frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 30 May 2025
DOI 10.3389/frai.2025.1507922

OPEN ACCESS

EDITED BY
Miodrag Zivkovic,
Singidunum University, Serbia

REVIEWED BY
Bosko Nikolic,
University of Belgrade, Serbia
S. Yuvaraj,
Sri Eshwar College of Engineering, India

*CORRESPONDENCE
Nasirul Mumenin
nmmouno@gmail.com
Farzan M. Noori
farzanmn@ifi.uio.no

RECEIVED 08 October 2024
ACCEPTED 21 April 2025
PUBLISHED 30 May 2025

CITATION
Mumenin N, Rahman MM, Yousuf MA, Noori FM and Uddin MZ (2025) Early diagnosis of autism across developmental stages through scalable and interpretable ensemble model. *Front. Artif. Intell.* 8:1507922. doi: 10.3389/frai.2025.1507922

Early diagnosis of autism across developmental stages through scalable and interpretable ensemble model

Nasirul Mumenin [1]*, Maisha Mumtaz Rahman[2], Mohammad Abu Yousuf [3], Farzan M. Noori[4]* and Md Zia Uddin [5]

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh, [2]Department of Electrical and Electronics Engineering, Islamic University of Technology, Dhaka, Bangladesh, [3]Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh, [4]Department of Informatics, University of Oslo, Oslo, Norway, [5]Department of Sustainable Communication Technologies, SINTEF (Norwegian: Stiftelsen for Industriell og Teknisk Forskning) Digital, Oslo, Norway

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental condition that challenges early diagnosis due to its diverse manifestations across different developmental stages. Timely and accurate detection is essential to enable interventions that significantly enhance developmental outcomes. This study introduces a robust and interpretable machine learning framework to diagnose ASD using questionnaire data. The proposed framework leverages a stacked ensemble model, combining Random Forest (RF), Extra Tree (ET), and CatBoost (CB) as base classifiers, with an Artificial Neural Network (ANN) serving as the meta-classifier. The methodology addresses class imbalance using Safe-Level SMOTE, dimensionality reduction via Principal Component Analysis (PCA), and feature selection using Mutual Information and Pearson correlation. Evaluation on publicly available datasets representing toddlers, children, adolescents, adults, and a merged dataset (Combining children, adolescents, and adults dataset) demonstrates high diagnostic accuracy, achieving 99.86%, 99.68%, 98.17%, 99.89%, and 96.96%, respectively. Comparative analysis with standard machine learning models underscores the superior performance of the proposed framework. SHapley Additive exPlanations (SHAP) were used to interpret feature importance, while Monte Carlo Dropout (MCD) quantified uncertainty in predictions. This framework provides a scalable, interpretable, and reliable solution for ASD screening across diverse populations and developmental stages.

KEYWORDS

Autism Spectrum Disorder, ensemble model, uncertainty analysis, explainable AI, Monte Carlo Dropout, SHAP

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by a wide range of symptoms and severity levels, affecting communication, behavior, and social interactions. The early and accurate diagnosis of ASD is crucial for initiating appropriate interventions that can significantly improve the quality of life for individuals with ASD and their families. However, the heterogeneous nature of ASD, combined with overlapping symptoms with other developmental disorders, poses substantial challenges to its diagnosis.

Several biomarkers have received attention because of their ability to detect individuals who are at increased risk of developing ASD. Questionnaire-based screening tools are among the most widely used methods for preliminary ASD assessment. These tools generate rich datasets that encode valuable information on individuals' behavioral and developmental characteristics. However, such data's complexity and high dimensionality necessitate sophisticated analytical techniques to extract meaningful patterns and insights. Comprehensive diagnostic assessments for ASD, such as ADOS (Bastiaansen et al., 2011) or ADI-R (De Bildt et al., 2004), can be time-consuming, often requiring multiple sessions over several days. Diagnoses are subject to interpretation by the clinician, which leads to potential variability between evaluators. This subjectivity can affect the consistency and reliability of the diagnosis. The need for specialized training to conduct assessments limits the availability of qualified professionals, especially in underserved or rural areas.

Recent years have witnessed an increase in the number of establishments exploring the use of ML techniques to aid in the early detection of ASD, with the aim of complementing traditional diagnostic processes with objective data-driven approaches (Hasan et al., 2022; Mumenin et al., 2023, 2024). ML algorithms can analyze complex and high-dimensional data from various sources, including genetic, neuroimaging, and behavioral data. Questionnaire-based tools, in particular, provide a valuable resource for ML models. Research in areas such as medical diagnosis, optimization, and pattern recognition has demonstrated the efficacy of these hybrid approaches (Talukder et al., 2023; Mumenin et al., 2025; Choudhury et al., 2025). While traditional ML methods have demonstrated success in ASD detection, they often fall short in three critical areas: (1) handling class imbalance effectively, (2) ensuring interpretability of predictions, and (3) providing reliable uncertainty estimates to support clinical decision-making. Furthermore, most existing methods focus on a single algorithmic approach, limiting their ability to fully exploit the diversity of complex data patterns inherent in questionnaire-based ASD assessments.

This paper introduces an innovative methodology that combines stacked ensemble model (EM), XAI, and Uncertainty Analysis (UA) to improve the precision and dependability of ASD classification. The EM that has been proposed not only enhances the accuracy of classification but also integrates mechanisms for evaluating and managing uncertainty. This facet is frequently disregarded in conventional models. The proposed model capitalizes on the advantages of multiple base classifiers, each providing distinct viewpoints in identifying ASD, thus generating a comprehensive, multidimensional feature space. Using a metaclassifier to integrate these base classifiers, the ultimate prediction is guaranteed to represent an exhaustive examination of the underlying patterns present in the data.

In addition, the proposed model places significant importance on interpretability, a critical aspect in medical applications where understanding the reasoning behind classifications is crucial to establishing trust and facilitating subsequent analysis. An essential advance of this research is to integrate uncertainty consciousness into the framework. Acknowledging that a specific sample might exhibit equivocal or deceptive characteristics, our model incorporates a confidence metric into its predictions to offer significant insights into its decision-making methodology. This functionality is critical for end-users and allows the model to be continuously enhanced by flagging instances with high uncertainty for additional investigation or manual review.

The main contributions of this study are:

- Development of a stacked EM model that effectively classifies ASD across multiple age groups. Safe-Level SMOTE ensures balanced data representation, improving model generalization and mitigating class imbalance issues.
- Incorporation of SHAP for model interpretability, enabling the identification of key features influencing predictions. SHAP plots provide insights into the factors responsible for particular classifications, fostering trust and understanding of the model's decisions.
- Utilization of MCD for uncertainty estimation, allowing the model to quantify its confidence in predictions. This enhances the reliability of the framework, addressing a critical need for dependable tools in clinical decision-making.
- The model is tested and validated on multiple publicly available datasets representing diverse developmental stages (toddler, child, adolescent, and adult) and an integrated dataset. Evaluation through standard metrics (accuracy, precision, recall, and F1-score) demonstrates the robustness and scalability of the proposed approach.

The rest of the paper is organized as follows: Section 2 presents a detailed literature review. Section 3 discusses the architecture and methodology of the study, along with the tools and techniques that were implemented. In Section 4, we analyze the experimental results and provide a performance comparison of the proposed method. Section 5 describes the implementation and findings of explainable artificial intelligence (XAI). Section 6 explains the use of Monte Carlo Dropout (MCD) and how it enhances the results. Finally, Section 7 concludes the paper with a concise discussion of the drawbacks and potential future directions of the work.

2 Literature review

A significant number of researchers have utilized ML-based models to diagnose ASD. Hasan et al. (2022) detected ASD in individuals of various age groups. The authors demonstrated that ML-based predictive models are effective instruments for this endeavor. Mukherjee et al. (2023) presented three frameworks with ML models to detect ASD among children, toddlers, and adults. They explored the facial image-based and questionnaire-based techniques for the detection of ASD. Bala et al. (2022) introduced an ML model that analyzes ASD data across various age groups and accurately identifies ASD. For such purpose, datasets were collected on ASD from toddlers, children, adolescents, and adults. Afterwards, various classifiers were implemented on these datasets, and evaluation metrics were used to assess their efficacy. Kamma et al. (2022) proposed a Light Gradient Boost (LGB) based model to classify ASD and a Random Search for hyperparameter optimization. A synthesis of three publicly accessible datasets comprising records of ASD in infants, adolescents, and adults was used. Devika Varshini and Chinnaiyan (2020) evaluated the efficacy

of a range of ML algorithms and preprocessing methods in the classification of medical datasets intending to forecast early autism symptoms in both toddlers and adults.

Stirling et al. (2021) examined the application of a Self-Organizing Fuzzy classifier and the UCI "Autism Screening Adult" dataset to predict whether an individual is more likely to have ASD and therefore merits a higher priority for subsequent testing and diagnosis. Using an efficient ensemble classification method, Haroon and Padma (2022) sought to detect and diagnose Parkinson's Disease and ASD in their early stages. A delayed or erroneous diagnosis could endanger the life of the patient. Consequently, early and accurate detection has been the primary objective of this research. In their study, Hasan et al. (2021) collected ASD data from both toddlers and adults, implemented seven distinct classification techniques, and evaluated the results. Using statistical and ML techniques, they computed the significant and associative features of both datasets. Additionally, they have identified characteristics that may be utilized to classify children with ASD. The ML architecture proposed by Uddin et al. (2023) was implemented to produce more accurate and efficient outcomes in the rapid diagnosis of ASD. FT techniques were implemented on the ASD samples and the modified dataset was evaluated to determine the effectiveness of numerous classifiers. The significant characteristics of normal and ASD individuals in Bangladesh were investigated (Satu et al., 2019). Individual samples were obtained from parents of children aged 16 to 30 months from various residents through the utilization of Autism Barta applications, both in the field and via the Internet. An evaluation was conducted on various tree-based techniques in order to determine their optimal classifier. Akter et al. (2021b) introduced an improved ML model that exhibits enhanced accuracy in autism detection. An examination was conducted on the correlation between individual and highly co-linear features in these datasets. To assess the symptoms of ASD, Thabtah et al. (2018) devised a rules-based ML (RML) methodology. They discovered that RML enhances the efficacy of the classifier. Abbas et al. (2018) combined ADI-R and ADOS ML methodologies in a unified assessment and resolved the scarcity, sparsity, and data imbalance challenges by implementing feature encoding techniques. In addition, an alternative investigation conducted by Thabtah et al. (2018) and Thabtah (2019) introduced Variable Analysis (VA), a computational intelligence (CI) method that used LR, decision trees (DT), and SVM to generate accurate prognoses and diagnoses for ASD. The VA method illustrated correlations between features and between features.

Researchers have also used various Deep Learning (DL) techniques to diagnose ASD. Mujeeb Rahman and Monica Subashini (2022) examined the accuracy with which DNN-based models identified autism in toddlers using the QCHAT datasets that had previously been collected. Two distinct DL models were developed to process the two iterations of the QCHAT and QCHAT-10 datasets. Mohanty et al. (2021) made an effort to integrate Principal Component Analysis (PCA) to reduce feature dimensions, after which DNN was utilized to classify the type of ASD. The results of the experiment suggest that the combination of PCA and DNN yields clinically acceptable results in accurately identifying ASD. Garg et al. (2022) introduced a hybrid methodology that combines XAI and DL to identify the most influential features for the timely and accurate prognosis of ASD. The suggested framework provides enhanced predictive capabilities and clinical recommendations for predicted outcomes, serving as a crucial tool for the early and improved identification of ASD traits in toddlers. Hajjej et al. (2024) proposed a two-stage framework: In the initial stage, a collection of ML models, such as a random forest ensemble and XGBoost classifiers, are utilized to accurately identify Autism Spectrum Disorders. Identifying appropriate teaching methods for children with ASD through an evaluation of their verbal, physical, and behavioral performance is the focus of the second phase of the research. Utilizing an EL approach, Kampa et al. (2022) developed a model for diagnosing ASD in datasets about children and toddlers. This method serves as a supplement to the traditional single-learning approaches. They achieved favorable performance outcomes by employing feature selection and an EM.

Among the numerous ML and DL methodologies, EM has demonstrated the most potential (Ganaie et al., 2022; Rincy and Gupta, 2020). Ensemble learning (EL) is a technique that combines multiple models to enhance the precision, resilience, and applicability of predictions. Despite progress in detection and classification methodologies, persistent challenges remain, specifically in managing the vast quantity and diverse range of newly discovered ASD samples. An increasing demand exists for models that possess the critical qualities of high accuracy, interpretability, and uncertainty tolerance. DL models, specifically those built upon ANNs, are frequently called "black boxes" due to their complex architectures and the opaque manner in which they produce results (Samek et al., 2017). XAI aims to address this disparity by offering stakeholders a deeper understanding of how these intricate models operate, thus cultivating confidence and empowering them to understand, rely on and efficiently administer AI solutions. Moreover, UA plays a critical role in DL as it enables the evaluation of the dependability and resilience of model predictions (Abdar et al., 2021). DL models may occasionally generate overly optimistic forecasts due to spurious correlations or an inadequate understanding of the data, which can result in decisions that are potentially risky and overly confident (Gawlikowski et al., 2023). Practitioners can enhance the prudence and knowledge of decision-making by identifying instances in which the model's output may be unreliable through the analysis and quantification of prediction uncertainty.

3 Proposed methodology

The workflow of proposed methodology is shown in Figure 1.

3.1 Dataset

The four ASD datasets (Toddlers, Adolescents, Children, and Adults) were obtained from repositories that are accessible to the public: UCI ML and Kaggle (Thabtah, 2018, 2017a,b; Tabtah, 2017). The ASDTests smartphone application, which employs the QCHAT-10 and AQ-10 for ASD screening in toddlers, children, adolescents, and adults, was developed by Thabtah et al. (2018) and Thabtah (2019). An affirmative diagnosis of ASD is indicated

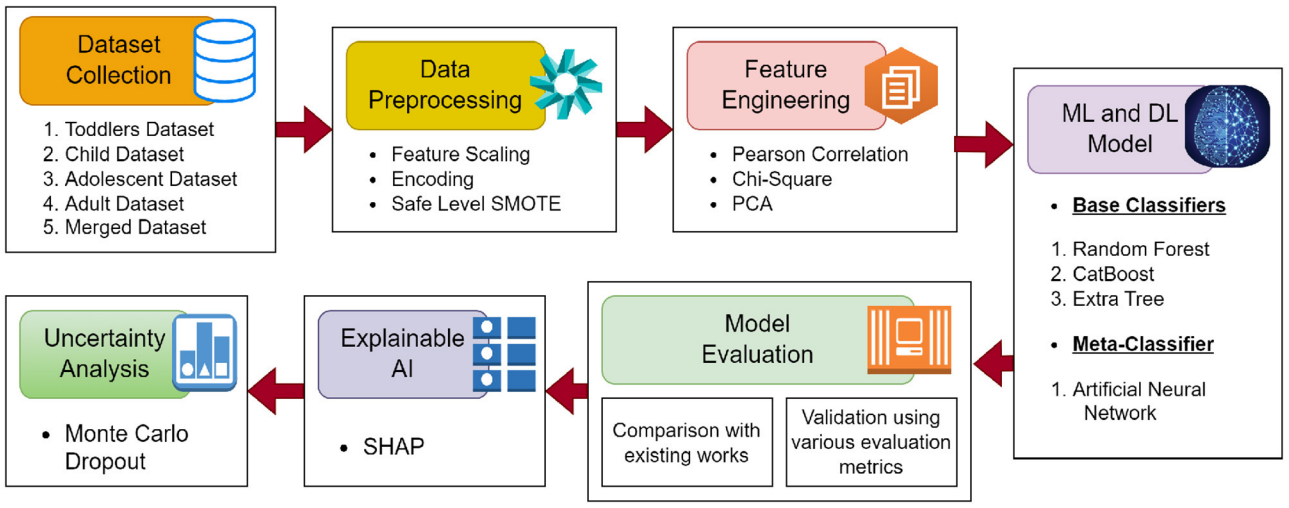

FIGURE 1
The workflow of the proposed methodology.

by an ultimate score of 6 out of 10 on a scale of zero to ten, which is calculated for each individual by application. Furthermore, the ASDTests application provides access to ASD data, and open-source databases are being created to aid in the investigation of this field. In conclusion, three datasets (child, adolescent, adult) have been merged to form a single dataset. The Adolescents, Children, and Adults datasets were merged primarily because they share identical feature sets, facilitating seamless integration into a unified dataset. Our primary motivation for merging these datasets was to build a robust model capable of generalizing across a broader developmental spectrum, rather than developing multiple separate age-specific models. While explicit statistical analyses (e.g., distributional comparisons) were not performed, the identical nature of the features and consistent data-collection procedures across these datasets and the experimental results proved that merging did not adversely impact the model's performance. The data sets used in this study have been classified as follows: toddler, child, adolescent, adult, and merged.

3.2 Data preprocessing

Data preparation is a critical phase in the ML pipeline, as it involves cleaning, transforming, and normalizing data to make it suitable for analysis and training models. Missing values have been found in "child (4)", and "adult (2)" the dataset's "age" columns. The missing values were replaced using the median imputation technique. Since most ML models are based on mathematical equations, categorical data must be converted to numerical data to avoid complications. So, we encoded the values in the column "ASD traits" that contained categorical data (No, Yes) into numerical values (0, 1).

To address the class imbalance, Safe-level-SMOTE was utilized. The Synthetic Minority Oversampling Technique (SMOTE) is a method used to address class imbalance in datasets, particularly in the context of supervised learning (Chawla et al., 2002). Figure 2 shows the size of the original datasets, and the size after applying SMOTE, and Safe-level-SMOTE. It works by creating synthetic samples from the minority class rather than copies, which helps overcome the overfitting problem of random oversampling. When constructing predictive models, it is critical to comprehend the significance of each feature in relation to the target variable. A productive approach to assess this level of importance is the computation of Mutual Information (MI) scores. This method proved particularly advantageous in our particular scenario, where our dataset comprised a combination of linear and non-linear relationships that the MI technique could accurately capture. The ranking of features according to their MI scores provided a distinct perspective on which features could serve as the most significant predictors of the target variable. The model could be simplified by identifying and retaining solely the most informative features, thereby mitigating the potential for overfitting and enhancing interpretability.

The Pearson correlation coefficient is calculated to quantify the linear association between two continuous variables within the dataset (Obilor and Amadi, 2018). Moreover, principal component analysis (PCA) was used to reduce the dataset's dimensionality by retaining principal components that explained 95% of the variance. Based on PCA, 12 optimal features were selected. The selected features are: "A1_Score", "A2_Score", "A3_Score", "A4_Score", "A5_Score", "A6_Score", "A7_Score", "A8_Score", "A9_Score", "A10_Score", "ethnicity", "contry_of_res". These components are linear combinations of the original features, which pose challenges for direct interpretability. To address this, we projected the SHAP values of the principal components back into the original feature space using PCA loadings. This approach allowed us to identify the contribution of each original feature to the retained components and, by extension, to the model's predictions.

To ensure unbiased evaluation and generalizability of the proposed model, the data was split into two subsets: 80% for training and 20% for testing, using a random stratified sampling approach to maintain the class distribution in both subsets. The

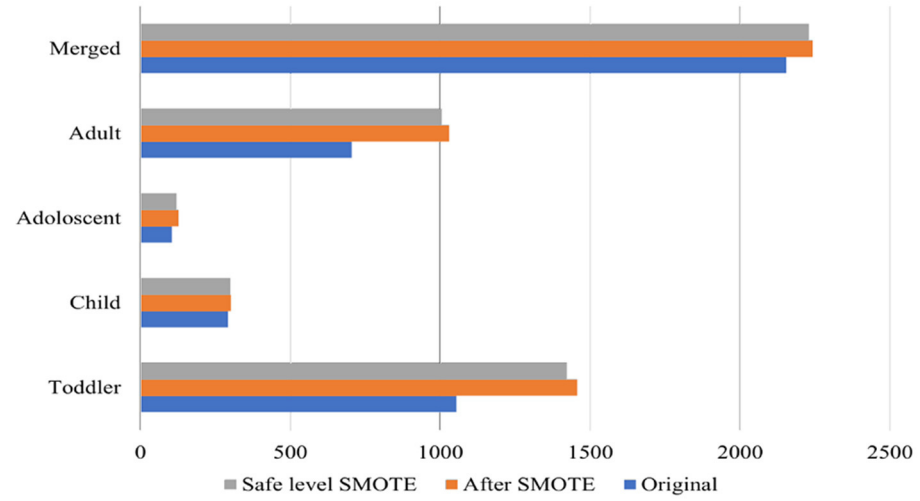

FIGURE 2
The number of data in original datasets, after applying SMOTE, and safe-level-SMOTE.

training data was further subjected to 5-fold cross-validation to validate the model's performance and tune hyperparameters. During cross-validation, the training data was divided into five folds, with four folds used for training and the remaining fold for validation in each iteration. This process was repeated five times, ensuring each fold was used once as the validation set. The final model was trained on the entire training set using the best hyperparameters obtained during cross-validation and evaluated on the independent test set. This approach mitigates the risk of overfitting and ensures reliable estimates of the model's performance on unseen data.

3.3 Stacked ensemble model

Stacking, also referred to as stacked generalization using the ensemble method, operates on a straightforward principle: rather than relying on basic functions like the voting ensemble, all predictors' predictions are combined. One advantage of stacking is its ability to leverage the performance of several high-performing models in a classification or regression task, resulting in predictions that surpass the performance of any individual model within the ensemble (Sesmero et al., 2015; Naimi and Balzer, 2018). The primary objective is to incorporate the benefits of distinct and discrete models into the hybrid ensemble model while minimizing their drawbacks. The architecture of the proposed stacked EM is shown in in Figure 3.

3.3.1 Base-classifier

3.3.1.1 Random forest

Random forest is a frequently implemented supervised ML algorithm used to address classification and regression issues. The algorithm comprises numerous DTs, each of which processes a distinct subset of the dataset and calculates the mean to improve the prediction's precision. RF is an EL technique that reduces overfitting and outperforms a single DT by aggregating the results (Biau and Scornet, 2016).

3.3.1.2 CatBoost

CatBoost (CB) (Prokhorenkova et al., 2018) is a gradient boost algorithm designed to efficiently handle categorical features. Incorporating innovative techniques to achieve high performance and robustness, particularly in scenarios with heterogeneous data types and large-scale datasets. CB minimizes a differentiable loss function $L(y_i, F(x_i))$, where y_i is the target variable, $F(x_i)$ is the predicted value for the i^{th} instance, and x_i is the characteristic vector. CB sequentially builds an ensemble of DTs to minimize the loss function.

3.3.1.3 Extra trees

The ET algorithm, also known as Extremely Randomized Trees, is an EL method that belongs to the family of tree-based algorithms. The system initially generates many DTs during the training phase. Subsequently, it determines the output class based on the mean prediction (regression) or mode of the classes (classification) of the individual trees. The fundamental concept underlying the ET algorithm is the incorporation of randomization, which augments the model's variance to mitigate the risk of overfitting (Alsariera et al., 2020).

3.3.2 Meta-learner

3.3.2.1 Artificial Neural Network

A feed-forward Neural Network (FFN) is a type of network that creates a directed graph with nodes and edges. Data are transmitted along these edges from one node to the next without forming a cycle. The ANN is a variant of FFN with three or more layers: an input layer, one or more hidden layers, and an output layer. Researchers utilize a hyperparameter optimization approach to ascertain the optimal number of concealed layers for an ANN. The process of information transfer between layers does not consider

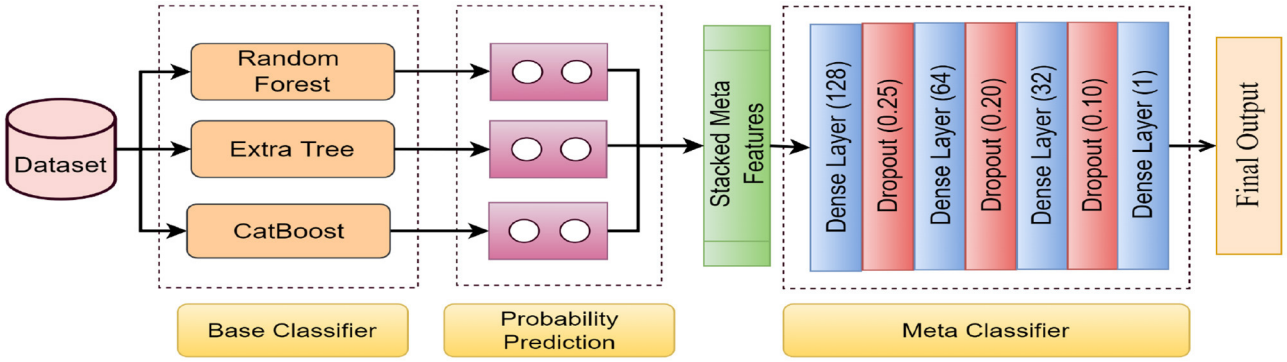

FIGURE 3
The architecture of the proposed stacked EM. The data is passed through the four base classifiers separately at first. Then, the base classifiers' predictions are stacked together to generate the meta-features. Finally, the meta-features are passed through the meta classifier to generate the final output.

previous values, and all neurons in each layer are interconnected, as supported by the sources (Goodfellow et al., 2016).

3.3.3 Proposed ensemble model

In this study, we have proposed a deep EL framework that synergistically combines multiple base classifiers with an ANN-based metaclassifier. The overarching goal is to leverage the diverse strengths of various classifiers to enhance the model's predictive performance. In the base classifier part of the EM, three different classifiers are employed, each processing the input data and providing outputs that will be used to create meta-features for the meta-classifier.

Let the input dataset be represented as X=[x_1, x_2, ..., x_n], where each x_i is a feature vector representing an individual sample, and n is the total number of samples. Correspondingly, the target labels are denoted by Y=[y_1, y_2, ..., y_n], where each y_i is the binary class label associated with x_i. Each base classifier C_j is trained on the dataset X with the goal of learning a mapping function, which predicts the probability that a given sample x_i belongs to the positive class. The training process involves optimizing the parameters of C_j to minimize the discrepancy between the predicted labels $y^i(j)$ and the actual labels y_i.

$$F_i : X_{\text{train}} \rightarrow Y^{(i)}_{\text{predicted}} \tag{1}$$

Upon training, each classifier C_j generates a predictive probability for each sample in the training set X_{train}, validation set X_{val}, and test set X_test. For a given sample x_i, the output is a probability score $P_i^{(j)}$ indicating the likelihood that x_i belongs to the positive class, according to classifier C_j. This can be formally represented as:

$$P_i^{(j)} = P(y_i = 1 | x_i; \theta_j) \tag{2}$$

which denotes the conditional probability that the label y_i is 1 given the feature vector x_i and the parameters θ_j of the classifier C_j. For each sample, the predictive probabilities of all base classifiers are aggregated to form a new feature vector x_{meta}, which serves as input to the meta-learner. The aggregation for a sample x_i across m base classifiers can be represented as:

$$x_{i_{\text{meta}}} = [p_i^{(1)}, p_i^{(2)}, \ldots, p_i^{(m)}] \tag{3}$$

where m is the number of base classifiers, the base classifiers effectively transform the original feature space into a meta-feature space of predictive probabilities. These meta-features, encapsulating the predictions from diverse algorithms, are then utilized by the meta-learner to make the final classification decision. This two-tier approach aims to capitalize on the strengths of individual classifiers and enhance overall predictive performance by synthesizing their predictions. In the EL framework, the meta-classifier operates as the second or final layer in the model hierarchy, synthesizing the outputs of the base classifiers to make a final prediction. The meta-classifier receives as input the meta-features X_{meta}, composed of the predictive probabilities or decisions made by the base classifiers. For a given instance x_i, the input to the meta-classifier can be represented as:

$$X_{\text{meta}}(i) = [p_i(1), p_i(2), \ldots, p_i(m)] \tag{4}$$

where $P_i^{(j)}$ is the probability that x_i belongs to the positive class as predicted by the j^{th} base classifier, and m is the total number of base classifiers.

The meta-classifier, denoted as C_{meta}, is trained on this transformed dataset X_{meta} to learn a mapping function $f_{\text{meta}} : X_{\text{meta}} \rightarrow Y$ which aims to predict the final class label y_i for each instance x_i. The function f_{meta} is optimized to minimize the discrepancy between its predictions $\hat{y}^{\text{meta}}$ and the actual class labels y_i. The output of the meta-classifier for each instance x_i is a final prediction $\hat{y}^{\text{meta}}$, which is based on the aggregated insights from the base classifiers' predictions. This prediction can be a class label for classification tasks or a continuous value for regression tasks. For binary classification, the output can also be a probability score $\hat{p}^{\text{meta}}$ representing the likelihood that x_i belongs to the positive class.

For class labels:

$$\hat{y}^{\text{meta}} = C_{\text{meta}}(X_{\text{meta}}(i))$$

For probability estimates:

$$\hat{p}^{\text{meta}} = P(y_i = 1 | X_{\text{meta}}(i); \theta_{\text{meta}})$$

where θ_{meta} represents the parameters of the meta-classifier.

The first layer is a Dense layer with 128 neurons, using the ReLU activation function. It is set to receive input data corresponding to the meta-features generated by the base classifiers. After that, 2 hidden layers are used, having 64 and 32 nodes. A dropout layer was used after each layer, which helped reduce the overfitting issue. Dropout rate has been set to 25%. Lastly, an output layer with 1 node corresponded to 1 output class. The number of layers and nodes was set after much experimentation to find the best possible outcome. ReLu and Sigmoid have been used as the input and output activation functions. We have used Adam as the optimizer and Binary_Crossentropy as loss function. The learning rate was to 0.001 and number of epochs to 10.

4 Evaluation

4.1 Evaluation metrics

Several evaluation metrics have been utilized to evaluate the effectiveness of the proposed model, i.e., *Precision*, *Recall*, F_1 *Score*, *Accuracy*, and *AUC-Score*.

4.2 Result analysis

The model demonstrates robust performance across various metrics, indicating its effectiveness in classifying ASD. The proposed model achieved an accuracy of 99.86%, 99.68%, 98.17%, 99.89%, and 96.96% in the Toddler, Child, Adolescent, Adult, and Merged datasets, respectively. Figures 4a, b present the box-and-whisker plot of accuracy and swarm plot of AUC for all five datasets used in this study. The confusion matrix of all these performance measures are shown in Figure 5. Table 1 presents the results obtained through the experiment of the proposed model. It can be deduced that the model can effectively identify a given sample as ASD or non-ASD.

The model achieved a perfect AUC score of 99.98%, indicating an exceptional ability to distinguish between ASD and non-ASD cases among toddlers. This result is particularly significant given the challenges of early diagnosis of ASD and the importance of timely intervention. With an AUC of 99.89%, the model demonstrated near-perfect performance in the Child dataset. This high score underscores the model's robustness and its potential utility in supporting clinicians and caregivers in the early detection of ASD in children. The model achieved an AUC of 98.16%, showcasing its strong discriminative power in identifying ASD among adolescents. This highlights the applicability of the model in a broad age range, addressing the varying presentation of ASD symptoms as children grow. Mirroring its success with the Toddler dataset, the model once again achieved a perfect AUC score of 99.99% for the Adult dataset. This remarkable consistency across the developmental spectrum emphasizes the model's comprehensive applicability and reliability in ASD screening for all age groups. The performance of the model on the Merged dataset, which amalgamates data across all age categories, resulted in an AUC of 96.04%. While slightly lower than the age-specific datasets, this score is still exceptionally high. It illustrates the model's effectiveness in handling a diverse and complex dataset that reflects the broad variability in ASD presentations across different ages.

The ROC-AUC graphs for the Toddler, Child, Adolescent, Adult, and Merged datasets are depicted in Figures 6a–e respectively. The ROC curve is a plot with the TPR on the y-axis and the FPR on the x-axis at various threshold settings. Figure 6a depicts a ROC curve with an AUC of 1.00. This value quantifies the overall ability of the model to discriminate between the positive and negative classes. The TPR (sensitivity) is constant at 1.0 across all levels of the FPR. This means that the model correctly identifies all positive cases regardless of the number of false positives. The FPR changes from 0.0 to 1.0 without affecting the TPR, which remains perfect throughout. The ROC curve is a horizontal line at the top of the plot area, indicating a perfect classifier. The AUC of 1.00 confirms this, as it suggests that the model has a perfect separability measure, meaning it can distinguish between positive and negative classes without error. Figure 6d is similar as Figure 6a. In Figure 6b ROC-AUC value is 0.98, which is very close to 1. The curve approaches the left-hand side and the top of the ROC space, indicating high sensitivity (TPR) and high specificity (low FPR). The model maintains a high TPR even as the FPR increases slightly, showing that the classifier is robust across different threshold settings. In Figures 6c, e ROC-AUC value is 0.96, which is very close to 1. The curve approaches the left-hand side and the top of the ROC space, indicating high sensitivity (TPR) and high specificity (low FPR). The model maintains a high TPR even as the FPR increases slightly, showing that the classifier is robust across different threshold settings.

4.3 Impact of feature selection

The PCA feature reduction method enhances the model's assessment precision. Using PCA reduces the number of parameters by exclusively selecting the significant features that account for the explained maximum variance. By implementing this method, the quantity of parameters is drastically reduced. Additionally, the accuracy of testing is marginally enhanced by excluding non-significant features. The comparison of test accuracy performance before and after PCA implementation is presented in Table 2 (feature importance bar graphs, Supplementary Figures S2–S6).

4.4 Impact of balancing class

The bias can result in poor predictive accuracy for the minority class, which is often of greater interest in medical and psychological research, including the diagnosis of ASD. Using the safe-level-SMOTE to balance classes in this study addresses the inherent challenges of imbalanced datasets. Safe-level-SMOTE, an advanced oversampling technique, generates synthetic samples

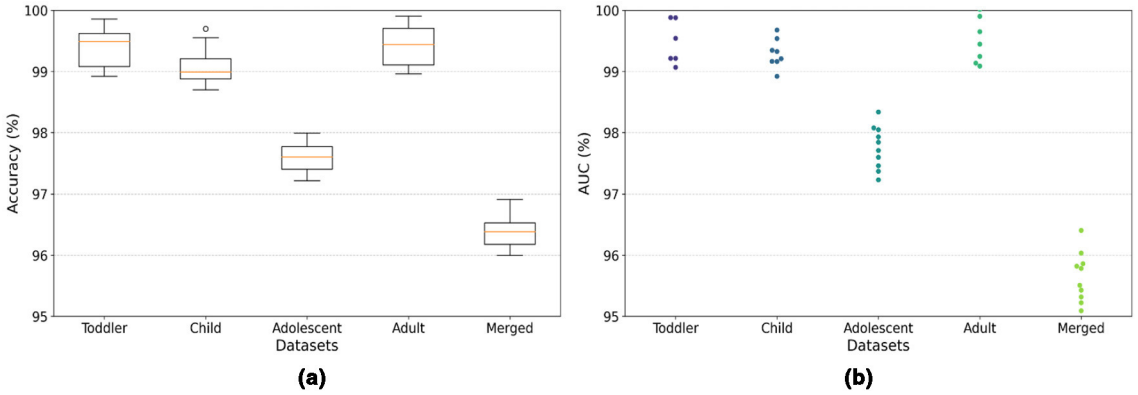

FIGURE 4
(a) Box-and-whisker plot of accuracy, and **(b)** swarm plot of AUC for all five datasets.

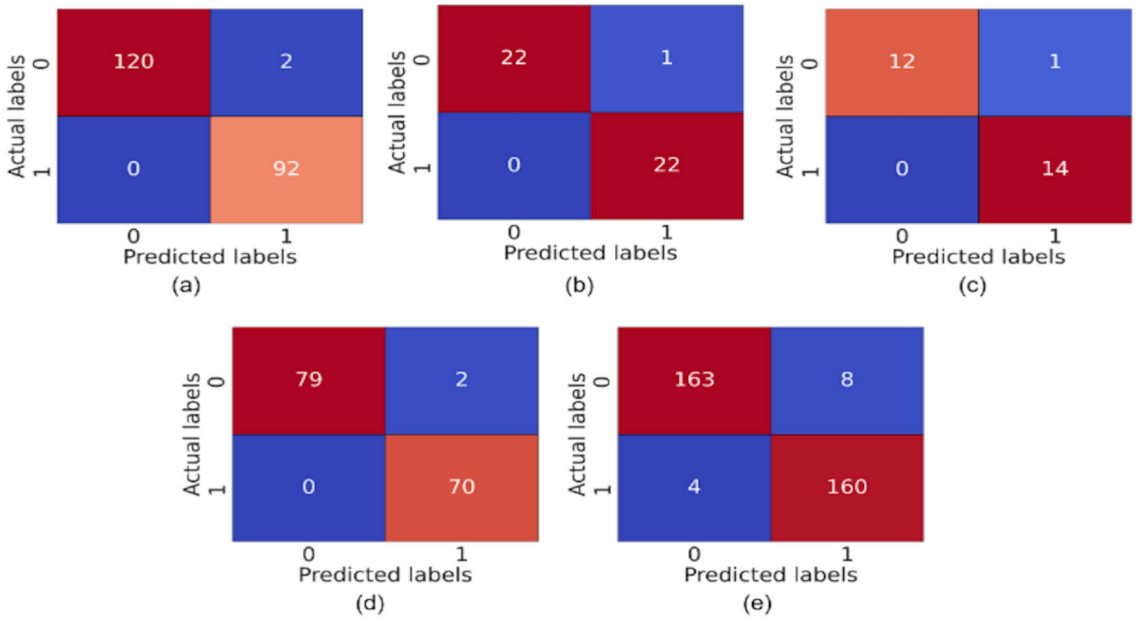

FIGURE 5
Confusion matrix of **(a)** toddler, **(b)** child, **(c)** adolescent, **(d)** adult, and **(e)** merged dataset.

for the minority class based on "safety" levels, which considers the data distribution to create more realistic and representative samples.

Table 3 presents the performance of the proposed model before and after applying Safe-level-SMOTE in the datasets. This enhancement is attributed to the algorithm's ability to mitigate the bias toward the majority class by enriching the dataset with synthetic, yet plausible, minority-class samples. This balanced class distribution allows for a more equitable learning environment, where the classifier can learn to recognize patterns and characteristics of both classes without being overwhelmed by the majority class.

4.5 Statistical analysis

To validate the statistical significance of the proposed model's performance over baseline machine learning models, a Wilcoxon signed-rank test was conducted. This test compared the proposed model against various baseline ML models. The test was applied to two key metrics: Accuracy, and F1-Score, across five datasets. The results, summarized in Table 4, indicate that the p-values for all comparisons are above the significance threshold (0.05). This suggests that while the proposed model consistently achieved high performance across all metrics and datasets, the improvements were not statistically significant compared to the baseline models.

TABLE 1 Results obtained from the proposed model.

Dataset	Class	Precision	Recall	F1-score	Validation accuracy	Validation loss	Accuracy	AUC
1	Class 0	99.97	100.00	99.93	99.95	0.0014	99.86	99.98
	Class 1	99.88	99.97	99.91				
2	Class 0	99.99	98.97	98.38	99.85	0.0412	99.68	99.89
	Class 1	99.99	98.29	98.21				
3	Class 0	99.99	98.38	98.89	99.11	0.0845	98.17	98.16
	Class 1	99.94	95.11	97.47				
4	Class 0	99.98	99.97	99.98	99.96	0.0036	99.89	99.99
	Class 1	99.99	99.97	99.98				
5	Class 0	98.15	97.68	97.87	98.74	0.0078	96.96	96.04
	Class 1	98.34	96.61	97.35				

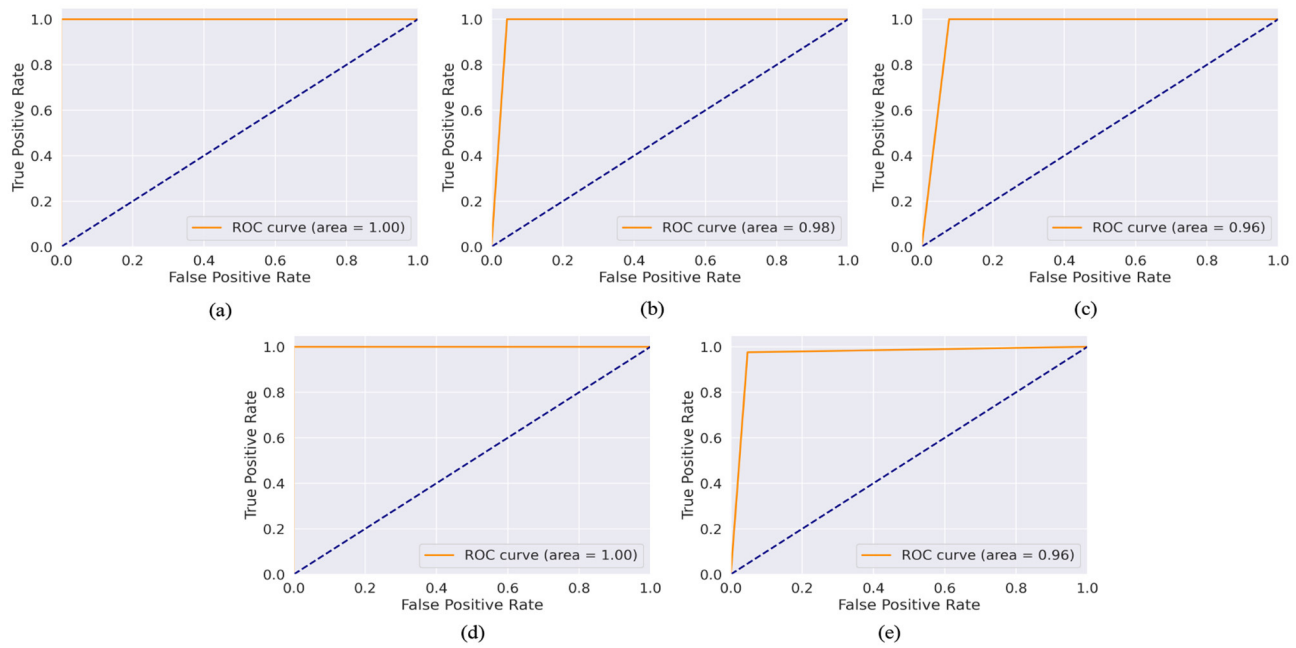

FIGURE 6
AUC-ROC curve of **(a)** toddler, **(b)** child, **(c)** adolescent, **(d)** adult, and **(e)** merged dataset.

TABLE 2 Performance of the proposed model before and after feature selection.

Dataset	Before feature selection					After feature selection				
	Acc	Pre	Rec	F1	AUC	Acc	Pre	Rec	F1	AUC
Toddler	98.12	99.04	97.26	98.58	98.03	99.86	99.92	99.98	99.92	100.00
Child	96.58	98.29	95.27	96.88	97.63	99.68	99.99	98.28	99.27	99.89
Adolescent	96.69	95.97	96.58	95.98	96.89	98.17	99.97	97.07	98.02	98.16
Adult	98.33	97.98	99.06	98.65	97.48	99.89	99.98	99.97	99.98	100.00
Merged	92.88	93.25	91.32	92.47	93.58	96.96	98.25	97.13	97.59	96.04

TABLE 3 Performance of the proposed model before and after addressing class imbalance.

Dataset	Before safe-level SMOTE					After safe-level SMOTE				
	Acc	Pre	Rec	F1	AUC	Acc	Pre	Rec	F1	AUC
Toddler	97.53	99.33	97.51	98.51	97.73	99.86	99.92	99.98	99.92	100.00
Child	97.43	98.64	97.46	96.88	98.90	99.68	99.99	98.28	99.27	99.89
Adolescent	97.39	96.67	97.58	96.78	96.94	98.17	99.97	97.07	98.02	98.16
Adult	97.68	98.59	98.54	98.59	97.77	99.89	99.98	99.97	99.98	100.00
Merged	94.72	94.15	92.37	93.67	94.78	96.96	98.25	97.13	97.59	96.04

TABLE 4 Wilcoxon signed-rank test results for accuracy comparing the proposed model's performance with baseline machine learning models across five datasets.

Dataset	SVM	LR	DT	RF	GB	XGB	CB	ET	KNN	NB	ANN	LDA
Toddler	0.822	1.000	0.118	0.625	0.445	0.605	0.482	0.750	0.220	0.060	0.568	0.215
Children	0.765	0.950	0.102	0.530	0.410	0.589	0.460	0.701	0.200	0.054	0.520	0.198
Adolescent	0.740	0.880	0.150	0.600	0.455	0.620	0.500	0.770	0.210	0.070	0.550	0.210
Adult	0.700	0.910	0.130	0.610	0.420	0.590	0.470	0.740	0.180	0.050	0.530	0.205
Merged	0.750	0.930	0.140	0.620	0.430	0.610	0.480	0.720	0.190	0.055	0.540	0.210

4.6 Handling overfitting

Overfitting, where a model performs well on the training data but poorly on unseen data, is a critical challenge for machine learning models. In this study, we adopted multiple strategies to mitigate overfitting and ensure the generalizability of the proposed ensemble learning framework. First, dimensionality reduction techniques, including PCA and MI-based feature selection, were applied to remove redundant and irrelevant features. This streamlined the model by retaining only the most informative features, reducing the risk of learning noise from the data. Second, within the ANN meta-classifier, dropout layers were employed with a dropout rate of 25%. Dropout is an effective regularization method that prevents overfitting by randomly deactivating neurons during training, which forces the network to learn more generalized patterns. Additionally, we implemented k-fold cross-validation ($k = 5$) to validate the model's robustness. Cross-validation splits the data into multiple training and validation subsets, ensuring the model is evaluated across different data partitions, which reduces variance and improves generalization to unseen data. MCD was applied during inference to estimate uncertainty in predictions to ensure reliability further. MCD helps evaluate the consistency and confidence of the model, enabling the detection of potential overfitting by analyzing output variations on test data. Finally, Safe-Level SMOTE addressed the class imbalance in the datasets, enhancing the model's ability to learn representative patterns from the minority class without overfitting to majority class samples. These combined techniques ensure that the proposed model remains robust, reliable, and free from overfitting, thus enhancing its applicability in real-world scenarios.

4.7 Comparison with existing works

In this research paper, we implemented an innovative methodology for the screening of ASD utilizing a stacked ensemble-based model. This model combines several ML algorithms to analyze the data obtained from the ASD screening questionnaires. The method we have developed is notable for its resilience and precision, supported by the exhaustive comparative analysis provided in the Table 5. The research is situated within the context of prior investigations that have sought to improve the precision and dependability of ASD screening instruments. In brief, the comparative analysis highlights the progress that our stacked EM contributes to ASD screening. The proposed work enhances the continuous endeavor to develop ASD screening tools that are clinically valuable, dependable, and accurate by resolving certain constraints identified in prior studies, including overfitting and the trade-off between recall and precision. The results of our research support the incorporation of advanced ML methods into the evaluation of ASD, which has great potential to advance the detection and treatment of individuals on the spectrum.

5 Explainable AI

SHapley Additive exPlanations (SHAP) is an innovative method within the domain of XAI that provides valuable insights into the results produced by ML models (Lundberg and Lee, 2017). The model-agnostic nature of SHAP is one of its assets. This feature enables XAI to employ diverse models, such as ensemble methods such as RFs and complex architectures like ANNs. SHAP focuses primarily on local explanations; however,

10.3389/frai.2025.1507922

TABLE 5 Comparison with existing works on the used datasets.

Dataset	Reference	Acc	Pre	Rec	F1	AUC
Toddler	Uddin et al., 2023	99.85	**1.00**	**1.00**	99.85	99.85
	Akter et al., 2021a	98.77	-	-	-	**99.98**
	Bala et al., 2022	97.82	-	-	97.8	99.7
	Hasan et al., 2021	99.25	99.89	98.45	99.1	
	Priyadarshini, 2023	99.64	96	94	91	-
	Vakadkar et al., 2021	-	-	-	98.00	-
	Mohanty et al., 2021	85.24	-	-	82.00	-
	Proposed	**99.86**	99.94	99.86	**99.90**	99.95
Child	Talabani and Engin, 2018	92.26	88.09	96.52	-	-
	Abitha et al., 2022	94.1	-	-	-	-
	Omar et al., 2019	92.26	-	-	-	-
	Haroon and Padma, 2022	95.5	98	97	96	-
	Kamma et al., 2022	95.82	-	-	-	-
	Gupta et al., 2022	-	92.59	97.09	94.71	-
	Thabtah, 2019	97.80	-	98	-	-
	Bala et al., 2022	99.61	-	-	**99.60**	99.60
	Akter et al., 2019	97.20	-	-	-	99.89
	Mohanty et al., 2021	84.21	-	-	84.21	-
	Garg et al., 2022	98.00	-	-	-	-
	Hasan et al., 2021	97.95	96.16	97.72	97.02	99.73
	Proposed	**99.68**	**99.97**	**99.03**	99.58	**99.89**
Adolescent	Talabani and Engin, 2018	93.78	89.85	98.4	-	-
	Omar et al., 2019	93.78	-		-	-
	Thabtah, 2019	94.23	-	92.20	-	-
	Kamma et al., 2022	95.82	-	-	-	-
	Gupta et al., 2022	-	93.25	74.15	84.21	-
	Akter et al., 2019	93.89	-	-	-	**98.61**

(Continued)

TABLE 5 (Continued)

Dataset	Reference	Acc	Pre	Rec	F1	AUC
	Bala et al., 2022	95.87	-	-	95.90	99.00
	Mohanty et al., 2021	85.71	-	-	88.52	-
	Hasan et al., 2021	97.12	97.25	**97.36**	97.69	99.72
	Proposed	**98.17**	**99.52**	97.18	**98.46**	98.16
Adults	Priyadarshini, 2023	98.89	94	91	93	
	Shuvo et al., 2019	95.71	-	85.71	-	-
	Kamma et al., 2022	95.82	-	-	-	-
	Gupta et al., 2022	-	97.46	91.27	94.26	-
	(Talabani and Engin, 2018)	96.91	90.07	96.87	-	-
	Akter et al., 2019	98.36	-	-	-	99.95
	Omar et al., 2019	97.10	-	-	-	-
	Abitha et al., 2022	98	-	-	-	-
	Bala et al., 2022	99.82	-	-	99.90	99.80
	Thabtah, 2019	99.85	-	99.90	-	-
	Mohanty et al., 2021	89.26	-	-	85.39	-
	Hasan et al., 2021	99.03	98.16	**100.00**	99.11	**99.99**
	Proposed	**99.89**	**99.98**	99.92	99.95	**99.99**

The bold values indicate the highest results obtained among all.

by aggregating these explanations, one can obtain global insights regarding the model. This facilitates comprehension of the model's overall behavior, including the features that exert the greatest influence and their interrelationships. SHAP ensures explanations remain consistent; if a model undergoes modification resulting in an increase or maintenance of a feature's contribution, its SHAP value will not diminish. This guarantees that the explanations accurately represent the model's behavior.

Several stages are required to implement SHAP in a binary classification task utilizing a stacked EM consisting of multiple ML models as the base classifier and an ANN as the meta-classifier. Predictions were initially generated by each of the fundamental classifiers, namely the RF, ET, and CB. For each class, these predictions are presented as probability distributions. Subsequently, the predictions generated by the base classifiers were employed as meta-features for the ANN meta-classifier. Each feature in the intermediate dataset corresponds to the output of one of the base models. After the meta-classifier has been trained, its predictions are interpreted using SHAP. Being model-independent and compatible with ANNs, SHAP may be implemented directly on the meta-classifier. Currently, the SHAP values for the predictions produced by the meta-classifier can be generated. The magnitude to which the final decision for each class was influenced by the output (now a meta-feature) of each base classifier will be denoted by these values. A SHAP summary plot was utilized to provide clarification.

SHAP was implemented individually on each base classifier to ascertain how the input features impact their respective predictions. This dual-level explanation of the stacked EM (at both the base classifier and meta-classifier levels) provides a comprehensive understanding of the model. Furthermore, SHAP has been used for both local and global interpretation. For each specific class, a summary plot has been generated in local interpretation. Four graphs, one for each of the four classes, have been produced in total for the classifier. A summary plot has been produced for global interpretation by aggregating SHAP values. This plot provides valuable insights into the model's overall behavior.

Figures 7–11 presents the explanations generated using SHAP on both meth classifier and base classifier for Toddler, Child, Adolescent, Adult, and Merged dataset respectively. Each figure is sepearted in two major parts that includes explaining the meta

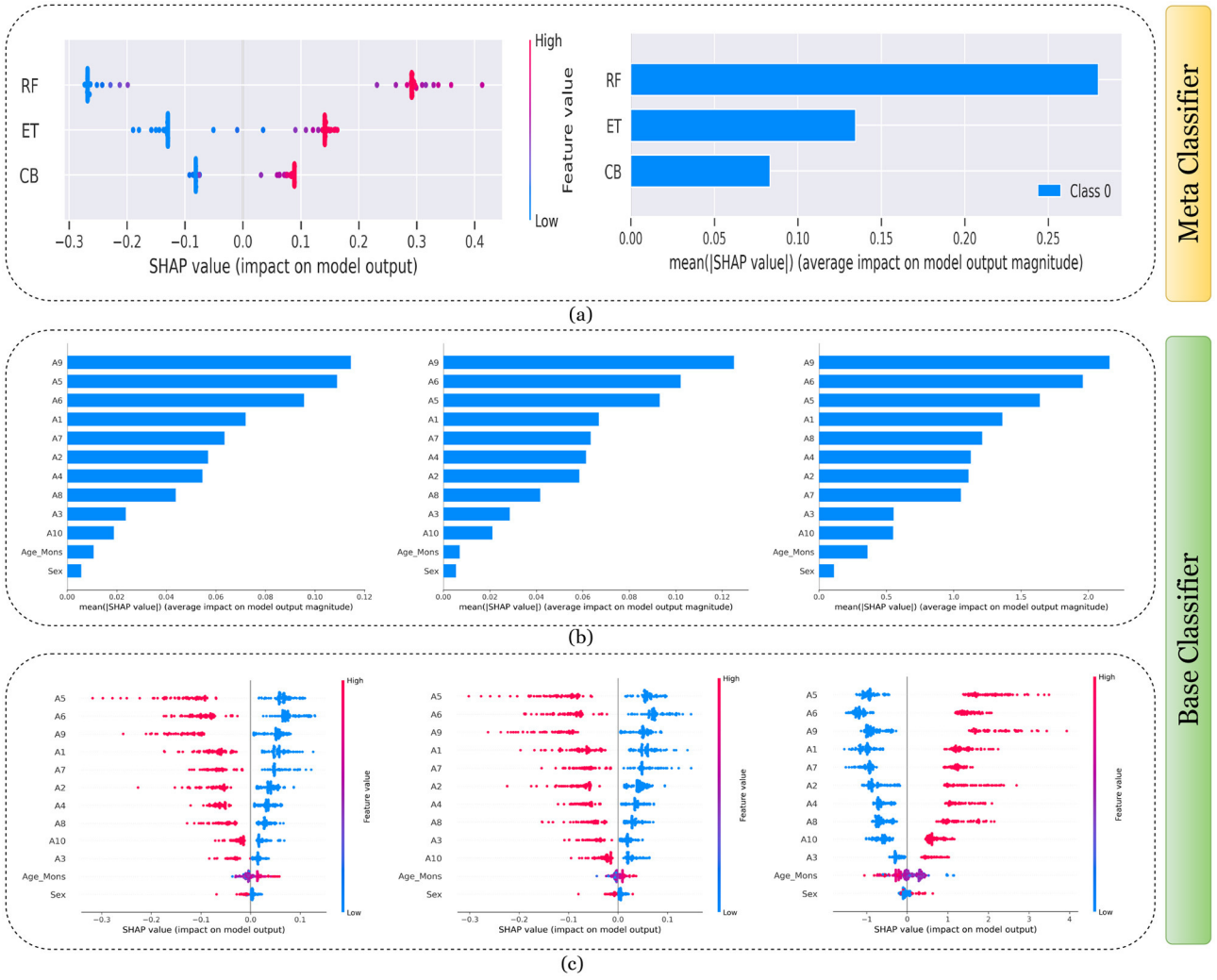

FIGURE 7
Explanation of models using SHAP on Toddler dataset. **(a)** Explaining meta-classifier using dot plot (left), and summary plot (right). Explanation of base classifiers, RF (left), ET (middle), CB (right) using **(b)** summary plot, and **(c)** dot plot.

classifier using classwise summary plot (left), and overall summary plot (right), and then explaining the base classifiers utilizing overall summary plot (middle) and classwise summary plot (below). For example, a classwise summary plot (left) and overall summary plot(right) generated from meta classifier for Toddler dataset has been provided in Figure 7a. From the classwise summary plot (left), it can be observed that most SHAP values are centered around zero but show a spread on both the negative and positive sides, indicating that features both positively and negatively influence the model's output for RF model.

The ET model's SHAP values show a similar pattern to those of the RF model, with a distribution around zero and a spread to both sides, again suggesting a mix of positive and negative feature impacts, while the CB model's SHAP values are more concentrated around zero compared to the RF and ET models, with fewer extreme positive or negative values. This suggests that individual features may have a more uniform influence on the CB model's output. The presence of SHAP values above and below zero across the models indicates that features increase and decrease the likelihood of a positive class prediction. Meanwhile, in the classwise summary plot (left), SHAP values quantify the impact of features to the model's prediction, and the mean provides an overall measure of the different features' impact on the model's output. It can be observed that the RF model has a greater mean, indicating that, on average, its features have a substantial influence on the model output. The ET model has the second-highest mean, which is slightly less than that of the RF model, suggesting that its features have a strong but slightly lesser impact on model output than the RF model. The CB model has the lowest mean value among the three models, indicating that its features, on average, have less impact on the model output. The term "Class 0" in the legend suggests that these SHAP values are associated with the impact on a specific class in a classification problem, likely the negative class if we assume a binary classification task.

Findings from this chart imply that the RF model relies more heavily on individual features for making predictions or it has a

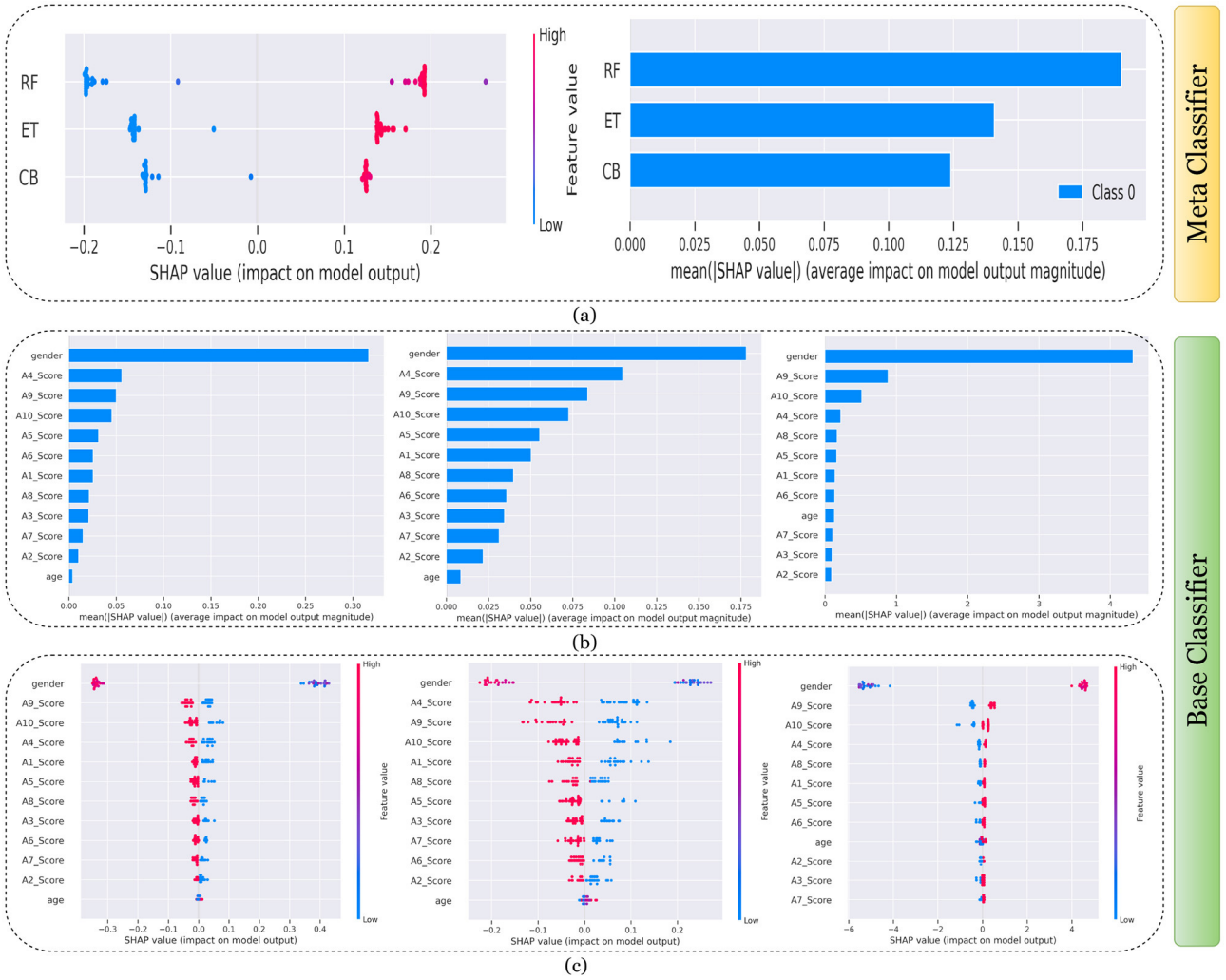

FIGURE 8
Explanation of models using SHAP on Child dataset. **(a)** Explaining meta-classifier using dot plot (left), and summary plot (right). Explanation of base classifiers, RF (left), ET (middle), CB (right) using **(b)** summary plot, and **(c)** dot plot.

few features with powerful impacts. In contrast, the CB model's predictions seem to be influenced less by individual features or have a more distributed influence across features. This could reflect differences in how the models handle feature interactions or their inherent algorithmic biases. Similarly, the classwise summary plot(middle) and overall summary plot (below) generated from the base classifier for the Toddler dataset have been provided in Figures 7b, c. In Figures 7b, c the predictions from RF (left), ET (middle), and CB (right) is provided. Figure 7b is a horizontal bar chart depicting the mean values for various features in a predictive model generated from RF (left), ET (middle), and CB (below). Figure 7c shows a detailed SHAP value scatter plot for various features in the base classifiers RF (left), ET (middle), and CB (below). SHAP values depict the impact of a given feature on the model's output for a prediction, with positive values indicating an increase in the likelihood of a particular outcome and negative values indicating a decrease. Similarly, Figures 8–11 can be interpreted to find the explanation provided by the models on the datasets.

The SHAP analysis provides critical insights into the significance of various features in predicting ASD across different age groups. The analysis highlights how feature importance shifts with developmental stages, reflecting the evolving nature of ASD markers over time. For example, A8, a behavioral trait, is less influential in toddlers but becomes a pivotal feature in adults, likely due to its association with advanced cognitive and social functions, such as abstract reasoning, self-awareness, and complex social interactions. These traits typically emerge in later developmental stages, making A8 more relevant in understanding ASD in adults. In contrast, features like A1 and A3 dominate the toddler dataset, as they are linked to early developmental markers such as sensory processing, responsiveness to stimuli, and essential social engagement, which are critical indicators of ASD at a younger age.

In children and adolescents, features such as A9 and A10 gain prominence, suggesting that adaptive behaviors and developmental milestones increase as children age and begin navigating more structured environments like school and peer interactions. The adolescent dataset further emphasizes features

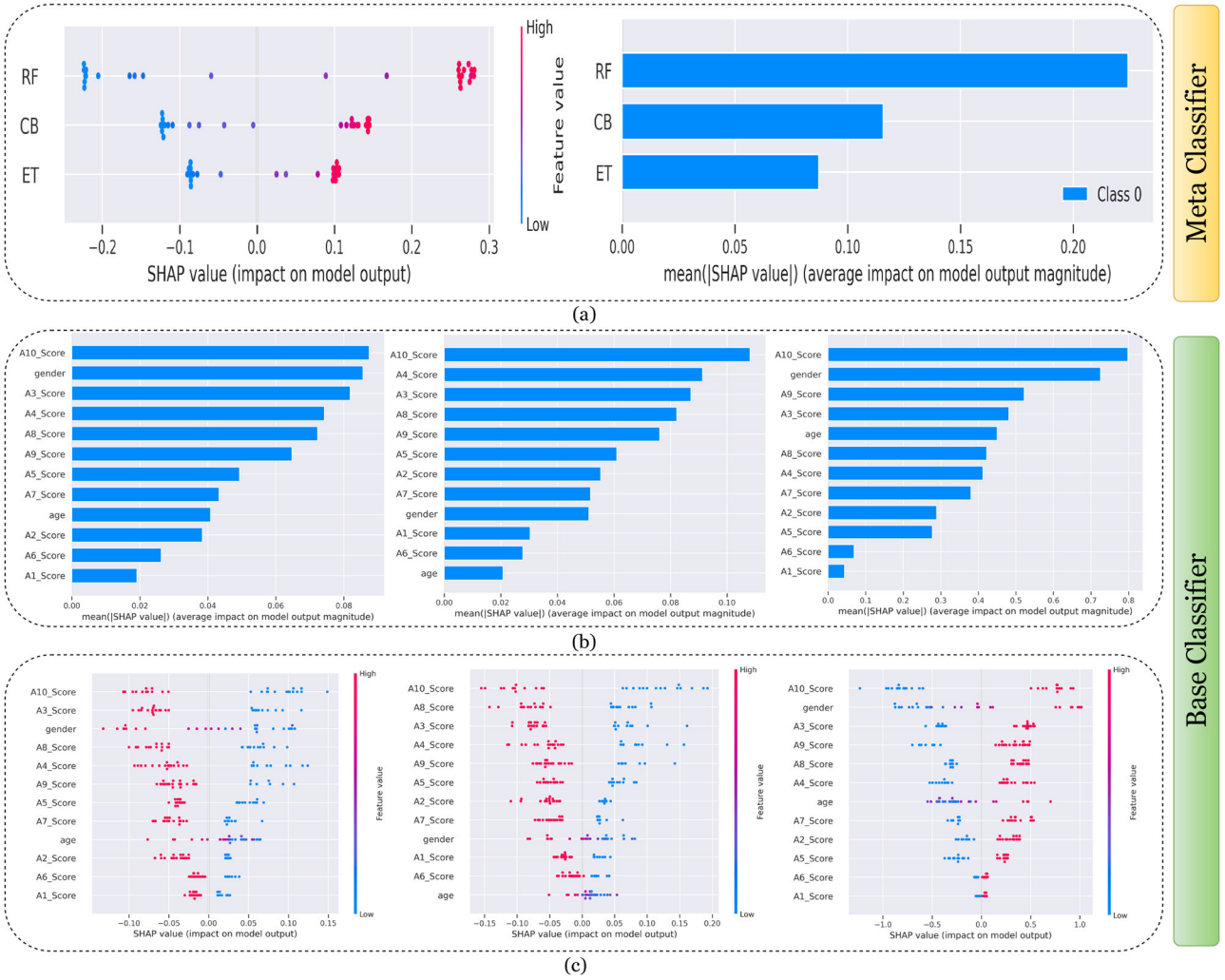

FIGURE 9
Explanation of models using SHAP on Adolescent dataset. **(a)** Explaining meta-classifier using dot plot (left), and summary plot (right). Explanation of base classifiers, RF (left), ET (middle), CB (right) using **(b)** summary plot, and **(c)** dot plot.

like A3 and A8, reflecting the developmental emergence of independence and higher-order cognitive abilities that are relevant during this transitional stage. For adults, behavioral and cognitive traits, represented by features like A8, become more critical as they pertain to advanced social, occupational, and emotional functioning, which are often areas of challenge for adults with ASD. The merged dataset presents a balanced representation of features, such as A1, A8, and A9, broadly significant across all age groups.

These variations underscore the dynamic nature of ASD manifestations and the necessity for diagnostic models tailored to specific age groups as the traits and behaviors associated with ASD evolve significantly over time. For instance, toddlers might display ASD-related traits primarily through sensory responses and early social behaviors, while adults may exhibit challenges in abstract reasoning, emotional regulation, and nuanced social interactions. This contextual understanding of feature importance across developmental stages provides actionable insights for clinicians and researchers and reinforces the importance of considering age-specific diagnostic markers.

6 Uncertainty analysis

UA is a technique utilized in the domain of ANNs to quantify a network's confidence level in its predictions (Bachstein, 2019). Neural networks, particularly DL models, are often regarded as opaque models that generate predictions without disclosing their level of certainty. To address this concern, the UA incorporates a confidence level into predictions. This notion is of notable significance in domains where decisions based on model predictions entail substantial implications. The Monte Carlo dropout (MCD) method (Gal and Ghahramani, 2016) is a technique used to quantify uncertainty in DL models. Estimating model uncertainty is a computationally efficient and practicable process that holds significant importance in numerous applications where confidence in model predictions is critical for decision-making. Initially, dropout was implemented in ANNs as a regularization technique to avert overfitting (Ahmed et al., 2023). MCD can be utilized to detect instances in which the predictions generated by the model are deemed unreliable. We have utilized

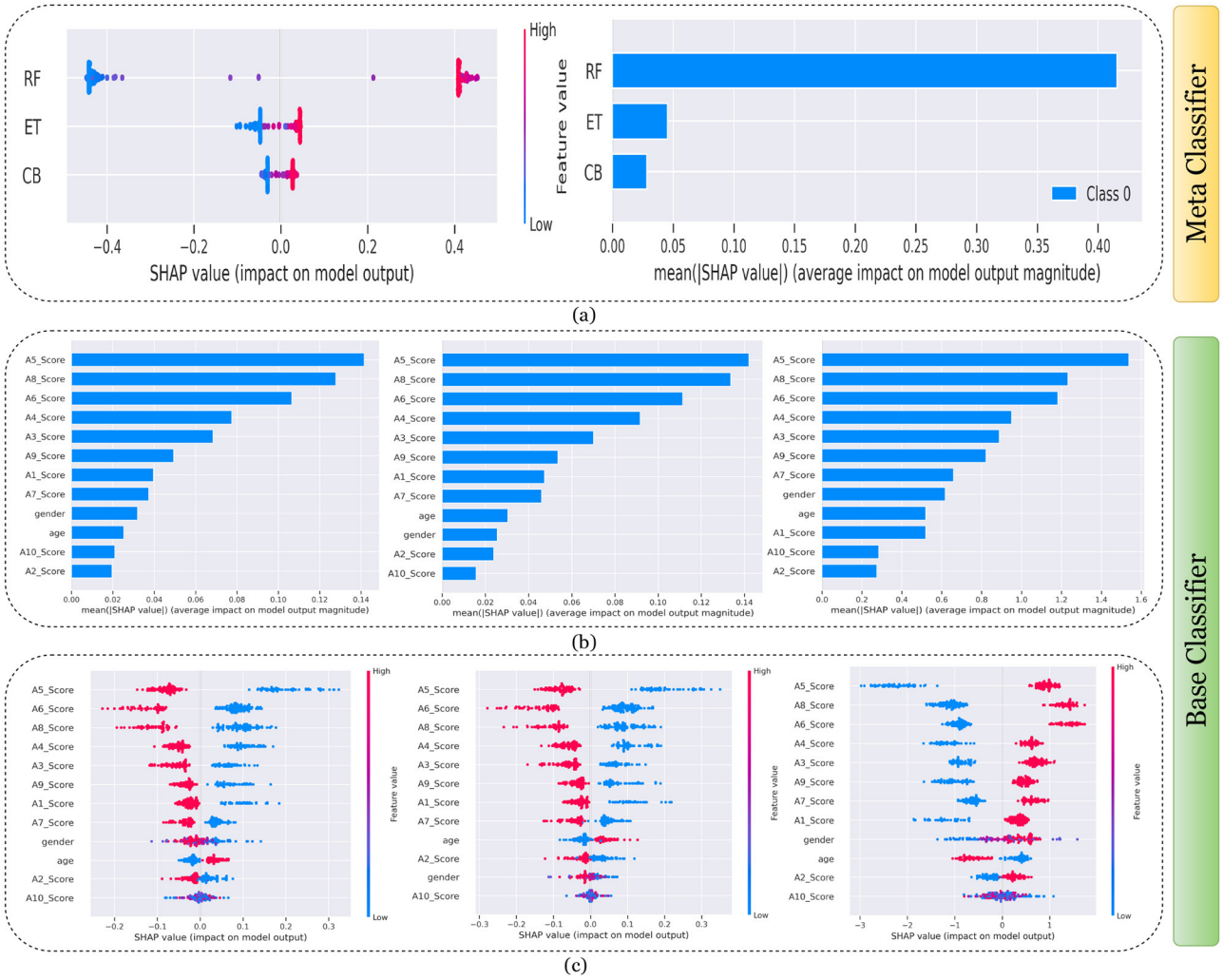

FIGURE 10
Explanation of models using SHAP on Adult dataset. **(a)** Explaining meta-classifier using dot plot (left), and summary plot (right). Explanation of base classifiers, RF (left), ET (middle), CB (right) using **(b)** summary plot, and **(c)** dot plot.

MCD inference on individual samples, a calibration curve, variance analysis, and standard deviation analysis.

Figures 12–16 present the UA generated using MCD on both meth classifiers for Toddler, Child, Adolescent, Adult and Merged dataset, respectively. Each figure is comprised of six subplots, including (a) the calibration curve, (b) the predicted class probability plot, (c) the scatter plot with an error bar, (d) the standard deviation distribution, (e) the predictive variance distribution, and (f) the predictive entropy distribution. A calibration curve is a graphical representation that compares the mean predicted value of a probabilistic classifier with the actual fraction of positives. This plot is typically used to assess the reliability of probabilistic predictions made by a model. A well-calibrated model means that if the model predicts an event with a probability of "p", then "p" percent of the time that event should occur. If the model is perfectly calibrated, the plot of the model's predictions would lie on the diagonal line representing the "Perfectly calibrated" classifier. Deviations from this line indicate a model whose probabilities are either over- or under-confident.

For the Toddler dataset, the UA is presented in Figures 12. Firstly, the calibration curve is provided in Figure 12a. The reference line, labeled "Perfectly calibrated", is a dotted line that forms a 45-degree angle, indicative of a hypothetical model where the predicted probabilities perfectly match the observed proportions. The "Meta Classifier" calibration curve closely follows the "Perfectly calibrated" reference line, suggesting that the classifier's predicted probabilities are well-calibrated. The squares on the "Meta Classifier" line potentially represent binned average predictions compared to the actual outcomes within those bins. Secondly, the predicted class probability is provided in Figure 12b. There are two prominent peaks, one at the extreme left (near 0.0) and one at the extreme right (near 1.0), which means the model is often very confident in its predictions, assigning probabilities close to 0 or 1. The frequency of predicted probabilities near 0.0 is slightly higher than those near 1.0. This suggests that the model predicts the negative class more frequently than the positive class or that there are more instances of the negative class in the dataset. A noticeable absence of predicted probabilities in the

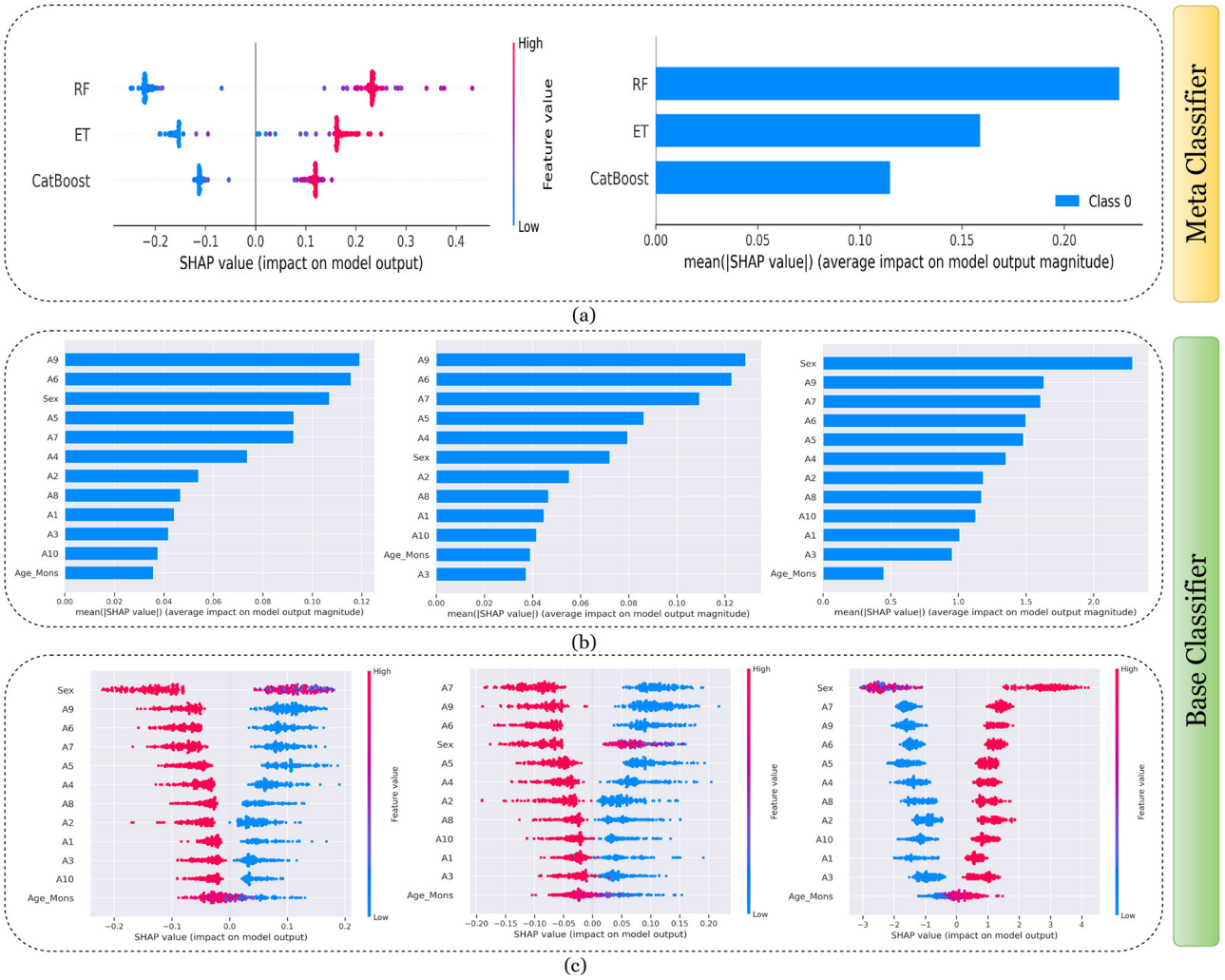

FIGURE 11
Explanation of models using SHAP on Merged dataset. **(a)** Explaining meta-classifier using dot plot (left), and summary plot (right). Explanation of base classifiers, RF (left), ET (middle), CB (right) using **(b)** summary plot, and **(c)** dot plot.

middle range (from 0.2 to 0.8) indicates that the model rarely assigns intermediate probabilities and is generally certain about its predictions. Figures 12c presents the scatter plot with error bars. The scatter plot illustrates the predictions and associated uncertainties for a series of samples. Predominantly, the predictions align with absolute certainty at 0.0 and 1.0, implying high confidence in these outcomes. A few predictions demonstrate considerable uncertainty, as evidenced by the longer error bars. These instances of increased uncertainty are interspersed without a clear pattern across the sample index. The visualization highlights the model's confidence in most predictions while acknowledging uncertainty in a subset of cases. This underscores the importance of accounting for error margins in predictive analysis, especially when utilizing these predictions for further decision-making processes. Lastly, Figures 12d–f present the distribution of predictive standard deviation, variance, and entropy. In Figure 12d, most of the data points have a very low standard deviation, close to 0.00. This is indicated by the tall bar at the extreme left of the histogram. There is a rapid decrease in frequency as the standard deviation increases. After the initial tall bar, subsequent bars are significantly shorter, showing that higher standard deviations are much less common in this dataset. The distribution is heavily skewed to the left, meaning that there is a higher concentration of lower standard deviation values and very few high standard deviation values. In Figure 12e, the histogram displays the distribution of predictive variances from a set of model predictions. A pronounced peak at a predictive variance of 0.0 indicates that nearly all predictions have no variance, implying a high degree of certainty or consistency in the model's output. The absence of visible frequencies for non-zero variances suggests either an absence or an insignificant number of predictions with any measurable uncertainty. In Figure 12f, there is a significant concentration of predictions with an entropy close to 0.0, as evidenced by the tall bar at the beginning of the histogram. This suggests that for many predictions, the model is very certain about the outcome. The frequency of predictions decreases sharply as entropy increases. There are very few predictions with higher entropy values, which would indicate uncertainty in the model's predictions. The histogram does not show any occurrences of

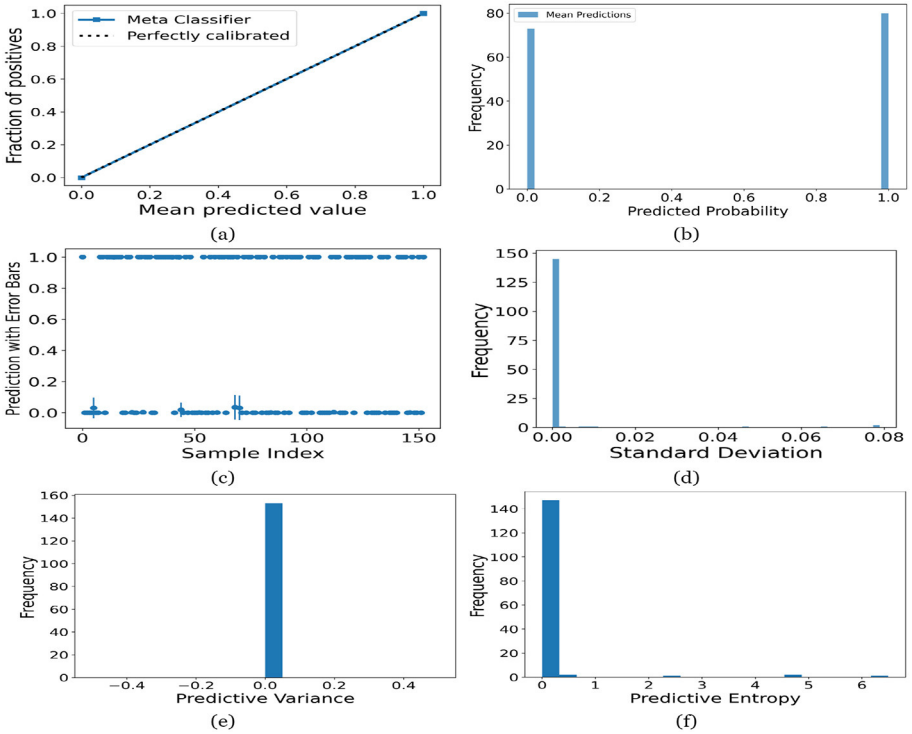

FIGURE 12
UA on Toddler dataset utilizing **(a)** calibration curve, **(b)** mean probability bar graph, **(c)** predictive error bar graph, **(d)** predictive standard deviation, **(e)** predictive variance, and **(f)** predictive entropy.

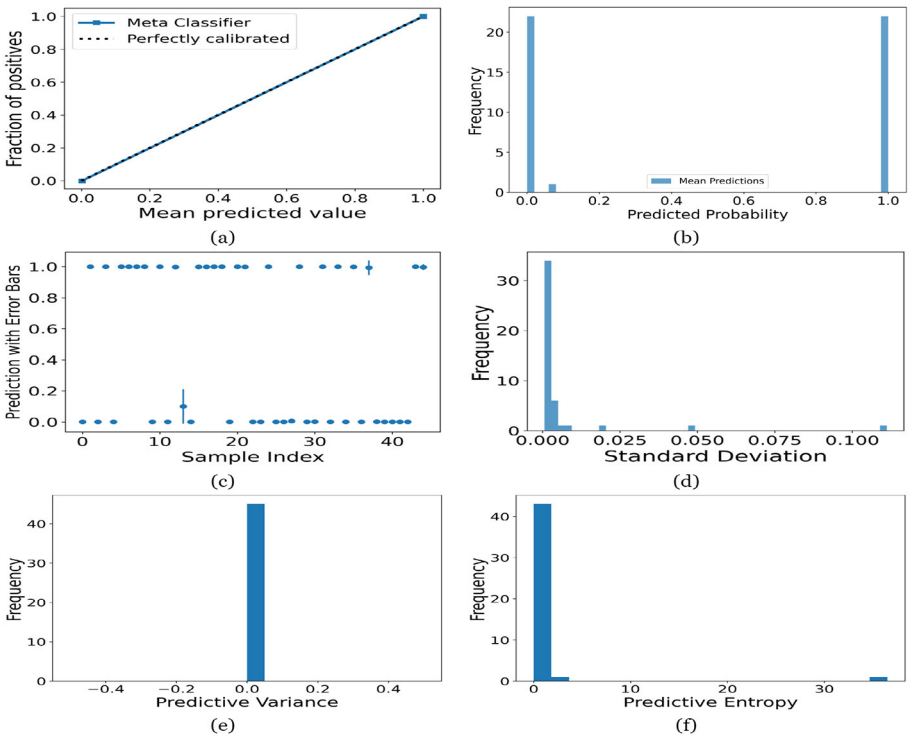

FIGURE 13
UA on Child dataset utilizing **(a)** calibration curve, **(b)** mean probability bar graph, **(c)** predictive error bar graph, **(d)** predictive standard deviation, **(e)** predictive variance, and **(f)** predictive entropy.

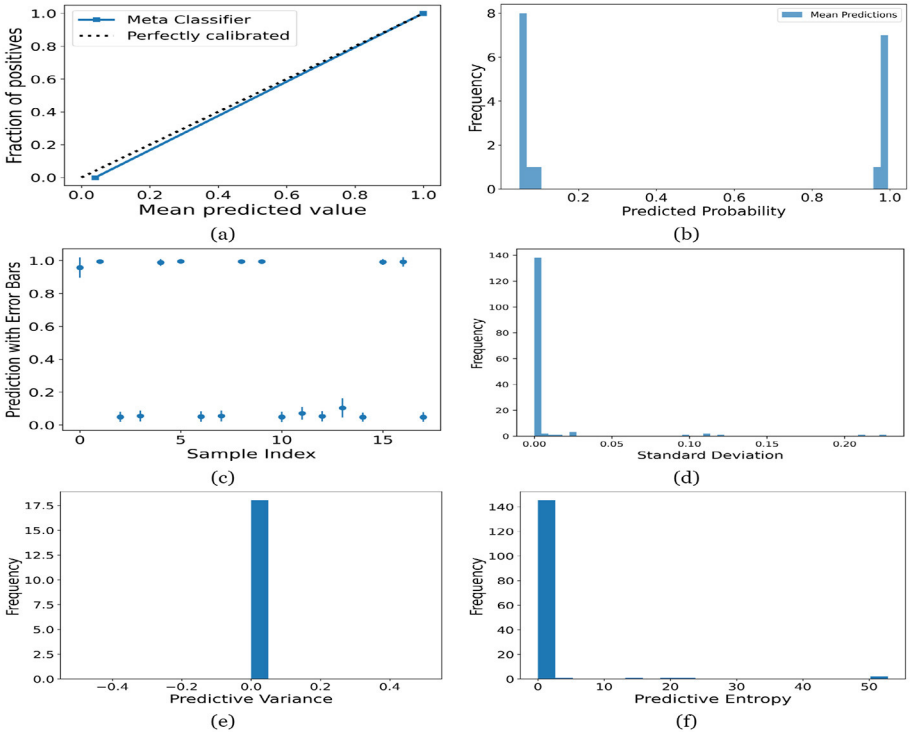

FIGURE 14
UA on Adolescent dataset utilizing **(a)** calibration curve, **(b)** mean probability bar graph, **(c)** predictive error bar graph, **(d)** predictive standard deviation, **(e)** predictive variance, and **(f)** predictive entropy.

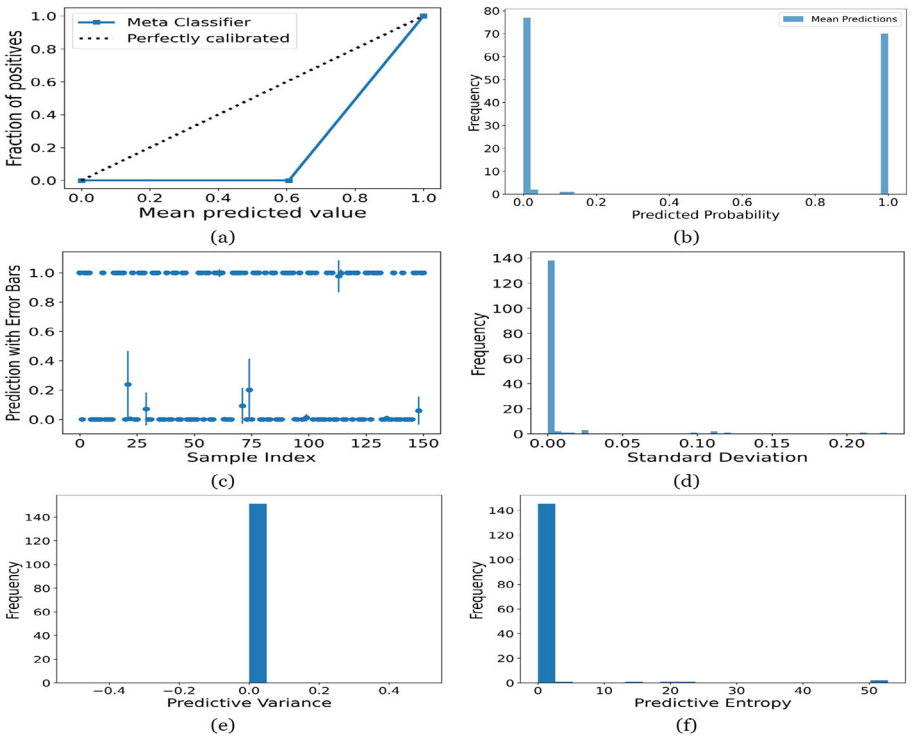

FIGURE 15
UA on Adult dataset utilizing **(a)** calibration curve, **(b)** mean probability bar graph, **(c)** predictive error bar graph, **(d)** predictive standard deviation, **(e)** predictive variance, and **(f)** predictive entropy.

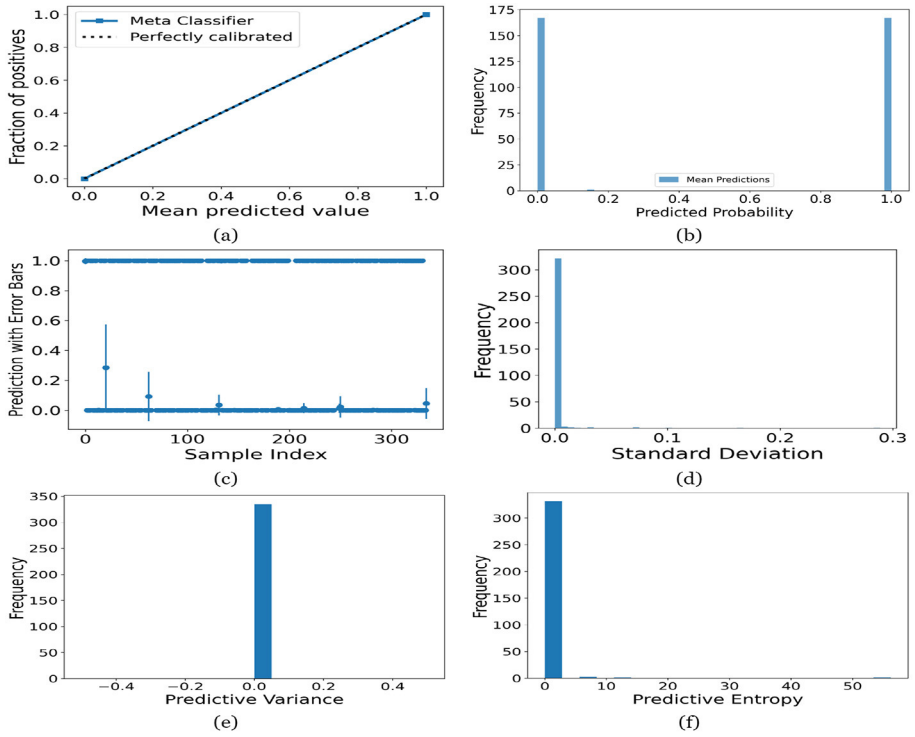

FIGURE 16
UA on Merged dataset utilizing **(a)** calibration curve, **(b)** mean probability bar graph, **(c)** predictive error bar graph, **(d)** predictive standard deviation, **(e)** predictive variance, and **(f)** predictive entropy.

predictions with entropy near 1.0. This would mean there are no instances where the model is completely uncertain about the outcome. Figures 13–16 can be interpreted in the similar way.

We assessed the impact of MCD, which quantifies prediction uncertainty. Removing MCD resulted in reduced reliability, as the model failed to estimate confidence in its predictions. This highlighted the importance of incorporating uncertainty analysis to make the framework robust and trustworthy for clinical applications.

7 Conclusion

In conclusion, this study proposed a novel ensemble learning framework for the classification of ASD using questionnaire data, integrating Safe-Level SMOTE for class imbalance handling, SHAP for model interpretability, and MCD for uncertainty estimation. Compared to baseline ML methods, the model demonstrated exceptional performance across five publicly available datasets representing multiple developmental stages, with high accuracy, transparency, and reliability.

However, this study has certain limitations that warrant further consideration. From a theoretical perspective, the reliance on questionnaire data introduces subjectivity and potential reporting biases, which may affect the accuracy and generalizability of the model. Additionally, while the stacked ensemble learning approach leverages the strengths of multiple classifiers, it requires significant computational resources for both training and hyperparameter tuning. This may limit its deployment in environments where computational power is constrained. Another theoretical limitation lies in the binary classification setup, as this study does not yet address the multi-classification of ASD severity, which could provide more granular insights into the disorder.

From a practical standpoint, the datasets utilized, while publicly accessible and diverse across age groups, may not comprehensively represent the full heterogeneity of the ASD population, particularly across different geographical and socio-demographic backgrounds. Furthermore, the lack of real-world clinical validation is a significant limitation, as the model's robustness and reliability have not yet been tested within practical healthcare workflows or clinical environments.

To address these limitations, future work will explore integrating hybrid metaheuristic optimization techniques, such as swarm intelligence, to enhance feature selection and model performance. Incorporating multi-modal data, including neuroimaging and behavioral assessments alongside questionnaire data, can further strengthen the predictive capability and generalizability of the framework. Additionally, collaboration with clinicians will enable real-world validation and facilitate the model's integration into clinical decision-making systems. Extending the current binary classification model to handle multi-class classification tasks will allow for more refined predictions, such as ASD severity levels or subtypes. Finally, efforts will be made to optimize the model for lightweight deployment, ensuring accessibility in real-time systems and resource-limited environments.

By addressing these theoretical and practical limitations through the outlined future directions, the proposed framework has the potential to evolve into a reliable, scalable, and clinically applicable tool for early ASD detection, contributing to improved diagnostic capabilities and better outcomes for individuals with ASD.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://archive.ics.uci.edu/dataset/426/autism+screening+adult and https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children.

Author contributions

NM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MR: Investigation, Validation, Writing – original draft, Writing – review & editing. MY: Supervision, Writing – review & editing. FN: Validation, Writing – review & editing, Resources. MU: Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1507922/full#supplementary-material

References

Abbas, H., Garberson, F., Glover, E., and Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* 25, 1000–1007. doi: 10.1093/jamia/ocy039

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inform. Fusion* 76, 243–297. doi: 10.1016/j.inffus.2021.05.008

Abitha, R., Vennila, S. M., and Zaheer, I. M. (2022). Evolutionary multi-objective optimization of artificial neural network for classification of autism spectrum disorder screening. *J. Supercomput.* 78, 11640–11656. doi: 10.1007/s11227-021-04268-4

Ahmed, S., Yousuf, M. A., Monowar, M. M., Hamid, M. A., and Alassafi, M. (2023). Taking all the factors we need: a multimodal depression classification with uncertainty approximation. *IEEE Access.* 11, 99847–99861. doi: 10.1109/ACCESS.2023.3315243

Akter, T., Ali, M. H., Satu, M. S., Khan, M. I., and Mahmud, M. (2021a). "Towards autism subtype detection through identification of discriminatory factors using machine learning," in *International Conference on Brain Informatics* (Cham: Springer), 401–410.

Akter, T., Khan, M. I., Ali, M. H., Satu, M. S., Uddin, M. J., and Moni, M. A. (2021b). "Improved machine learning based classification model for early autism detection," in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (Dhaka: IEEE), 742–747.

Akter, T., Satu, M. S., Khan, M. I., Ali, M. H., Uddin, S., Lio, P., et al. (2019). Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access* 7, 166509–166527. doi: 10.1109/ACCESS.2019.2952609

Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., and Alazzawi, A. K. (2020). Ai meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE Access* 8, 142532–142542. doi: 10.1109/ACCESS.2020.3013699

Bachstein, S. (2019). *Uncertainty Quantification in Deep Learning* (Master Thesis). Ulm: Ulm University.

Bala, M., Ali, M. H., Satu, M. S., Hasan, K. F., and Moni, M. A. (2022). Efficient machine learning models for early stage detection of autism spectrum disorder. *Algorithms* 15:166. doi: 10.3390/a15050166

Bastiaansen, J. A., Meffert, H., Hein, S., Huizinga, P., Ketelaars, C., Pijnenborg, M., et al. (2011). Diagnosing autism spectrum disorders in adults: the use of autism diagnostic observation schedule (ADOS) module 4. *J. Autism Dev. Disord.* 41, 1256–1266. doi: 10.1007/s10803-010-1157-x

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test* 25, 197–227. doi: 10.1007/s11749-016-0481-7

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intellig. Res.* 16, 321–357. doi: 10.1613/jair.953

Choudhury, M., Tanvir, M., Yousuf, M. A., Islam, N., and Uddin, M. Z. (2025). Explainable AI-driven scalogram analysis and optimized transfer learning for sleep apnea detection with single-lead electrocardiograms. *Comput. Biol. Med.* 187:109769. doi: 10.1016/j.compbiomed.2025.109769

De Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., et al. (2004). Interrelationship between autism diagnostic observation schedule-generic (ADOS-G), autism diagnostic interview-revised (ADI-R), and the diagnostic and statistical manual of mental disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *J. Autism Dev. Disord.* 34, 129–137. doi: 10.1023/B:JADD.0000022604.22374.5f

Devika Varshini, G., and Chinnaiyan, R. (2020). Optimized machine learning classification approaches for prediction of autism spectrum disorder. *Ann. Autism Dev. Disord.* 1:1001.

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York: PMLR), 1050–1059.

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: a review. *Eng. Appl. Artif. Intell.* 115:105151. doi: 10.1016/j.engappai.2022.105151

Garg, A., Parashar, A., Barman, D., Jain, S., Singhal, D., Masud, M., et al. (2022). Autism spectrum disorder prediction by an explainable deep learning approach. *Comp. Mater. Continua* 71, 1459–1471. doi: 10.32604/cmc.2022.022170

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al. (2023). A survey of uncertainty in deep neural networks. *Artif. Intellig. Rev.* 56, 1513–1589. doi: 10.1007/s10462-023-10562-9

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Gupta, U., Gupta, D., and Agarwal, U. (2022). "Analysis of randomization-based approaches for autism spectrum disorder," in *Pattern Recognition and Data Analysis with Applications* (Cham: Springer), 701–713.

Hajjej, F., Ayouni, S., Alohali, M. A., and Maddeh, M. (2024). Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques. *IEEE Access* 12, 35448–35461. doi: 10.1109/ACCESS.2024.3349988

Haroon, A. S., and Padma, T. (2022). An ensemble classification and binomial cumulative based PCA for diagnosis of parkinson's disease and autism spectrum disorder. *Int. J. Syst. Assuran. Eng. Managem.* 15, 216–231. doi: 10.1007/s13198-022-01699-x

Hasan, M., Ahamad, M. M., Aktar, S., and Moni, M. A. (2021). "Early stage autism spectrum disorder detection of adults and toddlers using machine learning models," in *2021 5th International Conference on Electrical Information and Communication Technology (EICT)* (Khulna: IEEE), 1–6.

Hasan, S. M., Uddin, M. P., Al Mamun, M., Sharif, M. I., Ulhaq, A., and Krishnamoorthy, G. (2022). A machine learning framework for early-stage detection of autism spectrum disorders. *IEEE Access* 11, 15038–15057. doi: 10.1109/ACCESS.2022.3232490

Kamma, S. P., Bano, S., Niharika, G. L., Chilukuri, G. S., and Ghanta, D. (2022). "Cost-effective and efficient detection of autism from screening test data using light gradient boosting machine," in *Intelligent Sustainable Systems: Proceedings of ICISS 2021* (Cham: Springer), 777–789.

Kampa, L., Yamini, K., Basavaraju, A., and Anoop, K. (2022). A stack based ensemble learning method for diagnosing autism spectrum disorder. *Mathem. Statist. Eng. Appl.* 71, 237–251.

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 30.

Mohanty, A. S., Parida, P., and Patra, K. (2021). "Identification of autism spectrum disorder using deep neural network," in *Journal of Physics: Conference Series* (Bristol: IOP Publishing), 1921.

Mujeeb Rahman, K., and Monica Subashini, M. (2022). A deep neural network-based model for screening autism spectrum disorder using the quantitative checklist for autism in toddlers (QCHAT). *J. Autism Dev. Disord.* 52, 2732–2746. doi: 10.1007/s10803-021-05141-2

Mukherjee, P., Sadhukhan, S., Godse, M., and Solutions, V. I. (2023). A review of machine learning models to detect autism spectrum disorders (ASD). *WSEAS Trans. Comp.* 22, 177–189. doi: 10.37394/23205.2023.22.21

Mumenin, N., Islam, M. F., Chowdhury, M. R. Z., and Yousuf, M. A. (2023). "Diagnosis of autism spectrum disorder through eye movement tracking using deep learning," in *Proceedings of International Conference on Information and Communication Technology for Development: ICICTD 2022* (Cham: Springer), 251–262.

Mumenin, N., Yousuf, M. A., Alassafi, M. O., Monowar, M. M., and Hamid, M. A. (2025). DDNet: A robust, and reliable hybrid machine learning model for effective detection of depression among university students. *IEEE Access* 2025, 155–159. doi: 10.1109/ACCESS.2025.3552041

Mumenin, N., Yousuf, M. A., Nashiry, M. A., Azad, A., Alyami, S. A., Lio', P., et al. (2024). ASDNet: A robust involution-based architecture for diagnosis of autism spectrum disorder utilising eye-tracking technology. *IET Comp. Vision* 18, 666–681. doi: 10.1049/cvi2.12271

Naimi, A. I., and Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. *Eur. J. Epidemiol.* 33:459–464. doi: 10.1007/s10654-018-0390-z

Obilor, E. I., and Amadi, E. C. (2018). Test for significance of pearson's correlation coefficient. *Int. J. Innovat. Mathem. Statist. Ener. Policies* 6, 11–23.

Omar, K. S., Mondal, P., Khan, N. S., Rizvi, M. R. K., and Islam, M. N. (2019). "A machine learning approach to predict autism spectrum disorder," in *2019 International conference on electrical, computer and communication engineering (ECCE)* (Cox'sBazar: IEEE), 1–6.

Priyadarshini, I. (2023). Autism screening in toddlers and adults using deep learning and fair ai techniques. *Future Intern.* 15:292. doi: 10.3390/fi15090292

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 31.

Rincy, T. N., and Gupta, R. (2020). "Ensemble learning techniques and its efficiency in machine learning: a survey," in *2nd International Conference on Data, Engineering and Applications (IDEA)* (Bhopal: IEEE), 1–6.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* [preprint] arXiv:1708.08296. doi: 10.48550/arXiv.1708.08296

Satu, M. S., Sathi, F. F., Arifen, M. S., Ali, M. H., and Moni, M. A. (2019). "Early detection of autism by extracting features: a case study in bangladesh," in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (Dhaka: IEEE), 400–405.

Sesmero, M. P., Ledezma, A. I., and Sanchis, A. (2015). Generating ensembles of heterogeneous classifiers using stacked generalization. *Wiley Interdiscipl. Rev.* 5, 21–34. doi: 10.1002/widm.1143

Shuvo, S. B., Ghosh, J., and Oyshi, A. S. (2019). "A data mining based approach to predict autism spectrum disorder considering behavioral attributes," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (Kanpur: IEEE), 1–5.

Stirling, J., Chen, T., and Adamou, M. (2021). "Autism spectrum disorder classification using a self-organising fuzzy classifier," in *Fuzzy Logic: Recent Applications and Developments* (Cham: Springer), 83–94.

Tabtah, F. (2017). *Autistic Spectrum Disorder Screening Data for Adolescent*. Noida: UCI Machine Learning Repository.

Talabani, H., and Engin, A. (2018). "Performance comparison of svm kernel types on child autism disease database," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (Malatya: IEEE), 1–5.

Talukder, M. A., Hasan, K. F., Islam, M. M., Uddin, M. A., Akhter, A., Yousuf, M. A., et al. (2023). A dependable hybrid machine learning model for network intrusion detection. *J. Inform. Secur. Appl.* 72:103405. doi: 10.1016/j.jisa.2022.103405

Thabtah, F. (2017a). *Autism Screening Adult*. Noida: UCI Machine Learning Repository.

Thabtah, F. (2017b). *Autistic Spectrum Disorder Screening Data for Children*. UCI Machine Learning Repository.

Thabtah, F. (2018). Autism screening data for toddlers. *Int. J. Med. Inform.* 117, 112–124.

Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. *Inform. Health Soc. Care* 44, 278–297. doi: 10.1080/17538157.2017.1399132

Thabtah, F., Kamalov, F., and Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *Int. J. Med. Inform.* 117, 112–124. doi: 10.1016/j.ijmedinf.2018.06.009

Uddin, M. J., Ahamad, M. M., Sarker, P. K., Aktar, S., Alotaibi, N., Alyami, S. A., et al. (2023). An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder. *Computers* 12:92. doi: 10.3390/computers12050092

Vakadkar, K., Purkayastha, D., and Krishnan, D. (2021). Detection of autism spectrum disorder in children using machine learning techniques. *SN Comp. Sci.* 2, 1–9. doi: 10.1007/s42979-021-00776-5