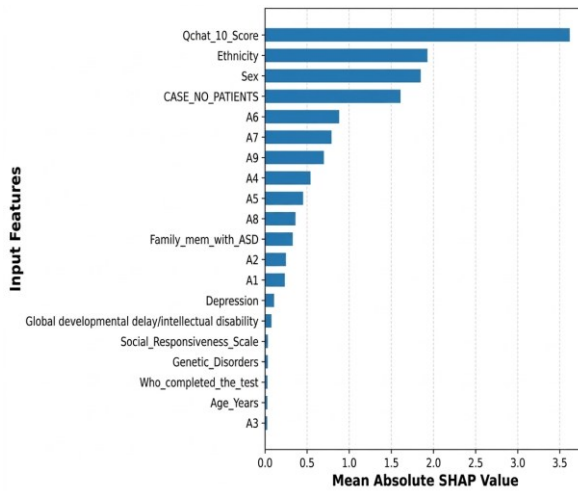


Figure 2. Feature importance in dataset-I

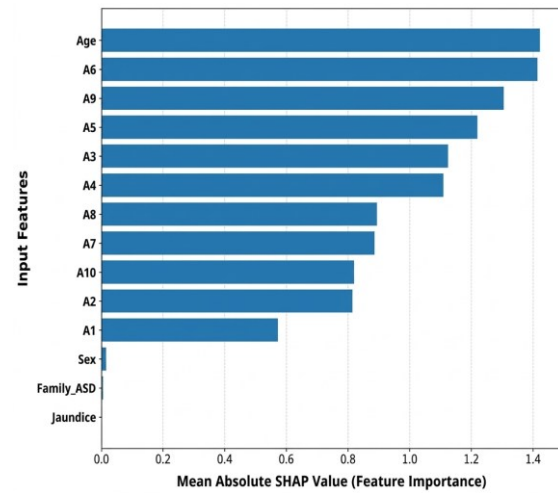


Figure 3. Feature importance in dataset-II

3.6. Model development

In this study, an artificial neural network (ANN)-based stacking ensemble framework was employed, integrating the predictions of three optimized base classifiers. These classifiers produce outputs after analyzing the input data, which are then used to create meta-features for the meta-classifier. The main objective is to expand the performance prediction of the model by utilizing the different strengths of different classifiers. The stacking ensemble model depicted in Figure 4, is a two-tier stacking ensemble in which 3 base classifiers LighGBM, RF and LR are trained on the original dataset to generate predictive probabilities for each input instance.

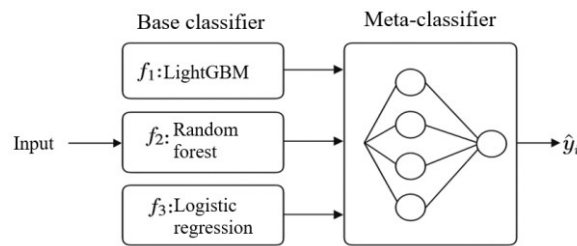


Figure 4. Stacking ensemble model

Consider the training dataset be represented as $X = [x_1, x_2, \dots, x_n]$ with labels $Y = [y_1, y_2, \dots, y_n]$, where each x_i is a feature vector and $y_i \in \{0, 1\}$ is the corresponding class label. For a given base classifier C_j , the output probability for sample x_i as in (1).

$$P_{ij} = P(y_i = 1 | x_i; \theta_j) \quad (1)$$

Where θ_j are the learned parameters of classifier C_j . The probability estimates from all m base classifiers form a meta-feature vector, as in (2).

$$x_i^{meta} = [P_{i1}, P_{i2}, \dots, P_{im}] \quad (2)$$

An ANN [29] acts as the meta-classifier, learning a mapping: $f_{meta}: X_{meta} \rightarrow Y$. The ANN architecture consists of two hidden layers (64 and 64 neurons, rectified linear unit (ReLU) activation), an input layer

(size m), and a sigmoid output layer. The final prediction for a sample x_i is: $\hat{y}_{meta} = C_{meta}(x_i^{meta})$ or, as a probability estimate, as in (3).

$$\hat{p}_{meta} = P(y_i = 1 | x_i^{meta}; \theta_{meta}) \quad (3)$$

By synthesizing predictions from multiple heterogeneous models and leveraging the non-linear learning capacity of the ANN, the proposed ensemble aims to improve accuracy, robustness, and generalization. Adam optimizer was used with its binary cross entropy as the loss function, a batch size of 20, default learning rate of 0.001, and trained for 50 epochs. AutoGluon, which is an AutoML framework was also used to create an ASD predictive model. The framework preprocesses the dataset for missing values and encoding of categorical variables, optimizes hyperparameters, and builds ensembles of diverse base learners. Its weighted ensembling strategies like WeightedEnsemble_L2 are used to stack gradient boosted trees, LightGBM, and RF to form a powerful predictor. The benefit of AutoGluon is that it can find good models with little human intervention and performs well on various datasets. Weighted ensemble prediction can be expressed as in (4).

$$\hat{y} = \sum_{i=1}^M w_i \cdot f_i(x) \quad (4)$$

Where M indicates the total count of base learners. The i -th model's prediction is represented by $f_i(x)$. The weight for the i -th model is w_i and thus, $\sum_{i=1}^M w_i = 1$. AutoGluon optimizes weights for the models to minimize the validation loss, ensuring a greater contribution from stronger models.

4. RESULTS AND DISCUSSION

4.1. Classification performance

The proposed methodology was assessed using two standard datasets to assess its effectiveness and robustness. A number of evaluation metrics have been employed to assess the efficacy of the proposed model i.e., accuracy, precision, recall, sensitivity, specificity, F1-score, and area under the curve (AUC)-score. According to the experimental results, the suggested model successfully balances sensitivity and specificity, guaranteeing accurate ASD identification while lowering false-positive and false-negative predictions.

Tables 3 and 4 show comparative performance of stacking + ANN meta model with various machine learning classifiers for dataset-I and dataset-II, as illustrated in Figure 5. The outcome shows that the proposed stacking + ANN meta classifier has the best performance for both datasets. For dataset-I, it achieved a high accuracy of 99.50%, a high sensitivity of 99.53%, a high specificity of 99.45%, and a receiver operating characteristic-AUC of 99.95%, which signifies its high accuracy in classification. The model showed better performance than all machine learning models with 97.20% accuracy, 95.84% sensitivity, 97.78% specificity, and 99.74% ROC-AUC on dataset-II.

Table 3. Performance comparison of machine learning models on dataset-I

Model	Accuracy	Precision	Recall	Sensitivity	Specificity	False negative rate	F1-score	AUC
Stacking + ANN meta (proposed model)	99.50	99.53	99.53	99.53	99.45	0.47	99.53	99.95
LightGBM	99.24	99.07	99.53	99.53	98.90	0.47	99.30	99.98
XGBoost	99.24	99.53	99.07	99.07	99.45	0.93	99.30	99.97
CatBoost	99.24	99.53	99.07	99.07	99.45	0.93	99.30	99.98
RF	98.74	99.07	98.60	98.60	98.90	1.40	98.83	99.97
LR	96.98	97.65	96.74	96.74	97.25	3.26	97.20	98.13

Table 4. Performance comparison of machine learning models on dataset-II

Model	Accuracy	Precision	Recall	Sensitivity	Specificity	False negative rate	F1-score	AUC
Stacking + ANN meta (proposed model)	97.20	94.79	95.84	95.84	97.78	4.16	95.32	99.74
LightGBM	96.95	93.09	96.95	96.95	96.96	3.05	94.98	99.67
CatBoost	96.79	92.59	96.95	96.95	96.72	3.05	94.72	99.48
XGBoost	96.54	91.64	97.23	97.23	96.25	2.77	94.35	99.58
RF	96.21	91.12	96.68	96.68	96.02	3.32	93.82	99.35
LR	90.37	79.19	91.69	91.69	89.81	8.31	84.98	95.61

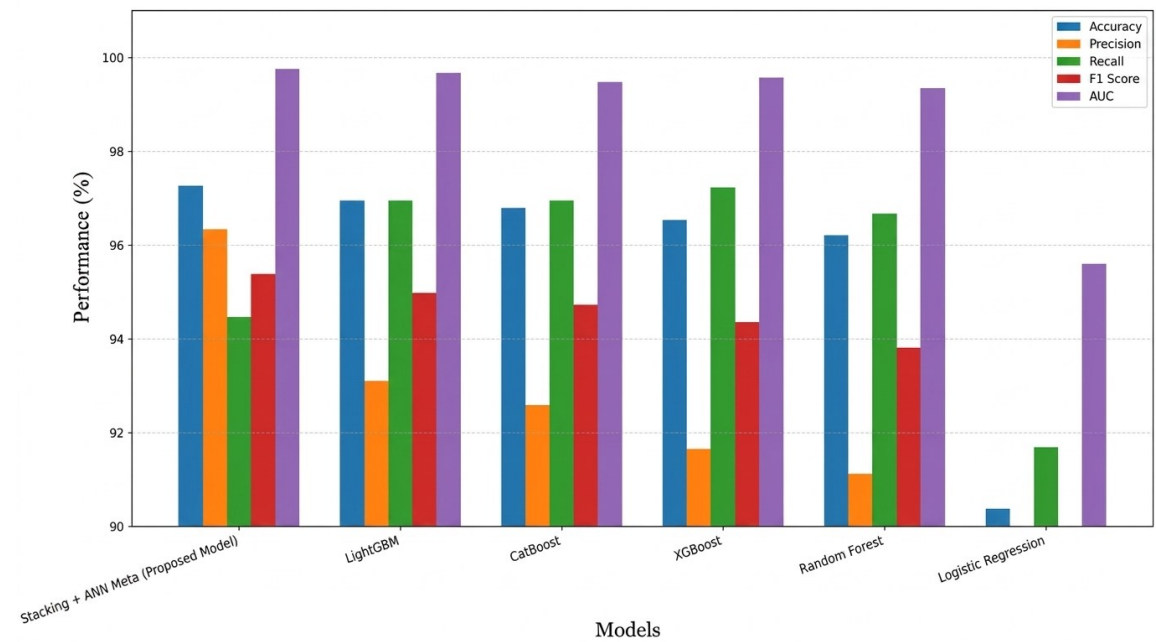


Figure 5. Comparison of machine learning models for ASD detection

Figure 6 shows a comparison of the ROC curves, confirming the discriminative ability of the proposed model across both datasets. Moreover, the proposed framework was consistently better than AutoGluon [17] on both datasets as summarized in Table 5. Figure 7 further depicts this comparison, demonstrating the strength, reliability, and effectiveness of the framework for ASD prediction.

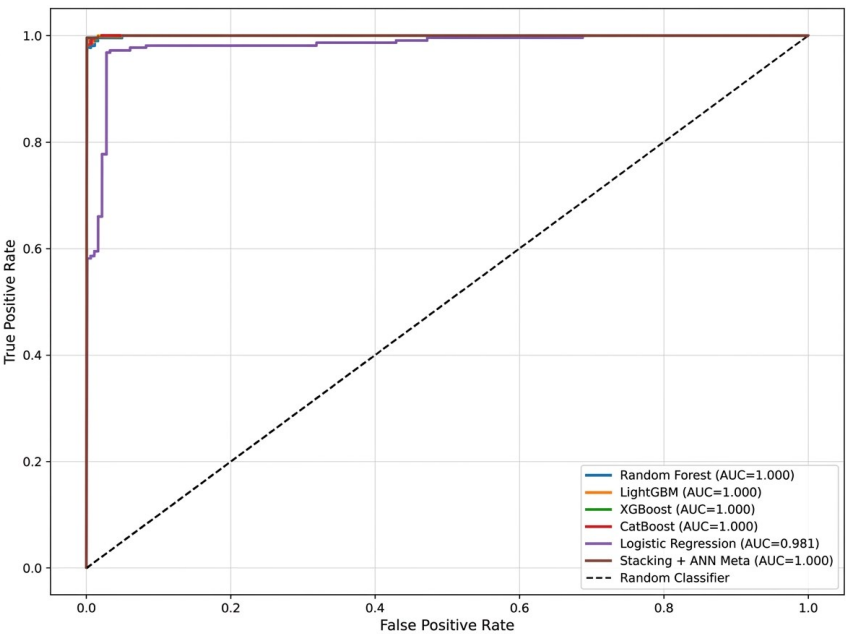


Figure 6. ROC curve comparison of ASD detection models

Table 5. Performance comparison of machine learning models on dataset-I and dataset-II

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Dataset-I	Stacking + ANN meta	99.50	99.53	99.53	99.53	99.95
	AutoGluon	98.74	96.80	98.60	97.70	99.72
Dataset-II	Stacking + ANN meta	97.20	94.79	95.84	95.32	99.74
	AutoGluon	82.63	68.38	77.29	72.56	89.46

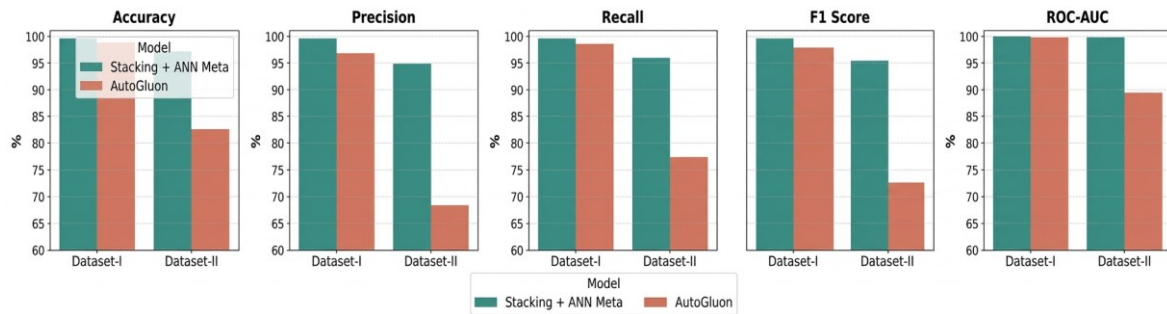


Figure 7. Performance comparison of proposed stacking + ANN meta model and AutoGluon

In ASD screening, sensitivity and false negative rates are especially crucial in because the clinical effects of misclassification go beyond traditional performance indicators. A false negative prediction may delay diagnosis and early intervention, potentially affecting developmental and behavioral outcomes, while a false positive prediction can lead to unnecessary clinical evaluations, increased healthcare costs, and psychological burden on families. The proposed stacking + ANN meta model is capable of attaining high sensitivity values of 99.53% for dataset-I and 95.84% for dataset-II, which shows that it has high capability to correctly identify the ASD cases. Additionally, the model had low false negative rates of 0.47% and 4.16% on dataset-I and dataset-II, respectively, reducing the risk of missed ASD diagnoses.

The overall accuracy (97.20%) and AUC (99.74%) of the proposed stacking ensemble was the highest, but the false negative rate (4.16%) was slightly higher than that of extreme gradient boosting (XGBoost) (2.77%) and LightGBM/categorical boosting (CatBoost) (3.05%). This means that there is a slight compromise in the overall classification accuracy and the number of missed cases of ASD. However, the model achieved a good sensitivity of 95.84% which means that it identified most of the cases of ASD, with better overall predictive performance. Combined with its high accuracy and ROC-AUC scores, these results demonstrate the clinical utility of the framework as a dependable decision-support tool for early ASD screening, where maximizing case detection while reducing diagnostic errors is of paramount importance.

Table 6 compares recent ASD detection models developed using tabular datasets and ensemble or stacking-based learning techniques. Most existing studies report accuracy values between 86% and 99%; however their diagnostic reliability is limited by the absence of comprehensive evaluation criteria including precision, recall, F1-score, and AUC. While some ensemble systems achieve great accuracy, they often rely on traditional voting or homogeneous learners and do not fully utilize meta-learning strategies.

Table 6. Comparative study with existing studies

Reference	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	Specificity (%)
[1]	97.20	—	—	—	—	—
[2]	98.30	98.10	98.40	98.25	—	—
[3]	96.85	—	95.70	—	97.10	—
[4]	97.90	—	—	—	—	—
[5]	98.18	—	—	68.75	—	98.16
[7]	99.15	98.77	99.38	99.07	—	—
Proposed model	99.50	99.53	99.53	99.53	99.95	99.45

The performance comparison of the various meta-learners depicted in Table 7 indicated that stacking with ANN meta-learner exhibited excellent performance with high accuracy (99.50%), high precision (99.53%), high F1-score (99.30%), best cross-validation mean (99.71%), and least standard deviation (0.39%), which demonstrated high stability and generalization. The recall (99.53%) and the same accuracy (99.24%) of the stacking with XGBoost meta-learner were slightly worse in the cross-validation but its F1-score (99.30%) was better. The stacking with gradient boosting meta-learner had a little lower accuracy (98.99%) and a bigger overfitting gap (0.0101), although achieving the greatest AUC (99.98%). Overall, the ANN-based meta-learner was the recommended for detection framework since it offered the best overall balance of robustness, generalization, and predictive performance.

Table 7. Meta learner performance comparison

Meta learner	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	CV mean (%)	CV Std (%)	Overfit gap
Stacking + ANN Meta	99.50	99.53	99.53	99.53	99.98	99.71	0.39	0.0056
Stacking + XGBoost Meta	99.24	99.07	99.53	99.30	99.97	99.30	0.51	0.0076
Stacking + gradient boosting meta	98.99	99.07	99.07	99.07	99.98	99.19	0.79	0.0101

4.2. External validation

To evaluate the generalization capability of the proposed framework, a leave-one-dataset-out external cross-validation strategy was conducted using four ASD datasets, including children, adolescent, adult, and custom dataset-I. In each iteration, one dataset was used exclusively for testing, while the remaining datasets were combined for training, ensuring evaluation on completely unseen data distributions without retraining. The main dataset-I was used to train the model, and it was tested on untested external datasets without retraining. External validation performance of the proposed ASD detection model on three independent age-group datasets: adult, child, and adolescent are described in Table 8. In order to determine the model's robustness and practical usefulness, external validation assesses the model's capacity to generalize to unseen data. The adult dataset achieved the highest overall performance, with an accuracy of 97.59% and an AUC of 99.73%. The model also obtained strong precision (95.74%), recall (95.24%), and F1-score (95.49%), indicating a balanced and reliable classification capability. For the child dataset, the model achieved an accuracy of 93.84% and an AUC of 99.05%. The precision value of 96.95% indicates that the majority of positive predictions were correct, while the recall of 90.07% shows that most ASD cases were successfully detected. The F1-score of 93.38% confirms the overall effectiveness of the model on child data. The adolescent dataset produced comparatively lower results, with an accuracy of 85.58% and an AUC of 98.30%. Although precision remained very high at 98.00%, the recall decreased to 77.78%, resulting in an F1-score of 86.73%. This suggests that while the model makes highly accurate positive predictions for adolescents, it misses a larger proportion of ASD cases compared with the Adult and Child datasets. The lower performance may be attributed to age-specific behavioral variations, dataset imbalance, or differences in feature distributions. Regarding computational efficiency, the training time increased from 39.29 seconds for the adult dataset to 51.66 seconds for the adolescent dataset. Despite this increase, the training times is still feasible for real-world deployment. Overall, the external validation results demonstrate that the proposed framework generalizes effectively across different age groups, maintaining 92.34% average accuracy and average AUC values above 99.03% for all datasets. The robustness and generalization capabilities of the suggested ASD detection model are confirmed by the high performance on the adult and child datasets and the acceptable results on the adolescent dataset.

Table 8. External validation results on independent age-group datasets

Test dataset	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-score (%)	Train time (s)
Adult	97.59	99.73	95.74	95.24	95.49	39.29
Child	93.84	99.05	96.95	90.07	93.38	45.63
Adolescent	85.58	98.30	98.00	77.78	86.73	51.66

4.3. Learning curves and overfitting analysis

The generalization ability of the proposed ASD recognition framework was analyzed using learning curve and overfitting methods. According to the analysis results shown by the learning curve, an incremental enhancement in the performance on validation is obtained with an increase in the number of samples used for training. With 87 training samples, the accuracy on validation reached 87.91%, while 95.51% accuracy was achieved with 873 training samples. Simultaneously, the standard deviation of validation accuracy decreased from 1.75 to 0.97, suggesting enhanced stability and reduced sensitivity to data variations. The convergence of training and validation performance at larger sample sizes confirms that the model learned meaningful ASD-related patterns rather than memorizing the training data.

The overfitting assessment further supports the robustness of the proposed framework. The model achieved a training accuracy of 94.99%, while the average cross-validation accuracy was 93.04%, resulting in an overfitting gap of only 1.94%. Such a small gap indicates excellent generalization and suggests that the model does not suffer from significant overfitting. The incorporation of SHAP-based feature selection, PCA dimensionality reduction, SMOTE-based class balancing, dropout regularization in the neural network meta-learner, and early stopping collectively contributed to controlling model complexity and improving generalization performance.

4.4. Fold-wise variance analysis and reproducibility assessment

To assess model stability and reproducibility, a fold-wise variance analysis was presented using five-fold cross-validation as depicted in Figure 8. The values of the accuracy obtained from each fold were found to be 91.78%, 92.69%, 92.66%, 93.12%, and 94.95% respectively, resulting in a mean cross-validation accuracy value of 93.04%, having a standard deviation of 1.05%. Such a low standard deviation clearly implies consistency in performance regardless of the train-test data split and shows that the proposed framework is not very sensitive to any particular choice of training and testing data.

This narrow band of variance in terms of accuracy between various folds highlights the consistency and reliability of the stacking ensemble design. The presence of low variance, along with the small gap in overfitting, indicates high levels of reproducibility of results. Such observations suggest that the proposed framework generalizes well on ASD-related datasets and is therefore fit for implementation in practical screening applications.

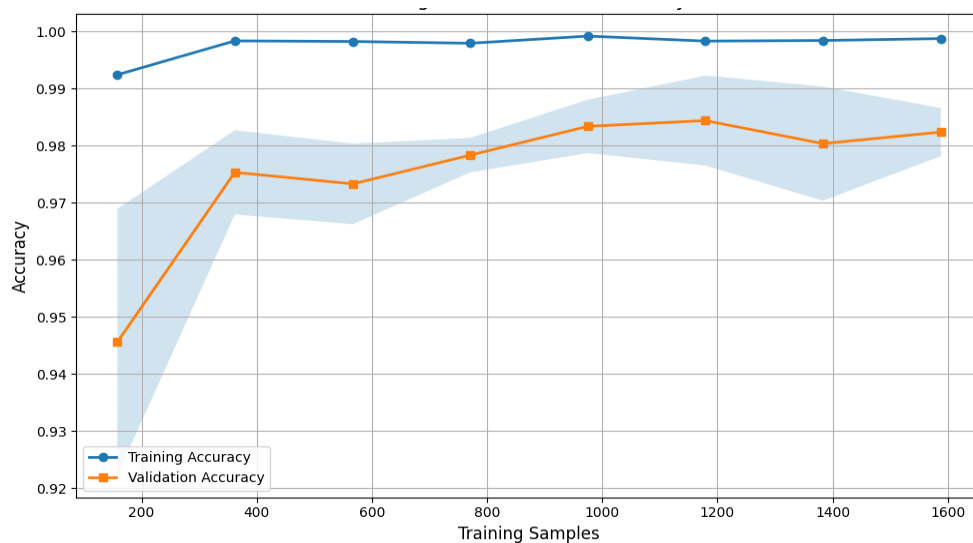


Figure 8. Learning curve and variance analysis

5. CONCLUSION

The proposed stacking ensemble with an ANN meta-classifier, SHAP feature selection, and data balancing outperformed traditional machine learning models and AutoGluon on various ASD datasets. The model had good predictive accuracy, sensitivity, specificity, and AUC showed high level of generalization through cross validation, overfitting test and external validation. Early diagnosis allows early intervention through behavioral therapy and other educational and therapeutic methods, which improve the growth and standard of living for autistic individuals and their families. Another advantage is that the system requires only structured tabulated data input, making it cost-effective and implementable even in settings that do not have access to advanced imaging technology and specialized knowledge. Future research will focus on optimizing efficiency for practical deployment in low-resource healthcare environments. Although no significant age- or gender-related bias was observed, missing socio-economic data and demographic differences may still affect model performance. Future studies should concentrate on verifying the suggested framework using bigger and more varied datasets. Additional objective biomarkers may further enhance screening performance through multimodal learning approaches. Advanced dimensionality reduction techniques such as autoencoders and alternative imbalance-handling methods including adaptive synthetic sampling approach (ADASYN) and cost-sensitive learning may also improve feature representation and model robustness. Furthermore, integrating fuzzy logic, optimization algorithms, reinforcement learning, and natural language processing can improve model performance, support adaptive decision-making, and generate clear clinical reports to assist healthcare professionals. Collectively, these directions offer promising opportunities for enhancing the fairness, adaptability, and clinical applicability of AI-assisted ASD screening systems.

FUNDING INFORMATION

No funding was received to conduct this study.

AUTHOR CONTRIBUTIONS STATEMENT:

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Savita Rani	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓			
Chandra Shakher Tyagi		✓	✓	✓						✓		✓		
Devendra Kumar	✓			✓						✓		✓	✓	

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

No relevant interests to disclose.

DATA AVAILABILITY

The data used in this work is available on Kaggle at ASD children traits: <https://www.kaggle.com/datasets/uppulurimadhuri/dataset> and autism screening data for toddlers: <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>.

REFERENCES

[1] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Frontiers in Neuroinformatics*, vol. 13, Nov. 2019, doi: 10.3389/fninf.2019.00070.

[2] R. Ayub, S. I. Ansarullah, F. Samdani, A. Akhtar, M. Afzal, and S. M. Ahmed, "Enhancing early detection of autistic spectrum disorder in children using machine learning approaches," *Journal of King Saud University-Science*, vol. 36, no. 10, Nov. 2024, doi: 10.1016/j.jksus.2024.103468.

[3] N. Mumenin, M. M. Rahman, M. A. Yousuf, F. M. Noori, and M. Z. Uddin, "Early diagnosis of autism across developmental stages through scalable and interpretable ensemble model," *Frontiers in Artificial Intelligence*, vol. 8, May 2025, doi: 10.3389/frai.2025.1507922.

[4] T. Nishat, T. Juvvaladinne, and D. S. Ambatipudi, "Advancing autism spectrum disorder diagnosis through ensemble learning," *i-manager's Journal on Artificial Intelligence and Machine Learning*, vol. 2, no. 1, 2024, doi: 10.26634/jaim.2.1.20500.

[5] I. Nafisah *et al.*, "Deep learning-based feature selection for detection of autism spectrum disorder," *Frontiers in Artificial Intelligence*, vol. 8, Jun. 2025, doi: 10.3389/frai.2025.1594372.

[6] S. S. Kumar, K. Selvakumar, and V. S. Murugan, "Identification of autism spectrum disorder using modified convolutional neural network (MCNN) and feature selection techniques," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 2s, pp. 678–691, 2024.

[7] K. Dick *et al.*, "Transformer-based deep learning ensemble framework predicts autism spectrum disorder using health administrative and birth registry data," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-90216-8.

[8] M. Temiz, B. B.-Gungor, N. Ersoz, and M. Yousef, "Machine learning-based prediction of autism spectrum disorder and discovery of related metagenomic biomarkers with explainable AI," *Applied Sciences*, vol. 15, no. 16, Aug. 2025, doi: 10.3390/app15169214.

[9] S. Alsubai, "Deep transfer learning and explainable AI framework for autism spectrum disorder detection across multiple datasets," *Frontiers in Neurology*, vol. 16, Jan. 2026, doi: 10.3389/fneur.2025.1617446.

[10] Z. Yin *et al.*, "Early autism diagnosis based on path signature and Siamese unsupervised feature compressor," *Cerebral Cortex*, vol. 34, no. 13, pp. 72–83, May 2024, doi: 10.1093/cercor/bhae069.

[11] A. Panchal, K. Khan, and R. Katarya, "Empirical analysis of nature-inspired algorithms for autism spectrum disorder detection using 3D video dataset," 2025, *arXiv: 2501.01202*.

[12] B. Paneru, B. Paneru, K. B. Shah, A. Shrestha, R. Poudyal, and K. N. Poudyal, "Autism spectrum disorder prediction using hybrid deep learning model and a recommendation system application for autistic patient," *Journal of Computer Science*, vol. 20, no. 9, pp. 1040–1050, Sep. 2024, doi: 10.3844/jcssp.2024.1040.1050.

[13] L. Kampa, K. Yamini, A. Basavaraju, and K. Anoop, "A stack based ensemble learning method for diagnosing autism spectrum disorder," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 3, pp. 237–251, 2022.

[14] Y. Alashban, "Enhanced detection of autism spectrum disorder through neuroimaging data using stack classifier ensembled with modified VGG-19," *Acta Radiologica*, vol. 66, no. 9, pp. 955–963, Sep. 2025, doi: 10.1177/02841851251333974.

[15] S. Rai, "A novel hybrid deep learning model using MEResNext for autism spectrum disorder detection," *Computational Biology and Chemistry*, vol. 120, Feb. 2026, doi: 10.1016/j.compbiolchem.2025.108619.

[16] R. A. A. Romero *et al.*, "Benchmarking AutoML frameworks for disease prediction using medical claims," *BioData Mining*, vol. 15, no. 1, Dec. 2022, doi: 10.1186/s13040-022-00300-2.

[17] N. Erickson *et al.*, "AutoGluon-tabular: robust and accurate AutoML for structured data," Mar. 2020, *arXiv: 2003.06505*.

[18] H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, "Automated machine learning with interpretation: a systematic review of methodologies and applications in healthcare," *Medicine Advances*, vol. 2, no. 3, pp. 205–237, Sep. 2024, doi: 10.1002/med4.75.

- [19] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Apr. 2016, doi: 10.1098/rsta.2015.0202.
- [20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [21] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.
- [22] V. O. Eguavoen, E. Nweli, and A. Onyenokwe, "Explainable AI framework for early autism spectrum disorder detection: integrating ensemble learning with clinical interpretability," *Fudma Journal of Sciences*, vol. 9, no. 10, pp. 209–216, 2025.
- [23] Z. Cai, H. Huang, G. Sun, Z. Li, and C. Ouyang, "Advancing predictive models: unveiling LightGBM machine learning for data analysis," in *2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI)*, Dec. 2023, pp. 109–112, doi: 10.1109/ICCBD-AI62252.2023.00027.
- [24] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
- [25] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *Journal of Data Analysis and Information Processing*, vol. 7, no. 4, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.
- [26] X. He, K. Zhao, and X. Chu, "AutoML: a survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, Jan. 2021, doi: 10.1016/j.knosys.2020.106622.
- [27] L. Scott M. and L. S.-In., "A unified approach to interpreting model predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.
- [28] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.
- [29] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, 2018, doi: 10.1016/j.heliyon.2018.e00938.

BIOGRAPHIES OF AUTHORS



Savita Rani    holds an M.Tech. degree in Computer Science and Engineering from Maharshi Dayanand University, Haryana, India and is pursuing her Ph.D. in the Department of Computer Science and Engineering at SRM Institute of Science and Technology. Her research interests include machine learning, artificial intelligence, and fuzzy logic. She can be contacted at email: ss0004@srmist.edu.in.



Dr. Chandra Shakher Tyagi    is an assistant professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Delhi NCR campus, Modi Nagar, Uttar Pradesh. He has been affiliated with SRM Institute of Science and Technology since 2012. Ph.D. completed in 2022 from Noida International University, Greater Noida, M.Tech. from SRM University Chennai, B.Tech. in Computer Science and Engineering from SRM University Chennai. His research area are artificial intelligence and machine learning, WSN and VLSI design. He has published various papers in reputed journals. He can be contacted at email: chandrah@srmist.edu.in.



Dr. Devendra Kumar    holds Ph.D. degree in Computer Science and Engineering. He is working as a professor and head of the Department of Computer Applications at ABES Engineering College, Ghaziabad, India. His areas of interest include data mining, software engineering, IoT, and information/database/system/network security. He can be contacted at email: devendra.arya@gmail.com.