# Optimized Random Forest for High-Accuracy Autism Spectrum Disorder Detection via Phenotypic Data

Mohamed Gawish, Nada S. El-Askary, Mohamed Mabrouk Morsey,
Abeer M. Mahmoud, Mostafa Aref, Taha Ibrahim El-Arif
Computer Science Department-Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt 11566

***Abstract*—Autism Spectrum Disorder (ASD) is a mental disorder with a neurological condition noticed in patients by their persistent deficits in social communication and interaction, along with the possibility of the presence of repetitive motor behaviors or activities. Early diagnosis of this disorder is crucial for improving patients' cognitive, emotional, and social development. Numerous studies on detecting autism exist; however, data limitations and imbalance affect their model generalization. This study proposes a new intelligent computational model for mental healthcare in individuals with ASD, utilizing machine learning (ML) to address these shortcomings. The proposed model enhances the random forest (RF) algorithm by setting optimal parameters and encompasses two key pipelines: 1) the data pipeline and 2) the learning pipeline. We first gathered a multi-source dataset, implemented integration and preprocessing via ML algorithms. The phenotypic data used were collected from 19 different sites and merged to ensure the diversity of the data used. The resulting dataset is subsequently fed into the learning pipeline, where a supervised ML algorithm is employed to create a trained computational model for detecting ASD. The model is based on tuning the RF algorithm by finding the optimal values for five key hyperparameters. After tuning the model, the accuracy of detecting ASD from phenotypic data reached 96.86%, with a sensitivity of 97.14% and a false positive rate of 3.39%. Comparing the tuned RF model with different ML models verified that tuning and optimizing RF achieves a preeminent classification accuracy for ASD detection using phenotypic data, as the accuracy of RF without tuning is 95.06%. In addition, to validate the tuned RF model's real-world applicability, a separate qualitative study was conducted on five independent, narrative-based case studies, where the model accurately classified four of them by translating descriptive language into quantitative features.**

***Keywords*—Mental healthcare; ASD; phenotypic data; ABIDE-II; random forest; hyperparameter optimizations**

## I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects the social communication and interaction between a patient and other people and may result in limited and recurring interests, activities, or behavioral habits. According to the World Health Organization [1], the prevalence of ASD has been steadily rising over the past few decades, affecting approximately 1 in 100 children in the world's population and about 3 in 100 children in the United States alone [2]. The causes for this neurological disorder are still not clearly defined, but many scientific researchers have suggested that it may be caused by genetic or clinical factors. For prompt intervention and better results for those impacted, early identification and diagnosis of ASD are essential.

Traditionally, ASD diagnosis has relied on comprehensive clinical assessments, which involve extensive evaluations by trained professionals, including psychologists, psychiatrists, and speech therapists [3]. However, this process can be time-consuming, costly, and subject to inter-rater variability, leading to delayed diagnosis and intervention. Therefore, there is a growing need for more efficient and accurate methods to detect ASD in children. Recent advancements in machine learning (ML) and data mining techniques have shown promise in enhancing the early detection of ASD. In particular, the utilization of phenotypic data has emerged as a valuable resource for developing predictive models.

Phenotypic data refer to the observable characteristics or traits of individuals with ASD, such as communication and social skills, repetitive behaviors, and sensory sensitivities. It can be collected through various sources, such as parental questionnaires, clinical assessments, electronic health records, direct observation, caregiver reports, and wearable sensors [4]. The breadth of behaviors and traits linked to ASD can be better understood using phenotypic data, which can also be utilized to create interventions that focus on particular symptoms. In the context of systems powered by artificial intelligence (AI), phenotypic data can be used to train ML algorithms to recognize patterns that indicate the presence of autism.

The Autism Brain Imaging Data Exchange (ABIDE) dataset, a publicly available dataset [5], is a significant resource for ASD research. It comprises two large-scale collections, ABIDE-I and ABIDE-II, each providing functional magnetic resonance imaging (fMRI) data along with a corresponding phenotypic dataset. ABIDE-II, collected from over 19 diverse sites, introduced more detailed phenotypic characterization not found in ABIDE-I [6]. This dataset is crucial for our research as it provides a comprehensive view of the phenotypic characteristics of individuals with ASD, which is essential for developing accurate ML models for ASD detection. ABIDE-II's phenotypic dataset comprises 347 columns encompassing demographic, diagnostic status, and an extensive array of behavioural assessments [6]. Additionally, patients' basic information, including sex, age, handedness index score, and handedness category, is utilized [6].

One ML algorithm that has gained popularity in the field of medical data is the Random Forest (RF) algorithm. RF is an ensemble learning technique that generates predictions by combining several decision trees. It has been widely used in

various domains due to its ability to handle high-dimensional data, handle missing values, and provide robust predictions [7]. The use of phenotypic data with ML techniques, particularly RF, holds promise for improving the precision and efficiency of ASD detection. Studies have shown that the nature of phenotypic data can enhance the discriminative power of ML models, enabling them to capture subtle patterns indicative of ASD. By leveraging a diverse array of features derived from behavioral observations, cognitive assessments, and physiological measures, the amalgamation of phenotypic data and RF can contribute to a more comprehensive and accurate diagnostic framework.

This research study presents a novel and promising approach to developing a computational model for mental healthcare, focusing on the early detection of ASD. The proposed model aims to distinguish ASD from typical control (TC) individuals using the ABIDE-II phenotypic dataset. We hypothesize that by leveraging a comprehensive set of tuned hyperparameters for RF along with a set of phenotypic features, we can develop a reliable and accurate RF model for early detection of ASD. The proposed model has the potential to highlight a set of hyperparameters for building RF model to guide researchers in their studies and also using the model by healthcare professionals to assist them in making informed decisions and improving the efficiency of ASD screening processes. The main contributions of this study are as follows:

- We proposed an enhanced integrated framework that is based on the random forest algorithm for optimal parameter setting.
- We employed two pipelines, the first of which seeks to overcome the data limitation issue by integrating more than 19 sources of phenotypic data, and then extensive preprocessing is performed.
- We developed a learning pipeline that seeks to find the optimal parameters and address the model generalization issue.
- We validated the proposed model using narrative-based real-world case studies to provide an aided diagnostic tool in the medical domain.

The article is organized as follows: Section II presents a comprehensive literature review, highlighting the limitations and challenges. Section III outlines the methodology of the proposed model and provides detailed descriptions of the techniques employed in the two proposed pipelines. Section IV presents the experimental results, data integration, data preprocessing and the performance evaluation of the developed model. Section V discusses the results of the developed model and presents a comparative analysis. Section VI presents the predictive capabilities of the model using real-world cases. Section VII concludes the findings and potential future research directions.

## II. Literature Review

In the era of AI, numerous researchers have explored the use of ML techniques to detect ASD in children. The researchers provided studies that highlighted the potential of utilizing various types of data, including genetic, neuroimaging- ing, and phenotypic data, to develop models that can predict, detect, and classify ASD. Nada S. et al. [8] published a comprehensive survey of studies that have developed ML models to detect brain neurological disorders. The survey included a list of various ML classifiers with different data types, which achieved varying levels of accuracy. It highlighted that in recent years, the utilization of phenotypic data has gained significant attention in the field of ASD detection due to its ability to deliver comprehensive behavioral and clinical information. Additionally, it addressed the challenges, including the availability of large datasets and the need for proper data preprocessing. Focusing on phenotypic data, here are some noteworthy examples of studies and applications that utilize AI techniques to detect ASD with phenotypic data.

*1)* Perochon et al. [9] developed and validated a mobile digital behavioral phenotyping application called SenseToKnow. The app utilizes computer vision and ML to analyze children's behavioral responses during pediatric well-child visits. The study involved a cohort of toddlers aged 17-36 months, including those diagnosed with ASD (49 children), developmental delay without autism (98 children), and TC (328 children). Children's behavior was recorded using an iPad, and 23 digital phenotypes were extracted, including gaze patterns, facial expressions, head movements, response to name, blink rate, and motor behaviors. The model was trained using extreme gradient boosting (XGBoost), a method that utilizes 1,000 tree-based classifiers, and was evaluated through five-fold cross-validation. It finally achieved a sensitivity of 87.8%.

*2)* Mazumdar et al. [10] presented a study on the early detection of ASD in children using an approach that depends on gathering data from tracking the patient's eye and then feeding it to an ML classifier. A total of 24 children (14 ASD and 10 TC) were included in the analysis, which utilized 400 natural images sourced from the Saliency4ASD [11] dataset. Using a TreeBagger (bagged DT) classifier trained on the Saliency4ASD dataset, it achieved an accuracy of 59%.

*3)* Alkahtani et al. [12] developed a facial landmark-based system using DL and ML algorithms to detect ASD in children aged 2 to 8 years. A total of 2,940 images (1,327 ASD and 1,613 TC) were used for analysis, obtained from Kaggle [13]. The study applied eight ML classifiers to the ASD dataset along with their respective accuracy rates: support vector machine (SVM) (81.46%), linear regression (LR) (82.14%),

*4)* RF (78.06%), k-nearest neighbor (KNN) (52.38%), decision tree (DT) (66.15%), neural network multi-layer perceptron (MLP) (77.04%), gradient boosting (GBoost) (75.51%), and convolutional neural networks (CNN) (92%).

*5)* Raj et al. [14] explored the effectiveness of ML and DL models in predicting ASD across three population groups: children, adolescents, and adults. The study utilized three publicly available datasets containing behavioral features and demographic information related to ASD from the UCI

machine learning repository: children (292 cases), adolescents (104 cases), and adults (704 cases). The study applied six ML classifiers on the ASD datasets: SVM, LR, Naıve Bayes (NB), KNN, NN, and CNN. They all achieved classification accuracy ranging from 80.95% to 99.53%. CNN consistently outperformed all other classifiers across datasets, demonstrating higher accuracy and robustness.

*6)* Akter et al. [15] developed ML models capable of detecting ASD at an early stage by leveraging behavioral data, thereby facilitating early interventions. In addition to the datasets utilized by Raj et al. [14], the authors utilized other publicly available datasets related to ASD from the Kaggle repository and combined them based on population group. The aggregated datasets comprise 2,009 cases (toddlers = 1,054, children = 248, adolescents = 98, and adults = 609) with 21 features. Three feature transformation methods were employed to reduce skewness, normalize the data to a range of -1 to 1, and enhance the performance of the ML model. These methods are Log transformation, Z-score normalization, and Sine transformation. A total of 250 classifiers were applied to these datasets, with 80 demonstrating good performance. Classifiers with accuracies under 70% were excluded. The final selection consists of 9 algorithms: adaboost (AB), flexible discriminant analysis (FDA), DT, boosted generalized linear model (GLMBoost), linear discriminant analysis (LDA), mixture discriminant analysis (MDA), penalized discriminant analysis (PDA), SVM, and classification and regression tree (CART). Optimal classifier per age group: SVM (mean accuracy = 98.77%) for the toddlers' dataset, AB (mean accuracy = 97.20%) for the children dataset, AB (mean accuracy = 98.36%) for the adults' dataset, and GLMboost (mean accuracy = 93.89%) for the adolescents' dataset.

*7)* Erkan et al. [16] investigated the utility of supervised ML models- kNN (k=3), SVM (radial basis function), and RF (60 trees)—for classifying ASD using three age-specific datasets (children, adolescents, and adults) from the UCI repository with a total of 1,100 samples. Each dataset, comprising 20 features including standardized screening questions and demographic variables, underwent rigorous preprocessing including missing value imputation and numeric encoding. Models were evaluated using repeated stratified train/test splits ($a$ = 50%, 60%, 70%, 80%, 90%). Results demonstrate that RF achieved 100% performance across all metrics and datasets, outperforming SVM (accuracy ranged from 91%–100%) and kNN (Accuracy ranged from 83%–95%). The findings underscore the potential of ML, particularly ensemble-based models, in supporting fast, accurate, and scalable ASD diagnosis across age groups. This work supports the integration of ML into clinical ASD screening protocols, especially where large and well-structured data is available.

*8)* Using the same datasets as those in Erkan et al. [16], Bawa et al. [17] implemented six supervised learning models—LR, SVM, RF, DT, NB, and KNN. Results showed that LR and SVM performed optimally across binary classification (yes/no for ASD) tasks, achieving 94.30 – 99.00% accuracy in individual groups and achieving 97.20% on merged data using LR. Extending to multiclass classification (e.g. child_ yes, adolescent_ no) on the merged dataset, SVM attained a peak accuracy of 99.55%, while LR achieved 95.90% accuracy, RF achieved 94.10%, and NB achieved 82.50%. Further enhancement by grid search hyperparameter tuning yielded SVM with 99.55% accuracy, LR with 99.09%, and RF with 96.82%. The study concludes that multiclass classification models tailored to age subgroups substantially improve diagnostic accuracy and could serve as a foundation for scalable ASD screening tools.

*9)* Kunda et al. [18] proposed an ML model to classify autism using the ABIDE-I dataset, which comprises a total of 1,035 cases (505 ASD, 530 TC) of resting-state fMRI and phenotypic data. The study proposed the tangent Pearson embedding method for extracting second-order features. It also implemented an approach for domain adaptation to enhance classification accuracy. The model was trained using Ridge, LR, and SVM, and its performance was evaluated through ten-fold and leave-one-out cross-validation. Ultimately, it achieved an accuracy of 73% with the Ridge classifier.

The reviewed studies collectively highlight significant advancements in AI-driven ASD detection and diagnosis. They show the potential of AI to significantly enhance and facilitate the early detection and diagnosis of ASD through the application of ML algorithms and dimensional reduction techniques. Table I summarizes the discussed studies for ASD detection from an AI perspective, and the symbol "*" indicates ML algorithm(s) that achieved the highest accuracy.

TABLE I. COMPARATIVE TABLE FOR ASD STUDIES USING PHENOTYPIC DATA

| Study | Dataset size | ML algorithms | The highest accuracy |
|---|---|---|---|
| Perochon et al. [9] | 475 | XGBoost* | 87.8% |
| Mazumdar et al. [10] | 400 | TreeBagger* | 59% |
| Alkahtani et al. [12] | 2,940 | SVM, LR, RF, KNN, DT, MLP, GBoost, CNN* | 92% |
| Raj et al. [14] | 1,100 | SVM, LR, NB, KNN, NN, CNN* | 99.53% |
| Akter et al. [15] | 2,009 | AB, FDA, DT, Glmboost, LDA, MDA, PDA, CART, SVM* | 98.77% |
| Erkan et al. [16] | 1,100 | KNN, SVM, RF* | 100% |
| Bawa et al. [17] | 1,100 | LR, DT, RF, NB, KNN, SVM* | 99.55% |
| Kunda et al. [18] | 1,035 | LR, SVM, Ridge* | 73% |

Although numerous datasets, studies, and approaches, ML approaches demonstrated high potential for identifying ASD from behavioral, physiological, and neurocognitive data. There

are several common limitations:

- Data limitations:
  - Small sample sizes: Many studies were conducted on limited-scale datasets, particularly those focusing on specific subpopulations (e.g. children, females, or multi-site studies). This scarcity of data reduces the statistical power and representativeness of models, undermining the reliability of their results and classification accuracy.
  - Dataset imbalance: ASD datasets are often highly imbalanced, leading to bias in ML models.
  - Variability across sites: Multi-site neuroimaging studies (e.g. fMRI) showed high inter-site variability due to differences in scanner protocols, demographic diversity, and sample heterogeneity.
  - Lack of standardized datasets: Many ML models were trained on custom datasets, making it challenging to compare results across studies.
  - Lack of detailed phenotypic data, especially regarding ASD severity and cognitive variability.
  - A lack of diverse participants (e.g. in age, gender, or ethnicity) further means models may be overfit to narrow populations.
  - Need for real-world testing: real-world data variability is untested.
- Model generalization:
  - Overfitting on small datasets: ML models tend to overfit when trained on limited or highly structured datasets.
  - Poor generalization to real-world clinical settings: Many ML models were evaluated only on academic datasets without validation in clinical environments.
  - Feature selection dependence: Some studies heavily relied on specific feature selection methods, which might not generalize well to different datasets.
- Computational challenges:
  - High computational cost: Some ML models require significant computing resources, making them infeasible for real-time clinical applications without further optimization.
  - Interpretability of AI models: Most ML models for ASD detection operate as "black boxes," offering little explanation for their decisions. This opacity makes it difficult for clinicians to trust and implement them in real-world diagnostic settings.

## III. Intelligent Mental Healthcare Model

The model consists of two main pipelines, as shown in Fig. 1. The model initiates by collecting the datasets and then processing them through the data pipeline using ML methods. Subsequently, the model inputs the resulting dataset into the learning pipeline, in which a supervised ML algorithm is applied to develop a trained computational mental model for ASD detection. The following sections focus on the details of the data and the learning pipelines, respectively.

### A. Data Pipeline

The data pipeline includes four primary sequential steps: data collection, integration, inspection, and preprocessing, as shown in Fig. 1. First, a unified dataset was created by merging data sources collected from many international sites that dealt with autism and employed various examinations and procedures for making diagnoses. Afterwards, data preprocessing techniques were employed on the dataset to enhance its quality and format after a thorough inspection. Furthermore, examine whether the dataset is balanced.

*1) Data collection:* The model utilized the phenotypic data in ABIDE-II [19] collected from over nineteen international sites, as it provides more detailed phenotypic data than any other dataset. As defined, the phenotypic data in ABIDE-II present 1,114 real-world cases aged 5 to 64 years. The dataset contains 521 ASD cases and 593 TC cases. Each case is represented by 347 structured characteristics and clinical data, such as IQ, medical history-related questions, demographic variables, diagnostic status, and an extensive array of behavioral assessments. Additionally, patients' basic information, including gender, age, handedness index score, and handedness category, is utilized.

*2) Data integration and inspection:* Upon downloading the phenotypic data from all nineteen sites participating in the ABIDE-II dataset, they were comprehensively analyzed. The individual datasets were concatenated into a unified file to ensure consistency and uniformity across the various sites. And then, inspecting the data to find structural issues, leading us to the next stage of applying data preprocessing techniques.

*3) Data preprocessing:* Crucial preprocessing techniques were applied to clinch the subsequent analyses quality and reliability and maintain and improve data balance.

*a)* Data cleaning is essential to ensure the data quality. Data cleaning operations need to be done very carefully and precisely so as not to end up with too little data or too much data with useless values. These operations were carried out in several steps as illustrated in the experimental results section.

*b)* Data balancing is crucial when the number of instances in different classes is significantly skewed as this will lead to biased model performance; where the model may be biased towards the majority class and the minority class may be significantly underrepresented making it challenging for the AI model to learn patterns effectively, resulting in poor performance on the minority class. In the ABIDE-II dataset, after applying data cleaning and removing some observations, the ground truth in all remaining data was checked. The ratio between ASD and TC observations is 1:1.08, which means they are nearly equal in size and the data is balanced.

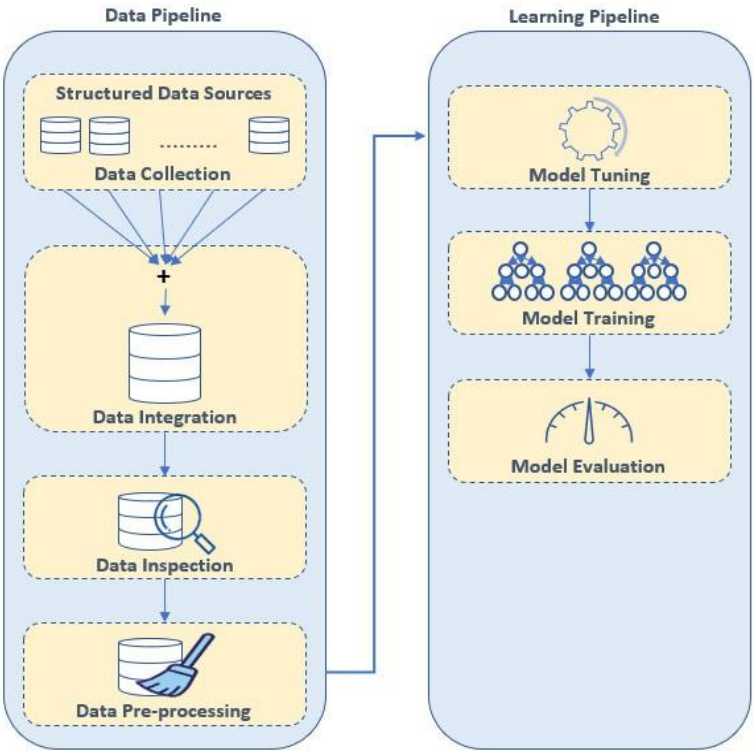


Fig. 1. The intelligent mental healthcare model for ASD detection.

### B. Learning Pipeline

Model tuning, model training, and model evaluation are the three successive stages that comprise the learning pipeline to develop a trained model for detecting ASD, as shown in Fig. 1. The model tuning is an iterative process that plays a substantial role in identifying the optimal values of the essential parameters associated with the selected supervised learning algorithm. Each iteration involves setting the other parameters to a constant value and testing the targeted parameter over a predetermined range of values. This process continues until all parameters have reached the optimal values. After that, the model is trained and evaluated using the supervised learning algorithm based on the optimal values of its parameters.

*1) Model tuning:* For the model, the RF algorithm was selected for its promising results in various healthcare studies with high-dimensional data. RF is an ensemble learning method that combines the predictions of multiple decision trees and aggregates their predictions, providing robustness and enhancing accuracy in the required task [5]. To achieve optimal performance, it is essential to tune the values of five distinct parameters:

- Two of them are for the preprocessing and dimensional reduction, which are:
  - The percentage of missing cells in a column at which this column will be deleted if it exceeds this percentage.
  - The percentage of testing data and training data to divide the dataset.
- The other three parameters that need to be tuned for setting the RF algorithm are:
  - In-bag-fraction, which determines the fraction of observations that are randomly selected with replacement for each bagged tree.
  - The number of trees in the forest.
  - The number of predictors to sample, which is the number of feature variables used to pick at random for each decision split at each node.

Five rounds were conducted to tune these parameters and select their optimal values, as shown in Fig. 2. A round was conducted to decide the best value for each parameter. In each round, the targeted parameter is tested with a given range of values while the other four parameters are given a fixed value. At the end of each round, the value that scored the highest accuracy is used to set this parameter in the following rounds, and so on over all the rounds till the best-scoring value for each parameter is determined.

*2) Model training and evaluation:* The model input training data comprises a feature matrix that includes instances representative of individuals with ASD as well as instances of typical control subjects; each instance is characterized by meticulously selected features. Subsequently, a testing phase is conducted utilizing a dataset that has not been previously encountered by the model, during which the model's performance is assessed through appropriate evaluation metrics. Ultimately, when presented with any new instance, the model is expected to identify individuals with ASD based on their respective features.

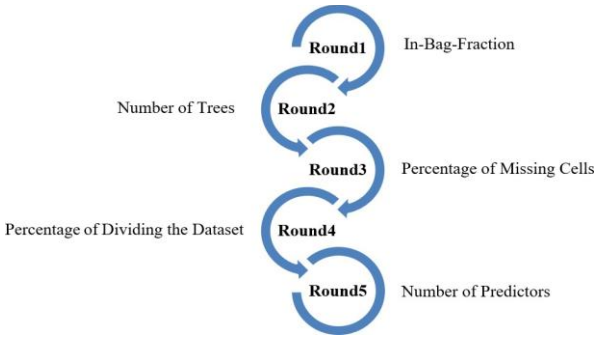


Fig. 2. Five sequential rounds of model tuning.

## IV. Experimental Results

### A. Dataset

The proposed model utilized ABIDE [5], an extensive electronic database of brain scans of individuals with ASD. More than 20 different sites shared in collecting this database to create two large-scale collections, ABIDE-I [20] and ABIDE-II [19]. Both collections provide resting state functional MRI (R-fMRI) along with phenotypic data for individuals with ASD and individuals with typical controls. The ABIDE-I [20] collection was established across 17 international brain imaging sites. ABIDE-I contains 1,112 cases, ranging between 539 ASD cases and 573 typical control cases.

The second collection, ABIDE-II [19], was publicly released in June 2016. Nineteen different international sites collaborated to donate 1,114 cases, ranging between 521 ASD cases and 593 typical control cases. After using ABIDE-I, many challenges and obstacles were found in it, such as the complex structure of the brain connectivity and the major heterogeneity of ASD; those issues were mapped for the essential need of a larger dataset that contains samples with more distinct characteristics [5], [21]. ABIDE-II was released to resolve these challenges and includes samples with better characteristics, features, and psychiatric variables [19]. Table II presents all 19 sites with their size and configuration.

*1) Phenotypic data:* The phenotypic data, in the ABIDE-II release, include a wide array of demographic, diagnostic, and behavioral measures for each participant. Many features are standardized scores from validated instruments for comparison across ages or test versions. In addition, patients' demographics and basic information, such as age, sex, handedness category, and handedness index score, were used in ABIDE-II. Around sixteen assessments were used in diagnosing ASD. They are as follows, highlighting the most essential features and their type and range that were used to build the dataset:

- Cognitive Ability (Intelligence Quotient – IQ): Verbal IQ, Performance IQ, and Composite IQ tests are administered and result in continuous scores in the range of 45-156.
- Assessments of social interaction, communication, restricted and repetitive behaviors, abnormal early development, stereotyped behaviors and restricted interests, imagination/creativity, social affect, and executive functions:
  - Autism Diagnostic Interview–Revised (ADI- R)
  - Autism Diagnostic Observation Schedule (ADOS)
  - Social Responsiveness Scale (SRS)
  - Social Communication Questionnaire (SCQ) and Autism Quotient
  - Repetitive Behavior Scale–Revised (RBS-R)
  - Behavior Rating Inventory of Executive Function (BRIEF)
  - Child Behavior Checklist (CBCL)
  - Adaptive Behavior (Vineland-II).

All scores resulting from these assessments are continuous numerical values. Higher scores indicate more ASD-related symptoms.

- Anxiety (Multidimensional Anxiety Scale for Children – MASC): MASC is a self-report questionnaire (child or adolescent) assessing anxiety symptoms. All scores are T-scores (mean=50, SD=10) for the MASC sub-scales: Physical Symptoms, Harm Avoidance, Social Anxiety, and Panic, plus an overall total and an Anxiety Disorder Index.
- Depression (Beck Depression Inventory – BDI): BDI examines depression symptoms and gives a score in the range (0-63) for each symptom.
- Language ability (Clinical Evaluation of Language Fundamentals – CELF): The CELF is a standardized language ability test. ABIDE II includes standard scores for Core Language, Receptive, and Expressive language abilities. Higher scores indicate better language skills.
- Other medical/behavioral details: Medication status indicates if the participant was on any psychotropic medications around the time of the scan and what they were (commonly ADHD meds, SSRIs, etc.). Body Mass Index (BMI) is included as an essential health metric.

### B. Data Integration, Inspection, and Preprocessing

The ABIDE-II individual datasets were concatenated into a unified file to ensure consistency and uniformity across the various sites. After the investigation, the first issue was the large amount of missing data. The data contains 347 features; the first two columns are unique identifiers. The third is the ground truth or the label for each entry that is filled with value 1 for ASD or 2 for TC (which changed after that to 0, Fig. 3 shows the difference before and after changing labels), and by which the dataset is divided into two classes ASD with 521 records and TC with 593 records. The remaining 344 columns are different features that are responses from the patient to the doctor's questions. By inspecting the data, many structural issues were found, leading us to the next stage of applying other data preprocessing techniques.

TABLE II. PHENOTYPIC DATA ON ABIDE-II FROM DIFFERENT SITES

| Site name | Samples Total size | Size of ASD | ASD Age Range | Size of TC | TC Age Range |
|---|---|---|---|---|---|
| Barrow Neurological Institute | 58 | 29 | 18 - 62 | 29 | 18 - 64 |
| Erasmus University Medical Centre | 54 | 27 | 6 - 11 | 27 | 6 - 10 |
| Neural Control of Movement Lab | 37 | 13 | 15 - 27 | 24 | 14 - 31 |
| Georgetown University | 106 | 51 | 8-13 | 55 | 8-13 |
| Indiana University | 40 | 20 | 17-54 | 20 | 19-37 |
| Institut Pasteur and Robert Debre Hospital | 56 | 22 | 6-27 | 34 | 8-47 |
| Katholieke Universiteit Leuven | 28 | 28 | 18-25 | - | - |
| Kennedy Krieger Institute | 211 | 56 | 8-13 | 155 | 8-13 |
| NYU Langone Medical Center Sample 1 | 78 | 48 | 5-34 | 30 | 5-23 |
| NYU Langone Medical Center Sample 2 | 27 | 27 | 5-8 | - | - |
| Olin Neuropsychiatry Research Center | 59 | 24 | 18-31 | 35 | 18-31 |
| Oregon Health and Science University | 93 | 37 | 7-15 | 56 | 8-14 |
| Trinity Centre for Health Sciences | 42 | 21 | 10-20 | 21 | 15-20 |
| San Diego Stat University | 58 | 33 | 7.4-18 | 25 | 8.1-17.7 |
| Stanford University | 42 | 21 | 8-13 | 21 | 8-13 |
| University of California Davis | 32 | 18 | 12-18 | 14 | 12-17 |
| University of California Los Angeles | 32 | 16 | 8-15 | 16 | 8-14 |
| University of Miami | 28 | 13 | 7-13 | 15 | 7-13 |
| University of Utah School of Medicine | 33 | 17 | 9-39 | 16 | 11-16 |
| Total | 1,114 | 521 | 5-62 | 593 | 5-64 |

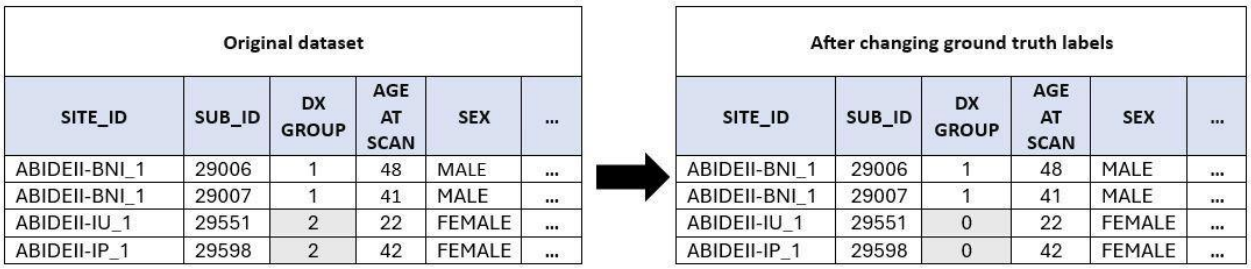


Fig. 3. Sample for the dataset after changing ground truth labels.

*1) Removing duplicate observations*. By observing the data, no duplicate records were found.

*2) Handling missing data*. Either by filling with appropriate values or removing the entire observation; the choice depends on the nature of the missing data and the specific requirements of the model.

*3) Filter undesired outliers*. Some data ranged from -100 to 100; we manipulated these values by shifting them by 100 to acquire non-negative values ranging from 0 to 200.

*4) Manipulating nonnumeric values*. The raw data contains six columns with a string data type. All string data has to be converted into numeric values to fit the data into the model. The Levenshtein distance algorithm was utilized for this task. In each column, the largest string value was picked and used as the reference string in the algorithm, and each string was replaced with the distance evaluated by the Levenshtein algorithm. Using the MATLAB tool, a function named convert_txt2num was written that takes a vector, which is a string column from the dataset, and it finds the longest string, then calls the implemented Levenshtein algorithm, which takes the string vector and the longest string as the reference value for the algorithm. The Levenshtein function returns a numeric vector that is used to replace the string vector. Algorithm 1 shows the pseudocode for the implemented convert_txt2num function that leverages the Levenshtein algorithm.

**Algorithm 1:** Function convert_txt2num that uses the Levenshtein algorithm

```
lenV ec ← length of each string in the vector (v)
maxInd ← max(lenV ec)
        longestStr ← v(maxInd) for each s in v do
numericVector(index(s)) = Levenshtein(s, longestStr)
end for
```

## C. Model Tuning

After performing all the steps of the data pipeline, as mentioned previously, five different parameters have to be tuned, two of them for data cleaning and preprocessing steps,

which are the percentage of missing cells in a column at which this column will be deleted if it exceeds this percentage, and the percentage of testing data for training data to divide our dataset. Regarding the other three parameters need to be tuned are for the building the RF model which are: number of used trees in the forest and the in bag fraction which is determining the fraction of observations that are randomly selected with replacement for each bagged tree and the last parameter is the number of predictors to sample which is the number of feature variables used to pick at random for each decision split at each node. To tune these parameters and select their best values, we created five rounds: a round to decide the best value for each parameter. In each round, the targeted parameter is tested with a given range of values while the other four parameters are given a fixed value. At the end of each round, the value that scored best is used to set this parameter in the following rounds, and so on over all the rounds till the best-scoring value for each parameter is determined.

*5) Round 1:* The first parameter to tune is the In-Bag-Fraction parameter for the RF model. This parameter refers to the number of instances that will be used in each bag/tree in the forest to train with them (sample size drawn randomly for the training of each bag/tree). The value is a fraction of the whole number of available instances in the training data. Setting this parameter with a fraction less than one means that each tree will train with a subset of the training data that is selected randomly. So, decreasing the covariance and avoiding the over-fitting problem. Fig. 4 shows the effect of changing this fraction on the model accuracy. The other parameters were set to their default values, which are 64 trees, 20% of the data for testing, and 80% for training. If the column contains more than 50% empty cells, then delete this column/feature. And finally, the number of predictors to choose from each node was set by the square root of the number of features, which is the default value given by Breiman [22]. By observing the accuracy in Fig. 4, the highest value for the accuracy was achieved by using the fraction 0.6. So, it is the optimal value to set this parameter in the model.

*6) Round 2:* The second parameter to tune is the number of trees used to build the forest. A study was published to determine how many trees should be found in the RF, and it stated that the best values will be found in the range (64-128) [23]. Training multiple trees is one of the most essential features of RF; a few trees can result in bad predictions, and a vast number can cause performance issues. To tune this parameter, the value of all other parameters was set as follows: 20% of the dataset for testing, the column is deleted if it contains 50% or more empty cells, and the number of predictors to choose from on each node is the square root of the number of features. Finally, the in-bag fraction was set with the best value from round 1, which is 0.6. The model was tested with a broader range, which is for a minimum of 30 trees and a maximum of 200 trees. By observing the rate of change of the model accuracy by changing the number of trees in the RF in Fig. 5, the optimal value to tune the number of trees parameter is 100 trees.

*7) Round 3:* The third parameter to tune is a dataset-related parameter, which is the percentage of empty cells that can be found in a column. This round aims to determine the acceptable percentage limit for empty cells within each feature, without losing significant information in the dataset. After the first two rounds, the in-bag fraction and the number of trees were tested for their best values, which were 0.6 and 100, respectively. In this round, the data will be split into testing and training sets in a 20:80 ratio, with the number of predictors set to the square root of the total predictors identified in the input data. The required parameter was tested using percentages ranging from 100% to 10%. Fig. 6 shows the accuracy assessed at each percentage value. As depicted in Fig. 6, the peak accuracy occurs at 70%, indicating that any column with over 70% of its cells empty will be removed, consequently decreasing the number of feature vectors in the dataset.

*8) Round 4:* The fourth parameter to tune is the percentage needed from the dataset to be used for testing and the rest for training. Determining the optimal percentage for this division directly impacts the performance of the model, as it needs a suitable amount of data that yields robust results and a scalable model. Ensuring enough data for training and testing can prevent problems such as overfitting -if too much data is used for training- and underfitting -if too little data is used for training- and so balance is achieved between model generalizability and its predictive power. The round started with a ratio of 15:85, then increased it to 20:80, then 25:75, and finally 30:70. Fig. 7 shows the variation in accuracy through these different ratios. The other parameters were set by the following values: 64 trees, in-bag-fraction was 0.6, and the number of predictors selected randomly at each node was 71. If the column contains more than 70% empty cells, then delete this column/feature. The experiment showed that allocating 20% of the data for testing and 80% for training gives the model the highest accuracy, so the ratio 20:80 was chosen to tune this parameter.

*9) Round 5:* The fifth parameter to determine is the number of predictors selected randomly at each node during tree construction to choose the best split from them, denoted by variable $m$. It is a critical parameter as it controls the randomness of feature selection and influences the diversity of the model, ensuring robustness and reliability. This parameter should be carefully adjusted due to its effect on the model performance, as high values of m might cause high computation cost, but low values might also reduce model diversity or even cause overfitting. The model was tested with m values ranging from 2 to 20, and its accuracy was observed. Fig. 8 clearly determined that $m$=5 scored best, so it will be used in the final model training phase.

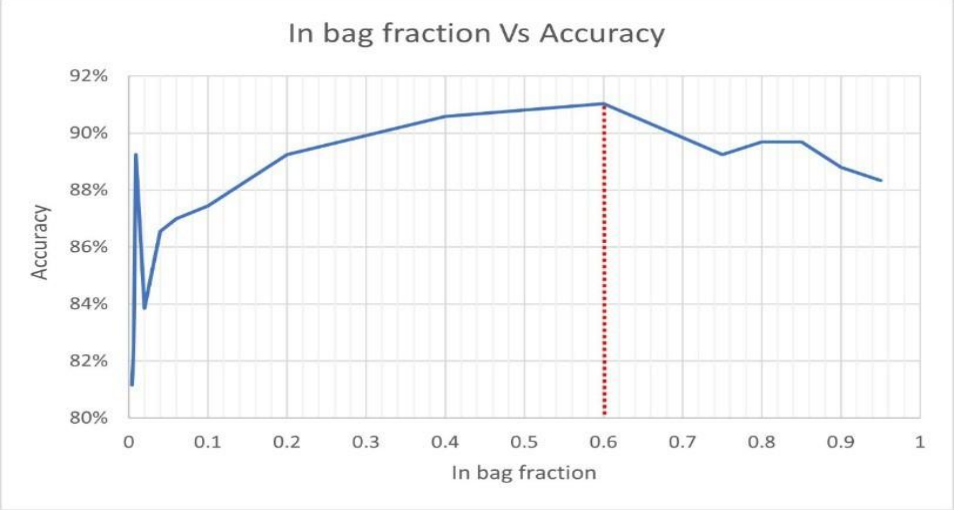


Fig. 4. Tuning the InBagFraction parameter for the random forest.

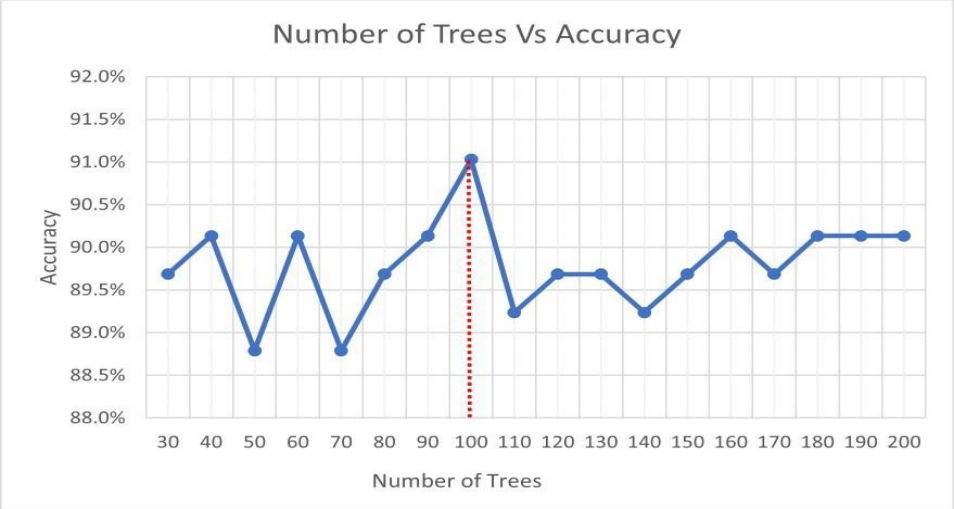


Fig. 5. Tuning the number of trees required to build the forest.

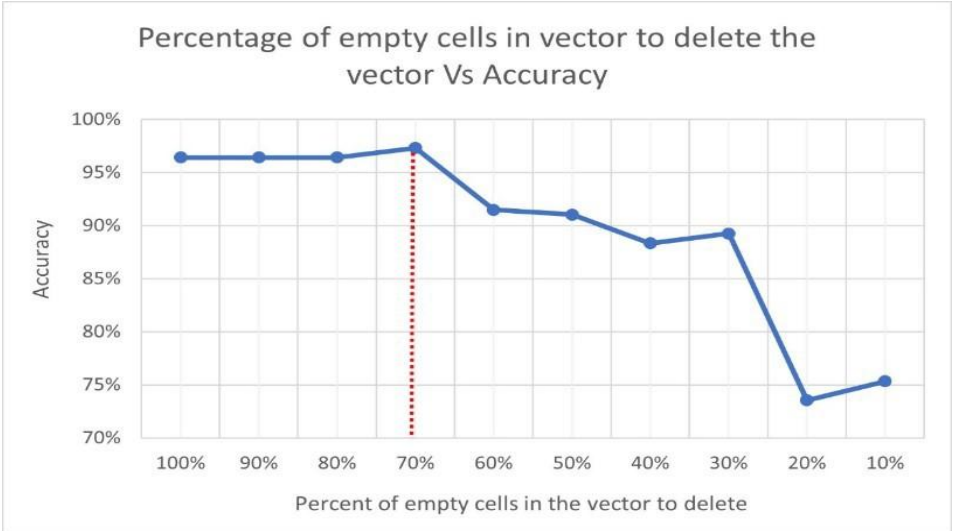


Fig. 6. Tuning the percentage of empty cells.

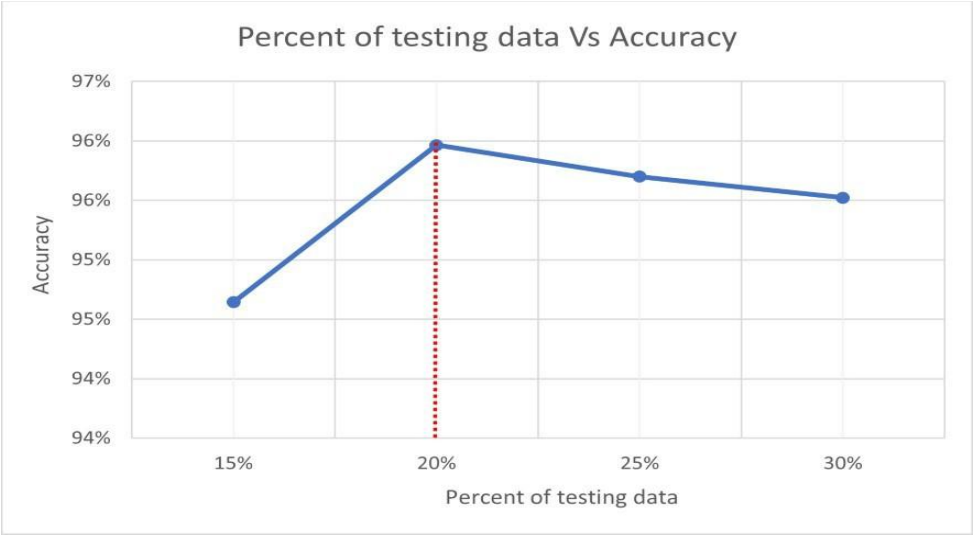


Fig. 7. Tuning the percentage of testing data from the original dataset.

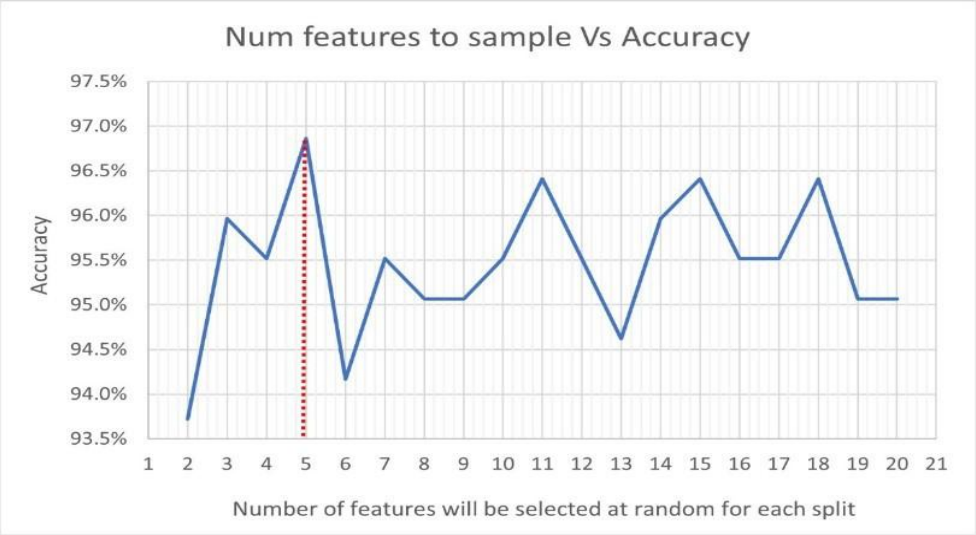


Fig. 8. Tuning the number of predictors selected randomly at each node.

## V. Results and Discussion

This research proposed an intelligent model with a tuning phase based on the ML pipeline and systematically optimized five key RF parameters: the number of trees, in-bag-fraction, percentage of training and testing data, number of features selected randomly at each node, and percentage of empty cells in which a column will be deleted. This optimization phase proved critical, as it significantly enhanced the classifier's performance compared to default configurations. The experimental results demonstrated that the optimized RF classifier performed with high accuracy after applying the tuning phase. Furthermore, to assess the RF model against other classifiers' accuracies, multiple ML classifiers were employed to evaluate the classification performance on the ABIDE-II phenotypic dataset. The models used include Random Forest RF, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Fine Gaussian SVM, Weighted KNN, and Random Forest integrated with Principal Component Analysis. MATLAB served as the primary programming platform for data handling, model development, and performance evaluation. The configurations for each model are detailed below:

*1)* Random forest: The Random Forest classifier was constructed using thirty decision trees. At each decision node, a random subset of predictors was selected, with the number of predictors set to the square root of the number of observations. The maximum number of splits per tree was set equal to the total number of observations, allowing for fully grown trees. These parameters were selected to balance model complexity with computational efficiency while maintaining robust performance across the training and testing sets.

*2)* Support vector machine: A linear kernel function was employed to construct the separating hyperplane. This configuration was chosen for its simplicity and efficiency in high-dimensional spaces.

*3)* Fine K-Nearest neighbors: This model used a single nearest neighbor ($k$=1) with the Euclidean distance metric. The distance weight was set to automatic, allowing the algorithm to select the most suitable weighting scheme based on the dataset.

*4)* Decision tree: The decision tree classifier was limited to a maximum of 100 splits. Gini's diversity index was used as the split criterion to evaluate the quality of each node partition.

*5)* Fine Gaussian support vector machine: A Gaussian (radial basis function) kernel was utilized to allow for nonlinear decision boundaries in the feature space.

*6)* Weighted K-Nearest neighbors: This variant of the KNN algorithm used ten nearest neighbors ($k$=10) with the Euclidean distance metric. Distance weights were computed using a squared inverse function, giving greater influence to nearer neighbors during classification.

*7)* Random forest with principal component analysis: Dimensionality reduction was performed using PCA prior to training the RF model. PCA components were selected to retain 95% of the total variance in the dataset. The RF classifier was configured identically to the standard RF model, with thirty trees, the square root of the number of observations as the number of predictors sampled at each node, and a maximum of splits equal to the number of observations.

The models attained accuracies of 95.06% with RF, 93.27% with SVM, 91.29% with KNN, 94.61% with DT, 77.47% with fine Gaussian SVM, 91.11% with weighted KNN, and 84.02% with RF+PCA, as shown in Fig. 9. The optimized RF model achieved the highest classification accuracy of 96.86% across all models. It consistently outperformed other classifiers. This finding highlights the importance of the model tuning stage and aligns with and reinforces observations from the literature, which often report superior performance of ensemble methods in similar high-dimensional healthcare datasets.

Moreover, this research underscored the pivotal role of data preprocessing, particularly data cleaning methods and dimensionality reduction, in improving model accuracy and generalization. This finding is consistent with previous studies that have highlighted how poor data preparation can severely degrade model performance, regardless of the classifier used.

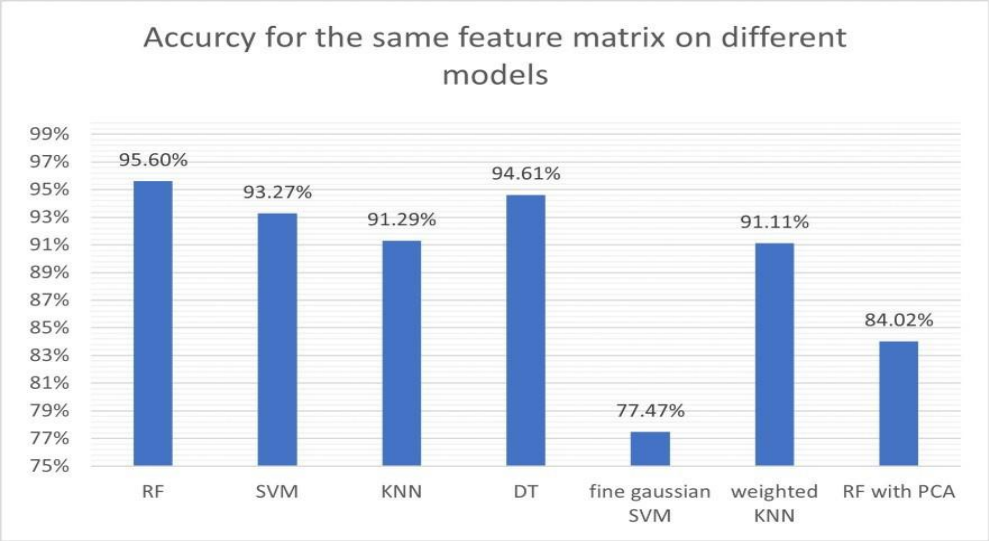


Fig. 9. Accuracy of different ML models on the same feature matrix.

## VI. Model Validation and Empirical Confirmation

To further validate the predictive capabilities of the model and confirm its real-world applicability, an empirical validation study was conducted. While the model demonstrated high accuracy metrics on the ABIDE-II dataset, a separate qualitative approach was employed to assess its performance on a set of independent, non-structured data points. This methodology was designed to simulate a real-world diagnostic scenario, moving beyond the quantitative features of the training data.

A search for anecdotal accounts and personal narratives from parents describing their children's autistic traits was performed. These qualitative descriptions, sourced from publicly available forums and support group websites, provided a rich, narrative-based perspective on the diagnostic process. Five distinct case studies were curated, each representing a unique constellation of behaviors and characteristics commonly associated with ASD.

The narratives were meticulously analyzed to extract values for the key features utilized by the model. This process involved a careful interpretation of the qualitative data, mapping descriptive language to the quantitative feature scales used during model training. For example, a parent's description of a child's intense focus on specific objects and resistance to change was translated into a high value for the "restricted and repetitive behaviors" feature.

The conversion of unstructured, qualitative narrative data into the quantitative, structured feature vector required by the trained RF model was accomplished using Gemini [24], a Large Language Model (LLM), as an advanced research agent

for automated feature extraction. The process involved a highly specific Prompt Engineering strategy:

- Input: The de-identified parental narrative was supplied as the primary unstructured text.
- Instruction set: The prompt explicitly instructed Gemini to act as a clinical coder and to map the descriptive language onto the precise quantitative features used in the ABIDE-II training data.
- Scale definition: The prompt provided the LLM with a defined, ordinal scale (e.g. 1-5 or a binary 0/1) for each feature, compelling the model to translate subjective descriptions (e.g. "rarely makes eye contact") into a numerical rating (e.g. "Social Reciprocity =4"). After that, the extracted numerical values were manually validated to avoid the potential for LLM Hallucinations or the ambiguous nature of subjective human language.
- Output format: The LLM was instructed to output the extracted feature values in a structured format (e.g. a delimited list) to facilitate direct integration into the classification model.

This utilization of an LLM allowed for the rapid and systematic transformation of complex qualitative data into a format suitable for computational analysis, a process that traditionally requires extensive manual human coding.

The extracted feature values for each of the five cases were then input into the trained model. The model's classification output for each case was subsequently compared against the known diagnosis (ASD vs TC) from the original narrative. The model correctly classified four out of the five cases, yielding an accuracy rate of 80% on this small, independent dataset.

This successful classification rate on qualitative data provides a robust, real-world confirmation of the model's accuracy and effectiveness. It demonstrates the model's ability to generalize beyond the structured data of the ABIDE-II database and accurately interpret complex, narrative-based descriptions of ASD symptoms. While this validation is not a substitute for a large-scale clinical trial, it provides strong evidence of the model's potential utility in a preliminary screening context.

## VII. Conclusion

This research aimed to advance the field of ASD detection by proposing a data-driven, ML-based framework, with a specific focus on the optimization of the RF classifier. It contributes to the growing body of work exploring the role of AI in the early and accurate identification of ASD, a complex neurodevelopmental condition characterized by a wide spectrum of symptoms. The proposed model highlighted the potential of leveraging a comprehensive set of phenotypic features and applying an ML algorithm to develop a reliable and accurate model, which can guide researchers in their studies. Healthcare professionals can also use it to assist them in making informed decisions and improving the efficiency of ASD screening processes.

Phenotypic data, encompassing a wide range of observable traits and characteristics, serve as valuable indicators for understanding the heterogeneity of ASD. Behavioral, cognitive, and physiological features have been extensively studied to identify patterns that differentiate individuals with ASD from their typically developing counterparts. These phenotypic markers often include social communication abilities, repetitive behaviors, sensory sensitivities, and developmental milestones. Research in the field has emphasized the importance of a comprehensive and multidimensional approach to phenotypic characterization, considering the diverse manifestations of ASD across individuals.

This research systematically adjusted the key hyperparameters for RF optimization. The experimental results spot the optimal values to tune the model and its input data which were: 0.6 for in-bag-fraction, 100 trees to build the forest, delete columns with more than 70% empty cells, number of features selected at random for each split is 5 and finally divide the dataset with 20% for testing and 80% for training. By tuning the model with those values, it achieved promising results with a classification accuracy of 96.86%, a sensitivity of 97.14%, and a false positive rate of 3.39%. Achieving 96.86% accuracy on the ABIDE-II dataset suggests that there are detectable signatures of ASD in phenotypic data that ML can pick up. Furthermore, this study provides evidence that ML can be effectively applied to large multi-site phenotypic datasets to identify individuals with autism based on behavior and medical patterns. This model presents a scalable and expedited alternative to current models, thereby aligning effectively with the clinical demand for efficient diagnostic tools.

The empirical validation study successfully bridges the gap between the RF model's high quantitative performance on structured datasets and its real-world applicability. By successfully translating qualitative, narrative-based descriptions into actionable data, the study demonstrated the model's ability to generalize beyond its training data and accurately interpret complex, non-structured information. The resulting 80% accuracy on this small, independent dataset provides strong evidence of the model's potential utility as a preliminary screening tool. While not a substitute for comprehensive clinical evaluation, this validation confirms the model's robustness and its capacity to assist in the early identification of ASD in practical, diagnostic scenarios, thereby paving the way for future clinical integration and large-scale validation studies.

While this study demonstrates the efficacy of the proposed approach, it is important to acknowledge certain limitations. Firstly, the performance of the model heavily relies on the quality and availability of phenotypic data. Obtaining comprehensive and standardized data across different populations and settings may pose challenges. Additionally, the generalizability of the model to diverse populations needs to be explored further to ensure its applicability in different cultural and geographical contexts.

Several avenues for future research warrant consideration. The integration of additional data sources and the incorporation of multimodal data, such as genetic and neuroimaging data, to further enhance the accuracy and predictive power of the

model. Explore hybrid AI models that combine ML models (e.g., SVM + RF) with DL architectures (e.g. CNN + Transformers) for improved accuracy. Moreover, the development of user-friendly tools and applications based on the model's findings could facilitate widespread adoption and implementation in clinical settings. Continued research and innovation in this field are crucial to further advance our understanding and detection capabilities for ASD.

## DATA AVAILABILITY

The ABIDE dataset used in our current study is publicly available on the Autism Brain Imaging Data Exchange repository. https://fcon_1000.projects.nitrc.org/indi/abide/fornon-commercial research purposes.

## REFERENCES

[1] "World health organization." https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. (accessed Dec. 28, 2024).

[2] R. P. Walensky, R. Bunnell, J. Layden, C. K. Kent, A. J. Gottardy, M. A. Leahy, et al., "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018," tech. rep., United States, 2021.

[3] R. G. Schwartz, Handbook of child language disorders: Second edition. Exon Publications, 2017.

[4] J. F. Underwood, K. M. Kendall, J. Berrett, C. Lewis, R. Anney, M. B. Van Den Bree, et al., "Autism spectrum disorder diagnosis in adults: Phenotype and genotype findings from a clinically derived cohort," *British Journal of Psychiatry*, vol. 215, no. 5, pp. 647–653, 2019.

[5] "Autism Brain Imaging Data Exchange." https://fcon 1000.projects.nitrc.org/indi/abide/. (accessed Dec. 28, 2024).

[6] A. Di Martino, D. O'connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, et al., "Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii," *Scientific data*, vol. 4, no. 1, pp. 1–15, 2017.

[7] N. S. El-Askary, M. A. Salem, and M. I. Roushdy, "Feature extraction and analysis for lung nodule classification using random forest," *ACM International Conference Proceeding Series*, pp. 248–252, 2019.

[8] N. S. El-Askary, M. Mabrouk Morsey, A. M. Mahmoud, and T. I. El-Arif, "On Development of Computer Aided System for Detecting Brain Neurological Disorders," *Proceedings - 11th IEEE International Conference on Intelligent Computing and Information Systems, ICICIS 2023*, pp. 94–101, 2023.

[9] S. Perochon, J. M. Di Martino, K. L. Carpenter, S. Compton, N. Davis, B. Eichner, et al., "Early detection of autism using digital behavioral phenotyping," *Nature Medicine*, vol. 29, no. 10, pp. 2489–2497, 2023.

[10] P. Mazumdar, G. Arru, and F. Battisti, "Early detection of children with Autism Spectrum Disorder based on visual exploration of images," *Signal Processing: Image Communication*, vol. 94, no. January, p. 116184, 2021.

[11] J. Gutie´rrez, Z. Che, G. Zhai, and P. Le Callet, "Saliency4asd: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder," *Signal Processing: Image Communication*, vol. 92, p. 116092, 2021.

[12] H. Alkahtani, T. H. Aldhyani, and M. Y. Alzahrani, "Deep learning algorithms to identify autism spectrum disorder in children-based facial landmarks," *Applied Sciences*, vol. 13, no. 8, p. 4855, 2023.

[13] "Kaggle." https://www.kaggle.com/. (accessed Dec. 28, 2024).

[14] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[15] T. Akter, M. S. Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lio, et al., "Machine learning-based models for early stage detection of autism spectrum disorders," *IEEE Access*, vol. 7, pp. 166509–166527, 2019.

[16] U. Erkan and D. N. Thanh, "Autism Spectrum Disorder Detection with Machine Learning Methods," *Current Psychiatry Research and Reviews*, vol. 15, no. 4, pp. 297–308, 2019.

[17] P. Bawa, V. Kadyan, A. Mantri, and H. Vardhan, "Investigating multiclass autism spectrum disorder classification using machine learning techniques," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, no. May, p. 100602, 2024.

[18] M. Kunda, S. Zhou, G. Gong, and H. Lu, "Improving Multi-Site Autism Classification via Site-Dependence Minimization and Second-Order Functional Connectivity," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 55–65, 2023.

[19] A. Di Martino, D. O'Connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, et al., "Enhancing studies of the connectome in autism using the autism brain imaging data exchange II," *Scientific Data*, vol. 4, pp. 1–15, 2017.

[20] A. Di Martino, C. G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, et al., "The autism brain imaging data exchange: Towards a large- scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[21] C. M. Williams, H. Peyre, R. Toro, A. Beggiato, and F. Ramus, "Adjusting for allometric scaling in ABIDE I challenges subcortical volume differences in autism spectrum disorder," *Human Brain Mapping*, vol. 41, no. 16, pp. 4610–4629, 2020.

[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[23] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7376 LNAI, pp. 154–168, 2012.

[24] Google, "Gemini 2.5 flash, large language model." https://gemini.google.com/, 2025. (accessed May 2, 2025).