International Journal of Scientific & Engineering Research, Volume 14, Issue 7, July-2023
ISSN 2229-5518

Evaluation and Comparison of Machine Learning Models for Autism Spectrum Disorder Prediction

Narla Sai Vardhan Reddy, M.V.S Koushik, Gudipati Vikas

Abstract— Autism Spectrum Disorder (ASD), a complex neurodevelopmental disorder, causes behavioral, social interaction, and communication difficulties. For those with ASD, early evaluation and treatment can improve results. Our investigation scrutinized six machine learning models - Random Forest, MLP, Naive Bayes, XGBoost, K-Nearest Neighbor, and Support Vector Machine - to ascertain their accuracy in predicting ASD. The dataset used for the present study included about 701 examples and 21 different attributes. Upon meticulous evaluation, we ascertained that all six machine learning models evinced remarkable accuracy in predicting ASD. Notably, the Random Forest model outperformed its counterparts, achieving an impressive accuracy rate of 99.30. These results demonstrate the important potential of machine learning models for aiding precise ASD prediction and advancing early detection and intervention efforts.

Index Terms— Autism Spectrum Disorder, behavior, early evaluation, machine learning models, neurodevelopment disorder, prediction, social interaction, treatment,

1 Introduction

A crucial area of research is forecasting autism spectrum disorder (ASD) with the goal of enhancing early detection and treatment of ASD patients. It substantially impacts people's everyday lives and general functioning and affects people of all ages [1]. The Centres for Disease Control and Prevention (CDC) estimate that 1 in 54 American children have ASD [2].

ASD prediction and diagnosis using machine learning approaches have gained more attention recently. Large datasets can be analyzed by machine learning algorithms, which can spot links and patterns that may escape the attention of human observers [3]. The application of machine learning to ASD prediction has the potential to offer insightful information and help medical practitioners make well-informed choices about the diagnosis and course of treatment. It can aid in the early detection of people at risk for ASD, allowing for prompt interventions and better results [4]. We can reveal hidden patterns and links in the data by utilizing the strength of machine learning algorithms, perhaps leading to a deeper comprehension of the underlying mechanisms of ASD. Additionally, creating precise prediction models can enhance the implementation of individualized treatment programs for people with ASD and help allocate healthcare resources [5].

This study investigates how machine learning techniques might be employed for predicting ASD. By utilizing a comprehensive dataset comprising various attributes and features, we seek to develop accurate and reliable prediction models that can assist in the early identification of ASD.

The paper is structured into distinct sections to organise the content. Section II focuses on Related Works, where a comprehensive literature survey is conducted to explore existing research relevant to our specific problem statement. Our research methodology is presented in Section III, which covers a range of topics, including dataset description, data analysis, preprocessing approaches, and the procedures used for model creation, training, and testing. In Section IV, under Results and Conclusion, we go into more detail about our models' outcomes, discuss the conclusions we may draw from them, and offer some parting thoughts that highlight the most important lessons learned from the research.

2 LITERATURE REVIEW

Numerous studies have explored the use of machine learning algorithms to predict Autism Spectrum Disorder (ASD). These studies have focused on different aspects, including data analysis, feature selection, and classification techniques.

In order to determine the frequency of putative risk factors connected to pregnancy and the peri-postnatal period in autism, Grossi et al. [6] carried out a pilot study. The study included 24 relatives of autistic children as a control group, 68 generally developing kids, and 45 autistic kids. By choosing 16 out of 27 factors, scientists used specialised artificial neural networks (ANNs) to distinguish between autism and control participants with an overall accuracy of 80.19%. While only 46% accurate globally, logistic regression produced disappointing results.

In their research, Vakadkar et al. [7] used basic behaviour sets chosen from diagnosis datasets to construct an automated ASD prediction model. Predictive models were built using a variety of machine learning approaches, such as Support Vector Machines, Random Forest Classifier, Naive Bayes, Logistic Regression, and KNN. Their goal was to speed up the diagnosing process and identify ASD early. The findings demonstrated that in the dataset they chose, Logistic Regression had the highest accuracy.

While the study by Vakadkar et al. addresses the need for better ASD diagnosis using machine learning approaches, additional research is required to validate and generalise the results. Machine learning-based screening techniques that are completer and more robust can greatly improve early identification and intervention for people with ASD.

Using resting-state functional MRI (rs-fMRI) interconnection measures as diagnostic biomarkers for autism spectrum disorder (ASD), Plitt et al. [8] conducted a study. Utilising rs-fMRI data from people with ASD and typically developing people, they created machine learning classifiers. The study used rs-fMRI techniques to attain excellent classification accuracy but discovered that behavioural metrics regularly beat brain-based classifiers. The brain-based classifiers' most illuminating associations were tied to areas that are important for social interaction. The study concludes that although rs-fMRI scans by themselves can categorise people with ASD, this approach falls short of biomarker standards.

In a study by Sólon et al. [9], brain activation patterns from the ABIDE dataset were used to identify individuals with autism spectrum disorder (ASD) using deep learning algorithms. The study showed modified anterior-posterior brain connections in ASD patients and had a 70% accuracy rate in differentiating ASD patients from controls. The study uses deep learning methods to help comprehend the brain patterns connected to ASD.

Using functional magnetic resonance imaging (fMRI), a deep multimodal learning strategy was recently presented for the diagnosis of autism spectrum disorder (ASD) [10]. Using two different forms of connectomic data from fMRI scans, the model combined two different representations of brain activity. The multimodal technique outperformed single-modality methods with a classification accuracy of 74%, a recall of 95%, and an F1 score of 0.805.

A machine learning strategy was put up by Kazi Shahrukh Omar et al. [11] to forecast Autism Spectrum Disorder (ASD). The goal of this work was to design a mobile application for predicting ASD in people of any age as well as an efficient prediction model based on machine learning techniques. To create the autism prediction model, the researchers used the Random Forest-CART (Classification and Regression Trees) and Random Forest-Id3 (Iterative Dichotomiser 3) algorithms. The AQ-10 dataset and an actual dataset made up of 250 people with and without autistic symptoms were used to evaluate the model. The evaluation findings showed that the suggested prediction model outperformed the control dataset in terms of accuracy, specificity, sensitivity, precision, and false positive rate (FPR).

Li et al. [12] investigated kinematic parameters and machine learning as a method of diagnosing autism. They looked at 40 kinematic measurements from 8 mimic circumstances and determined which ones stood out. On a short sample set, they achieved 86.7% accuracy using SVM and Naive Bayes classifiers. The work emphasises the possibility of quantitative kinematic measurements and machine learning for making a preliminary diagnosis of autism and comprehending motor subgroups.

Comprehensive experimental research was conducted in this paper to demonstrate the performance of various ML algorithms and enable comparisons between them, effectively addressing the issues previously mentioned. The subsequent section will provide further details on the method utilized to create a model and achieve precise results.

3 Proposed Methodology

The main goal of the suggested approach is to build a model that can precisely identify whether an individual has ASD based on the given features. We will use a supervised learning approach to accomplish this and train the model using the labelled data that is readily available. The model will then be able to generalise data patterns and correlations to generate predictions about instances that have not yet occurred.

3.1 Dataset

This study utilised a dataset from the UCI Machine Learning Repository [13]. The dataset titled "Autism Screening Adult" was created by Fadi Tabtah [14]. It has 21 features, which include category, continuous, and binary variables, and 704 samples. To verify the quality of the data, preprocessing procedures were carried out to deal with incoherent and categorical features in the dataset.

3.2 Data Analysis

Several important conclusions have been drawn after the dataset was analysed. First off, the ASD Class target variable is unbalanced, with about 75% of the samples being classified as ASD negative as well as 25% as ASD positive. During the modelling phase, this imbalance should be taken into consideration.

The dataset's gender distribution is generally balanced, with roughly 55% of the people being female and 45% being male. Notably, females account for about 60% of instances among people with ASD.

Remarkably, most patients did not have jaundice. There were more people who tested ASD-negative than ASD-positive among those who did have jaundice. This implies that having jaundice may not be an accurate predictor of ASD.

Upon reviewing the ethnicity column, it is evident that there are 11 distinct ethnic groups listed, along with the category "others" and the symbol "?" indicating unreported or missing information. The most prevalent ethnicity among the group is White European, followed by Asian and Middle Eastern. When analyzing the ASD-positive individuals within the population, the majority are White Europeans, with Asians, Latinos, and "?" following closely behind.

ASD-positive cases are more prevalent in the age range of 20 to 30 years (more than 300), whereas they are less prevalent in the age range of 40 and beyond.

When considering whether autism runs in the family, it becomes clear that most patients do not have any close relatives diagnosed with the disorder. Patients with autistic family members, however, are more likely to be given an autism diagnosis.

The United States, United Arab Emirates (UAE), and New Zealand have the highest proportion of patients in the dataset when the patient's country of residence is examined. Notably, only the USA had identical patients in the ASD-positive and ASD-negative groups, pointing to probable differences in prevalence between various nations.

Examining the relation of patients who completed the test, it is observed that the majority completed the test themselves, while participation from healthcare professionals was relatively low.

3.3 Preprocessing

Several modifications were made to the dataset in the preprocessing setp of the proposed solution. For greater uniformity and clarity, the column names "austim" and "contry_of_res" have been changed to "autism" and "country_of_res," respectively.

When the dataset was examined, it was discovered that the "ethnicity" column included duplicate entries, which were represented by the word's "others" and "Others" (which are nearly identical). Furthermore, the "relation" and "ethnicity" columns both had erroneous entries signified by "?". To preserve the data's integrity, the most often occurring value was selected to replace the "?" in the "relation" and "ethnicity" columns.

Label encoding was done on several categorical columns in the dataset, including "age," "gender," "ethnicity," "jundice," "autism," "country_of_res," "used_app_before," "relation," and "Class/ASD." Label encoding gives each category its own special numeric code, simplifying additional analysis and modelling procedures.

Finally, it was decided that the columns "index" and "age_desc" were determined to be unnecessary for the proposed solution and were therefore dropped from the dataset.

These preprocessing processes make the dataset more organised and prepared for following analysis and modelling tasks in the proposed approach.

3.4 Modelling

After completing the preprocessing procedures, the next step is to train our models. We split the dataset into a training set, which makes up 80% of the data, and a testing set, which contains the remaining 20%. The training set is used to learn the parameters of the models, while the testing set assesses their performance. We used six different classification algorithms for training: Logistic Regression, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost. Each algorithm has its unique characteristics and approaches to classification. A summary of these algorithms is provided below.

Logistic Regression

Logistic regression is an essential statistical method for binary classification tasks. Analyzing independent variables determines the probability of an outcome variable belonging to a specific class. The model calculates the coefficients for the independent variables and uses the logistic function to translate them into probabilities. Logistic regression is widely used in various fields to forecast outcomes and understand the correlations between variables [15].

Naive Bayes

Naive Bayes classifiers are straightforward probabilistic classifiers that assume that features are independent. They are scalable and capable of achieving great accuracy. They have been effective in real-world applications and perform well with limited training datasets. They may sometimes outperform other algorithms, despite how straightforward they are. [16]

Support Vector Machine (SVM)

Based on statistical learning theory, Support Vector Machines (SVM) are extremely precise machine learning techniques. They can deal with real-world issues like short sample numbers, nonlinearity, and high dimensionality and are used for classification tasks. SVM uses support vectors to determine the best surface for separating various classes. A radial basis kernel function is used to attain excellent classification accuracy. [17]

K-Nearest Neighbors(kNN)

K-Nearest Neighbours (kNN) is a simple and widely used classification technique in machine learning. It allocates a new data item to the majority class among its k nearest neighbours based on a distance metric, such as Euclidean distance. The selection of k is crucial for balancing smoothness and local details. Although kNN is simple to use, it can be expensive to compute for large datasets.

Random Forest

An ensemble learning technique called Random Forest (RF) combines various Classification and Regression Trees (CART). It is strong and effectively manages nonlinear data, especially for the prediction of autism screening. To create predictions, RF employs bootstrapped samples and a voting system. Utilising the Gini impurity criterion index, it ranks feature importance. RF has advantages such as robustness, non-linearity handling, and parallel processing capabilities.[18].

XGBoost

XGBoost is a high-performance, scalable machine learning algorithm. It integrates many decision trees to produce a potent ensemble model using gradient boosting and sophisticated regularisation methods. High-dimensional data are handled by XGBoost, which also does feature selection and is resistant to outliers. In many different machine learning tasks, it is commonly employed. [19]

Multi-layer perception (MLP)

A multi-layer perceptron (MLP) is an artificial neural network comprising numerous layers of linked nodes or neurons. Each neuron takes input signals, computes them, and sends the result to the following layer. MLPs can recognise non-linear relationships in data and learn complicated patterns. They have been effectively used in various fields, including financial forecasting, natural language processing, and picture recognition [20].

4 RESULTS

The feature importance graph (Fig. 1) highlights that 'result' is the most influential feature in predicting autism. Key features such as 'A4_Score,' 'age,' 'A3_Score,' and 'A9_Score' also play significant roles. 'Contry_of_res' and 'ethnicity' have moderate importance, while 'jaundice,' 'autism,' 'gender,' 'A8_Score,' and 'used_app_before' have lower importance.

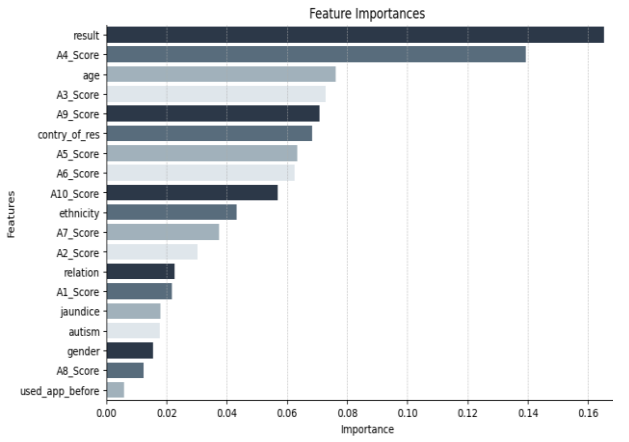

Fig. 1 Feature Importance of Attributes in the Dataset

In our study, we used six machine learning models to predict autism. Table 1 shows a comparison of these models. We evaluated their performance by analysing Precision, Recall, and accuracy metrics from the confusion matrix and classification report. These metrics helped us determine how effective the models were in predicting autism.

Based on the study, the Random Forest Classifier performed better than other classifiers regarding precision, recall, F1 score, and accuracy. It uses an ensemble technique and combines multiple decision trees to identify complex patterns and make accurate predictions.

XGBoost and MLP performed competitively but fell slightly short compared to the Random Forest Classifier. It is worth noting that XGBoost's boosting technique and MLP's ability to learn complex patterns did contribute to their performance, however.

On the other hand, Naive Bayes obtained a significantly lower score than the Random Forest Classifier, XGBoost, and MLP. This could be attributed to Naive Bayes' assumption of class conditional independence, which may have limited its ability to capture complex relationships in the data.

Finally, SVM and KNN achieved the lowest scores compared to the other classifiers. This can likely be attributed to SVM's sensitivity to hyperparameter selection and KNN's reliance on the number of neighbors (k) and the local structure of the data.

TABLE 1
EVALUATION METRICS

Model/ Metrics	Accuracy	Precision	Recall	F1 score
Naive Bayes	0.96	0.95	0.98	0.97
SVM	0.85	1.00	0.81	0.89
KNN	0.90	0.96	0.90	0.93
Random Forest Classifier	0.99	1.00	0.98	0.99
XGBoost	0.98	0.98	0.98	0.98
MLP	0.95	0.96	0.96	0.96

The differences in scores among the classifiers can be attributed to their underlying algorithms, assumptions, and their capability to handle the specific characteristics of the autism prediction task.

The confusion matrix for the top two performing models, Random Forest (Fig. 2) and XGBoost (Fig. 3) are shown below.

The confusion matrix is an essential tool for assessing the accuracy of a model's predicted classes compared to the actual classes. It comprises four values, namely true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). We can determine the performance of the Random Forest and XGBoost models by analysing the confusion matrix based on their precision, recall, accuracy, and other relevant metrics. This way, we can confidently evaluate the models' ability to classify instances accurately.

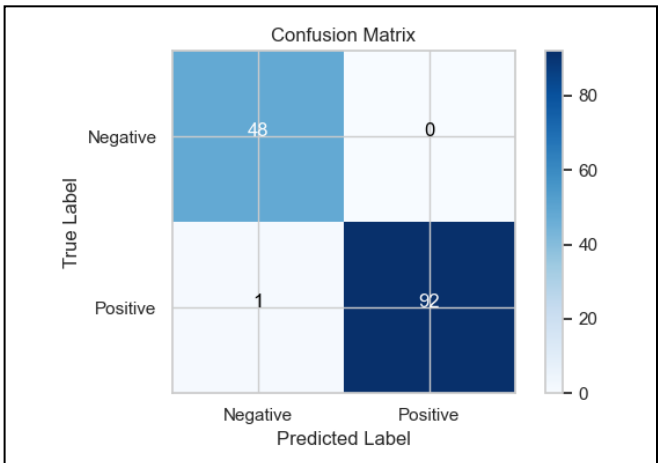

Fig. 2 Confusion Matrix of Random Forest Algorithm

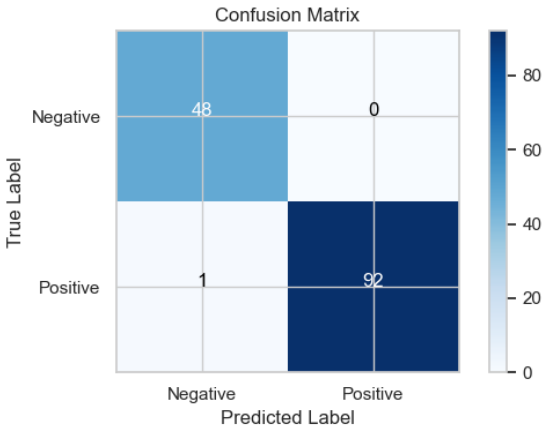

Fig. 3 Confusion Matrix of XGBoost

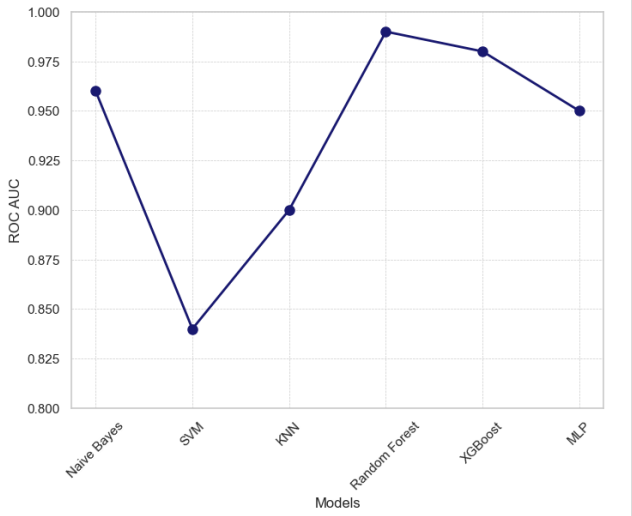

Fig. 4 Comparision of ROC AUC for each algorithm

The Receiver Operating Characteristic Area measures the efficiency of various classification models in distinguishing between classes Under the Curve (ROC AUC) scores. Among the models studied, Random Forest and XGBoost demonstrated high performance (Fig. 4), with Random Forest achieving the highest score. Despite its simplicity, Naive Bayes also performed well, whereas KNN and SVM scored considerably lower. The Multilayer Perceptron (MLP) exhibited competitive performance. This comparison highlights the strengths of each model and facilitates informed decisions in classifier selection.

5 Conclusion

In predicting autism, the Random Forest model outperformed all other classifiers with exceptional accuracy demonstrated through precision, recall, F1 score, and overall accuracy metrics. Its ability to identify intricate patterns and deliver precise predictions makes it the optimal choice. Moreover, by examining the significance of specific predictors such as 'result,' 'A4_Score,' 'age,' 'A3_Score,' and 'A9_Score,' this model provides critical insights for further research and comprehension of autism spectrum disorder. The Random Forest model's robust performance solidifies it as the superior choice for autism prediction.

References

[1] Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., ... & Durkin, M. S. (2018). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. MMWR Surveillance Summaries, 67(6), 1-23.

[2] Centers for Disease Control and Prevention. (2020). Data & statistics on autism spectrum disorder. Retrieved from https://www.cdc.gov/ncbddd/autism/data.html

[3] Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for a behavioral distinction of autism and ADHD. Translational psychiatry, 6(2), e732.

[4] Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., ... & Sigman, M. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. Pediatrics, 128(3), e488-e495.

[5] Zwaigenbaum, L., Bauman, M. L., Fein, D., Pierce, K., Buie, T., Davis, P. A., ... & McPartland, J. C. (2015). Early screening of autism spectrum disorder: Recommendations for practice and research. Pediatrics, 136(Supplement 1), S41-S59.

[6] E. Grossi, F. Veggo, A. Narzisi, A. Compare, and F. Muratori, "Pregnancy risk factors in autism: a pilot study with artificial neural networks," Pediatric Research, vol. 79, no. 2, pp. 339-347, 2016. doi: 10.1038/pr.2015.222.

[7] Vakadkar, K., Purkayastha, D., & Krishnan, D. (2021). Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN COMPUT. SCI., 2(6), 386. https://doi.org/10.1007/s42979-021-00776-5

[8] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," NeuroImage: Clinical, vol. 7, pp. 359-366, 2014. doi: 10.1016/j.nicl.2014.12.013.

[9] A. Sólon, A. Franco, C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," NeuroImage: Clinical, vol. 17, 2017. doi: 10.1016/j.nicl.2017.08.017.

[10] M. Tang, P. Kumar, H. Chen, and A. Shrivastava, "Deep Multimodal Learning for the Diagnosis of Autism Spectrum Disorder," Journal of Imaging, vol. 6, no. 6, p. 47, Jun. 2020, doi: 10.3390/jimaging6060047.

[11] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A Machine Learning Approach to Predict Autism Spectrum Disorder," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679454.

[12] B. Li, A. Sharma, J. Meng, S. Purushwalkam, and E. Gowen, "Applying machine learning to identify autistic adults using imitation: An exploratory study," PLoS ONE, vol. 12, no. 8, e0182652, 2017. doi: 10.1371/journal.pone.0182652.

[13] F. Tabtah, "Autism Screening Adult," UCI Machine Learning Repository, 2017. [Online]. Available: https://doi.org/10.24432/C5F019.

[14] Thabtah, F. An accessible and efficient autism screening method for behavioural data and predictive analyses. Health Informatics Journal 2019;25(4):1739–55. https://doi.org/10.1177/1460458218796636.

[15] Logistic regression. (2023, May 30). In Wikipedia, The Free Encyclopedia. Retrieved June 20, 2023, from https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=1157778320

[16] Naive Bayes classifier. (2023, June 10). In Wikipedia, The Free Encyclopedia. Retrieved June 20, 2023, from https://en.wikipedia.org/w/index.php?title=Naive_Bayes_classifier&oldid=1159533979

[17] Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. In: Liu, C., Wang, L., Yang, A. (eds) Information Computing and Applications. ICICA 2012. Communications in Computer and Information Science, vol 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34041-3_27

[18] Alessia, S., Antonio, C., & Aldo, Q. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience, 9, 329. doi: 10.3389/fnagi.2017.00329.

[19] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[20] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359-366, 1989. [Online]. Available: https://doi.org/10.1016/0893-6080(89)90020-8. [Accessed: June 24, 2023].