Journal of Computational Analysis and Applications VOL. 33, NO. 8, 2024
10.48047/jocaaa.2024.33.08.132

ASD Diagnosis Using Stacked Ensemble Learning with Neural Network Meta-Learner

Sahana Bisalapur[1a,1b], Geeta R Bharamagoudar[2], Shashikumar G. Totad[3]

[1a]Asst. Prof. Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi- 590010, Karnataka, India.

[1b]Visvesvaraya Technological University, Belagavi-590018, Karnataka, India.

Email: sahanab@sgbit.edu.in Orcid Id: 0000-0002-8030-4105

[2]Professor, Department of Computer Science and Engineering, KLE Institute of Technology, Hubballi-580030, Karnataka, India.

Email: geetatotad@yahoo.co.in Orcid Id: 0000-0003-3384-0201

[3]Professor, School of Computer Science and Engineering, KLE Technological University Hubballi 580030,Karnataka, India

Email: skumartotad69@gmail.com Orcid Id: 0000-0002-5132-8677

***Corresponsing Author:** geetatotad@yahoo.co.in

ABSTRACT

Autism Spectrum Disorder represents a continuous developmental brain condition which is characterized by causing difficulties and challenges in social interaction. It doesn't have proper treatment and has a wide range of impact on life therefore early and timely detection plays critical role to significantly improve the quality of life of individuals with ASD. The current diagnostic practices suffer from extended durations of assessment and subjective interpretation as well as patient needs a qualified expert for evaluation. Currently, data mining and machine learning based approaches have gained huge attention to make it automated and reliable approach for detection of ASD. This work presents a machine learning and data mining based approach for ASD diagnosis using behavioral along with demographic information. The proposed approach consists of several stages including missing values Imputation, outlier removal using normalization, feature selection as pre-processing phase. Generally, the raw data has class imbalance issue therefore Synthetic Minority Over-Sampling Technique (SMOTE) is introduced to overcome this issue. The proposed diagnostic model utilized stacked ensemble with Random Forest and Support Vector Machine (SVM) together with Gradient Boosting as base classifiers. The system trained the meta-learner through neural network processing probabilities generated by these classifiers. The proposed framework delivered outstanding diagnostic results to indicate that ensemble learning methods may create dependable and scalable ASD screening systems.

Keywords:

Autism Spectrum Disorder (ASD), Data Mining, SMOTE, Stacked Ensemble Learning, Classification.

1. INTRODUCTION

10.48047/jocaaa.2024.33.08.132

The Neurological development plays a significant role to human life, as it underpins the brain's ability to analyse the condition, process, and integrate information, ultimately shaping cognitive, emotional, and behavioral functions [1]. However, the disorder in this development leads to create several complexities. Autism Spectrum Disorder (ASD) is known as a group of neurodevelopment conditions that has severe impact on children which leads to raise the challenges in social interaction, communication, and behaviour [3].The signs of autism development start showing themselves for the first time at about six months old [4]. ASD impacts roughly 1% of global individuals as reported by the World Health Organization resulting in close to 75 million affected populations. Research shows that ASD results in developmental harm which makes studies show that 31% of kids with ASD also experience intellectual disability. The complex nature of ASD creates extra difficulties for children during learning processes while causing additional obstacles in their everyday activities. CDC data shows a substantial rise in ASD diagnoses among U.S. children to the point where 1 in 54 children meet these criteria.In the UK, the National Autistic Society estimates that approximately 700,000 individuals have ASD, impacting around 2.8 million people [5].

Therefore, early detection and identification of ASD has become the primary concern. Several methods have been used to detect the ASD including clinical assessment, behavioural screening, genetic and neurological analysis. The clinical assessment uses standard diagnostic tools such as Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) [6, 7] however, the accuracy of these methods rely on the skills of clinicians. Moreover, it is a time-consuming process and analysis can be subjective. The behavioural screen uses parent reported questionnaires, such as the Modified Checklist for Autism in Toddlers (M-CHAT), help in preliminary screening [8]. However, these rely on subjective responses and may lead to false positives or negatives. Similarly, the genetic and neurological analysis methods use advanced methods such as brain imaging (MRI, fMRI) and genetic testing, offer valuable insights but are expensive, require specialized equipment, and are not accessible in all healthcare settings [9].Despite these methods, challenges such as delayed diagnosis, high costs, limited accessibility, and the need for expert evaluation remain significant barriers to early ASD detection.

Current advancement in data processing and extraction techniques, the data mining approaches has emerged as a powerful operation of early and accurate detection of ASD. The data mining methods are used in extracting the meaningful patterns from large dataset, helping to identify early ASD indicators from behavioral, genetic, and neurological data [10]. The data mining approach for ASD detection is carried out in several stages and each stage ensures that the relevant patterns and features are extracted appropriately to enhance the overall accuracy of ASD diagnosis. The key steps of data mining includedata collection, data pre-processing, attribute pattern extraction, data splitting to processing in data mining models, selection of model for training i.e. supervised or unsupervised process and evaluation of model’s performance. The data mining process plays a crucial role in enhancing ASD detection by automating diagnosis, improving accuracy, and enabling

10.48047/jocaaa.2024.33.08.132

early intervention. By leveraging advanced machine learning models and data-driven techniques, the healthcare industry can move towards a more efficient and scalable ASD screening approach.

2. OVERVIEW OF DATA MINING

The data collection includes collection of data from different sources such as clinical data, behavioural data, genetic and neurological data. Generally, this is a raw data which is noisy and inconsistent. Therefore, pre-processing is applied to ensure the clean and structured data for data mining based Machine learning models. Several steps are performed in data pre-processing phase such as handling missing values where missing values are identified and imputed using suitable techniques, noise reduction where irrelevant and inconsistent data is discarded to enhance the accuracy, feature selection where most significant features are identified and data normalization is applied to ensure the uniformity across different features.

The next phase is data transformation phase where raw data is converted into a structured format that machine learning models can process. This includes data encoding, and dimensionality reduction. In next phase, the final dataset is divided into training, validation and testing set. Finally, the training is done by adopting supervised machine learning model and finally the performed of this trained model is evaluated by using performance measurement metrics such as precision, recall, F1-score, and ROC-AUC curve analysis. The data mining process plays a crucial role in enhancing ASD detection by automating diagnosis, improving accuracy, and enabling early intervention. By leveraging advanced machine learning models and data-driven techniques, the healthcare industry can move towards a more efficient and scalable ASD screening approach.

While traditional machine learning (ML) models have shown promising results in ASD detection, they also come with certain limitations. Moreover, these ML models often struggle with bias, overfitting, and variability in predictions, leading to inconsistent results. The challenges of these methods are

- **Overfitting to Training Data** – in some cases the ML models memorize patterns from training data but fail to generalize well on new, unseen data.
- **High Sensitivity to Noisy Data** – ASD datasets often contain missing or inconsistent values, making certain models prone to errors.
- **Class Imbalance Issues** – In ASD classification, the dataset often contains more non-ASD cases than ASD cases, leading to biased model predictions.
- **Low Stability Across Different Datasets** – A single ML algorithm may work well on one dataset but poorly on another due to variations in feature distributions.

To overcome these challenges, ensemble machine learning with data mining technique has been considered as a promising technique which combines multiple models to improve classification accuracy, robustness, and generalization.The ensemble learning is a powerful data mining approach which leverages different ML methods and enhances the predictive performance. Thus, instead of relying on a single model, ensemble methods aggregate multiple base learners to enhance accuracy, stability, and robustness. The ensemble learning techniques include bagging, boosting and stacking. The bagging approach

10.48047/jocaaa.2024.33.08.132

improves accuracy by training multiple models on different subsets of the dataset and averaging their predictions. The boosting approach focuses on misclassified cases by assigning higher weights to difficult instances, refining the model iteratively, and stacking model combinesdifferent ML models and uses a meta-model to make the final prediction.

Therefore, this work presents an ensemble machine learning based data mining approach to detect the autism spectrum disorder. The main contributions of this work are as follows:

- Pre-processing and Feature Selection Pipeline: the proposed model presents a robust pre-processing pipeline to handle missing values, encode categorical variables, normalize features using MinMaxScaler, and apply feature selection using filter based information gain feature selection process.
- Addressing Class Imbalance with SMOTE: generally, the raw data suffer from the issue of class imbalance problem therefore this approach incorporates the SMOTE to address the class imbalance problem
- Proposed Ensemble Model with Meta-Learning: A novel ensemble architecture is proposed where Random Forest, SVM, and Gradient Boosting act as base classifiers, and a Neural Network serves as the meta-learner. This stacking-based model outperforms individual models in terms of accuracy and ROC AUC

Rest of the manuscript is organized in following sections: section III presents the brief Related work about existing methods, section IV presents the proposed data mining based solution for ASD detection, section V presents the comparative analysis where outcome of proposed approach is compared with the state of art methods and finally, section VI presents the concluding remarks.

3. RELATED WORK

This section presents the literature review on existing methods of data mining and ML based approaches of ASD detection. Different types of data types are used to identify the ASD including EEG, video, images etc. EEG captures electrical brain activity as multi-channel time-series data, which is traditionally reviewed manually—a process that is often time-consuming and prone to errors. This section examines existing data mining methodologies together with ML strategies designed for ASD diagnosis. The ASD identification process employs various data types such as EEG and video and images. EEG allows researchers to capture electrical brain signals as multichannel time-series data through manual analysis that takes considerable time and shows potential human errors. Tawhid et al. [11] developed an automated classification system to process EEG signals through spectrogram image transformation with short-time Fourier transform. The spectrogram produces extracted texture features by undergoing noise removal along with signal normalization. PCA filters important data points which feed subsequent multiple ML models to perform efficient abnormality detection. Shinde et al. [12] created a multi-classifier recommendation model to enhance autism spectrum disorder (ASD) diagnosis by implementing Random Forest, Decision Tree, SVM, and Artificial Neural Network (ANN). The research study achieved enhanced accuracy rates through the utilization of Random Forest and Decision Tree classifiers as reported by its findings. Shrivastava et al. [13] conducted research on early autism spectrum disorder detection by utilizing data from

10.48047/jocaaa.2024.33.08.132

different age groups that included toddlers and children along with adolescents and adults. The technique comprised categorical encoding and kNN Imputer to handle missing values together with MinMaxScaler for data normalization. This research examines how SVM, K-Nearest Neighbors (KNN), Random Forest and ANN perform in separating ASD from typically developing (TD) subject groups.

Shashikumar and others suggested that Real-time anomaly detection using Facebook Prophet exemplifies a data mining application that models time series data [27] to identify significant deviations from expected trends, enabling timely recognition of unusual events.

Akter et al. [14] developed a strong ML model for autism detection through feature correlation analysis to eliminate duplicate data and performed both standardization and normalization processes. Researchers in this work applied both ANN and RNN along with Decision Tree and XGBoost as well as Logistic Regression classification modes. Logistic Regression yielded the most satisfactory results among all evaluated models across ASD data sets located in Kaggle and UCI repositories. Hasan et al. [15] developed a comparative prediction model through the integration of Quantile Transformer with Power Transformer along with Normalizer and Max Abs Scaler. The pre-processing methods came before researchers attempted the implementation of eight different ML algorithms which included AdaBoost along with RF and DT and KNN and GNB and LR and SVM and LDA. The experiments demonstrated through diverse datasets that correct feature scaling proves crucial for enhancing classification results across different age categories.
Geeta R B et al. suggested that the WPs-Hash-tree is a efficient Hash table with tree structure which allows access of selected WPs-Tree portions during the extraction task [26]

According to Chen et al. [16] early detection of ASD becomes achievable through analysis of real-world medical claim data. Predicting Autism Spectrum Disorder risk in 18 to 30 month old children formed the core subject of this research when it used over 38,000 database entries from MarketScan Health Claims Database. The study employed both logistic regression with LASSO regularization together with Random Forest models to analyze demographic factors and early medical history data for child ASD risk assessment. The method showed that medical records have the capability to serve as prediction tools for detecting autism spectrum disorder during its early stages. Shashikumar et al advised that Scalability and speedup enhancement of the SVM algorithm represents a data mining approach aimed at efficiently processing big data by optimizing computational performance for large-scale classification tasks.[28]. The research by Wu et al. [17] focused on behavioral sign detection for ASD diagnosis in infancy between 6 to 36 months based on video analysis. A two-stage machine learning framework was suggested by the authors which combined deep learning models with two sequential processes. The statistical assessment of behavioral indicators served as input to classification methods to determine ASD diagnosis with 82% accuracy. The research demonstrated that screening professionals should focus on behavioral indications captured from video files during initial ASD diagnosis assessments.
Jayaprakash et al. [18] conducted clinical research which evaluated developmental delayed children using the

Distributed computing frameworks have emerged as a transformative solution to the scalability issues faced by traditional machine learning algorithms. By partitioning data and computational tasks across multiple nodes [24], distributed systems enable parallel execution of machine learning algorithms, reducing computation time and overcoming memory constraints.

10.48047/jocaaa.2024.33.08.132

The authors introduced and implemented a method to classify brain MRI images as normal or abnormal, with abnormal images further segmented to detect the presence of brain tumors [25].

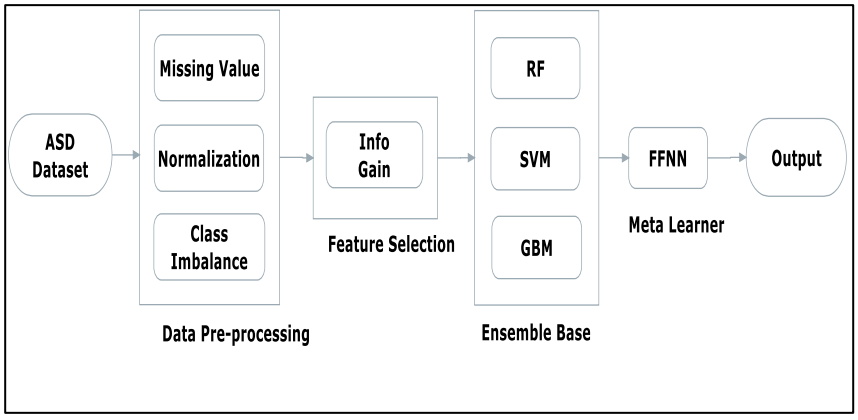
Childhood Autism Rating Scale (CARS). The development of multivariate logistic regression models served to detect behavioral patterns which are specific to autism by focusing on communication and social interaction. The Newton-CG solver optimization technique applied to MLR produced the best results with 97% accuracy thus validating the effectiveness of logistic regression in ASD clinical evaluations. Regarding better classification accuracy delivery Mohan et al. [19] established a method for selecting important features in this research. The researchers utilized features from the UCI ASD repository to extract weights from the SVMAttributeEval method which produced their rankings. The RFE algorithm removed unimportant features after the feature weighting step. The feature reduction reached 60% and reported better performance levels for classifiers LibSVM, IBk, and Naïve Bayes as a result of this process. The study demonstrates that reducing dimensions creates better results for both efficiency and accuracy rates in ASD analysis.

4. PROPOSED MODEL

In order to improve the accuracy of automated ASD detection with enhanced accuracy has been considered as a prime task for research community. Currently data mining based machine learning approaches have gained huge attention however single model based ML approaches face challenges due to data imbalance and generalization for real-time dataset due to their biasness. In order to overcome this issue, we propose a data mining-driven ensemble learning model that leverages multiple machine learning classifiers to create a robust predictive system. The model is designed to address challenges such as data noise, class imbalance, and feature complexity, which are common in ASD datasets.

4.1. Overview of proposed approach

The complete approach is carried out in various phases such as dataset collection, missing value imputation, normalization, feature selection, SMOTE (Synthetic Minority Over-Sampling Technique) to handling the class imbalance problem, and stacked ensemble learning where random forest, SVM and gradient boosting machines are used as base learners and neural network is used as meta learner. The figure 1 depicts the overall architecture of the proposed data mining based approach for ASD detection.



10.48047/jocaaa.2024.33.08.132

Fig.1. General architecture of proposed work

4.2. Data collection and pre-processing

The dataset for this task is obtained from the publicly available UCI repository. This dataset consists of clinical, behavioural, and sociodemographic information, ensuring diversity in features. However, this dataset has some missing attributes, therefore, we apply KNN based missing value imputation method to address the missing values. Let us consider that the dataset is represented as D which consists of m instances (patients or subjects), and each instance is represented by a feature vector in$\mathbb{R}^n$, where n is the number of features. This dataset can be presented as:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

Where $x_i = [x_{i1}, x_{i2}, \dots, x_{in}] \in \mathbb{R}^n$ represents the feature vector for i^{th} sample and $y_i\{0,1\}$ represents the class label where 0 is No ASD and 1 is ASD class. as discussed before, the dataset consist of missing values to represent that let $M \subset \{1,2, \dots., m\} \times \{1,2, \dots., n\}$ be the set of indices of missing values in the dataset

$$\left(x_{ij}\ is\ missing\ \right) \Leftrightarrow (i,j) \in M$$

The objective of this phase is to impute these missing values to obtain the complete matrix $\tilde{X} \in \mathbb{R}^{m \times n}$

In order to achieve this, we apply KNN based imputation mechanism. It is a distance-based imputation technique. The idea is to impute the missing value in a sample using the values from its K most similar (nearest) samples based on a distance metric.

- **Define the distance metric :** Let x_i be the sample with missing values, and x_j be a complete sample (i.e., no missing values in the same feature dimension). The distance between x_i and $x_j$based on the subset of features F_{ij} where both $x_i$and $x_j$have values, can be expressed as:

$$dist\left(x_i, x_j\right) = \sqrt{\sum_{k \in F_{ij}} \left(x_{ik} - x_{jk}\right)^2}$$

- **Identify the K nearest neighbours:** For a missing value x_{ip}, find the set $\mathcal{N}_K(i,p) \subset (1, \dots m)$of K nearest neighbors such that each neighbor $x_j$has a non-missing value at feature p:

$$\mathcal{N}_K(i,p) = \arg\min_j \left(dist\left(x_i, x_j\right) | x_{jp}\ is\ not\ missing\right)$$

- **Impute the identified missing values:** The missing value $x_{ip}$is imputed using a weighted average of the corresponding values from the K nearest neighbors. This can be expressed as:

$$x_{ip} = \frac{\sum_{j \in \mathcal{N}_K(i,p)} x_{jp}}{\sum_{j \in \mathcal{N}_K(i,p)} w_j}$$

where the weight w_j can be defined as:

10.48047/jocaaa.2024.33.08.132

$$w_j = \frac{1}{dist\left(x_i x_j\right)^2 + \epsilon}, \epsilon > 0$$

After imputing all missing values using KNN, we obtain the fully pre-processed dataset:

$$D = \{(\tilde{x}_1, y_1), (\tilde{x}_2, y_2), \dots, (\tilde{x}_m, y_m)\}$$

Later, we apply data normalizaiton on this dataset as follows:

$$\mathcal{N}(x_i) = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

4.3. Feature selection

Further, we apply feature selection process where important feaures are selected to enhance the overall performance of the system performance. for this, let us consider that the complete feature space is represented as

$$X = \{f_1, f_2, \dots, f_n\}$$

Where n represents the total number of attributes. The main aim of the feature selection process is to obtain the optimal number of subset which is represented as:

$$X^* \subset X \; such \; that \; |X^*| = k < n$$

In order to achieve this, we apply information gain based method for feature selection. According to this, for feature $f_j \in X$ the Information Gain with respect to the class label Y is expressed as:

$$IG(Y, f_j) = H(Y) - H(Y|f_j)$$

Where $H(Y)$ represents the entropy of class labels, $H(Y|f_j)$ represents the conditional entropy of Y given f_j. Once the feature importance is ranked, the top-k features are selected as input to the base classifiers in the ensemble and this can be expressed as

$$x_i^{sel} = [f_{i1}, f_{i2}, \dots, f_{ik}], x_i^{sel} \in \mathbb{R}^k$$

4.4.Ensemble classifier

The selected features are then processed through the ensemble classification approach. The proposed work uses random forest, SVM and gradient boosting machines as base learners and later neural network model is applied as meta learner. Let us consider that the base classifiers of ensemble classifier are represented as $h_1(x)$, $h_2(x)$ and $h_3(x)$ as random forest, SVM and GBM, respectively.

Random Forest is an ensemble of decision trees trained on bootstrapped samples with randomized feature selection. For each tree $t \in \{1, \dots, n\}$construct a tree classifier $f_t(x)$. Further, a random subset of features $F_t \subset \{1, .., n\}$ is used to split at each node. Later, majority voting is applied to obtain the output. The tree decision function is represented as:

$$f_t(x) = DecisionTree\ (x; \theta_t)$$

The final output of three can be expressed as:

10.48047/jocaaa.2024.33.08.132

$$h_1(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

Further, SVM classifier is applied which helps to find hyperplane that best separates classes in feature space, maximizing the margin. The primal optimization problem is presented as:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$$

Subject to : $y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i\,, \xi_i \geq 0$Where $\phi(x)$ is the feature transformation, and C represents the Regularization constant, and final prediction represented as

$$h_2(x)\sigma(w^T\phi(x) + b)$$

Finally, the GBM model is applied as base learner. GBM builds trees sequentially to minimize a differentiable loss function. The GBM model is represented as

$$F(x) = \sum_{t=1}^{T} \gamma_t f_t(x)$$

Where f_t represents a decision tree trained on pseudo residuals and expressed as:

$$r_i^{(t)} = -\left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)}\right]$$

The common loss function for this model is expressed as:

$$\mathcal{L}\big(y_i, F(x_i)\big) = -\Big[y_i; pg\ \Big(\sigma\big(F(x_i)\big)\Big) + (1 - y_i)\log(1 - \sigma(F(x_i)))\Big]$$

Based on this the final prediction function is expressed as

$$h_3(x) = \sigma\big(F(x)\big)$$

The final output of these base learner formulates the input vector to the meta learner. The final output can be expressed as:

$$z_i = [h_1(x_i), h_2(x_i), h_3(x_i)] \in \mathbb{R}^3$$

In this work, feed forward neural network is considered as meta learner which consist of hidden layers and output layer. The hidden layer is presented as

$$h^{(l)} = \phi\big(W^{(1)}z_i + b^{(1)}\big) \in \mathbb{R}^d$$

And, the final layer is presented as

$$y_i = \sigma\big(W^{(2)}h^{(1)} = b^{(2)}\big)$$

Where ϕ is the activation function σ represents the σ is the Sigmoid for binary classification, $W^{(1)}$ and $W^{(2)}$ are the weight matrices, $b^{(1)}$ and $b^{(2)}$ are the bias vectors. We train the meta-learner using binary cross-entropy loss which is expressed as:

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

The outcome of final ensemble classifier is presented as:

$$\mathcal{E}(x_i) = \begin{cases} 1, if\ \hat{y}_i \geq 0.5 \\ 0, otherwise \end{cases}$$

10.48047/jocaaa.2024.33.08.132

5. RESULTS AND DISCUSSION

This section presents the outcome of proposed approach and presents the comparative analysis where the obtained performance is compared with the state-of-art machine learning based data mining approaches. The first subsection presents the dataset details, next subsection presents the performance measurement parameters and finally comparative analysis is presented.

5.1. Dataset details

In this work, we have utilized the publically dataset for ASD screening. The obtained datasets are based on the AQ-10 screening technique which was created by Baron-Cohen et al. [20]. In order to assist the clinical implementation across varied settings, Allison et al. [21] introduced Q-CHAT-10 for toddler and AQ-10 child as the condensed version of original AQ. Based on these, we have used two publically available datasets which UCI data and Kaggle toddler dataset.

- UCI Child Dataset: The UCI Machine Learning Repository provides access to 292 sets of child data which covers children between 4 and 11 years old [22]. The dataset contains 21 variables which include data from the AQ-10 questionnaire along with gender information as well as ethnics data and results of ASD diagnoses. The dataset provides an excellent balance since the number of children who received ASD diagnosis matches the number of children without a diagnosis therefore becoming a dependable resource for model training and evaluation.
- Kaggle Toddler Dataset: A total of 1,050 toddler records aged 18 to 36 months were obtained from the Kaggle platform [23]. The database presents 18 features including both AQ-10 scores and parent observations together with ASD diagnostic markings. This dataset follows the UCI dataset by having a balanced ratio between ASD-positive records and ASD-negative records which ensures unbiased testing conditions.

5.2. Performance measurement parameters

The performance of this approach is measured with the help of confusion matrix which consist of true positive, false positive and false negative counts. Based on these parameters the performance can be obtained as:

- Accuracy: it measures the overall correctness of the model and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (also called Positive Predictive Value) indicates the proportion of correctly predicted positive instances out of all predicted positives. It can be expressed as:

$$Precision = \frac{TP}{TP + FP}$$

10.48047/jocaaa.2024.33.08.132

- Recall (also known as Sensitivity or True Positive Rate) measures the proportion of actual positives that were correctly identified. It can be obtained as:

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score** provides a balance between precision and recall, and is especially useful when the class distribution is uneven. This can be obtained as:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Specificity (True Negative Rate) measures the proportion of actual negatives correctly identified. It can be computed as:

$$Specificity = \frac{TN}{TN + FP}$$

5.3. Performance analysis

(a) Outcome for UCI dataset

In this phase, the ASD detection is performed on UCI dataset where state-of-art classification methods and proposed ensemble method is applied where feature selection, normalization, and oversampling (SMOTE) mechanisms are presented. Further, we evaluate the performance of Multiple machine learning algorithms using 5-fold stratified cross-validation, and the performance was assessed with two key metrics: Accuracy and ROC AUC (Area Under the Curve). This study includes Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Naive Bayes (NB), Multi-Layer Perceptron (MLP), and Proposed Stacked Ensemble.

According to this experiment, the proposed approach has reported 100% accuracy that demonstrates the ensemble's ability to harness the complementary strengths of its base learners. In this model the, RF ensures stability and resistance to overfitting, SVM contributes margin-based optimization for better generalization, GB reduces bias by iteratively improving weak learners and MLP Meta-Learner effectively learns how to weigh and integrate predictions from these diverse sources.

Table. 1. Comparative analysis for UCI dataset

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean Accuracy	Mean AUC
Logistic Regression	0.9333 / 0.9833	0.9667 / 1.0000	1.0000 / 1.0000	0.9831 / 0.9989	0.9661 / 0.9977	**0.9698**	**0.9960**
Random Forest	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	**1.0000**	**1.0000**
SVM	0.9000 / 0.9789	0.9500 / 1.0000	0.9667 / 0.9978	0.9322 / 0.9954	0.9322 / 0.9966	**0.9362**	**0.9937**

10.48047/jocaaa.2024.33.08.132

Gradient Boosting	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	1.0000 / 1.0000	**1.0000**	**1.0000**
KNN	0.9500 / 0.9933	1.0000 / 1.0000	0.9167 / 0.9778	0.9153 / 0.9897	0.9322 / 0.9977	**0.9428**	**0.9917**
Naive Bayes	0.9333 / 0.9867	0.9667 / 0.9967	0.9333 / 0.9933	0.9661 / 0.9931	0.9831 / 0.9943	**0.9565**	**0.9928**
MLP	0.8667 / 0.9656	0.9667 / 0.9944	0.9833 / 1.0000	0.9661 / 0.9977	0.9492 / 0.9943	**0.9464**	**0.9904**
Proposed Ensemble	**1.0000 / 1.0000**	**1.0000 / 1.0000**	**1.0000 / 1.0000**	**1.0000 / 1.0000**	**1.0000 / 1.0000**	**1.0000**	**1.0000**

The proposed ensemble model includes trees, linear models, margin classifiers, and boosting — each offering unique perspectives. Moreover, it uses MLP for meata learning where it learns complex patterns in base model outputs and can identify when to trust which model more. similarly, the data pre-processing includes SMOTE, normalization, and feature selection ensured the input was well-prepared, enabling models to perform at their best.

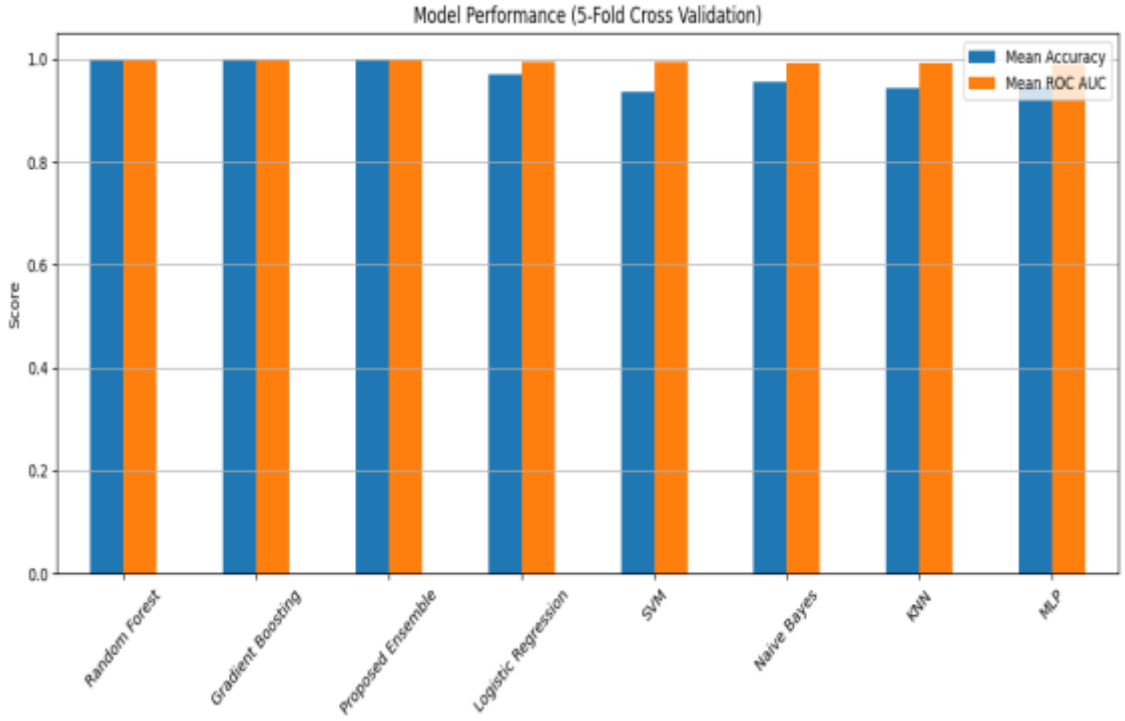


Fig .2. Model performance for UCI dataset

(b) Outcome for Kaggle dataset

This experiment compares the classification performance of several individual machine learning models and a proposed ensemble model. Each model is evaluated using 5-fold

10.48047/jocaaa.2024.33.08.132

cross-validation, with performance metrics reported in terms of Accuracy and AUC (Area Under the ROC Curve).

Table. 3. Comparative analysis for Kaggle dataset

Model	Fold 1 (Acc / AUC)	Fold 2 (Acc / AUC)	Fold 3 (Acc / AUC)	Fold 4 (Acc / AUC)	Fold 5 (Acc / AUC)	Mean Acc	Mean AUC
Logistic Regression	0.9479 / 0.9552	0.9520 / 0.9604	0.9502 / 0.9602	0.9409 / 0.9519	0.9467 / 0.9540	**0.9476**	**0.9563**
Random Forest	0.9485 / 0.9773	0.9590 / 0.9795	0.9514 / 0.9798	0.9479 / 0.9748	0.9496 / 0.9707	**0.9513**	**0.9764**
SVM	0.9544 / 0.9731	0.9620 / 0.9703	0.9561 / 0.9739	0.9532 / 0.9679	0.9526 / 0.9649	**0.9556**	**0.9700**
Gradient Boosting	0.9374 / 0.9611	0.9508 / 0.9645	0.9479 / 0.9660	0.9409 / 0.9587	0.9438 / 0.9594	**0.9442**	**0.9619**
KNN	0.9298 / 0.9570	0.9526 / 0.9689	0.9391 / 0.9623	0.9415 / 0.9592	0.9338 / 0.9570	**0.9394**	**0.9609**
Naive Bayes	0.9157 / 0.9517	0.9157 / 0.9562	0.9274 / 0.9593	0.9081 / 0.9481	0.9075 / 0.9506	**0.9149**	**0.9532**
MLP	0.9520 / 0.9778	0.9631 / 0.9792	0.9549 / 0.9809	0.9508 / 0.9755	0.9537 / 0.9731	**0.9549**	**0.9773**
Proposed Ensemble	0.9785 / 0.9776	0.9785 / 0.9805	0.9808 / 0.9764	0.9890 / 0.9895	0.9720 / 0.9673	**0.9798**	**0.9783**

As discussed before the complete experiment uses five-fold cross-validation to evaluate theperformance of all classification algorithms which included the ensemble model creation. According to the experiment, SVM delivered consistent performance in cross-validation tests with 95.56% accuracy and AUC of 0.9700 indicating that the SVMachieved notable performance in each iteration. TheMLP demonstrated equivalent performance to SVM by reaching 95.49% mean accuracy while achieving the best AUC value at 0.9773 from individual models because of its capability to detect difficult non-linear patterns in the data. Both Random Forest and Gradient Boosting exhibited reliable high accuracy performance through testing with 95.13% and 94.42% respectively. Their AUC values remained above 0.96 establishing ensemble-based learning methods as proficient tools for handling feature interactions and reducing overfitting.

The Proposed Ensemble model demonstrated superior performance than every single model for most evaluation metrics. The ensemble learned to combine base learner predictions and it reached mean accuracy levels at 97.98% with mean AUC performance at 0.9783 which proves its exceptional power for classification and discrimination. The ensemble optimized performance by harnessing the strengths of its base learners through a mechanism which minimized their individual weaknesses. The consistent high scores across all five folds, with particularly notable performance in Fold 4 (Acc: 98.90%, AUC: 0.9895), affirm the stability and reliability of the proposed ensemble.Simpler models such as Logistic Regression and Naive Bayes produced reduced mean accuracies of 94.76% and 91.49% and corresponding AUC values of 0.9563 and 0.9532 because they failed to identify complex patterns in the dataset.

10.48047/jocaaa.2024.33.08.132

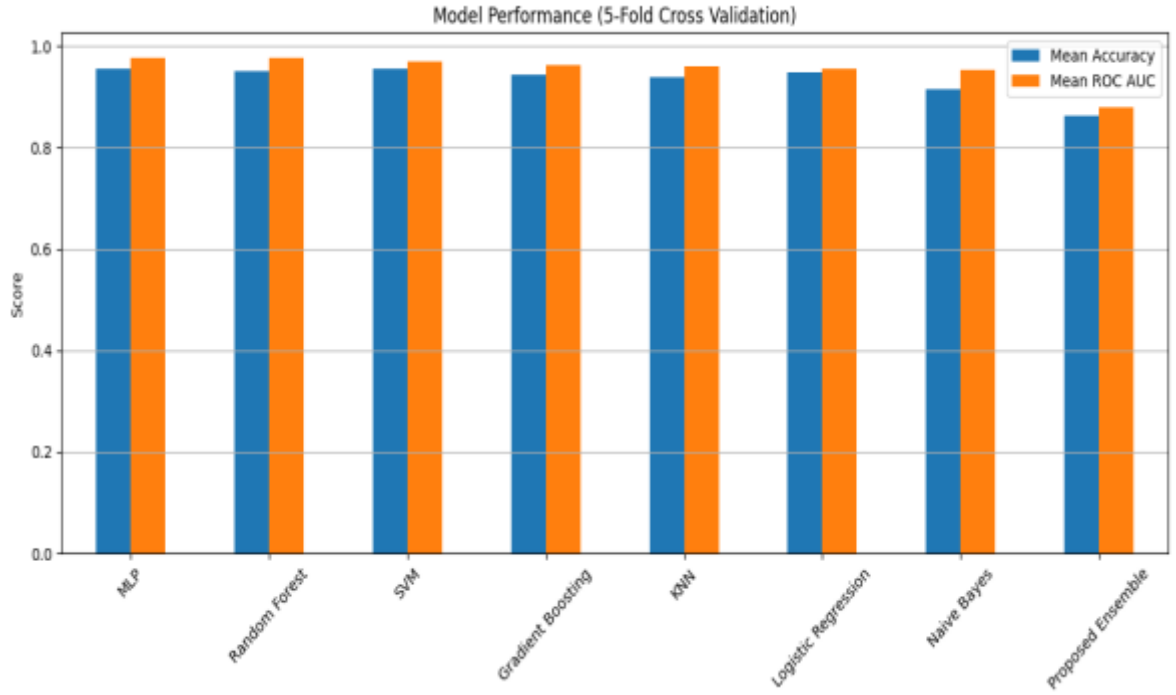


Figure.3. Model performance for Kaggle dataset

6. CONCLUSION

The Autism Spectrum disorder has severe impact on quality of life of human because it causes complexity in social interaction, communication. The early detection of ASD can be beneficial to enhance take the appropriate steps. Several methods have been introduced but traditional methods require human intervention thus it can be time consuming and reliability of these methods also remains a challenging issue. Therefore, researchers have introduced the concept of data mining where data are analysed and patterns are extracted to identify the specific patterns and classify the samples. In this work, we present an automated data mining based ensemble machine learning classification solution where several preprocessing steps are performed such as handling missing values, outlier removal, normalization, and feature selection. Further, SMOTE is also applied to handle the class uncertainty. Finally, the stacked ensemble model was employed, using Random Forest, Support Vector Machine (SVM), and Gradient Boosting as base classifiers. The meta-learner was a neural network trained on the probabilistic outputs of these classifiers. The proposed model showed high classification performance, demonstrating the potential of ensemble-based learning for reliable and scalable ASD screening.

REFERENCES

[1] Megari, K., Frantzezou, C. K., Polyzopoulou, Z. A., & Tzouni, S. K. (2024). Neurocognitive features in childhood & adulthood in autism spectrum disorder: A neurodiversity approach. International Journal of Developmental Neuroscience, 84(6), 471-499.

[2] Bertelli, M. O., Azeem, M. W., Underwood, L., Scattoni, M. L., Persico, A. M., Ricciardello, A., ... & Munir, K. (2022). Autism spectrum disorder. In *Textbook of*

10.48047/jocaaa.2024.33.08.132

psychiatry for intellectual disability and autism spectrum disorder (pp. 369-455). Cham: Springer International Publishing.

[3] Qin, L., Wang, H., Ning, W., Cui, M., & Wang, Q. (2024). New advances in the diagnosis and treatment of autism spectrum disorders. *European Journal of Medical Research*, *29*(1), 322.

[4] Rochat, P. (2021). Clinical pointers from developing self-awareness. *Developmental Medicine & Child Neurology*, *63*(4), 382-386.

[5] Urbistondo Cano, F. (2021). Autism and learning disabilities: developing a relationship-centred approach for support workers in social care.

[6] Lebersfeld, J. B., Swanson, M., Clesi, C. D., & O'Kelley, S. E. (2021). Systematic review and meta-analysis of the clinical utility of the ADOS-2 and the ADI-R in diagnosing autism spectrum disorders in children. *Journal of autism and developmental disorders*, 1-14.

[7] Adamou, M., Jones, S. L., & Wetherhill, S. (2021). Predicting diagnostic outcome in adult autism spectrum disorder using the autism diagnostic observation schedule. *BMC psychiatry*, *21*, 1-8.

[8] Ariffin, R. A., Ismail, J., Abd Rahman, F. N., Ismail, W. W., Ahmad, N., Ghafar, A. A., ... & Nor, N. K. (2024). Malay translation and validation of modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Frontiers in Pediatrics*, *12*, 1384292.

[9] Nisar, S., & Haris, M. (2023). Neuroimaging genetics approaches to identify new biomarkers for the early diagnosis of autism spectrum disorder. *Molecular psychiatry*, *28*(12), 4995-5008.

[10] Saleh, A. I., & Rabie, A. H. (2023). A new autism spectrum disorder discovery (ASDD) strategy using data mining techniques based on blood tests. *Biomedical Signal Processing and Control*, *81*, 104419.

[11] Tawhid, M. N. A., Siuly, S., Wang, K., & Wang, H. (2021). Data mining based artificial intelligent technique for identifying abnormalities from brain signal data. In Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I 22 (pp. 198-206). Springer International Publishing.

[12] Shinde, A. V., & Patil, D. D. (2023). A multi-classifier-based recommender system for early autism spectrum disorder detection using machine learning. Healthcare Analytics, 4, 100211.

[13] Shrivastava, T., Singh, V., & Agrawal, A. (2024). Autism spectrum disorder detection with kNN imputer and machine learning classifiers via questionnaire mode of screening. *Health Information Science and Systems*, *12*(1), 18.

[14] Akter, T., Khan, M. I., Ali, M. H., Satu, M. S., Uddin, M. J., & Moni, M. A. (2021, January). Improved machine learning based classification model for early autism detection. In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 742-747). IEEE.

[15] Hasan, S. M., Uddin, M. P., Al Mamun, M., Sharif, M. I., Ulhaq, A., & Krishnamoorthy, G. (2022). A machine learning framework for early-stage detection of autism spectrum disorders. IEEE Access, 11, 15038-15057.

[16] Chen, Y. H., Chen, Q., Kong, L., & Liu, G. (2022). Early detection of autism spectrum disorder in young children with machine learning using medical claims data. BMJ Health & Care Informatics, 29(1), e100544.

10.48047/jocaaa.2024.33.08.132

[17] Wu, C., Liaqat, S., Helvaci, H., Chcung, S. C. S., Chuah, C. N., Ozonoff, S., & Young, G. (2021, March). Machine learning based autism spectrum disorder detection from videos. In 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM) (pp. 1-6). IEEE.

[18] Jayaprakash, D., & Kanimozhiselvi, C. S. (2024). Multinomial logistic regression method for early detection of autism spectrum disorders. *Measurement: Sensors*, *33*, 101125.

[19] Mohan, P., & Paramasivam, I. (2021). Feature reduction using SVM-RFE technique to detect autism spectrum disorder. Evolutionary Intelligence, 14, 989-997.

[20] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*, 5-17.

[21] Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.

[22] Thabtah, F. F. (2017). Autistic spectrum disorder screening data for children. *UCI Machine Learning Repository*, *10*, C5659W.

[23] Thabtah, F. (2018). Autism screening data for toddlers. *Kaggle. Last checked on*, *26*(07), 2022.

[24] Joshi, Y., Totad, S.G., Geeta, R.B., Prasad Reddy, P.V.G.D. (2018). Mobile Agent-Based Frequent Pattern Mining for Distributed Databases Intelligent Computing and Information and Communication. Advances in Intelligent Systems and Computing, vol 673. Springer, Singapore. https://doi.org/10.1007/978-981-10-7245-1_9

[25] Madgi, M., Giraddi, S., Bharamagoudar, G., Madhur, M.S. (2021). Brain Tumor Classification and Segmentation Using Deep Learning. In: Satapathy, S.C., Bhateja, V., Favorskaya, M.N., Adilakshmi, T. (eds) Smart Computing Techniques and Applications. Smart Innovation, Systems and Technologies, vol 224. Springer, Singapore. https://doi.org/10.1007/978-981-16-1502-3_21

[26] Geeta R.B, Shashikumar G.Totad, and Prasad Reddy PVGD,"Mining Association Rules Using Hash-Index Technique" in International Conference on Digital Image Processing and Pattern Recognition held at Manonmaniam Sundaranar University, Tirunelveli on 23rd -25th September 2011, ISBN: 978-3-642-24055-3, Communications in Computer and Information Science book series (CCIS,) volume 205, pp 239-248, https://link.springer.com/chapter/10.1007/978-3-642-24055-3_25 https://doi.org/10.1007/978-3-642-24055-3_25, Springer, Berlin, Heidelberg.

[27] Shashikumar G Totad, Karibasappa K. G., Geeta R. B. Nithish T, "Detection of Anomaly in Streaming Dataset", 4th International Conference on Intelligent Computing and Communication (ICICC - 2020) 18th,19th & 20th Sept, 2020, Dayanand Sagar University, Bengaluru, Published in: Computer Communication, Networking and IoT, Part of Lecture Notes in Networks and Systems, LNNS Vol-197, pp 431-440, 19-06-2021,Springer, Singapore, Scopus Indexed, DOI: doi.org/10.1007/978-981-16-0980-0_38.

[28] Shashikumar G Totad., Geeta R.B., Satwik Belaldavar. (2019) "Scalability and Speedup Enhancement of SVM Algorithm for Big Data Processing ". AICTE

10.48047/jocaaa.2024.33.08.132

Sponsored International Conference (ICCEEE2020) 24th-26th August 2020 at St. Peter's Engineering College, Hyderabad.