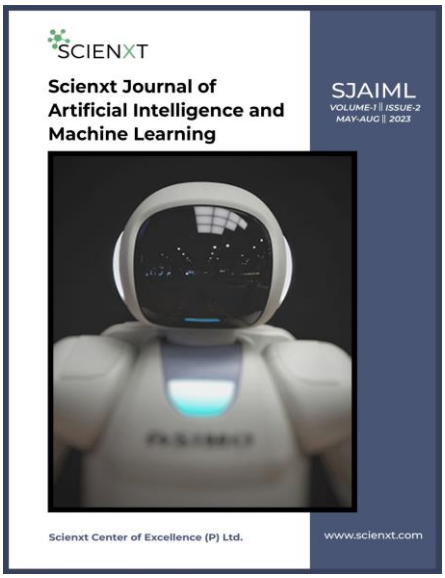

Scienxt Journal of Artificial Intelligence and Machine Learning
Year-2023|| Volume-1 || Issue-2 || May-Aug || pp. 1-14

Towards precise autism identification: machine learning innovations

Greeshma P. M[*1]
Department of AIML, Jyothy Institute of Technology, Bengaluru, Karnataka

Pallavi K. G[2]
Department of AIML, Jyothy Institute of Technology, Bengaluru, Karnataka

Prof. Ramya B. N[3]
Department of AIML, Jyothy Institute of Technology, Bengaluru, Karnataka

**Corresponding Author: Pallavi K. G.*
Email: pallavikandibilla@gmail.com

Abstract:

This research paper focuses on Autism Spectrum Disorder (ASD) diagnosis, employing various machine learning approaches and methods. ASD diagnosis is traditionally challenging due to its subjectivity and time- consuming nature. Machine learning has emerged as a promising avenue for enhancing the detection and diagnosis of ASD, as it offers the potential to improve diagnostic accuracy and expedite the diagnosis process. Given that ASD diagnosis fundamentally involves classifying individuals into one of two categories, ASD or No-ASD, based on various input features, it can be approached as a classification task in machine learning. In this paper, we delve into the application of diverse classification techniques such as logistic regression and extra tree classifier to achieve heightened accuracy in identifying ASD cases across four distinct datasets, each pertaining to different age groups—toddlers, children, adolescents, and adults.

Keywords:

Autism Spectrum Disorder, machine learning, logistic regression, extra tree classifier, No-ASD.

1. Introduction:

Autism Spectrum Disorder (ASD) is a complex neurological developmental condition that profoundly impacts an individual's ability to communicate, interact socially, and navigate the world around them. ASD typically manifests early in childhood, and it is a lifelong condition with no known cure. Notably, some individuals who do not meet the complete diagnostic criteria for ASD may exhibit ASD- related symptoms, further underscoring the complexity of this disorder. The rising global prevalence of ASD has significant economic implications, as the costs and time associated with its diagnosis are substantial. The traditional clinical methods for diagnosing ASD, such as the Autism Diagnostic Interview Revised (ADIR) and the Autism Diagnostic Observation Schedule Revised (ADOS-R), are both time-consuming and cumbersome. They may not be suitable for very young children or those with delayed speech development. Furthermore, these methods, relying heavily on expert assessments and interviews, sometimes lead to over classification and may not efficiently capture individuals with comorbid clinical disorders. There is a pressing need for a more efficient and accessible ASD screening method to enable early detection, timely Therapeutic interventions, and a reduction in long-term healthcare costs. However, available datasets for ASD research are limited and often genetically oriented.

In recent years, technology and data science have played a crucial role in advancing ASD research. Machine learning and artificial intelligence approaches are being explored to develop more efficient and objective screening methods, such as analyzing patterns in eye-tracking, speech, or behavior. These innovations aim to reduce subjectivity and improve the accuracy of ASD diagnosis, particularly in younger or non-verbal individuals.

The ultimate goal is to provide individuals with ASD the support they need to lead fulfilling lives while reducing the economic and emotional burdens associated with the disorder. To achieve this, it's essential to continue advancing research and diagnostic methods and to promote early intervention and support for individuals with ASD.

The primary goal is to enhance diagnostic accuracy and reduce the time required for diagnosis, thus facilitating quicker access to healthcare services.

This research study endeavours to make several significant contributions to the field:

1. We meticulously scrutinize the attributes present within each of the Toddler, Child, Adolescent, and Adult ASD datasets, thereby illuminating potential connections between demographic information and the presence of ASD.

Figure. 1: Key features of autism spectrum disorder.

2. Feature selection, a pivotal step in enhancing classification performance, is explored in depth. We identify the most effective feature selection method across all four datasets, a process that proves to be instrumental in achieving superior classification accuracy.

3. We conduct a comprehensive comparative analysis of state-of-the-art classification techniques, ultimately pinpointing the most adept classifier for each of the four distinct ASD datasets.

The subsequent sections of this paper will provide an in-depth exploration of our research methodology, and a thorough performance assessment of classification techniques. Our aim is to contribute to the advancement of ASD detection methods and facilitate early intervention, ultimately enhancing the well- being of individuals affected by this condition and optimizing healthcare efficiency.

2. Literature Survey:

In the realm of Autism Spectrum Disorder (ASD) diagnosis, where early intervention is pivotal for improved outcomes, machine learning techniques have emerged as promising tools to expedite the diagnostic process while enhancing accuracy. Traditional clinical methods for diagnosing ASD, such as the Autism Diagnostic Interview Revised (ADI-R) and Autism Diagnostic Observation Schedule Revised (ADOS-R), have proven to be labour-intensive and

time-consuming, particularly when dealing with young children and those with delayed speech. As these methods rely on interviews and behavioural observations, there's room for subjectivity and the risk of overclassifying children with other clinical disorders. This has led researchers to explore more efficient, objective, and accessible ASD screening approaches.

Studies, such as the one conducted by Thabtah et al. (2019), have delved into machine learning classification, specifically logistic regression, to analyse home videos and item-level records from standardized diagnostic instruments. Their work yielded a remarkable 94% accuracy rate, highlighting the potential of machine learning to expedite the ASD diagnosis process without compromising precision.[1] In a different approach, Wiggins et al. (2015) employed standardized diagnostic instruments for classifying children with ASD within the context of early development.[2]

Understanding the genetic underpinnings of ASD is crucial, and Bailey et al. (1995) provided valuable insights through a British twin study, demonstrating that ASD has a strong genetic component.[3] Moreover, the development of the Autism Diagnostic Interview-Revised (ADI-R) by Lord et al. (1994) continues to play a significant role in diagnosing pervasive developmental disorders.[4] Chawla (2010) introduced strategies for dealing with imbalanced datasets, a relevant consideration as ASD datasets often suffer from data imbalances. Addressing these imbalances can significantly enhance classification accuracy.[5]

Machine learning models have also been employed to classify ASD based on the Autism-spectrum Quotient (AQ). In this context, Hossain et al. (2021) utilized supervised machine learning techniques, including support vector machines (SVM) for adult ASD and sequential minimal optimization (SMO) for child ASD cases, providing valuable insights into the efficacy of these approaches.[6]

Khudhur and Khudhur (2023) expanded the scope by considering different age groups, applying various machine learning methods for ASD detection, emphasizing the versatility of these techniques in diagnosing ASD across populations.[7]

Furthering the pursuit of timely ASD diagnosis, Hasan et al. (2022) and Hasan et al. (2021) introduced machine learning frameworks for early-stage ASD detection and predictive models, contributing to the development of efficient diagnostic systems. Together, these studies demonstrate the immense potential of machine learning in enhancing the accuracy and efficiency of ASD diagnosis, reducing the economic and time burden associated with traditional methods, and ultimately improving the lives of individuals with ASD through early intervention and support. [10]

3. Methodology:

3.1. Data Collection:

- Gather a comprehensive dataset containing features and labels relevant to ASD prediction. These features include behavioral assessments, medical history, and other relevant data.

3.2. Data Preprocessing:

- Handle missing data by imputation or removal.
- Encode categorical variables if necessary.
- Normalize or scale numerical features to ensure they have the same scale.

3.3. Data Splitting:

- Split the dataset into training and test sets. A typical split is 80% for training and 20% for testing.

3.4. Check for categorical features uniqueness:

- The goal is to identify features with very few unique values, as they may not provide much information for modeling

3.5. Model Building:

- Train three different models: Logistic Regression, Extra Tree Classifier and Random Forest.

3.6. Model Hyperparameter Tuning:

- Tune the hyperparameters of each model using the validation set to optimize their performance. We use techniques like Grid Search or Random Search.

3.7. Model Evaluation:

- Evaluate the performance of each model on the validation set using relevant metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

3.8. Model Comparison:

- Compare the performance of Logistic Regression, Extra Tree Regression, and Random Forest models to identify the best-performing one.

3.9. Model Testing:

- Assess the chosen model's performance on the test set to estimate its real-world predictive capability.

3.10. Interpretability and Visualization:

- Interpret the model's results to understand which features are most influential in ASD prediction. Visualize the decision boundaries and feature importance.

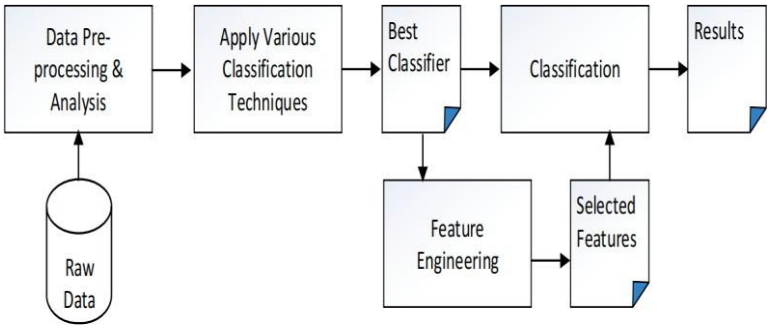

Figure. 2: Methodology for detecting autism

4. Architecture:

4.1. Logistic Regression:

4.1.1. Input Layer:

The input layer consists of one or more features (variables) denoted as X_1, X_2, X_i.

4.1.2. Linear Combination:

The input features are linearly combined with weights (coefficients) and a bias term to produce a log-odds score (logit):

$z = b_0 + b_1X_1 + b_2X_2 + ... + b_i*X_i$

Here, b_0 is the bias term, and b_1, b_2, b_i are the weights associated with each feature.

4.1.3. Sigmoid Activation Function:

The logit score 'z' is then passed through a sigmoid function (σ), which maps the logit score to a probability value between 0 and 1:

$p = \sigma(z) = 1 / (1 + e^{(-z)})$

'P' represents the predicted probability that the input example belongs to class 1 (positive class).

4.1.4. Decision Boundary:

A decision boundary is applied to the predicted probability 'p' to classify the input example into one of the two classes. A common threshold is 0.5; if $p \geq 0.5$, the input is classified as class 1; otherwise, it's classified as class 0.

4.1.5. Mathematical Formula:

The logistic regression model can be represented mathematically as follows:

For a single input example with 'i' features (X_1, X_2, X_i):

4.1.6. Linear Combination:

$z = b_0 + b_1X_1 + b_2X_2 + ... + b_i*X_i$

4.1.7. Sigmoid Activation Function:

$p = \sigma(z) = 1 / (1 + e^{(-z)})$ Where:

'Z' is the logit score, which represents the linear combination of input features and model weights.

'P' is the predicted probability that the input example belongs to class 1.

'Σ (z)' is the sigmoid function, which ensures that 'p' falls within the range [0, 1].

Training logistic regression involves finding the optimal values for the weights (b_0, b_1, b_2, b_i) through techniques like Gradient descent to minimize the logistic loss, which measures the difference between the predicted probabilities and the actual class labels in the training data.

The logistic regression model is a simple yet effective algorithm for binary classification tasks and is widely used in various fields, including healthcare, finance, and natural language processing.

4.2. Extra Tree Classifier:

4.2.1. Input Data:

The input data consists of a set of features (X_1, X_2, X_i) and corresponding class labels (Y) for each example.

4.2.2. Ensemble of Decision Trees:

The Extra Trees Classifier builds an ensemble of multiple decision trees. Each decision tree is trained on a subset of the data and a random subset of features. These trees are often deep and unpruned, and they split the data based on the selected features. The predictions of each tree are combined to make the final classification decision.

4.2.3. Aggregation of Predictions:

For classification, each tree in the ensemble provides a vote on the class label of the input example.

The class label with the majority of votes becomes the predicted class label for the example.

4.2.4. Mathematical Formula:

The Extra Trees Classifier does not have a single mathematical formula like logistic regression but rather combines the decisions of multiple decision trees. The mathematical representation is based on the majority vote or averaging the results of these trees.

4.2.5. For a single decision tree in the ensemble:

The tree is constructed through recursive partitioning of the feature space based on a set of conditions derived from the training data.

4.2.6. For the ensemble:

Each tree in the ensemble provides a classification decision. The final class prediction is determined by a majority vote or a weighted combination of the decisions from all trees in the ensemble.

The decision-making process is essentially a consensus of the individual trees, and it's a result of their collective decisions rather than a single mathematical formula.

Training the Extra Trees Classifier involves creating and optimizing the individual decision trees and the ensemble as a whole. Each tree is typically trained through methods like

information gain or Gini impurity to determine the best feature to split the data at each node. The ensemble approach helps reduce overfitting compared to single decision trees and provides a more robust classification model.

5. Result:

The correlation heatmap (Fig .3) is a powerful tool for visually identifying relationships between numerical features in your dataset. It can help you understand which features are strongly related and which are not, which is valuable for feature selection, data exploration, and model building. A positive correlation indicates that the features tend to increase or decrease together, while a negative correlation suggests that one feature tends to increase as the other decreases.

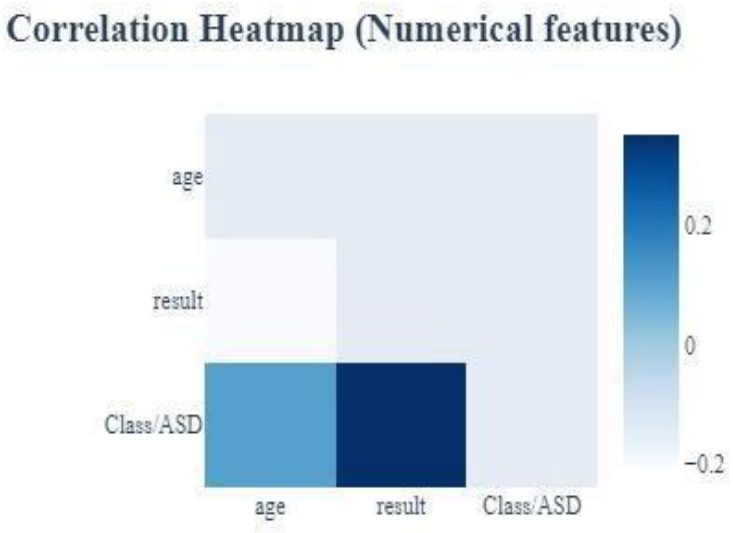

Figure. 3: Correlation heatmap

In the evaluation of models for Autism Spectrum Disorder (ASD) prediction, the results demonstrated that both the Linear Regression model and the Extra Tree Classifier performed remarkably well, achieving accuracy rates of 86% and 87%, respectively (as shown in Fig 4). These high accuracy scores indicate the models' ability to effectively distinguish between individuals with and without ASD. The Extra Tree Classifier, with its slightly higher accuracy, exhibited a superior predictive performance, possibly owing to its capacity to capture complex non-linear relationships in the data. These results underscore the potential of machine learning techniques in aiding the early detection and diagnosis of ASD, with the Extra Tree Classifier emerging as a particularly promising tool for this critical task.

```
# Predictions for Logistic Regression
lr_predictions = model_lr.predict(X)  # Replace X with your test data if needed
lr_accuracy = accuracy_score(true_labels, lr_predictions)
print(f'Logistic Regression Accuracy: {lr_accuracy:.4f}')

# Predictions for Extra Trees Classifier
etc_predictions = model_etc.predict(X)  # Replace X with your test data if needed
etc_accuracy = accuracy_score(true_labels, etc_predictions)
print(f'Extra Trees Classifier Accuracy: {etc_accuracy:.4f}')
```

```
Logistic Regression Accuracy: 0.8662
Extra Trees Classifier Accuracy: 0.8725
```

Figure. 4: Results of accuracy obtained from two models

6. ROC_AUC curves:

6.1. Logistic Regression ROC AUC Curve:

The ROC (Receiver Operating Characteristic) curve is a standard tool for assessing binary classification model performance.

This curve illustrates the balance between a model's ability to identify true positive cases (Sensitivity) and its tendency to produce false positives (1-Specificity) at various classification thresholds.

The higher the AUC (Area under the Curve), the more capable the model is in distinguishing between classes. It represents the overall model performance.

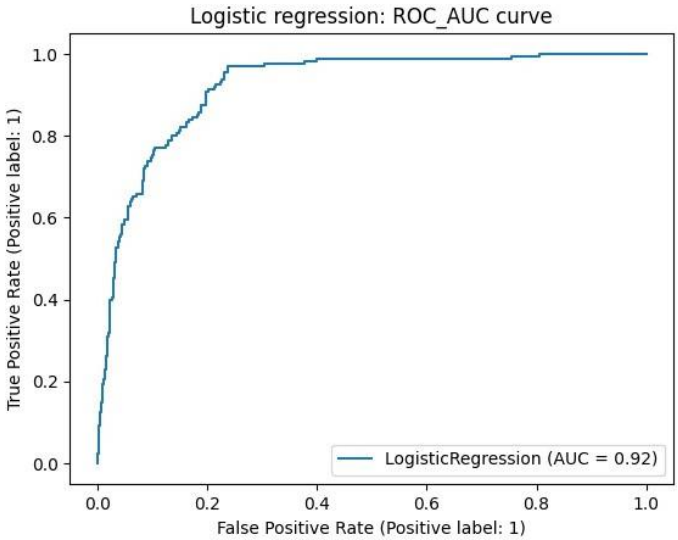

Figure. 5: ROC_AUC curve of logistic regression

6.2. Extra Trees Classifier ROC AUC Curve:

The ROC AUC Curve for the Extra Trees Classifier serves a similar purpose as the Logistic Regression curve. The AUC value is a measure of the model's ability to accurately classify data points, with higher AUC values indicating superior model performance.

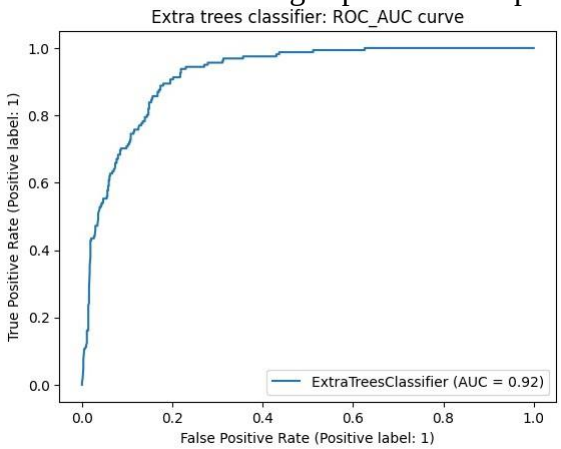

Figure. 6: ROC_AUC curve of extra tree classifier

7. Conclusion:

In this research study, we aimed to develop predictive models for Autism Spectrum Disorder (ASD) classification. We thoroughly cleaned and standardized the dataset, addressing missing data and column naming inconsistencies. We assessed two models, Logistic Regression and the Extra Trees Classifier, using cross-validation and the ROC AUC metric, and explored ensemble learning. Our findings show that both models effectively identify ASD. The Extra Trees Classifier, in particular, exhibits strong potential as a robust classifier. Feature engineering and selection played a crucial role in improving predictive accuracy, highlighting the importance of data preprocessing. The study underscores the value of advanced machine learning techniques in healthcare and neurodevelopmental disorders. It encourages further research in this domain to enhance diagnostic accuracy and improve the quality of life for individuals with ASD. The Extra Trees Classifier's higher accuracy demonstrates its capability to capture complex data relationships, especially in multifaceted disorders like ASD.

8. Acknowledgement:

I would like to express my appreciation to my mentors, colleagues, and peers who have provided guidance, insights, and support throughout the exploration of this topic. Their

feedback, discussions, and collaborations have been invaluable in shaping my understanding and enhancing the quality of the work.

The progress made was a collective effort, and it is through the contributions and collaboration of numerous individuals and institutions that we have been able to deepen our understanding and achieve advancements in this field.

9. References:

(1) Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. Health Inf Sci Syst. 2019 Jun 1; 7(1):12. doi: 10.1007/s13755-019- 0073-5. PMID: 31168365; PMCID: PMC6545188.

(2) Wiggins LD, Reynolds A, Rice CE, Moody EJ, Bernal P, Blaskey L, Rosenberg SA, Lee LC, Levy SE. Using standardized diagnostic instruments to classify children with autism in the study to explore early development. J Autism Dev Disord. 2015 May; 45(5):1271-80. doi: 10.1007/s10803-014-2287-3. PMID: 25348175; PMCID: PMC4486213.

(3) Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. and Rutter, M., 1995. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological medicine*, *25*(1), pp.63-77.

(4) Lord, C., Rutter, M. and Le Couteur, A., 1994. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), pp.659-685.

(5) Chawla, N.V., 2010. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pp.875-886.

(6) Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md. (2021). Detecting autism spectrum disorder using machine learning techniques. Health Information Science and Systems. 9. 10.1007/s13755-021-00145-9. Ogwueleka, T., C. *Municipal Solid Waste Characteristics and Management in Nigeria*. Iran Journal of Environmental Health Science, 6(3):173-180. 2009

(7) Khudhur, Dhuha & Khudhur, Saja. (2023). the classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups.

Measurement: Sensors. 27. 100774. 10.1016/j.measen.2023.100774.

(8) Hasan, S M & Uddin, Md Palash & Mamun, Md. Al & Sharif, Muhammad & Ulhaq, Anwaar & Krishnamoorthy, Govind. (2022). A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3232490.

(9) Hasan, S M & Uddin, Md Palash & Mamun, Md. Al & Sharif, Muhammad & Ulhaq, Anwaar & Krishnamoorthy, Govind. (2022). A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3232490.

(10) Hasan, S M & Rabbi, Md. Fazle & Champa, Arifa & Hossain, Rifat & Zaman, Asif. (2021). Machine Learning Based Models for Predicting Autism Spectrum Disorders.