Report

A Hybrid GA-PSO Optimized Machine Learning Framework with Strategic Data Augmentation for Early Autism Spectrum Disorder Detection and Risk Factor Analysis

Predictive Framework for Early Detection of Autism Spectrum Disorder Using Hybrid GA-PSO Optimization, CTGAN Data Augmentation, and Sensitivity Analysis

Submitted By

Argina Akter
BFH2211MSc106F
Session: 2021-22

Supervised By

Professor Dr. Md. Ashikur Rahman Khan

Professor & Chairman

Department of Information and Communication Engineering

Noakhali Science and Technology University

A Report for the Master of Science (Engg.) in Information and Communication Engineering

Noakhali Science and Technology University

Noakhali - 3814

DECLARATION

This paper is presented to the Department of Information and Communication Engineering, Noakhali, Science and Technology University, Noakhali-3814, in part fulfilling the requirement of the award of the M.Sc. degree in ICE. This is also required to certify that the research work falls under the M.Sc. degree in ICE under the course entitled ICE-5312. Hence, I hereby make known that this thesis paper has not been presented to any other institution under the requirement of any form of degree, diploma, or publication.

.....

Argina Akter

Roll No: BFH2211MSc106F

Year: 02, Term: 01

Department of Information and Communication Engineering

Noakhali Science and Technology University

Noakhali- 3814

ACCEPTANCE

This thesis report is accepted and submitted to the Department of Information and Communication Engineering (ICE), Noakhali Science and Technology University, Noakhali-3814, in partial fulfillment of the requirements for having the M.Sc. (Engg.) degree in ICE.

-

…….

Signature of Supervisor

Professor Dr. Md. Ashikur Rahman Khan

Professor & Chairman

Department of Information and Communication Engineering

Noakhali Science and Technology University

Noakhali- 3814

ACKNOWLEDGEMENT

I am grateful and would like to express my sincere gratitude to my supervisor Professor Dr. Md. Ashikur Rahman Khan for his germinal ideas, invaluable guidance, continuous encouragement and constant support in making this research possible. He has always impressed me with his outstanding professional conduct, his strong conviction for science, and his belief that a B.Sc. program is only a start of a life-long learning experience. I appreciate his consistent support from the first day I applied to graduate program to these concluding moments. I am truly grateful for his progressive vision about my training in science, his tolerance of my native mistakes, and his commitment to my future carrier. My sincere thanks to all my classmates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways. Last but not the least, I would like to thank my parents for encouraging me and standing with me throughout my life.

Abstract

This research presents a robust machine learning framework for the early detection of Autism Spectrum Disorder and the analysis of its associated risk factors. The study leverages a comprehensive dataset of 288 individuals, encompassing behavioral, developmental, familial, and socio-demographic attributes. Five machine learning classifiers—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression—were implemented and enhanced through a novel two-pronged approach. First, a true hybrid Genetic Algorithm-Particle Swarm Optimization technique was employed for superior hyper-parameter tuning, significantly boosting model performance. Second, strategic data augmentation was performed using Conditional Tabular Generative Adversarial Networks combined with boundary sampling to effectively mitigate data scarcity. The optimized Random Forest model achieved exceptional performance, with an accuracy of 98.85% and an Area Under the Receiver Operating Characteristic Curve of 0.9989. The synergistic combination of GA-PSO optimization and data augmentation proved transformative, enabling multiple models to achieve perfect classification metrics of 100% accuracy on augmented datasets. Model robustness was rigorously validated through stability analysis across multiple data splits (70:30, 75:25, 80:20) and, critically, on a completely unseen test set, where the best model maintained a high accuracy of 94.1%. Beyond predictive performance, a sensitivity analysis was conducted to identify and rank the most influential risk factors. "Limited Speech" emerged as the most critical predictor, followed by parental age at childbirth (i.e., "Mother age at childbirth" and "Father age at childbirth") and key developmental milestones such as "Ability to walk" and "Ability to respond when called by name". This study conclusively demonstrates that the integration of hybrid metaheuristic optimization and strategic data augmentation creates a highly accurate, reliable, and generalizable tool for early ASD screening. The framework provides a practical solution for overcoming data limitations in clinical settings and offers data-driven insights into key risk factors, thereby supporting earlier intervention and informed clinical decision-making.

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by social interaction challenges, restricted interests, and repetitive behaviors. Early detection is

critical for timely intervention, yet conventional diagnostic methods are often subjective, time-consuming, and inaccessible in low-resource settings. This research develops a robust machine learning-based framework for ASD detection using a multidimensional dataset of 288 individuals aged 3–27 years, encompassing behavioral, developmental, health, familial, and socio-demographic attributes. Five classifiers—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression—were implemented and further enhanced through hybrid Genetic Algorithm–Particle Swarm Optimization (GA-PSO) for hyper-parameter tuning. To overcome data scarcity, Conditional Tabular GAN (CTGAN) with boundary sampling was employed for synthetic data augmentation. The optimized and augmented models achieved significant performance improvements, with Random Forest enhanced by GA-PSO yielding near-perfect accuracy (>98%, AUROC ≈ 0.999). Rigorous validation, including stability analysis across multiple data splits and testing on a completely unseen dataset, confirmed the model's reliability and generalizability. Sensitivity analysis identified and ranked key risk factors, with limited speech, parental age, and developmental milestones emerging as dominant predictors. This integrated framework demonstrates the potential of hybrid optimization and augmentation strategies to deliver an accurate, scalable, and clinically meaningful screening tool for ASD, bridging the gap between machine learning innovation and real-world healthcare needs.

Keywords: Autism Spectrum Disorder (ASD), Machine Learning, Hybrid GA-PSO Optimization, Hyper-parameter Tuning, Data Augmentation, CTGAN, Risk Factor Analysis, Sensitivity Analysis, Early Detection.

TABLE OF CONTENTS

Page No

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
MLP	Multilayer Perceptron
LR	Logistic Regression
GA	Genetic Algorithm

PSO	Particle Swarm Optimization
CTGAN	Conditional Tabular **G**enerative Adversarial Network

CHAPTER 1

INTRODUCTION

1.1 Introduction

Autism spectrum disorder (ASD) is characterized by growth-related difficulties within the neurological system [1]. Autism spectrum disorder is also known as ASD. In ASD, the term "spectrum" implies that every child is unique and possesses certain characteristics [2]. Approximately 1.47% of all children worldwide suffer from ASD. The prevalence of the condition reached roughly 1 in 45 children according to the 2014 National Health Interview Survey, with the largest rates in children. One out of every 59 children in the United States is affected, which accounts for roughly one-third of all cases worldwide. According to the WHO, 1 in 160 children worldwide suffers from ASD [3]. Symptoms appear in early infancy, and the illness lasts the entirety of a person's life [4]. ASD is marked by limited and cyclical behavior [5]. With this illness, some will be able to live self-reliantly, while others will require lifelong nurturing and guidance [6]. Compared to the general population, people with ASD struggle much more with early development [7]. Repetitive behaviors are a result of ASD, which primarily affects social interactions and communication [8]. Among the many difficulties faced by people with autism are learning disabilities, concentration problems, mental health conditions like anxiety and depression, and problems with their motor and sensory systems [9]. Symptoms are different for each individual [10]. An IQ of less than 70 indicates an intellectual disability in 31% of children with ASD. Forty-four percent have IQs that are normal or above normal, above 85, while 25 percent have IQs that are on the edge of normal, between 71 and 85 [11]. Even though the precise cause of ASD remains unclear, studies indicate that biological indicators such as genetic malformations, brain inflammation, and perinatal abnormalities can contribute to ASD [3]. Having an aging parent, a sibling with ASD, and a low birth weight are some risk factors that may have an impact [12]. Researchers are looking into genetic, socio-demographic, and familial factors because autism has a significant hereditary component [13]. However, conventional behavioral research finds it extremely difficult to detect and diagnose ASD [14]. It involves a great deal of time and money to conduct screening studies to identify traits of autism [15]. Formal or semi-structured questions based on a child's behavior are used in ASD screening, and responses from parents, relatives, or other caregivers are examined for signs of autism [16]. It takes a great deal of time and effort to identify ASD early on using standard clinical procedures, which can complicate matters [17]. Also, it is difficult to diagnose ASD due to because of the lack of and effective medical or blood test to detect the disease [18]. By observing the child's behavior, doctors can begin a diagnostic to enhance their quality of life [19]. The DSM-V defines ASD as an individual

who has difficulties in two primary domains: social engagement and discourse, and redundant and limited actions [20]. Between the ages of 18 and 24 months, accurate results for the detection of ASD can be obtained [3]. Although nowadays many diseases may be identified early on, in childhood, there remains the case when the indicators are not identified prior to adolescence or grown up years [21]. ASD is over four times more in boys in compared to girls, and it has been noted in individuals of all racial, ethnic, and socioeconomic backgrounds [22].

This study utilized a dataset of 288 people between the ages of 3 and 27 to develop a predictive model for autism spectrum disorder detection. The data included behavioral characteristics, developmental milestones, physical health indicators, family history, and socio-demographic indicators. Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and Multilayer Perceptron are five machine learning algorithms that were implemented and improved through more advanced methods. A hybrid GA-PSO method optimized hyper-parameters, while CTGAN with boundary sampling was used for strategic data augmentation. The model was tested and stabilized with different training and testing ratios (70:30, 75:25, 8-:20). To ensure that the results were reproducible, final validation was performed on a completely independent set of 17 samples. A table of major factors and factor combinations was produced by sensitivity analysis, which ranked the individual or combined major factors based on their significance for each prediction month. An accurate and trustworthy ASD screening tool is offered by this hybrid optimization algorithms/augmentation/verification approach, which is incredibly resilient.

1.2 Justification of the Research

Autism Spectrum Disorder (ASD) is a intricate neurodevelopmental condition marked by challenges in social exchange, restricted interests, and recurrent behaviors. Timely detection and interference are vital to improve long-term consequences, yet many children remain undiagnosed until later childhood, delaying access to essential therapies. This research is justified by the urgent need for improved ASD detection methods and a deeper understanding of its risk factors to enhance early diagnosis, intervention, and public health strategies.

(i) Importance of Early Detection

Early identification of ASD allows for rapid action, which can significantly boost cognitive, linguistic, and adaptive functioning. However, current diagnostic tools often lean on interpretive behavioral appraisals, bringing about instability in diagnosis. Advanced detection methods, such as machine learning models analyzing behavioral, genetic, and neuroimaging data, can offer more objective and spot-on timely screening. This work endeavors to evolve and validate such tools to reduce diagnostic delays and improve accessibility, particularly in underserved regions.

(ii) Understanding Risk Factors for Prevention and Awareness

The etiology of ASD involves a fusion of inherited, environmental, and epigenetic factors. While genetic predispositions (e.g., mutations, family history) have a significant impact in ASD, emerging evidence highlights the significance of:

- **Prenatal and perinatal factors** – Health status of mother of child (e.g., diabetes, hypertension), infections, stress, sorrow, distress, and anxiety during pregnancy may influence fetal neurodevelopment.
- **Parental characteristics** – Advanced parental age, blood group compatibility (e.g., Rh factor mismatches), and significant age gaps between parents have been hypothesized to contribute to ASD risk, warranting further investigation.
- **Socio-demographic influences** – Socioeconomic status, maternal education, and access to healthcare may affect both risk exposure and diagnostic delays.

By systematically analyzing these modifiable and non-modifiable factors, this research will provide evidence-based insights for preventive strategies, genetic counseling, and public health policies.

(iii) Addressing Disparities in Diagnosis and Care

Disparities in ASD diagnosis persist across socioeconomic, racial, and geographic lines. Minority and low-income populations often face barriers to early screening and intervention. This research can reduce disparities in the assessment and treatment of ASD by integrating digital tools to detect ASD and address cultural sensitivity and exploring the risk factors among different populations

(iv) Public Health and Policy Implications

It is possible to improve the detection and risk factor analysis in the sphere of public health policies in order to allocate resources more efficiently and provide early intervention programs, specific screening programs, and other actions. Governments and medical professionals can use these results to develop affordable ASD management plans. In conclusion, this study can be crucial to advancing early diagnosis of ASD, understanding its multifaceted etiology, and promoting equal access to health services. The research aims to help in the global contribution towards the handling of cognitive developmental disorders and enhancing the outcome of individuals with ASD and their respective families through interdisciplinary methods and technology.

1.3 Problem Statement

Early detection of ASD using traditional clinical techniques is time-consuming and cumbersome and that can cause complexity [3]. It is also a problem that no proper medical test or blood test exists to diagnose ASD [4].It is suspected that biological factors such as genetic defects, brain inflammation, and prenatal abnormalities are involved in the etiology of ASD [5]. Autism Spectrum Disorder is a cognitive developmental disorder that is difficult to understand and have caused problems in social exchange, and recurrent behavior [6]. Convenient intervention associated with ASD severity evaluation and early diagnosis is essential to determine timely improvements in developmental outcomes [7]. But conventional diagnostic methods can be lengthy, expensive, and demands the availability of specialized clinical knowledge, not always present, particularly in low-resource or rural areas [8].

The lack of a biomarker or a standardized medical test only adds to the diagnostic situation and makes mass screening impossible. Therefore, the creation of an objective, reliable and accessible auxiliary tool to help with ASD early screening is highly in demand. This paper seeks to satisfy this major gap by using machine learning to develop a powerful predictive model to detect ASD. This study will utilize a multidimensional screening framework to create and test the predictive validity of a structurally diverse set of behavioral, developmental, familial, and socio-demographic variables of individuals aged 3-27 through the use of a comprehensive dataset.

The secondary ambition of this study is to recognize and rank the most influential risk factors contributing to ASD likelihood through a sensitivity analysis. This analysis seeks to provide actionable insights into modifiable and non-modifiable risks, potentially informing preventive strategies and guiding clinical counseling. By integrating advanced computational techniques into a unified predictive framework, this research ultimately seeks to enhance diagnostic accuracy, mitigate delays in intervention, and provide a scalable tool to support healthcare decision-making, thereby boosting the standard of life for individuals with ASD and their families.

1.4 Objectives

This research paper will work on autism spectrum disorder detection using several machine learning technique. There are several objectives of this research we have enlisted. Among all of them the main objectives of this research are given below:

1. To experiment on real-world autistic and traditionally developed children datasets

2. To implement and evaluate the performance of five machine learning classifiers for the binary classification of ASD.

3. To optimize classifiers hyper-parameters using a hybrid GA-PSO technique and quantitatively evaluate the performance enhancement.

4. To validate model robustness and generalizability through stability analysis across data splits and final testing on a completely unseen dataset.

5. To investigate CTGAN augmentation with boundary sampling at varying multipliers (2x, 4x, 7x) for overcoming data scarcity and improving classification results.

6. To perform a comparative analysis of standalone optimization, standalone augmentation, and their combined use (GA-PSO with CTGAN) to determine the optimal strategy.

7. To identify and rank the most influential ASD risk factors through sensitivity analysis

1.5 Contribution of this research

This study makes numerous notable contributions to the field of Autism Spectrum Disorder (ASD) detection and machine learning-based healthcare analytics. Among all of them, the key contributions are listed below.

1. Development of a High-Performance, Multidimensional Predictive Framework:
This research successfully developed and validated a robust machine learning framework for the early detection of ASD. By involving a comprehensive set of features—including behavioral attributes, developmental milestones, physical health indicators, family history, and socio-demographic information—the model provides a holistic and accurate screening tool that surpasses the limitations of unidimensional approaches.

2. Demonstrated Superiority of Ensemble Methods and Advanced Optimization:
The study provides empirical evidence that the Random Forest (RF) algorithm, especially when enhanced with a hybrid Genetic Algorithm-Particle Swarm Optimization (GA-PSO) technique, is exceptionally effective for ASD classification, achieving near-perfect performance (accuracy >98%, AUROC ≈ 0.999). This establishes a new benchmark for model selection and tuning in this domain.

3. Introduction of a Novel Hybrid Optimization and Augmentation Pipeline:
A key methodological contribution is the innovative application and comparative analysis of a combined GA-PSO and CTGAN augmentation strategy. The research demonstrates that this synergy is not merely additive but transformative, enabling multiple models to achieve perfect classification metrics (100% accuracy), a rare accomplishment in medical diagnostic research.

4. Rigorous Validation of Model Generalizability and Real-World Applicability:
Moving beyond standard k-fold validation, this research rigorously tested model robustness through stability analysis across multiple data splits and, most importantly, validation on a completely unseen dataset**.** The models maintained exceptional performance (94.1% accuracy on unseen data), proving their potential for reliable deployment in real-world clinical settings to support diagnosticians.

5. Strategic Data Augmentation for Overcoming Medical Data Scarcity:

The study supplies a practical strategy to the common challenge of limited medical datasets. It systematically demonstrates that strategic data augmentation using CTGAN with boundary sampling, even at a low multiplier (2x), can significantly boost performance to optimal levels, providing a highly efficient blueprint for other researchers working with small clinical samples.

6. Data-Driven Identification and Ranking of Key ASD Risk Factors:

Beyond prediction, this research provides valuable clinical insights through a sensitivity analysis. It identifies and ranks the most influential risk factors, with "Limited Speech" emerging as the strongest predictor, followed by parental age and key developmental milestones.

7. Closing the Gap between Clinical Practice and Machine Learning:

This work is ultimately a critical bridge between high-technology computational methods and clinical requirements. This research could directly decrease diagnostic delays, decrease costs, decrease healthcare disparities and interventions, and therefore have a positive effect on the living quality of persons with ASD and their families through provision of an accurate, automated, and accessible screening tool.

1.6 Research Hypothesis

The key idea behind this research is that a machine learning model, augmented with specialized hyper-parameter optimization and effective data augmentation, will achieve a meaningful improvement over traditional approaches and state-of-the-art algorithms in early Autism Spectrum Disorder (ASD) detection. The hypothesis is that a hybrid Genetic Algorithm-Particle Swarm Optimization (GA-PSO) algorithm will be able to explore the complex hyper-parameter space of different classifiers e.g., Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP) and Logistic Regression (LR) that will result in significant increase in predicting accuracy, precision and robustness. Moreover, it is hypothesized that generator-based data augmentation via Conditional Tabular Generative Adversarial Network (CTGAN) combined with a boundary sampling method will effectively address the constraints in dataset size and imbalance, thus further improving the generalizability of the models. Finally, the

hypothesis of the study is that the combined, synergistic use of these two sophisticated methods will result in a better and highly accurate predictive model.

This model will not only achieve exceptional accuracy on a completely unseen test set, demonstrating real-world applicability, but will also, through a sensitivity analysis, validly identify and rank key contributing factors such as parental age and developmental milestones, providing actionable insights for clinical practice.

1.7 Scope of the Research

The scope of this research encompasses the development and rigorous validation of a machine learning-based screening tool for Autism Spectrum Disorder using a specific multidimensional dataset of 288 samples. The study is delimited to the analysis of five selected machine learning algorithms: Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression. The methodological scope includes the application of a hybrid GA-PSO technique for hyper-parameter optimization and the use of CTGAN with boundary sampling for data augmentation at 2x, 4x, and 7x multipliers. The evaluation of these models is confined to specific performance metrics—accuracy, precision, recall, F1-score, and AUROC—and their validation across predefined data splits (70:30, 75:25, 80:20) and a final, completely unseen test set of 17 samples.

The research focuses exclusively on the factors contained within the acquired dataset, which includes behavioral attributes, developmental milestones, physical health indicators, family history, and socio-demographic information. The sensitivity analysis is scoped to rank the influence of these pre-existing features on the model's predictions. It is important to note that this study does not involve the collection of new primary clinical data, genetic sequencing, or neuroimaging. The findings are intended to provide a robust proof-of-concept for an auxiliary screening tool and insights into associative risk factors, not to establish new clinical protocols or elucidate the fundamental biological causes of ASD. The outcomes are therefore contextualized within the limitations of the dataset and the selected computational methods.

1.8 Thesis Structure

This thesis is organized into five core chapters that systematically present the research problem, methodology, findings, and implications of this study on machine learning-based early detection of Autism Spectrum Disorder. The structure is designed to provide a logical flow from the foundational background to the detailed results and concluding insights. The organization of the thesis is also depicted in Figure 1.1.

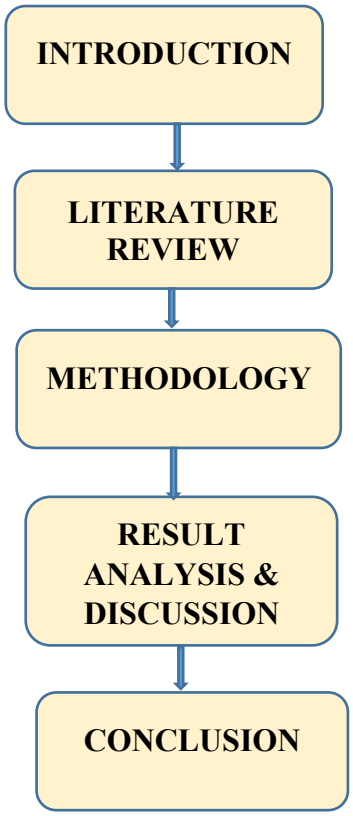


Figure 1.1: Structure of the thesis paper

Chapter 1: Introduction

Introduces Autism Spectrum Disorder (ASD), its global prevalence, and diagnostic challenges. Outlines the research justification, Problem statement, objectives, hypotheses, and scope of the work

Chapter 2: Literature Review

Critically reviews traditional ASD diagnostic methods and machine/deep learning approaches in neurodevelopmental disorder detection. Identifies gaps in existing research.

Chapter 3: Methodology

Details the dataset, preprocessing steps, and implementation of five machine learning models (SVM, DT, RF, MLP, LR). Explains the hybrid GA-PSO optimization and CTGAN-based data augmentation with boundary sampling. Describes evaluation metrics and validation protocols.

Chapter 4: Results Analysis and Discussion

Presents and critically discusses experimental results, including baseline model performance, impact of optimization/augmentation, validation on unseen data, comparative analysis of methods, and sensitivity analysis of risk factors.

Chapter 5: Conclusion and Future Work

Summarizes key findings, contributions to ASD screening, and limitations. Suggests concrete directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

In this section, we present a detailed review of prior studies focused on autism prediction, with an emphasis on machine learning and deep learning models applied to behavioral, clinical, and multimodal datasets. The section is organized into distinct thematic subparts to reflect the diversity and depth of the research landscape. We begin with a summary table that encapsulates key works, highlighting their methods, datasets, results, and future research directions. Following that, we explore literature under the following categories:

- **Behavioral Data-Based Autism Prediction**, which discusses early machine learning applications and classification models,
- **Motor Task-Based Detection**, focusing on kinematic and motion data,
- **Eye-Tracking Data Analysis**, which uses visual attention patterns to diagnose ASD,
- **Gene Expression Data**, emphasizing genetic markers and classification techniques,
- **EEG-Based Prediction**, leveraging brainwave activity for early diagnosis,
- **Facial Image Data**, highlighting automated and CNN-based approaches,
- **Multimodal Data Integration**, combining eye, facial, EEG, and behavioral signals, and
- **Neuroimaging Data**, which explores diagnosis using fMRI and structural brain imaging.

Each sub-section examines contributions from diverse researchers, compares algorithmic performance, and highlights limitations and opportunities for future improvements. The review also identifies the shift towards hybrid models, personalized diagnosis, and the integration of multimodal features—underscoring the growing impact of AI in transforming autism spectrum disorder (ASD) detection and intervention strategies. We also provide a summary table that encapsulates key works, highlighting their methods, datasets, results, and future research directions

2.2 Previous Related Work on Autism Spectrum Disorder Prediction

(i) Behavioral Data Based Prediction

B. Deepa et al. [2] applied Machine learning-based novel and early prediction techniques and all of them achieved an accuracy of 100% using the correlation technique. The accuracy of naïve bayes, and support vector machine was best to classify each individual who was affected by autism

spectrum disorder than decision table. The best result was produced by naive bayes and support vector machine compared to the decision table [3].

An effective framework proposed by S. M. Mahedy Hasan et al. employed four different feature scaling and features selection techniques. The data were preprocessed by applying mean value imputation, an oversampling technique, and other methods. AdaBoost obtained the highest prediction accuracy for Toddlers as well as Children datasets. The prediction accuracy of linear discriminant analysis was 97.12% and 99.03% for the adolescent and adult datasets respectively [5].

A. Vikram et al. [8] explored the six supervised algorithms for predicting autism. A mobile application was intended to be developed to provide healthcare facilities. Adaptive wrapper feature selection as well as backward elimination were employed. On the original dataset, random forest obtained an accuracy of 99.75% which is highest. The lowest accuracy of 95.65% was obtained by the k-nearest neighbor. Compared and evaluated five algorithms to find the optimal model for autism screening datasets. Logistic Regression attained an accuracy of 98.36% and 92% in the adult and adolescent datasets, respectively, making it the top performer. For the dataset involving children, SVM demonstrated a peak accuracy of 96.88% [9]. Prediction models based on convolutional neural networks outperform other approaches, showing superior accuracy levels of 99.53%, 98.30%, and 96.88% across datasets for adults, adolescents, and children, respectively [10].

Six classifiers were used by M. B. Mohammed et al. to perform binary classification on separate datasets and multiclass classification on the combined dataset. Using the combined dataset, the support vector machine achieved an accuracy level of 99.55% for multiclass classification, while logistic regression achieved 99% accuracy for binary classification of the teenage data. The use of innovative medications that offer enhanced effectiveness and can be customized to match the unique needs of each patient could aid in a better understanding of this complex illness [11].

T. Akter et al. [14] predicted autism spectrum disorder in children and adolescents, using artificial neural networks and the associated feature selection random forest algorithm. The results showed that the system produced an accuracy of 93.03%. Additionally, a 97.68% prediction accuracy was attained by an artificial neural network. The data set utilized was the adult questioner AQ10. To

improve the detection process and investigate how the treatment impacts the patient's everyday activities, an automated medication system and therapy suggestion model can be put into place.

Elizabeth Stevens et al. employed hierarchical clustering and Gaussian Mixture Models to analyze treatment response and discover behavioral phenotypes. The Skills system was utilized to track treatment progress and evaluate deficiencies in eight developmental domains using retrospective data.16 latent clusters were identified, using the Expectation-Maximization Algorithm, which was further aggregated into 5 high-level clusters (A through E) based on hierarchical relationships. Prioritizing standardized assessments, replication studies, and incorporating biological and genetic research is crucial for validating findings and understanding the disorder's underlying causes [16].

To automate the diagnosing procedure, S. Anitha Elavarasi et al. employed random forest, naïve bayes, alternating decision trees, and logistic regression. Continuous characteristics were discretized, and missing values were substituted. The accuracy of the alternating decision tree and regression models was 100% regarding the adult and child datasets[18]. To identify autism, logistic regression and support vector machine classifiers were locally trained via federated learning. The support vector machine's greatest accuracy of 99% was obtained on the children's dataset, while 81% was obtained on the adult dataset. The drawbacks of this study were the lack of openness, communication overhead, restricted model complexity, and heterogeneity of the data. Autism detection can also be done with transfer learning models like ResNet and MobileNet [19].
Sara Karim et al. proposed an innovative semi-supervised learning approach based on fuzziness, which uses a Neural Network with random weights. The classifiers used were OneR, PART, ZeroR, Bagging, Classification via Regression, J48, Random Tree, Fuzzy MinMax, Fuzzy Data Mining, and Fuzzy Semi-supervised. Among all of them, the proposed Fuzzy Semi-Supervised achieved an accuracy of 94.36%. Novel algorithms of deep learning could be created to identify ASD abnormalities [20].

It aimed to assess the accuracy of various algorithms, including decision tree, logistic regression, naive Bayes, support vector machine, and k-nearest neighbor, in autism detection. Logistic regression, naive bayes, support vector machine, and k-nearest neighbor gave an AUC of 0.977, 0.656, 0.976, and 0.844 respectively. Neural network systems can be developed to train models and select the most effective algorithms for predicting autism-supported image processing [21].

A. Novianto eta al. created an early predictive model to aid in diagnosing children with autism spectrum disorder, utilizing significant family and socio-demographic factors. The preprocessing steps included using a 1-Neural Network model to handle missing data, applying Chi2 and Relief methods and for adaptive data balancing Synthetic minority oversampling technique was used. A process and dataset can be created to identify the emergency level of autistic patients. A novel machine-learning model can be designed to predict autism using dependable clinical assessments and familial background factors [22].

The classifiers were compared using support vector machines, decision trees, neural networks, and random forests on the data of teenagers. Both the random forest and the decision tree classifiers attained 100% accuracy. This study can be extended to incorporate data from MRI scans, genetic sequences, EEG recordings, and other forms of autism spectrum disease [27]. Random sampling was used to balance the unbalanced data. Age, gender, and jaundice had a greater impact on autism spectrum disorder than other data characteristics. In order to determine the key characteristics that influence autism, dimensionality reduction was used. To increase accuracy, a variety of deep learning techniques can be used [28].

(ii) Motor Task Based Prediciton

Zhong Zhao et al. [29] used machine learning classifiers to investigate constrained kinematic features in the identification of autism spectrum disorder. With four kinematic features chosen through forward feature selection, a k-nearest neighbor with a k value of 1 had the highest accuracy of 88.37% among the classifiers. Children with autism who functioned well were chosen to complete motor tasks. Future study directions might include categorizing individuals with autism spectrum disorder into different categories, as those with attention-deficit hyperactivity disorder. Kinematic analysis of a job was utilized to accurately identify children with autism spectrum disorder using a support vector machine. Seven characteristics were chosen by the Fisher discriminant ratio, which led to a 96.7% accuracy rate by support vector machines [30].

Z. Zhao et al.[31] attempted to use characteristics of head movements to identify individuals with autism spectrum disorder. The OpenFace 2.0 head position tracking method was used to derive six features. Features were selected using a random feature combination approach. The decision tree classifier's RR_Pitch and ARPM_Yaw features yielded the best classification accuracy, 92.11%.

We used linear mixed-effect models. Larger and more varied cohorts of people with autism spectrum disorder, covering a variety of behavioral patterns, may be used in future studies.

(iii) Eye Tracking Data-Based Prediction

R. Carette et al.[32] evaluated non-neural network approaches and neural networks to diagnose participants with autism spectrum disorder into severity categories. Image augmentation, scaling, and grayscale conversion were performed. The best-performing model was a single-layer artificial neural network with 200 neurons, which provided high accuracy for binary classification. The number of participants and the duration of the video scenarios can be increased. The visualization of eye-tracking scan paths can possibly be utilized to assist in the diagnosis of similar psychological disorders. A dynamic saliency model was developed and trained using differences in fixation patterns for predicting ASD and ASD-related symptoms severity. The support vector machine model was trained to predict autism severity using Degrees of Freedom maps. An accuracy of 94.59% was obtained for autism spectrum disorder classification using a support vector machine. In addition, an accuracy of 94.74%, for ASD symptom severity prediction for the support vector machine. This study was limited by gender imbalance, sample size disparity, lack of developmental matching, inclusion of comorbid conditions, and absence of comparative static stimuli data [33].

(iv) Gene Expression Data-Based Prediction

In 2019, the author M. Go developed a classification model for predicting autism spectrum disorder risk genes utilizing lncRNA gene data was developed using the Bayes network learning method, discretization, and the Haar wavelet transform. The experimental results confirmed the efficacy of the model with sensitivity, area under the ROC curve, Matthews correlation coefficient, and F-measure scores of 0.902, 0.839, 0.583, and 0.806, respectively. To enhance classification performance, an ensemble of classifiers might be taken into consideration for the classification step. A web server that forecasts the risk gene classification for autism can be developed using the recommended methods [34].

In 2021, the authors, Eman Ismail et al., predicted disease genes and calculated gene functional similarity using Gene Ontology. Three methods of semantic similarity were used: Relevance, Resnik, and Wang. A range of classification approaches were used to calculate the performance of

the proposed strategy. The approaches were evaluated using five-fold cross-validation. Of the genes associated with autism spectrum disorder, only 20% to 25% have been found. The Random Forest classifier, particularly when utilizing the Resnik similarity metric, outperformed other classifiers trained on high-confidence genes [35].

In this study, thirty-one kids with ASD diagnoses participated. The researchers stressed that the HEATR1 gene is particularly relevant to the differentially expressed regions associated with social communication deficits in ASD. Based on regions of differential expression as biomarkers, the study clustered samples with autism spectrum disorder into subgroups using random forest and support vector machine algorithms. The existence of language impairment, a strong predictor of prognosis in subtypes of autism spectrum disorder, can be predicted by transcriptomic patterns linked to social communication issues and the manufacture and metabolism of cholesterol. In order to maximise results through customized interventions based on individual profiles and the need to improve overall treatment effectiveness [36].

(v) EEG Data-Based Prediction

T. Arunprasath et al. [37] utilized machine learning techniques, specifically logistic regression, to discern patterns within EEG data. The dataset creation involved meticulous extraction, cleaning, and transformation of EEG data using Anaconda Navigator's tools. With logistic regression, 100% accuracy was attained. It is possible to overcome the difficulties brought on by the variability of EEG data and to produce a bigger and more varied dataset. More advanced feature engineering can be explored to enhance predictive accuracy.

The study focused on identifying early biomarkers for Autism Spectrum Disorder through EEG connectivity analysis in clinical settings. EEG data collected during sleep, which differs from active paradigms used in many studies, was analyzed using phase lag index and corrected PLV methods, alongside spectral power features. Explainable machine learning methods, like Shapley Additive Explanations, were employed to interpret feature importance, showing alignment with statistical analyses. It employed a simple logistic regression model with L2 regularization for each type of feature. Additionally, layered cross-validation with six outer and five inner splits was employed. Machine learning identified a subgroup of autism spectrum disorder children with

lower certainty, possibly due to heterogeneity in EEG patterns across autism spectrum disorder subtypes. Additionally, recordings from several hospitals may introduce confounding factors [38].

(vi) Facial Image Data Based Prediction

In 2022, the author B. Ramdam et al. used the Hyperpot and tree-based pipeline optimization tool, an automatic machine learning model was suggested, and its accuracy was approximately 96%.Hyperparameter optimization was performed. Normalization, min-max scaling, and other preprocessing modules were among them. Pipeline Optimization was carried out using Randomized PCA. The top deep learning model improved accuracy by 7%, and automated machine learning improved accuracy by about 24% when compared to using traditional machine learning techniques. Combining image data with other data modalities, like behavioral and genetic data, allows automated machine learning to predict autism spectrum disease in youngsters [39].

(vii) Using Multimodal Data Based Prediction

Mengyi Liao et al. [40] suggested a hybrid fusion method that integrated facial expression, eye fixation, and electroencephalogram data. For this study, data from 80 youngsters was gathered. Eye fixation features were extracted using a k-means technique, face expression features were extracted using a convolutional neural network and a soft label, and EEG features depending on the strength of different brain areas were extracted. A hybrid fusion approach was presented for multimodal data fusion, achieving an 87.50% classification accuracy. Over time, the number of data samples and further studies on adopted detection models might be increased.

Data from eye tracking and electroencephalograms were fed into a support vector machine. Information was gathered from 97 kids ages 3 to 6. Areas of interest were established for the power spectrum analysis of the electroencephalogram and the facial gaze analysis of the eye-tracking data. Support Vector Machine classifiers in conjunction with the minimum redundancy maximum relevance feature selection method were used for the classification. Re-referencing, band-pass filtering, downsampling, notch filtering, independent component analysis for artefact removal, visual inspection for noise rejection, and segmentation were the preprocessing techniques used on the EEG data. The next study may involve younger children, those ages 2 and under [41].

In 2020, the authors J.I. Chen et al. proposed a multimodal framework that automatically detects children with autism spectrum disorder based on information on their ocular fixation, facial expression, and cognitive level. Following an optimized random forest strategy, temporal synchronization was performed using a hybrid fusion approach dependent on the data source. The classification accuracy rate attained by the framework was 91%. It is possible to expand the amount of data samples and include additional data modalities including electroencephalograms, body motions, and peripheral physiological signals [42].

(viii) Neuroimaging Data Based Prediction

Caroline L. Alves et al. [43] proposed an automated approach to autism diagnosis using functional brain imaging data from the Bootstrap Analysis of Stable Cluster map. The study used the preprocessed Autism Brain Imaging Data Exchange dataset from the Preprocessed Connectomes Project, which included artefact removal, motion correction, intensity normalization, and time correction. Data augmentation was also applied. Six machine learning models were evaluated: logistic regression, random forest, naïve bayes, support vector machine, multilayer perceptron, and an untuned convolutional neural network. Random forest and logistic regression had the greatest results, with 99% accuracy, precision, recall, and F1-score. The same techniques can be used to other fMRI datasets, such as those for ADHD or schizophrenia.

Using brain networks built from functional MRI data, Sakib Mostafa et al. created a feature-based approach for diagnosing autism spectrum disorder. Motion correction, skull stripping, standard space registration, and Gaussian smoothing were among the preprocessing procedures. Additional regression was used to account for physiological noise and motion distortions. In order to extract features, topological centralities and eigenvalues from Laplacian matrices have to be evaluated. The classification accuracy was 77.7% using a backward sequential feature selection technique with Linear Discriminant Analysis. The suggested approach produced an average classification accuracy of 98.5% across all sites of the dataset, with several sites achieving 100%. Logistic regression also showed high performance, with 97.3% accuracy [44].

Table 2.1: Summary of Previous Related Works

SL. No	Title	Method	Dataset	Findings	Limitations
1.	Dr. R. Surendiran et al.[19]	Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Artificial Neural Network, Correlated Feature selection based on Random Forest algorithm	The Adult Questioner (AQ10) data from the University of California, Irwin ML repository.	Artificial Neural Network : 97.68% , Random Forest algorithm : 93.03%	This research relies on self-reported behavioral questionnaires, which may introduce subjective bias, and the proposed model lacks validation on external or real-time clinical datasets.
2.	Muhammad Shoaib Farooq et al.[21]	Logistic Regression and Support Vector Machine	Kaggle and UCI repositories	on the children dataset Support Vector Machine: 99%, Logistic Regression:98%, On the adult dataset Support Vector Machine: 81%, Logistic Regression: 78%	This study's limitations include the heterogeneity of data across different sources, potential constraints on model complexity due to federated learning, and a relatively small dataset for children.
3.	A. Novianto et al.[22]	Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, AdaBoost, and Neural Network Multilayer Perceptron	Dataset on ASD is selected from the second-largest HMO, Maccabi Healthcare Services computer database	AdaBoost: 99.95%, Neural Network: 99.25%, K-Nearest Neighbour: 98.34%, Decision Tree: 97.86%, Support Vector Machine: 71.05%	The study relies on retrospective sociodemographic data, which may not capture real-time or clinical behavioral factors, and the high accuracy may be influenced by synthetic data balancing techniques like SMOTE.
4.	Puneet Bawa et al.[23]	Logistic Regression,Support Vector Machine, Random Forest,Naïve Bayes, K-Nearest Neighbors, Decision Tree	UCI Machine Learning Repository	For children dataset Logistic Regression: 94.3%, For adolescent dataset	Limited sample sizes per age group and reliance on a single preprocessed dataset may affect generalizability.

				Logistic Regression: 99.0%, For adult dataset Support Vector Machine: 98.5%, For binary classification for merged dataset Logistic Regression: 97.2%, For multi-class classification for the merged dataset Support Vector Machine: 98.7%,	
5.	Elizabeth Stevens et al.[24]	Gaussian Mixture Models and Hierarchical Clustering	Retrospective data from a large archival database was used	-	Lack of standardized assessment data and exclusion of core ASD symptoms like repetitive behaviors.
6.	Sara Karim et al.[25]	OneR, PART, ZeroR, Bagging, Classification Via Regression, J48, Random Tree, Fuzzy MinMax, Fuzzy Data Mining, Fuzzy Semi-Supervised	UCI database repository	Fuzzy Semi-Supervised: 94.36%	Restricted generalizability as a result of only one particular dataset and no comparison to a wider variety of contemporary deep learning baselines.
7.	M. S. J. K. T. Amalraj Victoire et al. [21]	Decision Tree Classifier, Logistic Regression, Naive Bayes, Support Vector Machine, and K-Nearest Neighbor	UCI repository	-	Limited to a single, small dataset (n=292) and lacks external validation or comparison to clinical standard screening tools.
8.	T. Lakshmi Praveena and N.V. Muthu Lakshmi [27]	Decision Tree (J48), Random Forest, Support Vector Machine,	UCI repository	Decision Tree: 100%, Random Forest: 100%,	Small dataset, potential overfitting indicated by 100% training accuracy, and lacks external validation.

		and Neural Networks		Support Vector Machine: 95.90%, Neural Networks: 95.20%	
9.	Murali Anand Mareeswaran and Kanchana Selvarajan[28]	Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network and Logistic Regression	-	Support Vector Machine: 96%, Logistic Regression: 90%, Decision Tree: 91%, Artificial Neural Network: 94%, Random Forest: 87%	Limited to a small, potentially imbalanced dataset and lacks clinical validation or comparison to standard diagnostic tools.
10.	Zhao et al.[29]	Support Vector Machine, Linear Discriminant Analysis, Decision Tree, Random Forest, and K-Nearest Neighbor	Data were collected from 43 participants	Support Vector Machine: 81.40% Linear Discriminant Analysis:81.40% Decision Tree: 67.44% Random Forest: 76.74% K-Nearest Neighbor: 88.37%	Small, specific sample (high-functioning autism only) and task-specific kinematic features limit generalizability to broader ASD population and real-world settings.
11.	Crippa et al. [30]	Support Vector Machine	Data were collected from pre-school aged children	Support Vector Machine: 96.7%	The findings' generalizability is limited by the small sample size and absence of validation on an independent cohort, and it is yet unknown whether the model is specific to ASD as opposed to other neurodevelopmental disorders.
12.	Zhao et al.[31]	Support Vector Machine, Linear Discriminant Analysis, Decision	Data were collected from 38 participants	Support Vector Machine: 81.58%,	Small sample size (n=38) and constrained experimental setting limit generalizability to real-

		Tree, Random Forest, and Ensemble learning with boosting(ENS)		Linear Discriminant Analysis : 84.21%, Decision Tree: 92.11%, Random Forest: 84.21%, Ensemble learning with boosting (ENS): 86.84%	world, naturalistic interactions.
13.	Romuald Carette et. al [32]	Naive Bayes, Logistic Regression, Support Vector Machine, and Random Forests, Artificial Neural Network	59 children data were collected	Neural networks provided a notable prediction accuracy that went beyond 90%	Short video durations and a very small participant sample are among the limitations that could restrict the eye-tracking data's generalizability and richness.
14.	Ryan Anthony J. de Belen et al.[33]	Support Vector Machine	-	Support Vector Machine: 94.59%, for ASD classification, Support Vector Machine: 94.74%, for ASD symptom severity prediction	A comparatively small and gender-imbalanced sample size limits the study's conclusions, and in order to guarantee generalizability, the model's performance needs to be validated on bigger, more varied datasets.
15.	Murat Gok [34]	Naive Bayes, Bayes network, K-Nearest Neighbor, Logistic, Random Forest, Linear Support Vector Machine, Radial Basis Function Support Vector Machine and polynomial Support Vector Machine, Proposed model (based on HWT,	BrainSpan dataset	Naive Bayes: 67.5%, Bayes network: 59.6%, K-Nearest Neighbor: 82.9%, Logistic: 74.2%, Random Forest: 83.4%, Linear Support Vector Machine: 79.9%, Radial Basis Function Support Vector Machine: 83.6%, and polynomial	The model's performance is limited by its reliance on a specific dataset of IncRNA genes and may not generalize to other genetic or phenotypic data types.

		discretization and Bayes network learning algorithm)		Support Vector Machine: 81.8%, Proposed model (based on HWT, discretization and Bayes network learning algorithm): 78.31%	
16.	Eman Ismail et al.[35]	Random Forest, Naive Bayes, K-Nearest Neighbors ,and Support Vector Machine	Simons Foundation Autism Research Initiative (SFARI) gene dataset	For Resnik similarity measures (HD +non mental genes) Random Forest: 80%, For Relevance similarity measures (HD +non mental genes) Random Forest: 76.8%, For Wang Similarity measures (HD +non mental genes) Random Forest: 74.5%	The study relies on the SFARI database and excludes low-confidence and syndromic genes, potentially overlooking relevant ASD-associated genes and limiting generalizability.
17.	Ping-I Lin et al.[36]	Random Forest and Support Vector Machine	Thirty-one children diagnosed with ASD were recruited, with clinical diagnoses made by a board-certified psychiatrist and	Predictors (191 probes) Random Forest-Partitioning Around Medoid: 96.90%, Support Vector Machine : 99.90% Predictors (10 PC*), Random Forest-Partitioning	The study is limited by a very small sample size (n=31), raising concerns about overfitting and the generalizability of the high classification accuracy results.

			confirmed through the ADI-R interview with parents	Around Medoid: 67.70%, Support Vector Machine: 93.30%	
18.	T. Arunprasath et al.[37]	Logistic Regression	-	Logistic Regression: 100%	The paper reports a 100% accuracy using logistic regression on EEG data, which is highly unusual and strongly suggests severe overfitting, a lack of proper validation, or an oversimplified/non-representative dataset.
19.	Jacek Rogala et al.[38]	Logistic Regression with L2 regularization	EEG data were recorded between November 2017 and March 2022 at the Institute of Mother and Child in Warsaw (Poland)	Logistic Regression with L2 regularization: 83–86%	The study is limited by a small sample size (n=36), potential overfitting due to high feature dimensionality, and the use of retrospective clinical EEG data lacking precise sleep staging, which may introduce confounding factors.
20.	Basma Ramdan Gamal Elshoky et al.[39]	Classification and Regression Trees, Logistic Regression, Linear Discriminant Analysis, Boosting, Support Vector Machine, Gradient Boosting Machine, Adaboost, Random Forest, Naive Bayes, K-Nearest Neighbors, and Extra Trees, Convolution Neural Network, VGG16, VGG19,	Facial Image data from Kaggle	Logistic Regression: 55.60%, Linear Discriminant Analysis: 57.32%, Classification And Regression Trees: 60.38%, Naïve Bayes: 69.42%, K-Nearest Neighbor: 62.42%,	The use of a dataset containing web scraped images is the study's main weakness. This could limit the model's generalizability to actual clinical settings by introducing biases related to image quality, demographic homogeneity, and the absence of controlled conditions.

		ResNet50, ResNet101, ResNet152, proposed AutoML		Support Vector Machine: 71.28%, AdaBoost: 67.08%, Gradient Boosting Machine: 70.82%, Random Forest: 70.94%, Extra Trees: 72.64%, Convolution Neural Network: 84%, VGG16: 89%, VGG19: 88%, ResNet50: 86%, ResNet101: 85.33%, ResNet152: 87%, proposed AutoML: 96%	
21.	Mengyi Liao et al.[40]	Hybrid Multimodal Fusion Model for Autism Spectrum Disorder Classification (HMFM-ASD)	Eye Fixation Data, Facial Expression Data, and EEG Data of eighty children with and without ASD were collected	Hybrid Multimodal Fusion Model (HMFM-ASD): 87.50%	Limitations include a small sample size (n=80) and nonstationary data characteristics due to varying recording conditions.
22.	Jiannan Kang et al.[41]	Support Vector Machine	EEG and eye-tracking data of 97 children were collected	Support Vector Machine: 85.44%	The study's sample was limited to children aged 3–6 years, excluding younger children for whom early diagnosis is most critical.
23.	Jingying Chen et al.[42]	Random Forest, Support Vector Machine, Discriminant Analysis, Improved	eye fixation, facial expression, and cognitive level data of	Decision Fusion: Random Forest: 82%, Support Vector Machine: 78%,	A comparatively limited sample size and the omission of additional data modalities, such as physiological signs or

		Random Forest algorithm, Averaged Classifier	50 ASD and 50 TD children were collected	Discriminant Analysis: 81%, Improved Random Forest Algorithm: 87%, Averaged Classifier: 82%, Hybrid Fusion: Random Forest: 86%, Support Vector Machine: 83%, Discriminant Analysis: 85%, Improved Random Forest Algorithm: 91%, Averaged Classifier: 86.25%	EEG, are two of the study's drawbacks.
24	Caroline L. Alves et al. [43]	Support Vector Machine, Random Forest, Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Untuned Convolutional Neural Network	Autism Brain Imaging Data Exchange (ABIDE) dataset	Random Forest: 99% Logistic Regression: 99% Support Vector Machine: 98% Naïve Bayes: 98% Multilayer Perceptron: 99% Untuned Convolutional Neural Network: 87%	The proposed methodology's generalizability to other disorders and data types (fMRI, EEG) remains unvalidated, and its performance on small datasets via transfer learning is untested.
25	Sakib Mostafa et al. [44]	Linear Discriminant analysis, Neural Network, K-Nearest Neighbor, Support Vector Machine, Logistic Regression	ABIDE 1 dataset	Neural Network: 71.7% K-Nearest Neighbor: 73.7% Support Vector Machine: 75.7% Logistic Regression: 77.0%	The study's generalizability is limited by its reliance on the specific ABIDE I dataset and the lack of validation on independent or multi-site datasets.

				Linear Discriminant Analysis: 77.7%	

2.3 Summary

The reviewed studies focus on the application of machine learning (ML) and deep learning (DL) models to predict Autism Spectrum Disorder (ASD) using diverse datasets and methodologies. Commonly used algorithms include Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbor, Naive Bayes, and Neural Networks. High accuracies were achieved in many works, such as ANN (97.68%) and SVM (99%) in child datasets, while ensemble techniques like AdaBoost (99.95%) and AutoML (96%) showed strong performance. Several studies utilized UCI and Kaggle repositories, while others used EEG, gene, and image datasets. Works like those of Crippa, Zhao, and Chen incorporated eye-tracking, facial expression, and cognitive-level data, enhancing prediction precision through multimodal and hybrid models.

Some researchers focused on subgroup classification (e.g., ASD vs. ADHD), while others aimed to improve ASD severity prediction. Approaches like fuzzy logic, hierarchical clustering, and hybrid fusion models contributed to better diagnostic insights. Proposed improvements include expanding datasets, integrating EEG/fMRI/genetic data, and adopting transfer learning and AutoML. Some of the useful applications proposed are real-time ASD detection system, individual treatment planning, and mobile/web interfaces. This body of research highlights the increasing importance of AI in the preliminary and precise diagnosis of ASD using both conventional classifiers and state-of-the-art deep-learning-based and data integration models with diverse data sources.

CHAPTER 3
METHODOLOGY

3.1 Introduction

This paper aims to determine and examine the multifaceted aspects of growth and development, along with its major emphasis on crafting forecasting models of conditions such as autism spectrum disorder or other developmental characteristics. A systematic and evidence-based approach is embraced to tackle this complexity systematically. This chapter summarizes the methodology applied to analyze the raw data into validated and interpretable outcomes.

In order to conduct a logical and comprehensive investigation, we have structured our method into four distinct steps. We will present the dataset and our initial preparations first. Second, we will immerse ourselves in the Exploratory Data Analysis, becoming acquainted with the nuances, patterns, and peculiarities of the data. Third, our cleaning of the data to prepare it to be modeled will be described. Finally, we will explain how we select, train, and evaluate our machine learning models to ensure they are both accurate and reliable.

This step-by-step workflow is designed to make our research rigorous and reproducible. Every choice—from how we handle a missing value to which algorithm we select—is made with intention, guided by the nature of our unique dataset and the specific questions we are asking. By clearly outlining our process here, we provide a trustworthy blueprint for our analysis, laying a solid foundation for the results we will present next.

3.2 Methodological Workflow

The following figure represents the workflow of our proposed approach. Including all of the steps, starting from data collection, data preprocessing, feature selection, model selection & development, model evaluation, comparative analysis, and result interpretation.

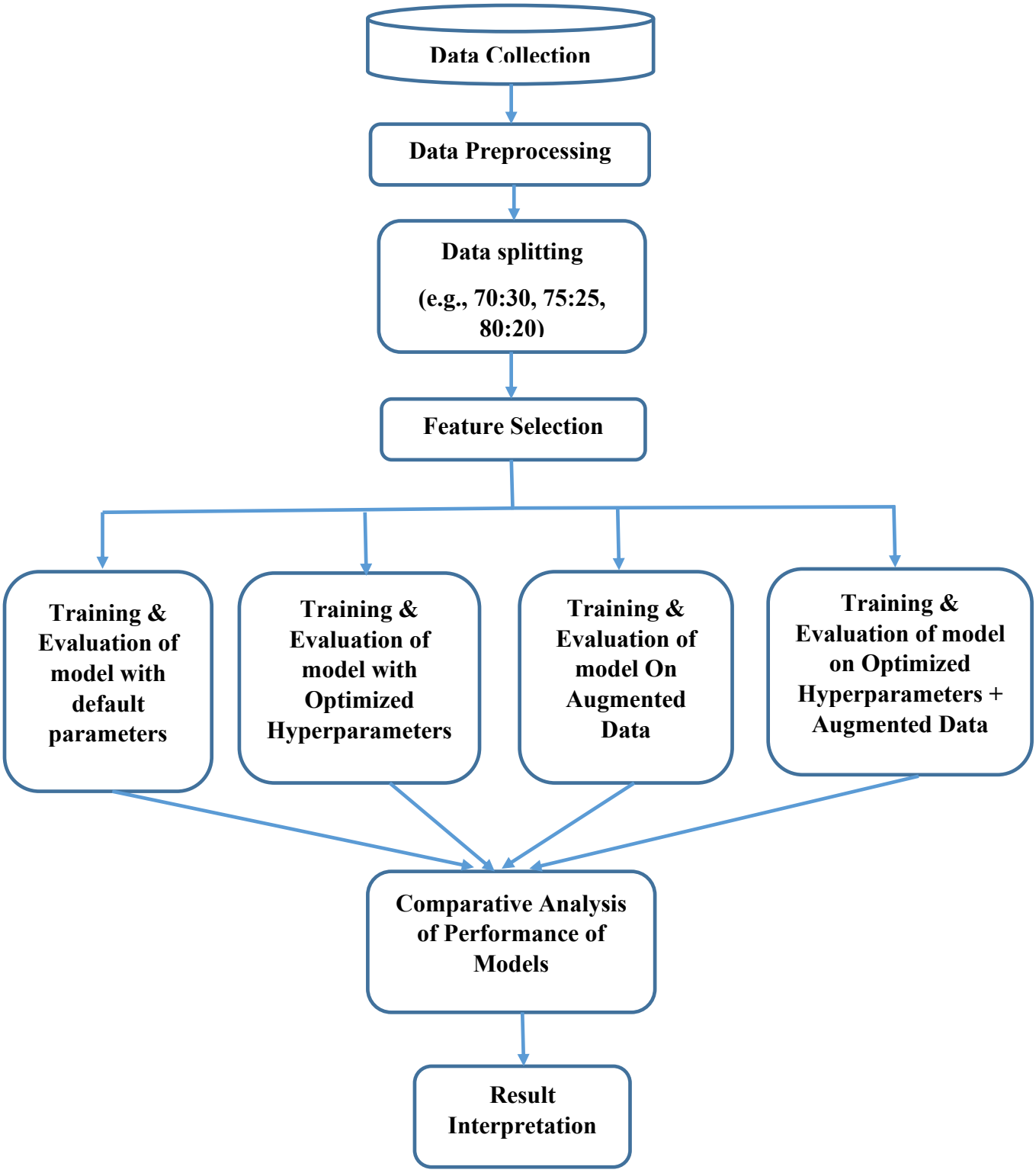


Figure 3.1: Proposed Methodological Workflow

3.3

……
……
…….

(i) Attributes

Table 3.1 represent all attributes of collected dataset and their type and value. Dataset contain around 46 attributes. Attributes are of nominal, numerical, ordinal types

Table 3.1: List of all attributes of collected dataset

S/L No.	Attribute Name	Type	Value
1	Name	Nominal	Unique Identifier
2	Age	Numerical	3-27
3	Gender	Nominal	Male, Female
4	Height (m)	Numerical	Continuous value
5	Weight (kg)	Numerical	Continuous value
6	BMI	Numerical	Continuous Value
7	BMI Category	Ordinal	Healthy, Overweight, Underweight, Obese
8	Aimless work with object	Binary Nominal	Yes, No
9	Choosing a choice	Binary Nominal	Yes, No
10	Mingles with peers	Binary Nominal	Yes, No
11	Ability to Express pain & joy	Binary Nominal	Yes, No
12	Needs held in Feeding	Binary Nominal	Yes, No
13	Needs help in dressing	Binary Nominal	Yes, No
14	Toilet Training	Binary Nominal	Yes, No
15	Ability to Follow Instruction	Binary Nominal	Yes, No
16	Curiosity in device	Binary Nominal	Yes, No
17	Congenital Malinformation	Binary Nominal	Yes, No
18	Ability to speak	Binary Nominal	Yes, No
19	Limited Speech	Binary Nominal	Yes, No
20	Ability to communicate with people	Binary Nominal	Yes, No
21	Ability to make eye contact	Binary Nominal	Yes, No
22	Relationship understanding with parents	Binary Nominal	Yes, No
23	Ability to respond when called by name	Binary Nominal	Yes, No
24	Produce meaningless sound	Binary Nominal	Yes, No
25	Uncertain laugh or cry	Binary Nominal	Yes, No
26	Ability to walk	Binary Nominal	Yes, No

27	Repititive motor movement	Binary Nominal	Yes, No
28	Ability to engage in play	Binary Nominal	Yes, No
29	Epilepsy	Binary Nominal	Yes, No
30	Mode of Delivery	Nominal	C-section, Vaginal
31	Premature Birth	Binary Nominal	Yes, No
32	Birth weight (kg)	Numerical	Continuous number
33	Occupation of Father	Nominal	Catergory labels
34	Occupation of Mother	Nominal	Category labels
35	Father age at child birth	Numerical	Continuous number
36	Mother age at child birth	Numerical	Continuous number
37	Parental Age gap	Numerical	Continuous number
38	Matches Parents Blood group	Binary Nominal	Yes, No
39	Blood Pressure of Mother at childbirth	Ordinal	Normal, Low, High, Excessive High
40	Mother child count	Numerical	Discrete count
41	Maternal anxiety during Pregnancy	Binary Nominal	Yes, No
42	Maternal Physical Illness	Binary Nominal	Yes, No
43	Birth order of the child	Numerical	Discrete count
44	Sibling with autism	Binary Nominal	Yes, No
45	Presence of Family Member with Autism	Binary Nominal	Yes, No
46	Residential area	Nominal	Urban, Rural

3.4 Data Collection and Data Preprocessing

Data collection is performed based on the provided attribute listed on table 3.2.1. we collected listed features from the parents and relatives of several autistic children from several institution and schools dedicatedly developed for autistic children.

……
……
……
……
……..

3.4.1 Data Preprocessing

This code performs comprehensive data cleaning and preparation. It strips whitespace from column names and string values, handles missing values in numeric columns by replacing them with the median, and drops irrelevant columns like 'Name'. Categorical variables are encoded using label encoding, and numeric columns are cleaned by removing non-numeric characters before conversion. The preparation ensures the dataset is structured and free of invalid entries for subsequent modeling

(i) Method for Imputing Missing Data

Imputation is a method for keeping the majority of the data or information in a dataset by substituting a value for the missing data. These methods are employed because it is impractical to remove data from the dataset repeatedly and because doing so may significantly reduce its size, which raises worries about biasing the dataset and perhaps resulting in inaccurate analysis. In mean imputation, the Missing values are filled in by using the arithmetic mean of the available data in the column. Since it preserves the average of the dataset, it is particularly useful when the numerical data are approximately symmetrically distributed. The mean imputation formula is as given below:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \ldots\ldots \quad (3.1)$$

Where

- n is the number of non-missing values.
- X_i are the individual non-missing values.

Median Imputation a reliable substitute for mean imputation, which substitutes the column median for missing values. Because the median is less susceptible to extreme values, this approach works better with skewed distributions or datasets that contain outliers. The following formula can be used to get the median if the total number of observations provided is odd:

$$Median = \left(\frac{n+1}{2}\right)^{th} observation$$

If the total number of observations is even, then the median formula is:

$$Median = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} observation + \left(\frac{n}{2}+1\right)^{th} observation\right]$$

where n is the number of observations.

Mode imputation dealing with categorical data, substitutes the most common value (mode) in the column for any missing values. Because it guarantees that the imputed value falls into the most prevalent category, this approach is perfect for nominal or ordinal data.

$$X_i = \begin{cases} X_i, & \text{if } X_i \text{ is not missing} \\ \text{Mode}(X), & \text{if } X_i \text{ is missing} \end{cases}$$

Where:

- X_i represents an individual value in the dataset.
- Mode (X) is the most frequently occurring value in the dataset.
- If multiple values have the highest frequency, any of them can be used for imputation.

(ii) Feature Encoding

- **One-Hot Encoding:** One-hot encoding was used to convert the categorical variables to binary columns that represented each category, thus converting the qualitative data into machine understandable format. This method avoids using categorical values as ordinal relationships where none exist. Encoding schema was only trained on the training data and the same category mapping was carried out on the validation and test data to make a comparison. This method preserves the integrity of categorical features, and renders them algorithmic processable.

- **Label Encoding:** The target variable was implemented with label encoding which converts the categorical class labels into numerical values which machine learning algorithms can use. This method correlates the distinct categories (e.g., Autism, Non-Autism) to an integer number (e.g., 0, 1) without altering the classification hierarchy. Encoding was only applied to training labels, meaning that the same encoding was always used both in validation and test sets. This ensures the integrity of the target variable as well as allowing the compatibility of the model with the classification algorithm.

(iii) Feature Scaling

- **Standardization:** To normalize numeric features, Z-score was used to put all the variables into a common scale with a zero mean and unit variance. This is done to make sure that features with an inherent magnitude are not overly represented in the model training. Standardization parameters (mean and standard deviation) were estimated solely on training data and used across validation and test samples, the integrity of data distribution was preserved, and the optimal algorithmic performance was achieved.

(iv) Feature Selection Techniques

- **Model-based Feature Selection (Embedded Method)**

Employed Random Forest importance as an embedded feature selection method using SelectFromModel with median thresholding. This approach leverages the inherent feature importance scores generated during Random Forest training to identify the most predictive variables. The median threshold automatically selects features with importance above the median value, providing an adaptive filtering mechanism that retains the most relevant features while eliminating noise. This method combines the predictive power of tree-based models with efficient feature selection, ensuring optimal feature subset identification based on actual model utility rather than standalone statistical measures.

3.2.3 Exploratory Data Analysis

Table in 3.3.2 shows various characteristics of our data set. This table it shows the distribution of Autism and Non-autism classes and in both the classes it shows distribution of male and female. Also it shows Variables with missing data, duplicate values number contains in our dataset.

Table 3.2: Characteristics and value of dataset

Characteristic	Value

Sample Size	287
Autism Cases Male Female **Total** **Non-Autism Cases** Male Female **Total**	 104 42 **146** 70 71 **141**
Variables with Missing Data	10
Duplicate Records	0

In this section visualization of several attributes is performed based on their distribution in different categories. Data analysis is performed on several attributes showing their distribution in different categories.

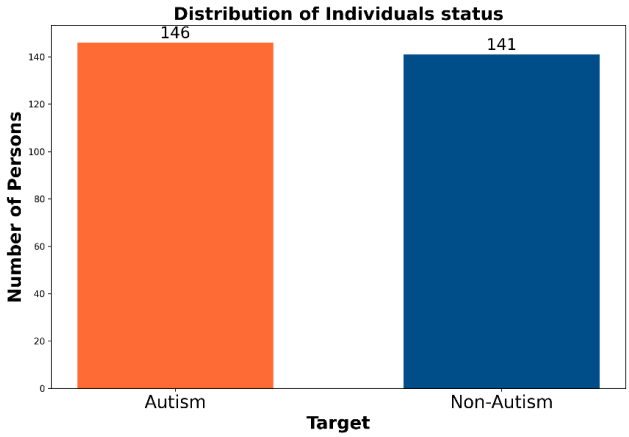


Figure 3.2: Target distribution

In age distribution bar plot it visualizes the female and male categories for both the autism and non-autism classes. This provides clear distribution of classes in different categories of this feature. In age distribution and BMI distribution plot it visualize number of persons in each range. As from the plot it is clearly visible how many numbers of persons with autism and non-autism falls in each category.

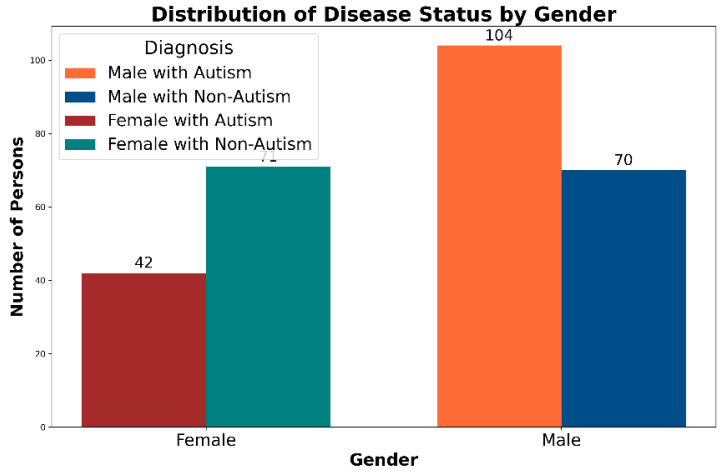


Figure 3.3: Gender Distribution based on target

The given figure it visualizes the distribution of individuals based on different age groups. From the visualization, it can be seen that in 6-10 age group range, there are most children of autistic and non-autistic individuals. It include

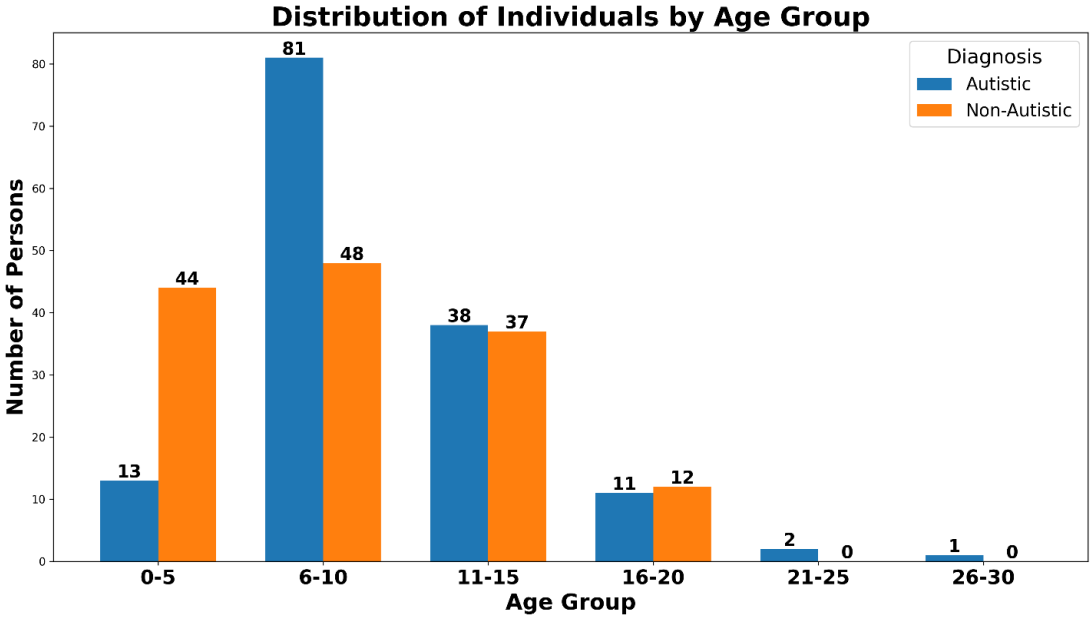


Figure 3.4: Number of individuals on each age group

It compares the number of autistic and non-autistic children across four BMI categories: Healthy, Obese, Overweight, and Underweight. The data indicate that for the Healthy BMI category, there

are 6 autistic children and 13 non-autistic children. In the Obese category, there are 14 autistic children and 10 non-autistic children. The values for the Overweight and Underweight categories are not fully specified in the text but would be represented by adjacent bars on the chart.

The visual comparison suggests a difference in BMI distribution between the two groups, with a notably higher proportion of non-autistic children in the Healthy weight category and a higher number of autistic children in the obese category.

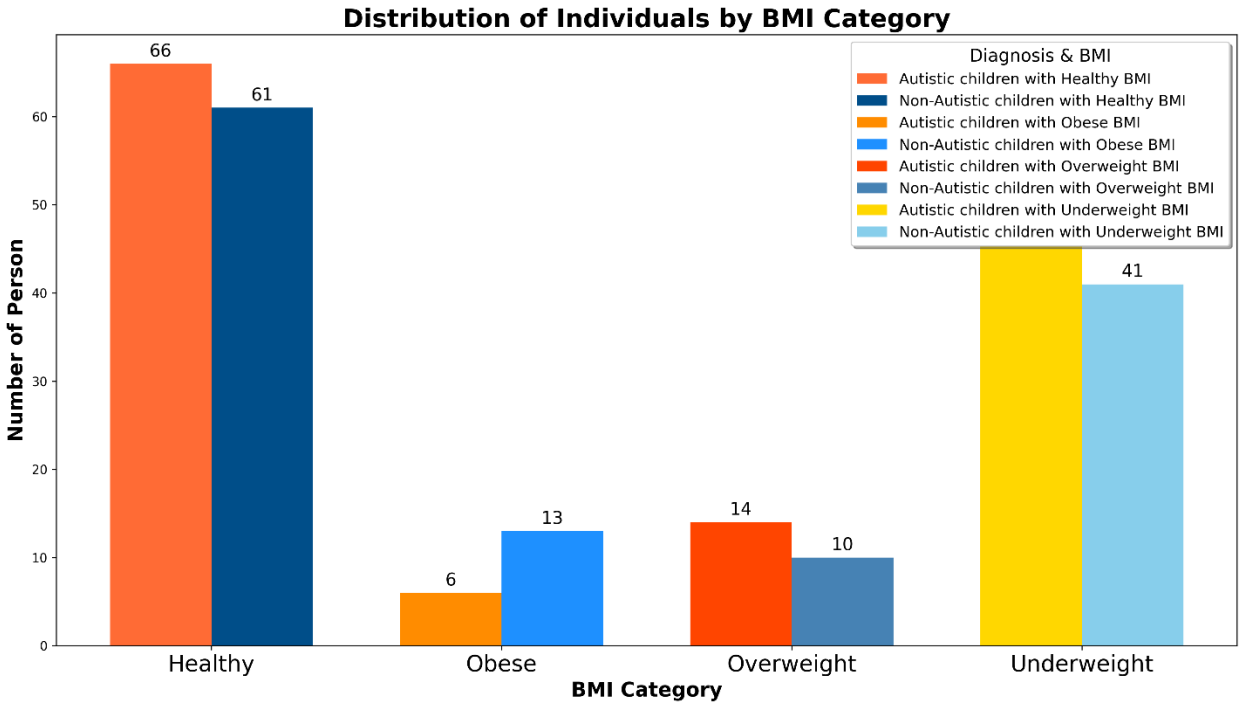


Figure 3.5: BMI category distribution

This visualizes the distribution of autistic and non-autistic individuals based on having epilepsy. Also, some autistic individuals have epilepsy. Epilepsy is very uncommon in healthy or non-autistic children.

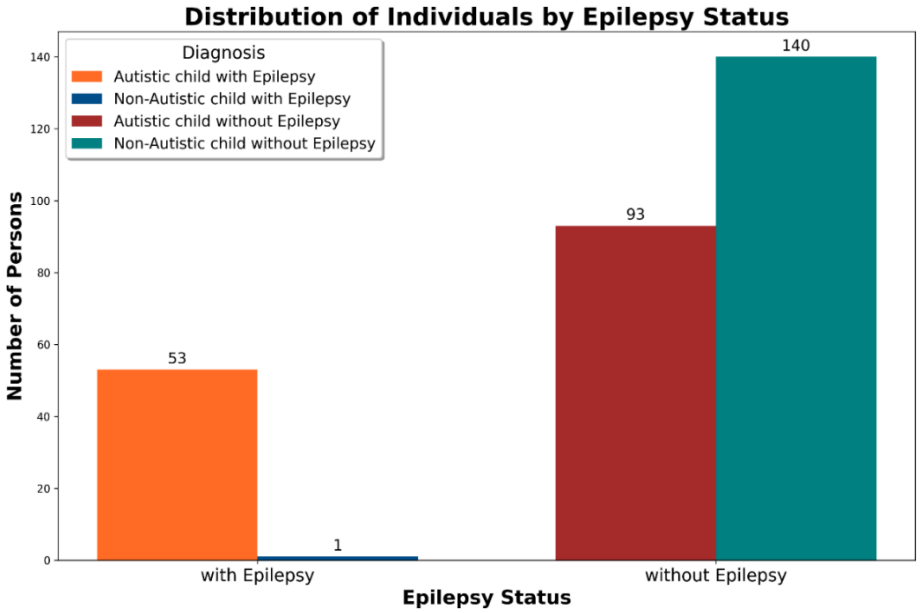


Figure 3.6: Epilepsy status Distribution

This image compares the prevalence of congenital malformations between autistic and non-autistic children. It shows the number of individuals with and without this condition for each diagnostic group. The tallest bar appears to represent non-autistic children without a malformation.

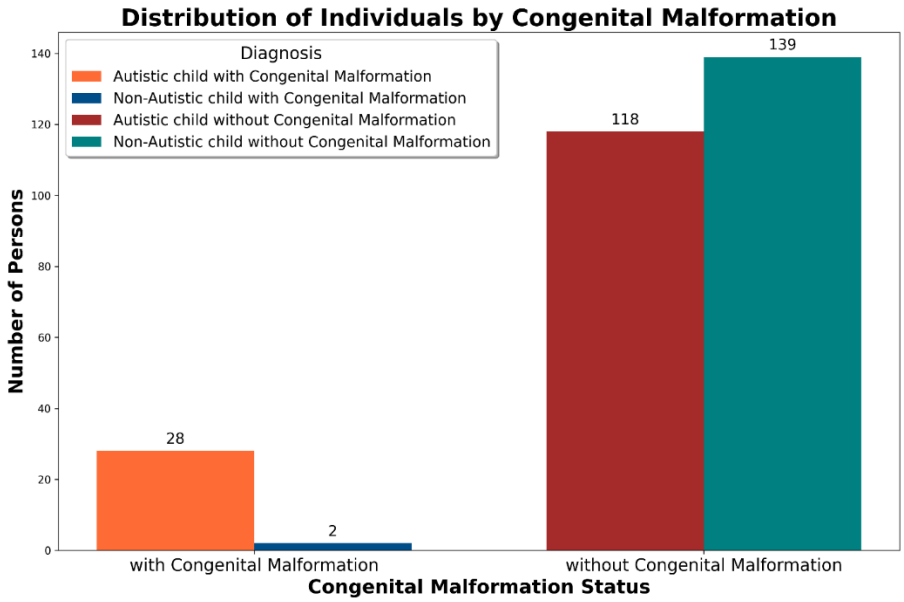


Figure 3.7: Congenital Malformation Status Distribution

This image visualizes the Distribution of Premature Birth among Individuals. It compares the number of full-term and premature births between autistic and non-autistic children. The data shows 13 premature autistic children and 5 premature non-autistic children.

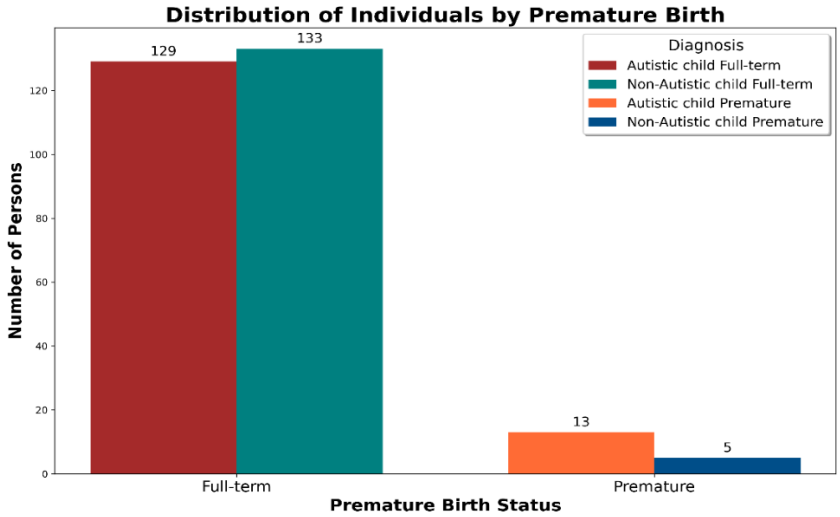


Figure 3.8: Premature Birth status Distribution

This figure visualizes the distribution of individuals based on mode of delivery. It shows that autistic children and non-autistic children are in both categories of delivery, i.e, C-section and vaginal.

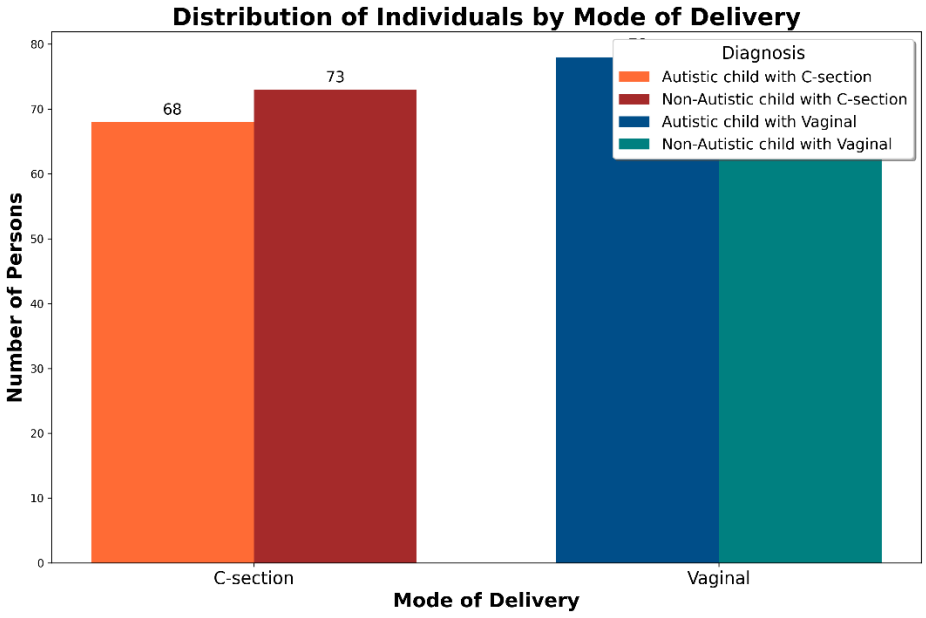


Figure 3.9: Distribution of Individuals based on Mode of Delivery

This figure shows the distribution of birth weights for autistic and non-autistic children, categorized into weight ranges from 2-2.5 kg up to 3.1-3.5 kg. The tallest bars indicate that the most common birth weight for both groups falls within the 2.5-3 kg range. The chart allows for a direct comparison of birth weight prevalence between the two diagnostic groups.

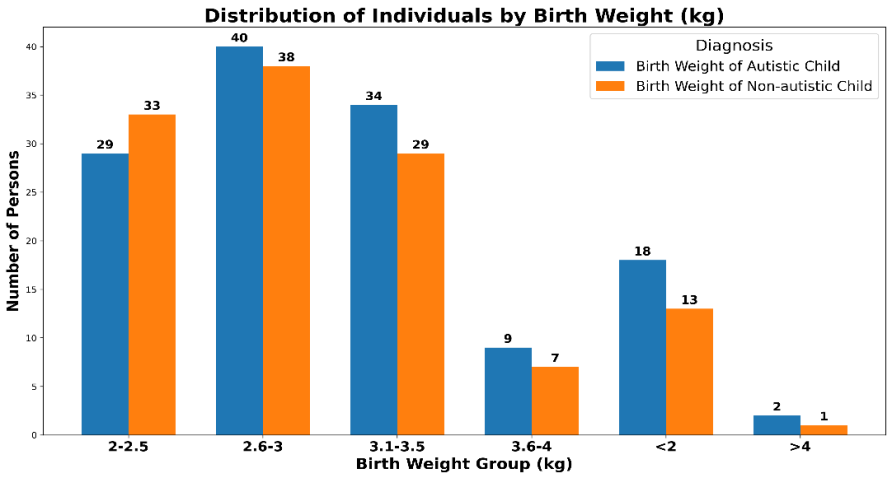


Figure 3.10: Distribution of Individuals By birth weight

The given figure visualizes the occupation of fathers of autistic and non-autistic children. From the data distribution bar plot, it can be seen that fathers' occupation is divided into 16 categories. Some are unemployed also. But it can be seen that a person from any background can have autistic children. The category included in the bar plot is doctor, engineer, electrician, businessman, teacher, government employee, private employee, and many more.

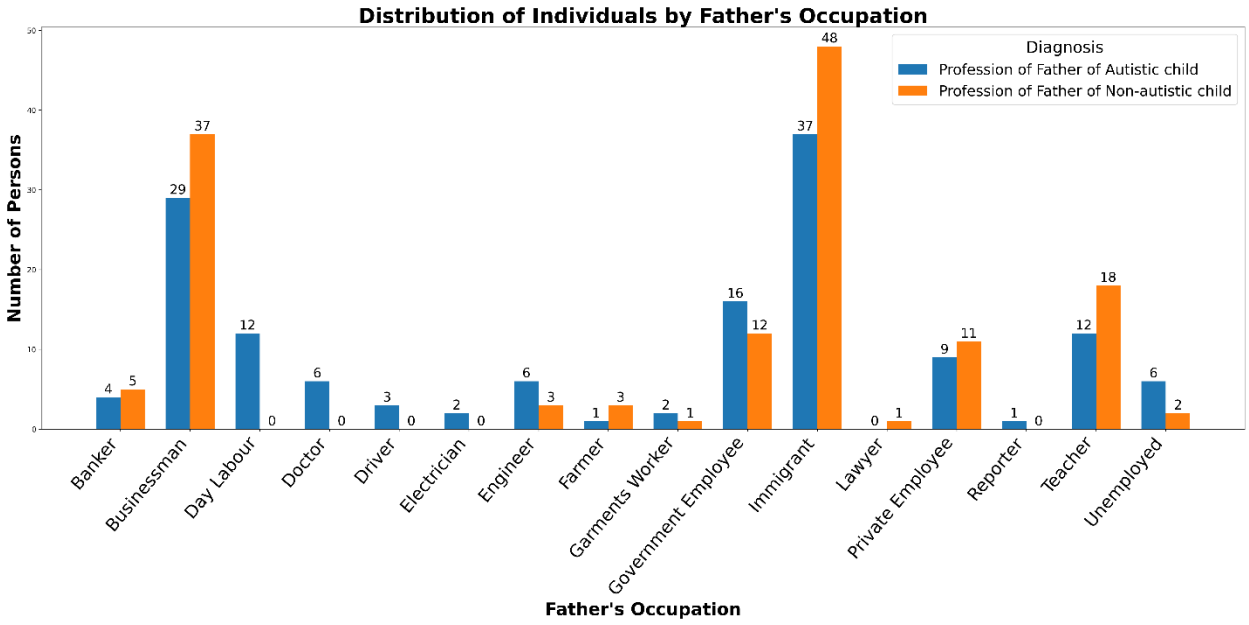


Figure 3.11: Father's Occupation Distribution among Individuals

It shows the number of children in categories like Doctor, Engineer, and Teacher. The data indicate a higher count of children whose mothers are listed as "Housewife" in both groups.

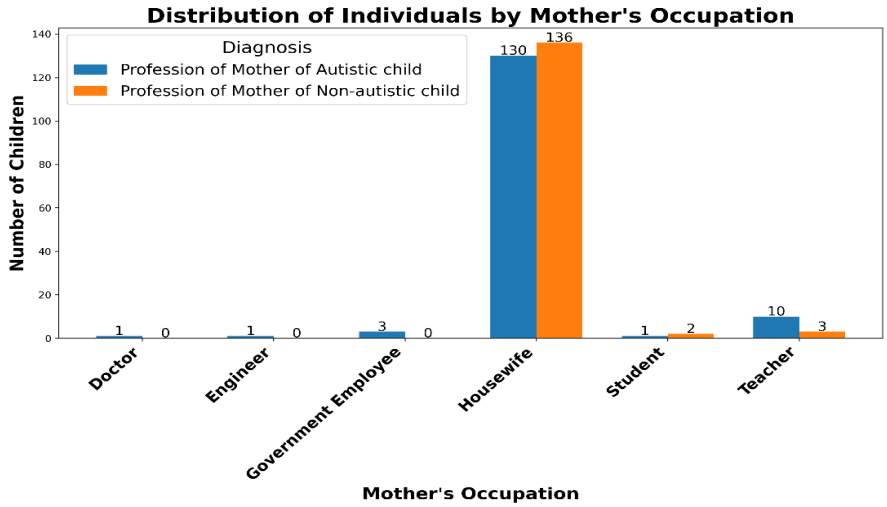


Figure 3.12: Mother's Occupation Distribution among Individuals

The given figure visualizes the mother's age in different age groups at the time of childbirth. The age group range of 21-25 contains many people compared to other age group ranges.

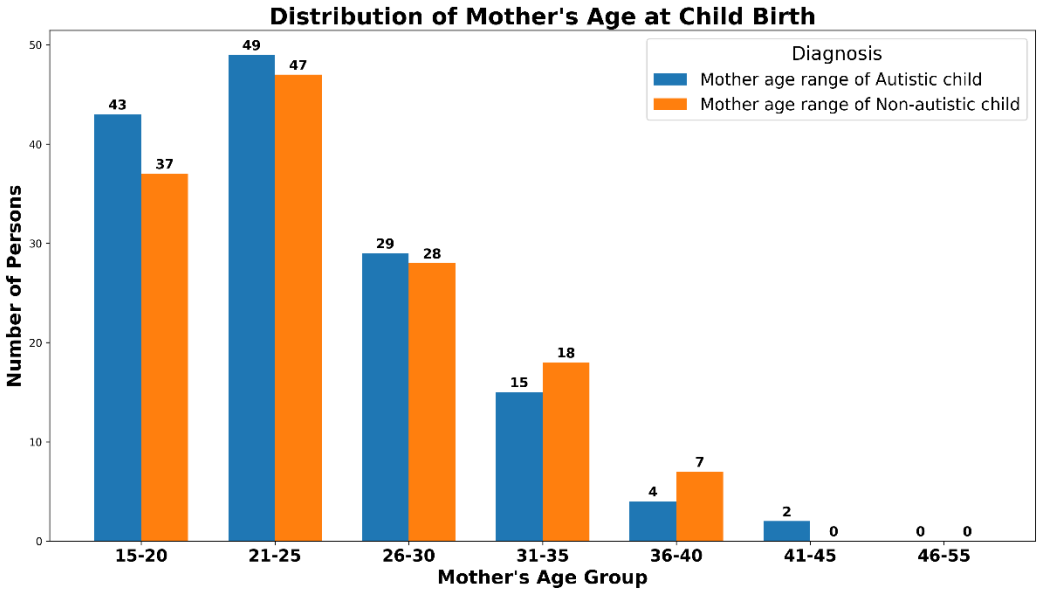


Figure 3.13: Mother's age at child birth distribution based on age group

This visualizes the Father's age range at the time of the child's birth for both autistic and non-autistic children. In the range of 31-35, there are many more people than in other ranges.

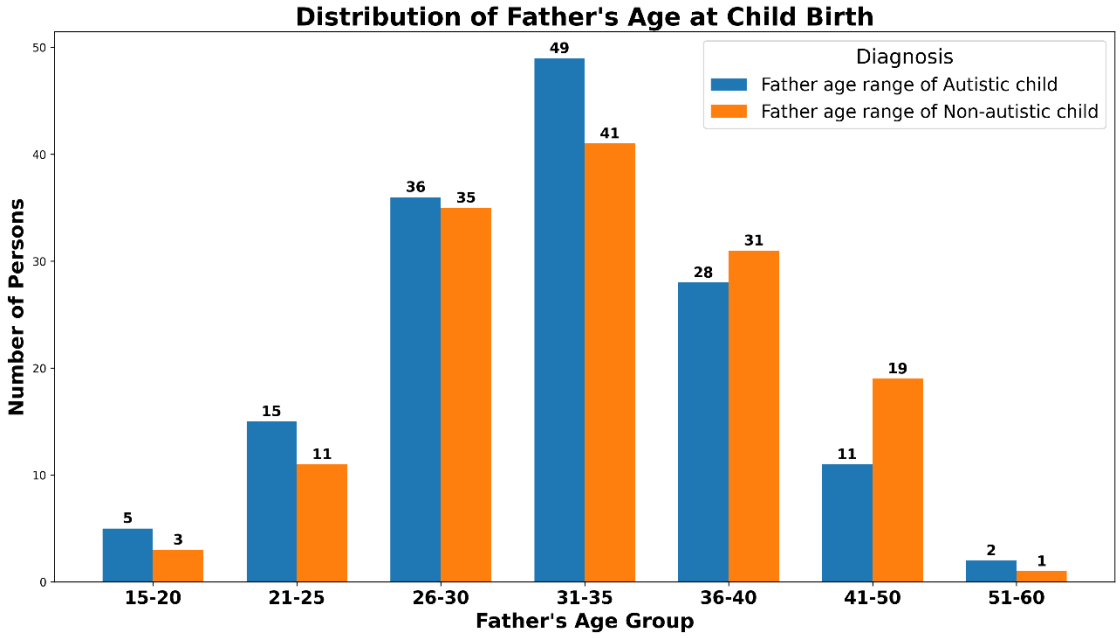


Figure 3.14: Father's age distribution based on age group

This figure visualizes the parental age gap of both autistic and non-autistic children. Age gap 11-15 shows more people than others

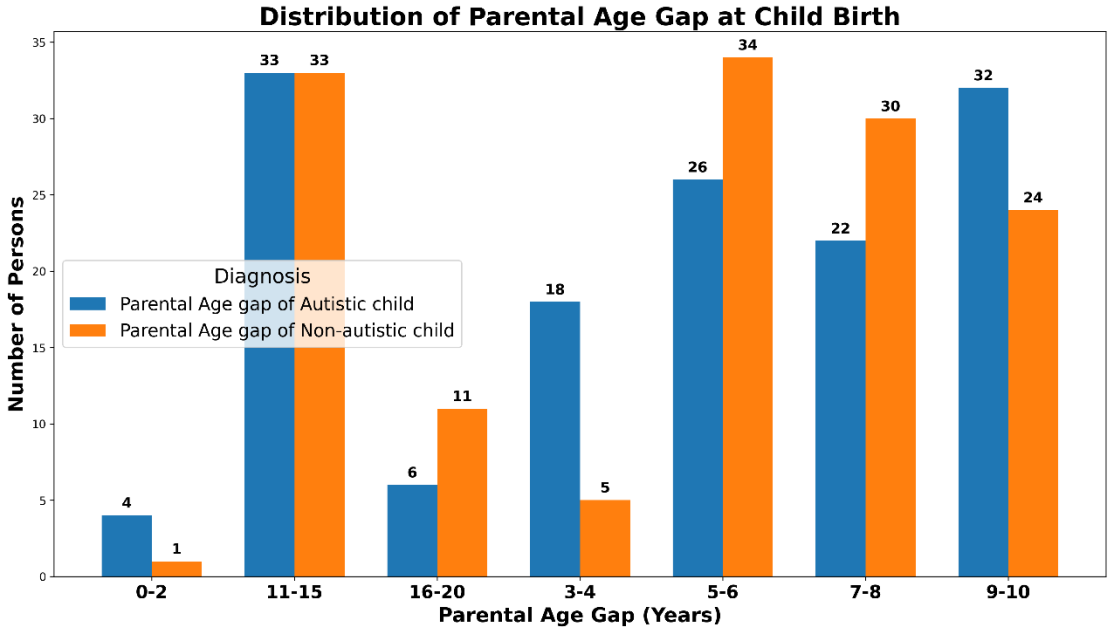


Figure 3.15: Distribution of Parental Age gap based on age group

This figure visualizes the mother's blood pressure at the time of childbirth. Many people have normal blood pressure at the time of their delivery.

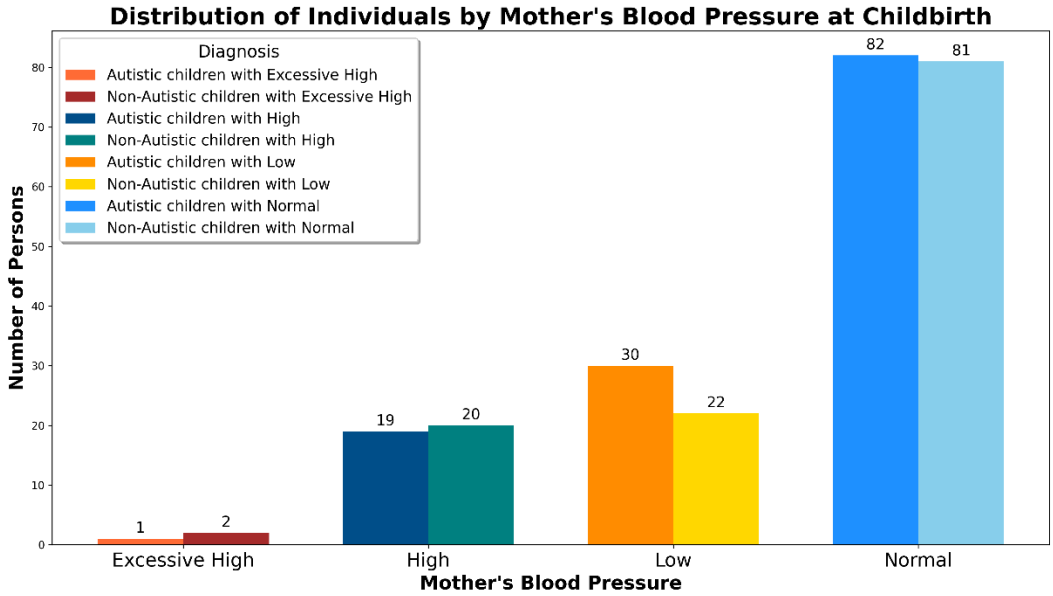


Figure 3.16: Mother's Blood Pressure distribution at childbirth

This chart compares the experience of maternal anxiety during pregnancy between mothers of autistic and non-autistic children. It shows that a higher number of children, both autistic and non-autistic, were born to mothers who did not experience anxiety. The bar for non-autistic children with no maternal anxiety is the tallest.

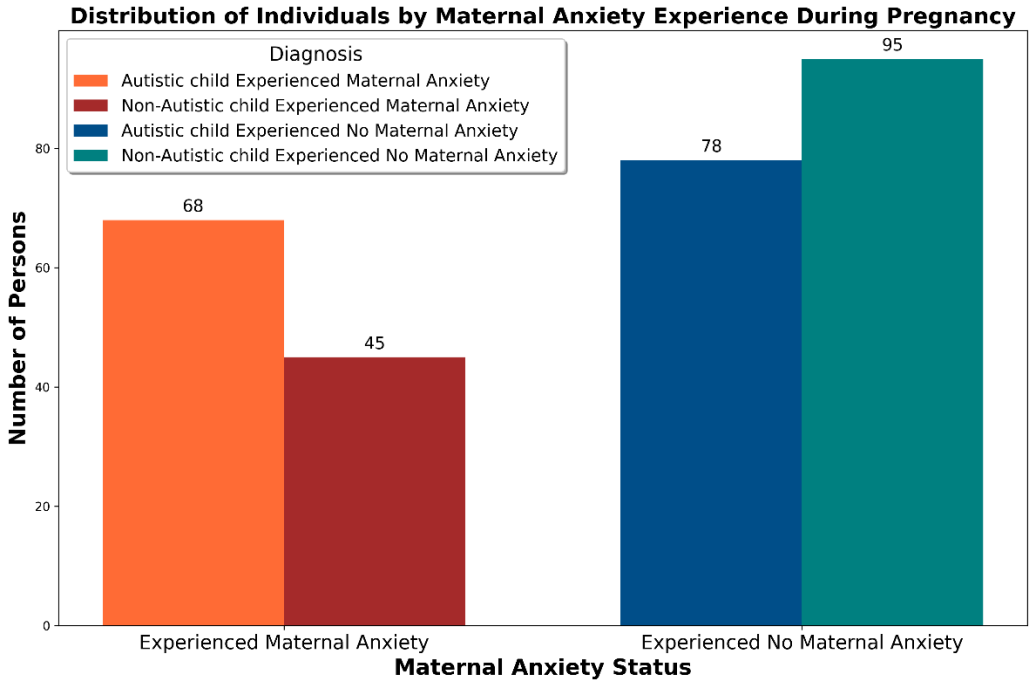


Figure 3.17: Maternal Anxiety experience distribution among individuals

This figure indicates that a majority of children in both groups were born to mothers who did not experience anxiety, with the largest group being non-autistic children without maternal anxiety.

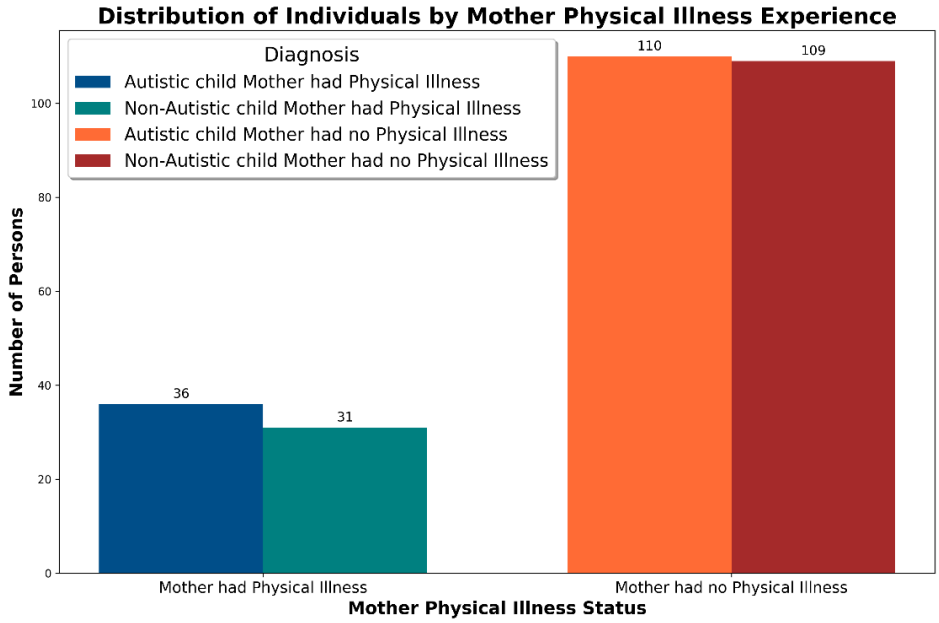


Figure 3.18: Mother Physical Illness Experience Distribution among Individual

There is a very small ratio of autistic people having siblings with autism. Many people from both categories don't have any siblings with autism.

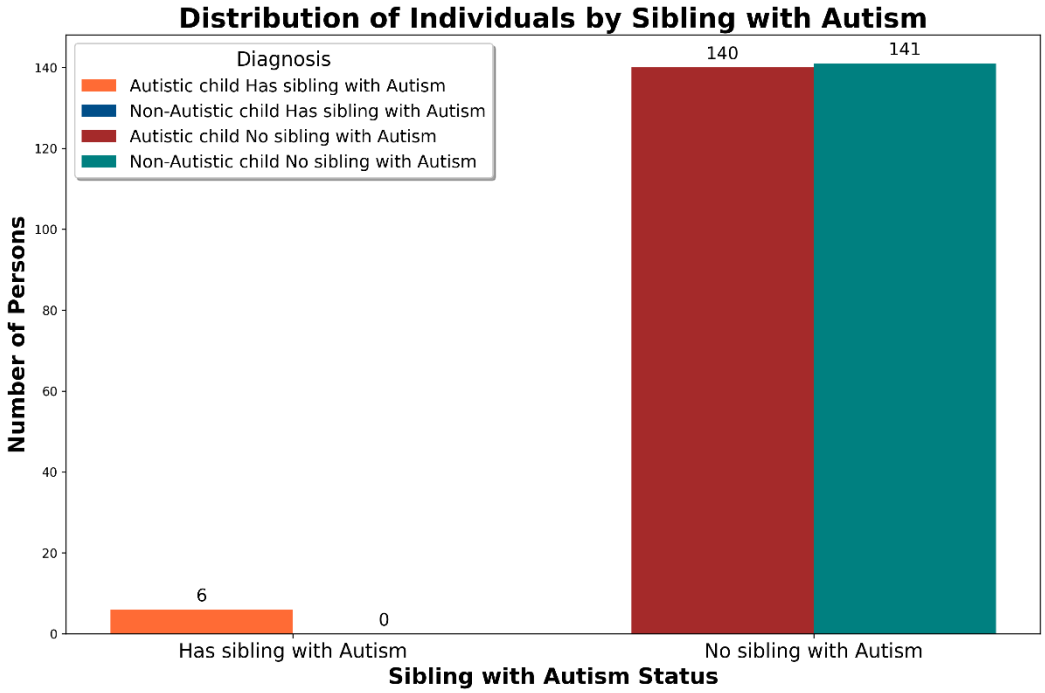


Figure 3.19: Sibling with autism status distribution among individuals

A significantly higher number of autistic children (130) don't have a family member with autism. Conversely, the number of children with a family history of autism is very small, and around 16.

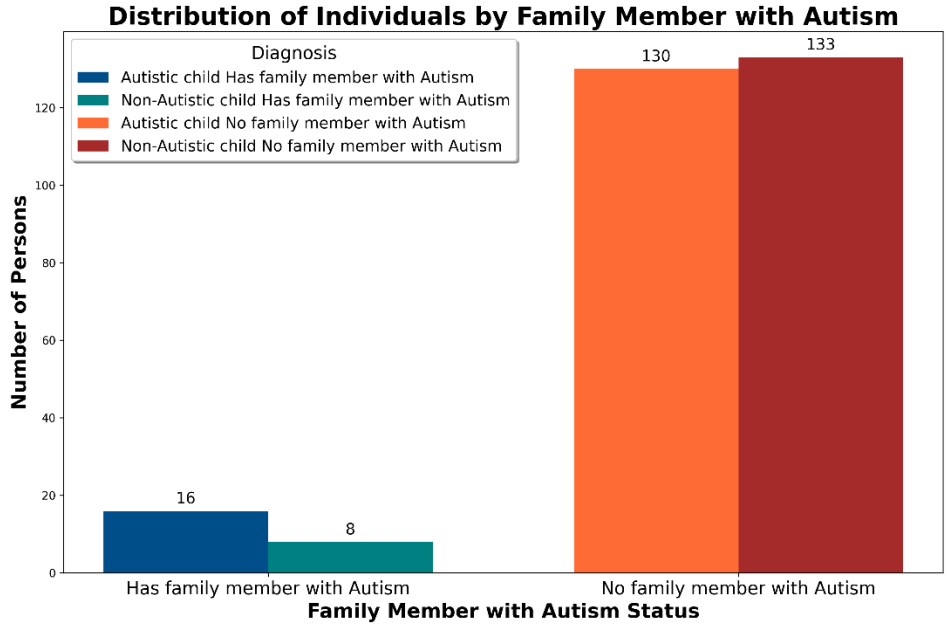


Figure 3.20: Family member with autism status distribution among individuals

This chart compares the residential area (rural vs. urban) of autistic and non-autistic children. It shows that a significantly higher number of children in both groups reside in Urban areas. But the ratio of autistic children is also alarming in rural areas.

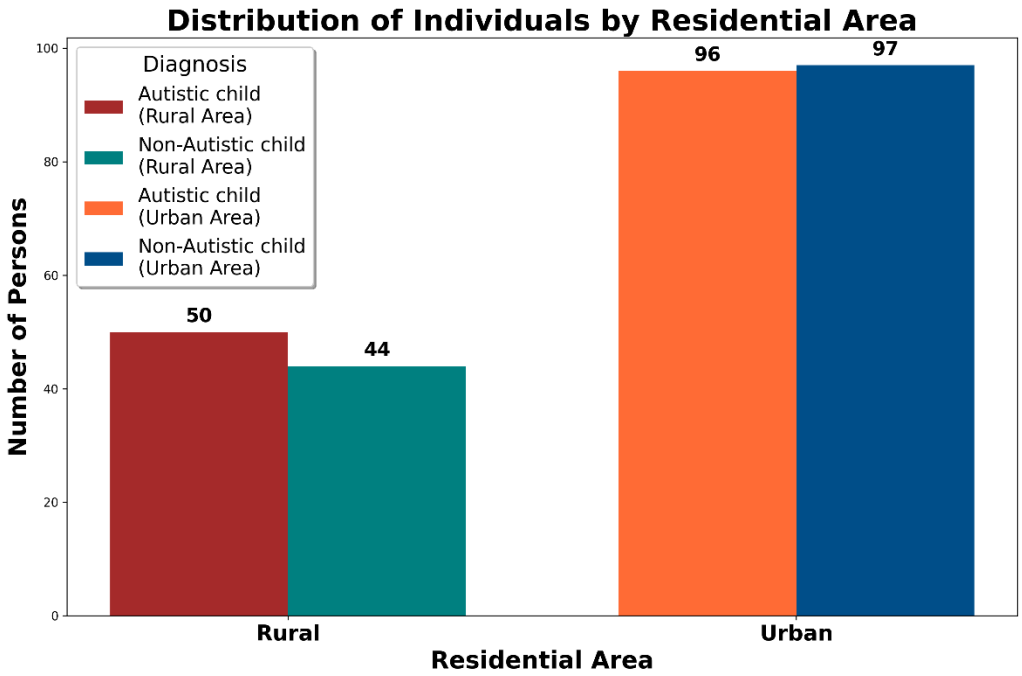


Figure 3.21: Distribution of Residential Area among individuals

Spearman's rank correlation coefficient method is used for generating the correlation matrix. And we visualize only the features with strong correlations. Strong positive correlations are observed between the target and features such as “Ability to communicate with people” (0.86) and “Toilet Training” (0.78), suggesting these are critical predictors. Conversely, features like “Mingles with peers” (−0.71) and “Limited Speech” (−0.54) show negative associations with the target, indicating inverse relationships. Several inter-feature dependencies are also evident, for instance, between “Needs help in dressing” and “Needs help in feeding” (0.73), reflecting overlapping behavioral pattern.

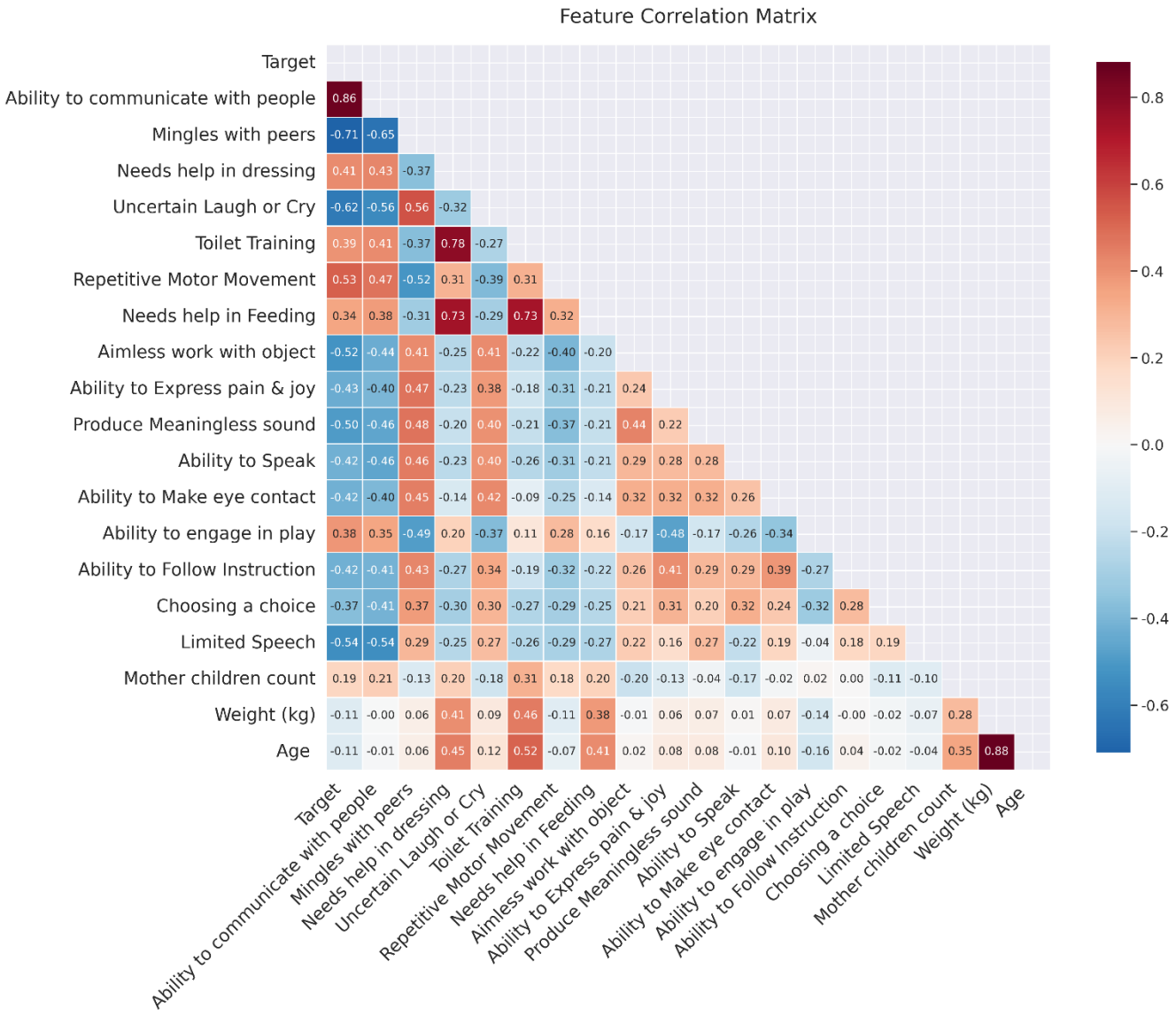


Figure 3.22: Correlation matrix of strongly correlated Features

3.3 Algorithms for Autism Prediction

For the task of autism prediction several machine learning algorithms are employed. The algorithms include Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron and Logistic Regression.

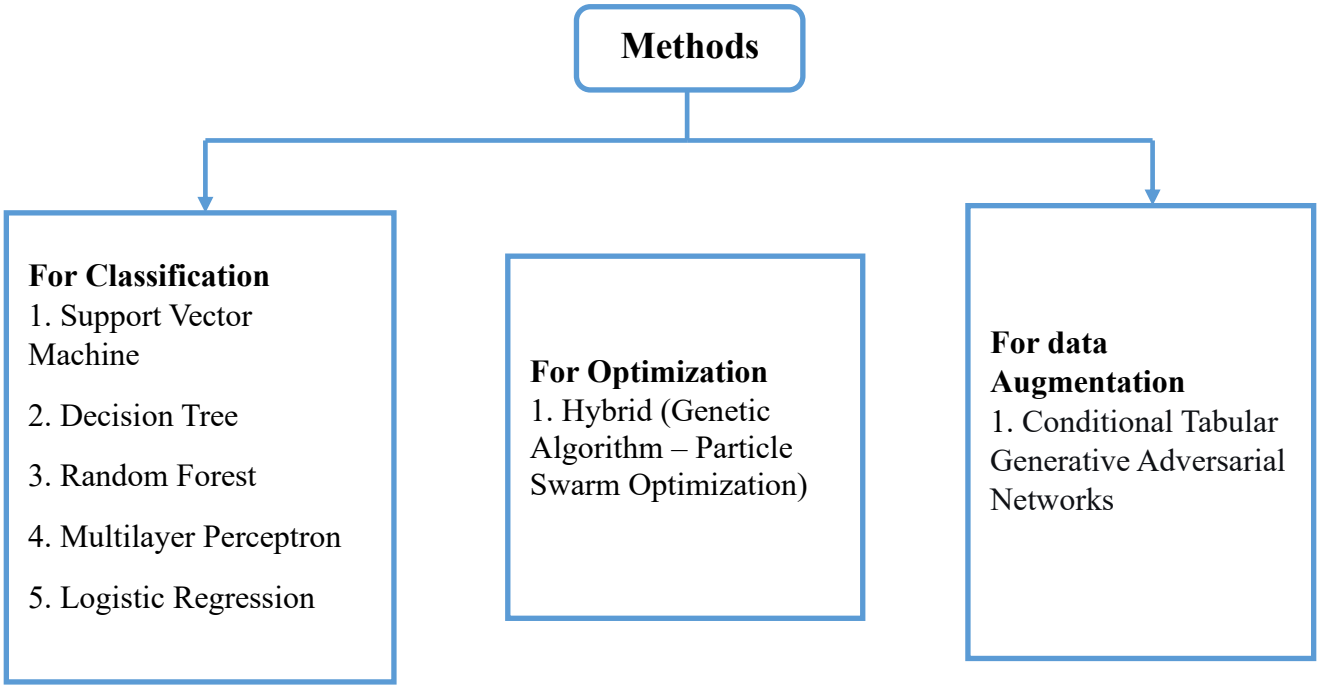


Figure 3.23: Several Methods for Classification, Optimization and Augmentation

(i) Support Vector Machine

The Support Vector Machine is a highly effective supervised learning technique for tasks involving regression as well as classification. However, its main application is in categorisation issue solving via machine learning. Its method aims to identify the decision boundary line that might divide the space with n dimensions into discrete groups so that further data points can be easily classified in the future. We refer to this ideal decision boundary as a hyperplane. SVM selects the hyperplane's extreme points and vectors. There's a reason the model is termed a Support Vector Machine: these severe conditions have been called support vectors. Examine the following graphic, which differentiates between two different categories using a type of hyperplane.

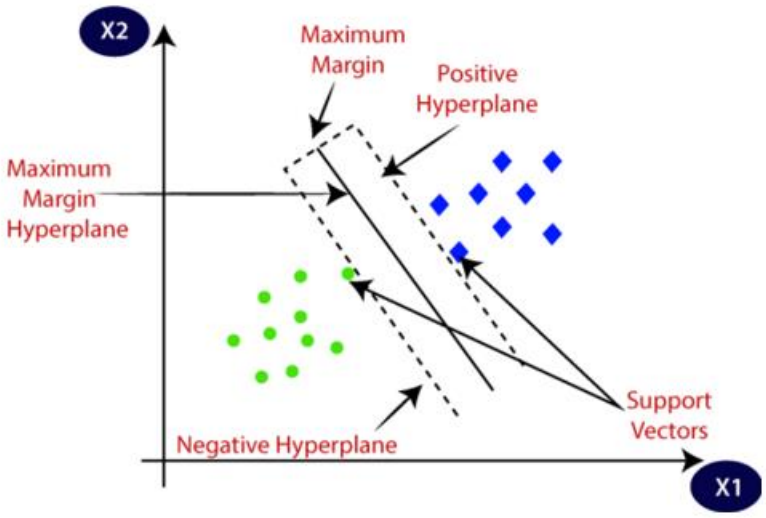


Figure 3.24: Support Vector Machine Classifier with Margin and Support Vectors [45].

(ii) Decision Tree

The decision tree is a part of the supervised learning process. In contrast to other models in this category, this approach can also be applied to regression and classification issues. By learning basic choice rules produced from previous data (training data), a decision tree can be used to develop a training model that can be used to predict the class or value of the target variable. When the target variable may accept continuous values, usually real numbers, regression trees are decision trees. DTR (Decision Tree Regression) allows regression models to be constructed as tree structures. A dataset is initially divided into progressively smaller subgroups, and then an associated decision tree is built incrementally, generating decision nodes or leaf nodes. A decision node typically has two branches, each of which represents a value for the tested attribute. if a judgement regarding the numerical aim is represented by the leaf node. The root node, the highest decision node in a tree, is referred to as the best predictor.

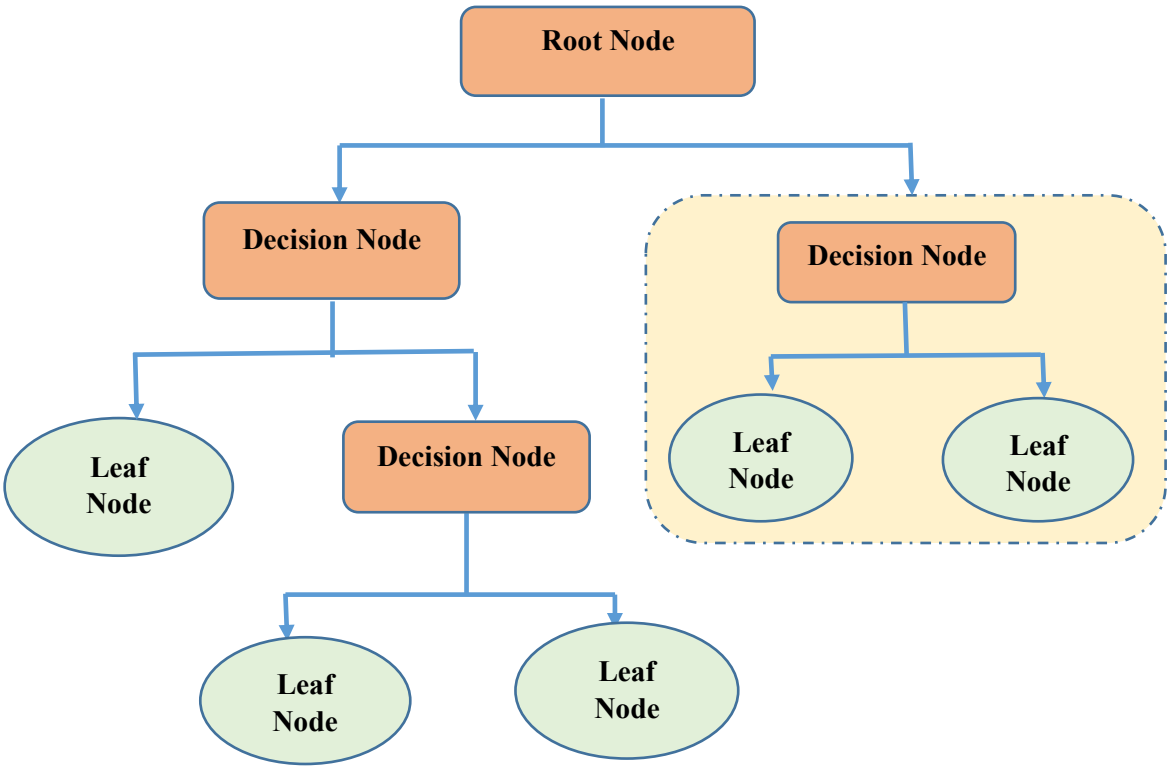


Figure 3.25: Structure of Decision Tree

(iii) Random Forest

One well-known computational learning method that is included in the guided learning category is the random forest approach. It can be applied to machine learning problems involving both regression and classification. Its foundation is the idea of collective learning, which integrates multiple classifiers to tackle a challenging problem while improving the model's performance. As the name implies, Random Forest functions as a learner that combines several decision trees depending on different subgroups of the dataset being utilised and selects the average to further improve the dataset's predicting accuracy. The random forest method calculates the majority vote of projections from each decision tree to decide the outcome rather than concentrating on just one.

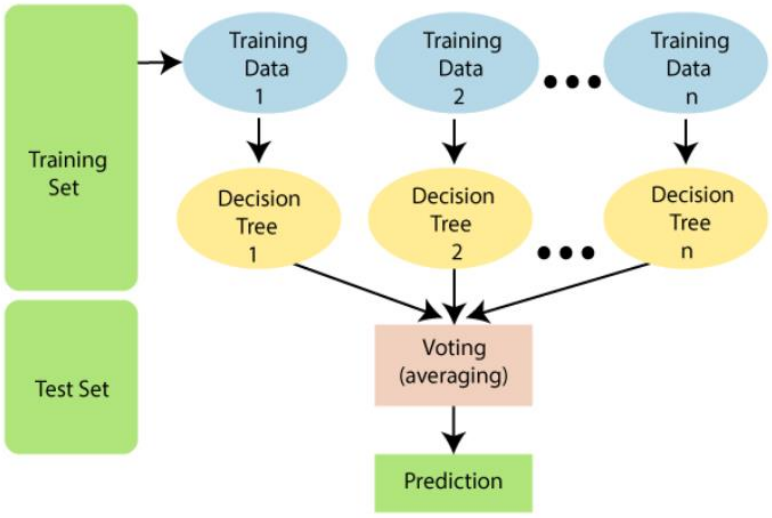


Figure 3.26: Random Forest's Architecture showing Multiple Decision Trees [46]

(iv) Multilayer Perceptron

A multilayer perceptron is a sort of feed-forward computational neural system that is very easy to understand. At least three basic node layers make up an MLP: an input layer, one or more hidden levels, and an output layer. The network is not isolated in any manner because all the nodes, or neurons, are connected to all the other nodes in the bottom layer. These hidden layers are what give MLPs their real power because they let them uncover complicated and non-linear relationships between the data. Each neural connection gets a different weight during training. Back-propagation is what makes learning happen. This algorithm compares the network output to the actual value after a prediction is made and figures out the error between the two. It then sends this error back through the layers. Each neuron has an output that is then placed through an activation function like Sigmoid or ReLU. This gives the function the non-linearity it needs to solve hard issues. MLPs are general-purpose universal approximators that can be utilized for many things, like processing natural language, recognizing images, and doing regression tasks. One of their strengths is that they can see complicated patterns in high-dimensional data.

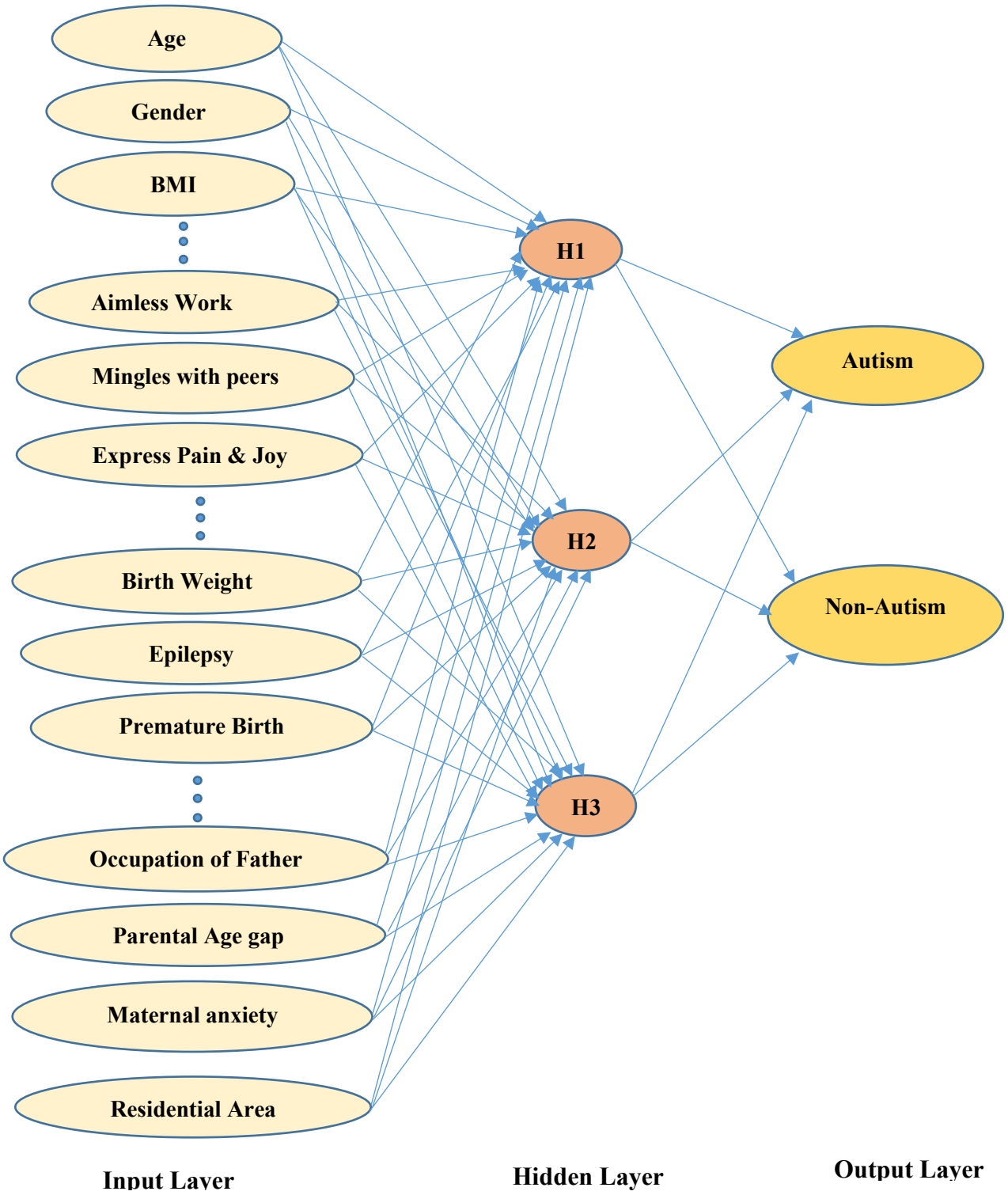


Figure 3.27: Architecture of Multilayer Perceptron Neural Network

(v) Logistic Regression

One approach for making predictions about binary classifications is logistic regression. Only two possible values can be assigned to the outcome or target variable. The term "dichotomous" describes the existence of only two classes. Finding canceer difficulties is one instance of this. It determines the likelihood of an event happening. Regression analysis is used to determine the parameter at which an event is likely to occur. A category is the desired parameter in this type of regression analysis. A logarithm of probability is used as the variable of interest. The process of employing a logistic function to estimate the likelihood of a binary event occurring is known as logistic regression.

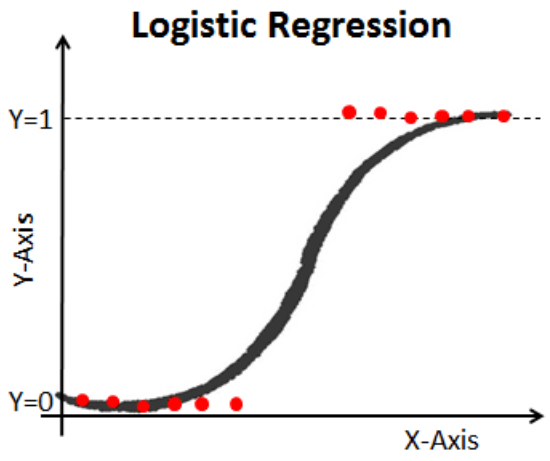


Figure 3.28: Logistic Regression [47]

Linear Regression Equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \quad \text{..... (3.2)}$$

Where y is the dependent variable and x1, x2,... and so on are the independent variables.

Sigmoid Function:

$$p = \frac{1}{1 + e^{-y}} \quad \text{..... (3.3)}$$

Apply the Sigmoid function on linear regression:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_n X_n)}} \qquad \ldots\ldots\ldots (3.4)$$

Logistic Regression Characteristics

- The variable that is dependent a in logistic regression described by a Bernoulli Distribution.
- Estimates are made by maximum likelihood.
- No such coefficient as R Square exists, to determine the goodness of fit of the model, concordance and KS-statistics are used.

(vi) Hybrid (GA-PSO) Optimization

A true hybrid GA-PSO optimization algorithm is a sophisticated metaheuristic that synergistically combines the strengths of Genetic Algorithms (GA) and Particle Swarm Optimization (PSO). It is designed to overcome the limitations of each individual method, creating a more robust and efficient search strategy. The core idea is to leverage PSO's strength in fine-tuning solutions through social learning and GA's power in exploring the search space via genetic operations. In this hybrid, a population of agents is treated both as a swarm of particles and a population of chromosomes.Each agent has a position, velocity (PSO traits), and a genetic representation (GA trait). The algorithm typically operates in cycles, alternating between PSO and GA phases. In the PSO phase, particles update their velocities and positions based on personal and global bests, promoting exploitation and local refinement. Subsequently, the GA phase injects exploration through selection, crossover, and mutation on the population's genetic material. This introduces new genetic diversity, preventing premature convergence to local optima.

The true hybrid seamlessly integrates these mechanisms, allowing information from the PSO's social bests to influence the GA's population and vice-versa. This creates a powerful balance, using PSO to quickly converge towards promising regions and GA to thoroughly explore and escape from them. The result is an algorithm with superior global search capability, faster convergence speed, and a higher likelihood of discovering the true global optimum for complex, high-dimensional problems that challenge standalone algorithms.

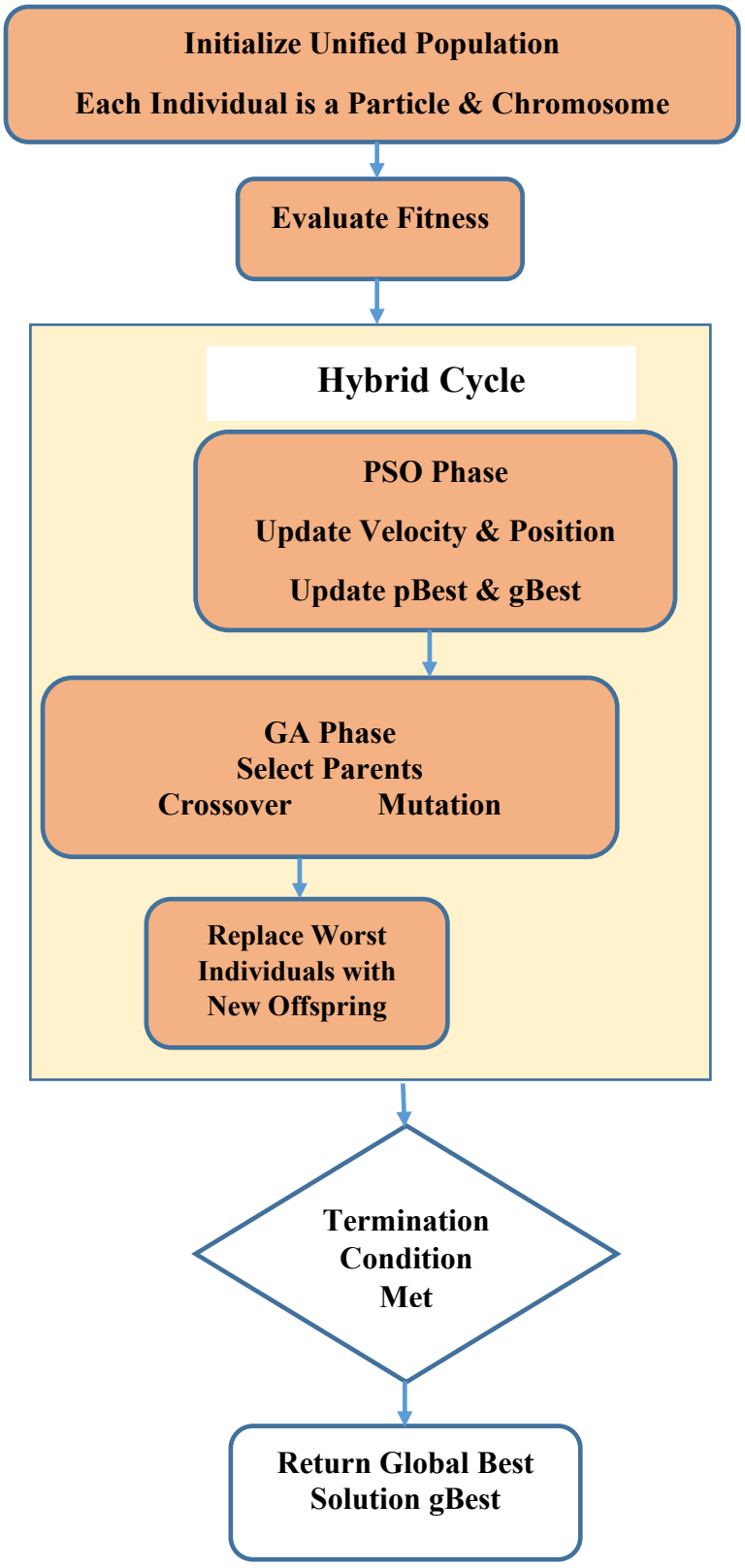
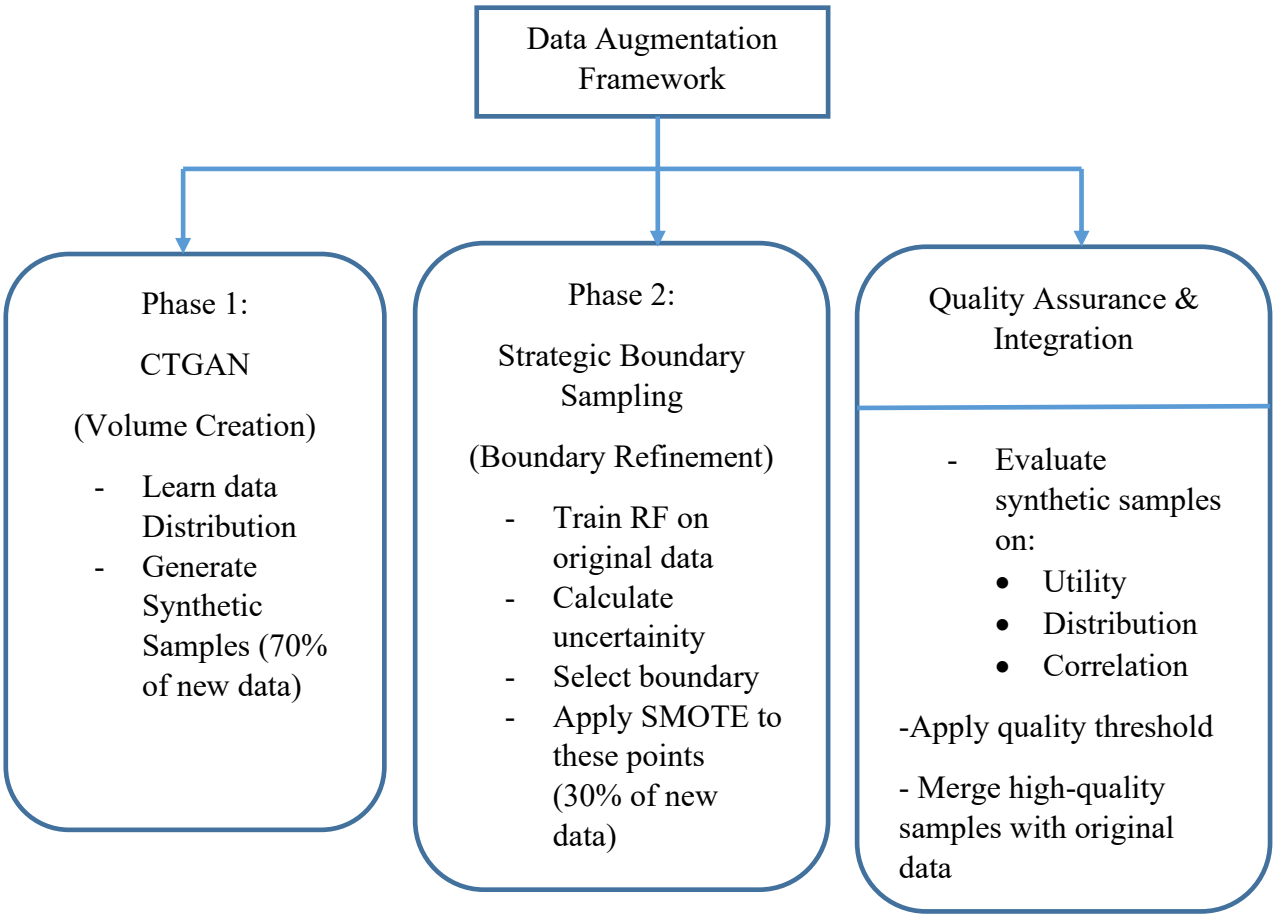


Figure 3.29: Architecture of (GA-PSO) Hybrid Optimization

3.4 Data Augmentation Framework

The primary augmentation technique used here is CTGAN (Conditional Tabular Generative Adversarial Network). This is supplemented by a secondary technique called Boundary Sampling. The pipeline first uses CTGAN to generate the majority (70%) of the new synthetic data and then uses boundary sampling to generate the remaining 30%. The quality of the generated data is rigorously assessed before it is used to augment the original dataset.



(i) Phase 1: Synthetic Sample Generation via CTGAN

CTGAN, which stands for Conditional Tabular Generative Adversarial Network, is a specialized generative model designed to create realistic synthetic tabular data. Unlike standard GANs that excel with image data, CTGAN addresses the unique challenges of mixed-type datasets containing both continuous and categorical variables. It employs mode-specific normalization to effectively handle non-Gaussian distributions in continuous columns, preventing the model from generating

unrealistic values. A key innovation is its use of conditional training through training-by-sampling, allowing it to generate high-quality samples for underrepresented categories by conditioning on specific features like the target variable.

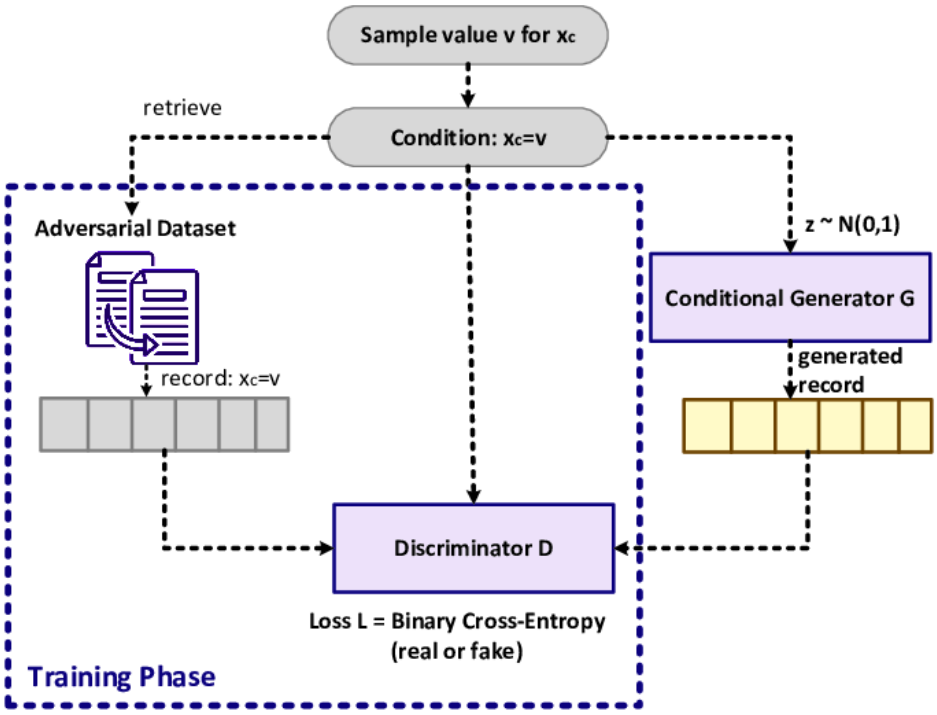


Figure 3.30: Architecture of CTGAN [48]

The architecture includes both a generator and a discriminator with deep neural networks, typically featuring multiple hidden layers for capturing complex data patterns. Its conditional vector implementation ensures generated data maintains meaningful relationships between features and specified conditions. CTGAN has demonstrated strong performance in preserving statistical properties and correlation structures from original datasets while generating entirely new samples.

The model includes privacy-aware mechanisms like PAC (Privacy-Aware Conditioning) to reduce the risk of memorizing and reproducing original data points verbatim. It has become a popular choice for data augmentation in research and applications where limited data availability constraints machine learning performance, particularly in healthcare for rare disease studies and financial services for fraud detection scenarios.

(ii) Strategic Boundary Sampling

Boundary Sampling is an advanced data augmentation technique that strategically generates synthetic samples in the most informative regions of the feature space—specifically, near the decision boundary of a trained classifier. By focusing on these ambiguous areas where the model exhibits uncertainty, the technique effectively targets the most challenging cases for classification.

The process begins by training a Random Forest classifier on the available data and using its prediction probabilities to identify samples where confidence is lowest. These uncertain points, typically located near the decision boundary, are selected for interpolation. New synthetic samples are created by linearly combining pairs of these boundary instances using a random weighting factor, which ensures diversity in the generated data.

(iii) Quality Assurance and Integration

After the creation of synthetic samples with the help of CTGAN and strategic SMOTE, a thorough, multi-dimensional, quality assessment was performed in order to warrant the fidelity and usefulness of the synthetic data prior to integration. This is an important step to avoid the entry of pernicious artifacts or distributional changes that will diminish model performance. All produced samples were tested on three main dimensions: classification utility (determined by the performance of a Random Forest classifier that was trained on synthetic data and tested on the held-out real data), distribution similarities (measured between a given feature under consideration and the original data via the Kolmogorov-Smirnov test), and correlation preservation (measured by comparing the inter-feature correlation matrices between the synthetic and the original data). Only synthetic samples that passed a pre-set quality threshold together, dynamically re-calibrated according to the complexity of the original data, were retained. This guaranteed that a selection of quality and plausible data was done. The ultimate augmented training set was then built by combining the approved synthetic samples of the two generative phases with the original, which led to much larger balanced and informative dataset that will be used to train a robust classifier.

3.5 Model Training and Testing

Train AI models using pre-processed meteorological data, dividing it into training and testing sets. The data is divided into k fold split, 70:30, 75:25, 80:20.The 70% of the total data will be used for training and 30% data will be used for testing for 70:30 ratio, similarly this will performed for

75:25 and 80:20. Create strong models using this split: 70%, 75%, 80% for learning from the data and 30%, 25%, 20% for checking how well the models predict new info. This way, models learn a lot but also get tested on new stuff. It makes sure they're good at predicting real-world weather situations. This split helps make accurate models that work well with different weather scenarios.

3.6 Performance Analysis

In order to compare performance, this study will employ a variety of evaluation metrics. In order to evaluate the effectiveness of the proposed model in the study, performance metrics are essential. Classification accuracy, recall/sensitivity, precision, and F1-score are the four performance metrics that will be used to qualify the study's results. To assess the value of these measurements, four terms will be used: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Matrix of Confusion (i): Confusion matrices are tables that highlight the performance of categorization techniques

(i) Confusion matrix: A confusion matrix is a summary and visual representation of the efficiency of an algorithm used for classification.

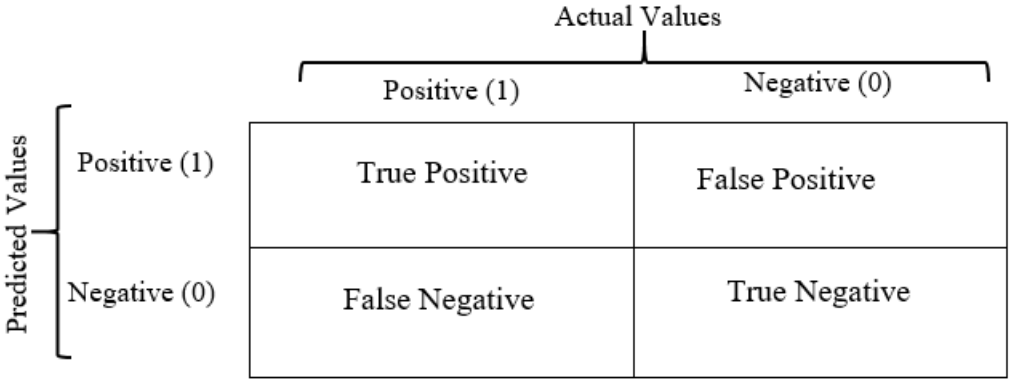


Figure 3.30: Confusion Matrix

(ii) Precision (P): A positive predictive value is what it is understood to be. It calculates the percentage of cases that are positive and anticipated to be positive relative to all cases that are anticipated to be positive.

$$P = \frac{TP}{TP+FP} \qquad \text{.....(3.5)}$$

(iii) Recall (R): From the total number of truly positive subjects, it calculates the percentage of positive and expected to be positive. Another name for it is an actual positive rate.

$$R = \frac{TP}{TP+FN} \qquad \text{.....(3.6)}$$

(iv) F1-score (F1): F1-score is calculated using precision and recall. It is very important to the experiment since it shows how accurate the test is.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \text{.....(3.7)}$$

(v) Accuracy (A): The number of correct prediction cases divided by the total number of cases is the ratio.

$$A = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{.....(3.8)}$$

(vi) AUROC: In binary classification problems, the Area Under the Receiver Operating Characteristic Curve, or AUROC, is a commonly used statistic to evaluate the performance of models used for classification. An illustration of a classifier's diagnostic ability, the ROC curve is produced by plotting the True Positive Rate (TPR), sometimes referred to as sensitivity or recall, across the False Positive Rate (FPR) (1 - specificity) at various levels of threshold. The area under this curve, or AUROC, is a single scalar statistic that calculates the model's ability to distinguish between the positive and negative classes.

(vii) ROC Curve

A ROC curve is a graph that plots the True Positive Rate (TPR) (Sensitivity) on the y-axis towards the False Positive Rate (FPR) (1-Specificity) on the x-axis to show how well a binary classification model performs across multiple thresholds. With the area under the curve (AUC) acting as a single indicator of model correctness, it aids in evaluating the model's capacity to discriminate either both positive and negative categories.

It works by Varying Thresholds: A score or probability is frequently the result of a classification model. A threshold is used to this score in order to categorize an item as either positive or negative.

TPR and FPR computation:

Determine the following at various threshold values:

The percentage of real positive instances that are accurately classified as positive is known as the true positive rate (sensitivity).
The percentage of true negative cases that are mistakenly classified as positive is known as the False Positive Rate (1 - Specificity).
Curve Plotting:
Plotting these TPR and FPR values for every threshold results in a series of points that make up the ROC curve.

3.7 Sensitivity Analysis

Sensitivity analysis is a financial and scientific modeling technique used to determine how variations in individual input variables impact a model's output, often referred to as a "what-if" analysis. By changing one independent variable at a time and observing the effect on the dependent variable, analysts can identify which factors have the greatest influence on the outcome, helping to predict future results, manage risk, and find opportunities.

(i) Non-Parametric Binned Sensitivity Analysis

To complement model-derived feature importance metrics and to better understand the fundamental relationship between each input variable and the target outcome, a non-parametric binned sensitivity analysis was employed. This technique, adapted from the work of, measures the sensitivity of the output variable to changes in each input feature independently, without assuming any underlying linear or parametric relationship. The procedure was executed as follows for each feature Xi:

Data Preparation: The dataset was sorted based on the values of the feature Xi under analysis, creating an ordered sequence of observations.
Binning: The sorted dataset was partitioned into k=20 quantile-based bins. This ensured an equal distribution of data points across bins and allowed for the analysis of localized effects across the entire range of the feature's values.

Local Mean Calculation: For each bin j_i, the mean value of the feature x_j and the corresponding mean value of the target variable y_j were calculated. This step effectively creates a piecewise linear approximation of the functional relationship $Y=f(X_i)$.

Sensitivity Calculation: The sensitivity between consecutive bins was computed to capture the localized rate of change. For each pair of consecutive bins j*j* and j+1*j*+1, the following equations were applied:

The percentage change in the input feature:

$$P = \frac{\bar{x}_{j+1} - \bar{x}_j}{|\bar{x}_j|} \times 100\% \quad \text{.....(3.9)}$$

The percentage change in the target output:

$$Q = \frac{\bar{y}_{j+1} - \bar{y}_j}{|\bar{y}_j|} \times 100\% \quad \text{.....(3.10)}$$

The individual sensitivity coefficient Si*Si* for that interval was then derived as:

$$S_i = \frac{Q}{P} \times 100\% \quad \text{.....(3.11)}$$

This coefficient represents the percentage change in the target output per percentage change in the input feature.

Aggregation and Filtering: For each feature, a set of sensitivity coefficients $\{S_1, S_2,..., S_{k-1}\}$ $\{S_1, S_2,...,S_{k-1}\}$ was generated. Extreme and non-finite values resulting from divisions by near-zero means were filtered out to ensure robustness. The final sensitivity score for the feature was defined as the mean of the absolute values of these individual coefficients:

Final Sensitivity = mean (|Si|)

This absolute mean provides a measure of the average magnitude of influence a feature has on the target, regardless of the direction (positive or negative).

Normalization and Ranking: The final sensitivity scores for all features were normalized on a scale of 0 to 100% relative to the highest-scoring feature to facilitate interpretation and comparison. Features were then ranked based on this normalized sensitivity score.

CHAPTER 4
RESULT & DISCUSSION

4.1 Introduction

In case of a binary classification task, there are five machine learning models analyzed in detail in this section: Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression. The individual models were evaluated through a set of key performance indicators like accuracy, precision, recall, F1-score, and ROC curves, and confidence matrices and ROC curves as well. The most confident classifier, based on the data, is the random Forest, and there is a clear hierarchy in the effectiveness of models. The pros and cons of each of these models are thoroughly analyzed, offering useful guidance to selecting an algorithm in practice. The impact of training-testing data splits, generalization on unseen data, and improvements using hybrid optimization and data augmentation strategies are also further discussed in this work.

4.2 Performance Analysis for distinct methods

In this section of a binary classification job, the authors provide a detailed performance comparison of five various forms of machine learning: Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression. Accuracy, precision, recall, F1-score and AUROC constitute the most significant measures, which were used to comprehensively evaluate any of the models; ROC curves and confusion matrices provided us with additional information. As Random Forest seems to be the most effective classifier, the data portrays a clear hierarchy in the efficacy of the model. The study offers a comprehensive analysis of the predictive capacity, strengths, and weaknesses of each model to give important insights to be taken when selecting the most efficient algorithm that can be applied in the real world.

(i) Support Vector Machine

The Support Vector Machine model was employed to classify the dataset and test its prediction strength against a variety of performance indicators. The measures of efficiency parameters are auroc, f1-score, recall, accuracy and precision.

Table: 4.1: Performance metrics of SVM

Model	Accuracy	Precision	Recall	F1-score	AUROC
SVM	0.9540	0.9550	0.9540	0.9540	0.9968

According to Table 4.1, SVM gave a high percentage of correct predictions across the dataset with an accuracy of 95.40. The accuracy of the model is 95.50% which shows the percentage of minimization of false positives and the recall is 95.40% which shows the percentage of the actual positive cases which the model can detect. The F1-score of 95.40% guarantees consistency since it confirms an excellent balance between precision and recall. Moreover, the outstanding discriminative ability of the model is indicated by the 0.9968 AUROC score indicating near perfect separation of classes.

(ii) Decision Tree

To classify the dataset and measure its predictive power using various performance metrics, the Decision Tree model was applied. The performance measures include accuracy, precision, recall, f1-score and auroc.

Table 4.2: Performance metrics of DT

Model	Accuracy	Precision	Recall	F1-score	AUROC
DT	0.919540	0.921606	0.919540	0.919476	0.919926

Table 4.2 showed the DT could correctly classify most of the cases with an accuracy of 91.95. The model showed an equal ability to identify true positives and minimize false positives since it got a precision and a recall of 92.16% and 91.95%, respectively. Its accuracy-recall reliability is supported by an F1-score of 91.95 percent. Moreover, the discriminative power of the model is also substantial as the AUROC value is 0.9199, which implies that the model can reliably discriminate classes.

(iii) Random Forest

To classify the data and evaluate the model predictive performance based on various performance measures, the Random Forest model was used. The analyzed performance indicators include accuracy, precision, recall, f1-score and auroc.

Table: 4.3: Performance metrics of RF

Model	Accuracy	Precision	Recall	F1-score	AUROC
RF	0.977011	0.978011	0.977011	0.976993	0.998943

Table 4.3 shows that this model was found to be good in detecting positive cases and minimizing misclassifications as illustrated by the good precision (97.80%) and recall (97.70%). The F1-score of 97.70 percent indicates an excellent combination of recall and precision. Moreover, the discriminative ability of the model is also great since the value of the AUROC score is 0.9989 and indicates nearly perfect discrimination.

(iv) Multilayer Perceptron

The Multilayer Perceptron model was used to classify the dataset and evaluate its predictive performance using the following performance metrics: accuracy, precision, recall, f1-score and auroc.

Table: 4.4: Performance metrics of MLP

Model	Accuracy	Precision	Recall	F1-score	AUROC
MLP	0.954023	0.955008	0.954023	0.954011	0.996829

The overall classification of the MLP was excellent with an accuracy of 95.40 as shown in Table 4.4. The model made positive identifications continually and reduced its error rates as shown by its precision and recall of about 95.50% and 95.40, respectively. The balance between recall and precision is verified by the F1-score of 95.40%. Moreover, the discriminative power of the model is excellent because the AUROC score is 0.9968.

(v) Logistic Regression

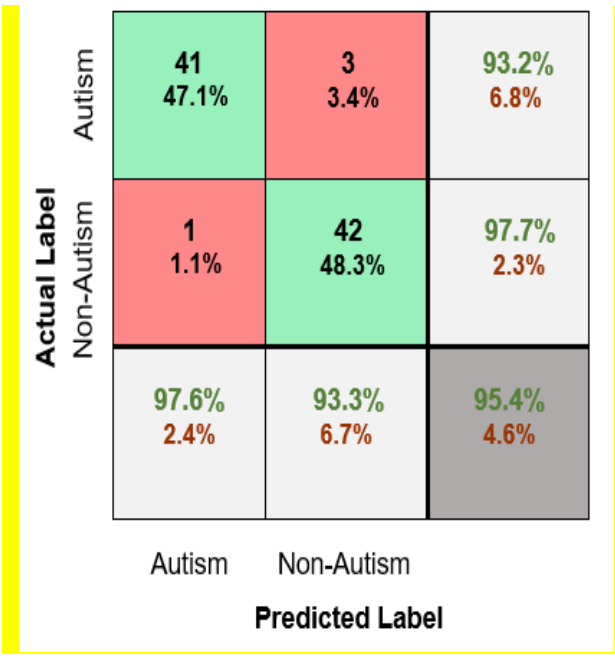
A Logistic Regression model was applied to predict the data and evaluate its prediction potential according to various performance indices including accuracy, precision, recall, f1-score and auroc.

Table: 4.5: Performance metrics of LR

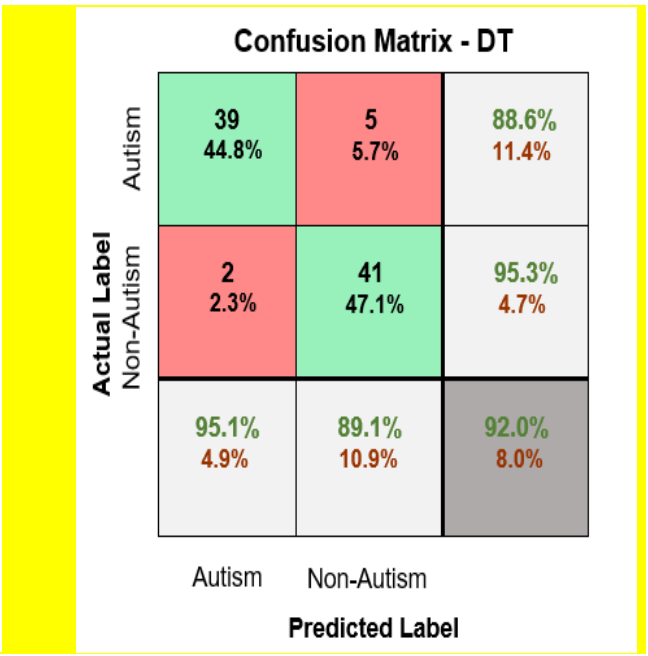
Model	Accuracy	Precision	Recall	F1-score	AUROC
LR	0.9425	0.9446	0.9425	0.9424	0.9941

As it is observed in Table 4.5, LR model accurately and reliably identified most of the cases with a percent accuracy of 94.25. Although the recall of 94.25% indicates that the technique is effective in identifying actual positive cases, the false positive which is 94.47% indicates that the technique is effective in reducing false positives. The F1-score of 94.25% ensures consistent prediction quality since there is a balanced performance of accuracy and recall. Additionally, the model's outstanding discriminative performance is shown by the AUROC score of 0.9942, which shows that it can successfully distinguish between the classes.

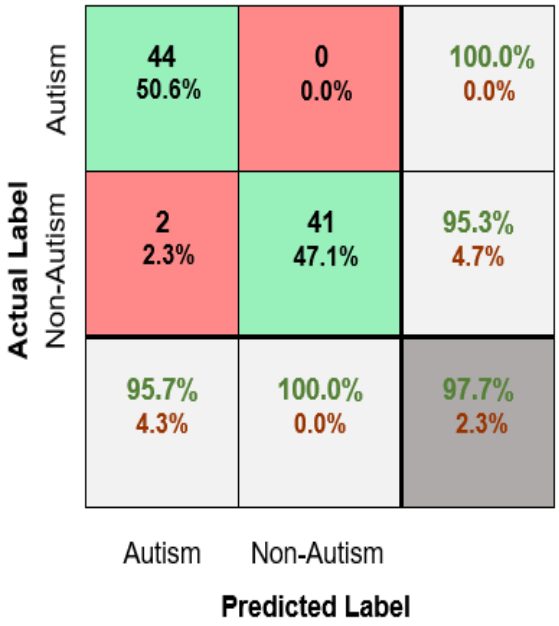
The performance of five different machine learning models—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression—was assessed for a binary classification task, most likely predicting autism, using the confusion matrices shown in Figure 4.1. Along with measurements like precision, recall (sensitivity), specificity, and accuracy, each matrix shows true positives, false positives, true negatives, and false negatives.



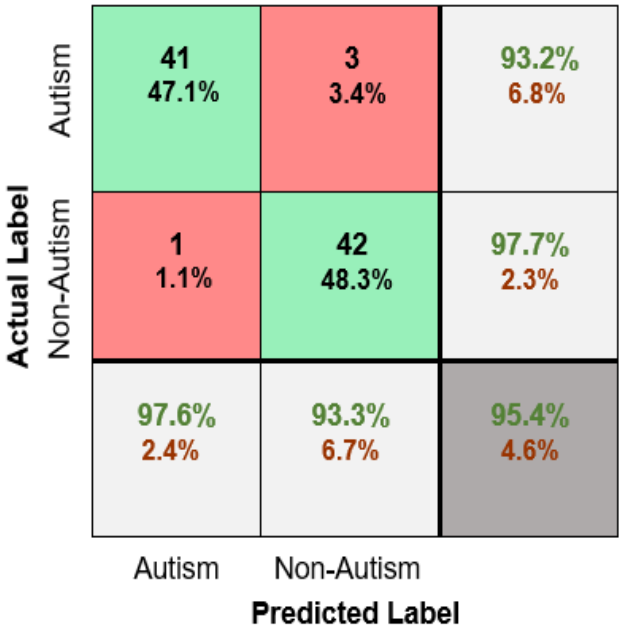
(a) SVM



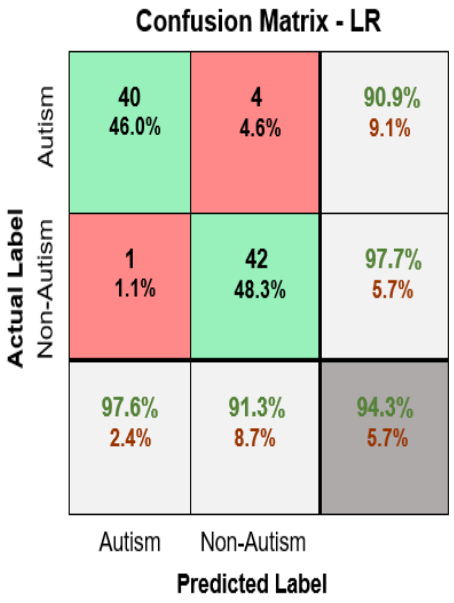
(b) DT



(c) RF



(d) MLP



(e)

Figure 4.1: Confusion Matrices of (a) SVM, (b) ……..all models

Random Forest demonstrates the highest performance, achieving perfect precision for the "Autism" class 100% with no false positives, indicating that every positive prediction was correct. It also has the highest overall accuracy 97.7% and a notably high recall for "Autism" 100%, meaning it correctly identified all actual positive cases. The model shows strong specificity 95.7% for "Non-Autism," with only two false negatives.

Support Vector Machine and Multilayer Perceptron exhibit identical results, each with an accuracy of 95.4%. Both models correctly classified 41 autism cases and 42 non-autism cases, with three false positives and one false negative. They show balanced precision and recall, with high specificity 97.6% and moderate sensitivity 93.2% for the "Autism" class.

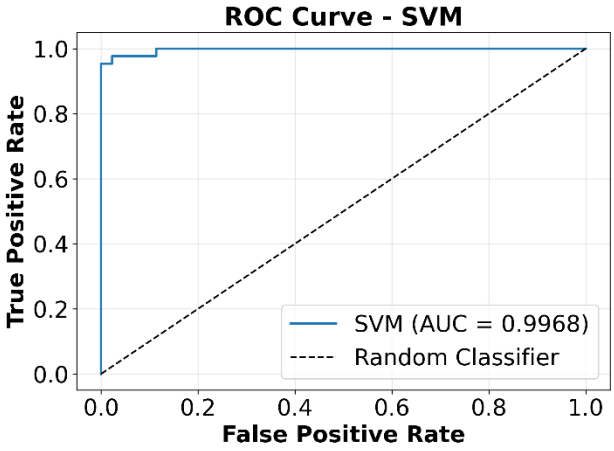
Logistic Regression follows closely with an accuracy of 94.3%. It misclassified four autism cases as false positives and one non-autism case as a false negative. While its sensitivity for "Autism" is lower 90.9%, it maintains high specificity 97.6% for "Non-Autism."

Decision Tree performs the weakest among the models, with an accuracy of 92.0%. It has the highest number of errors: five false positives and two false negatives. This results in lower

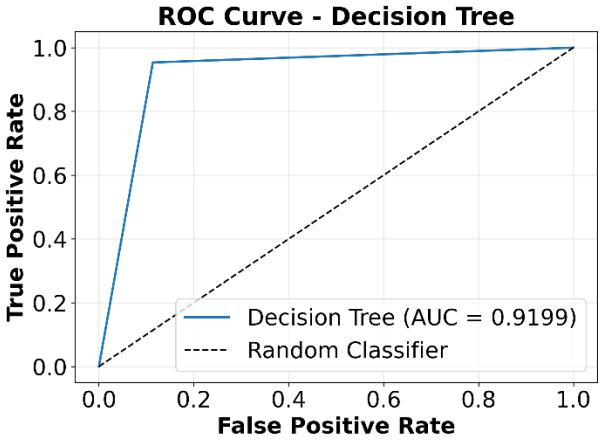
precision 89.1% and recall 88.6% for the "Autism" class, indicating a higher rate of misclassification compared to other models.

In summary, Random Forest stands out as the most robust model, offering optimal balance between precision and recall with the highest accuracy. Support Vector Machine and Multilayer Perceptron provide consistent and reliable performance, while Logistic Regression remains competitive. Decision Tree, though less accurate, still achieves reasonable results but with a higher propensity for errors. These findings highlight RF's superiority for this specific classification task.

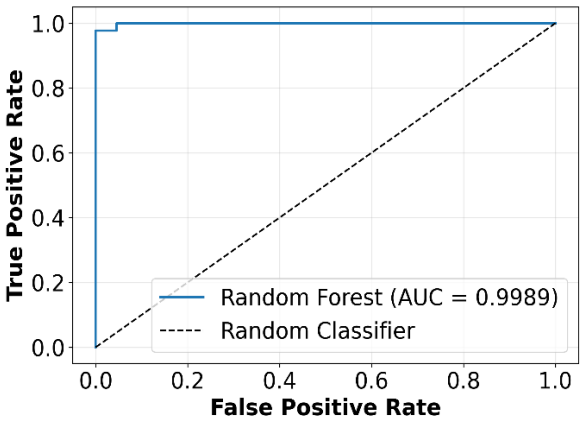
The provided ROC curves in Figure 4.2 illustrate the diagnostic performance of five machine learning models in classifying autism, with each curve's Area Under the Curve quantifying its ability to distinguish between classes.



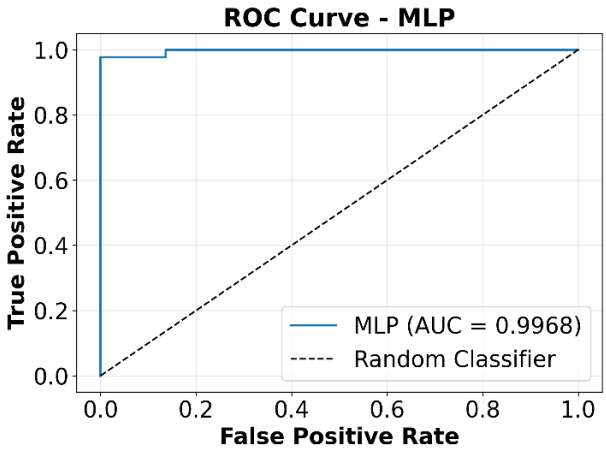
(a)



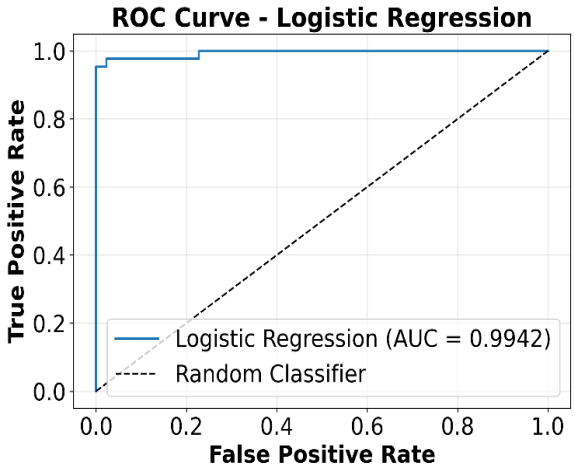
(b)



(c) RF



(d)



(e)

Figure 4.2: ROC curve of all Models

Near-perfect separation is achieved by Random Forest, which has the greatest AUC of 0.9989. With a high true positive rate (sensitivity) and a very low false positive rate across all thresholds, the model performs exceptionally well, as evidenced by its curve hugging the top-left corner. This demonstrates that it is the most reliable classifier.

Similarly, with an AUC of 0.9968, Support Vector Machine and Multilayer Perceptron also exhibit exceptional and similar performance. Additionally, their curves are situated much above the random classifier line, indicating a high degree of predictive accuracy and a well-balanced sensitivity and specificity.

Then comes Logistic Regression, which has a great AUC of 0.9942. At different probability thresholds, it maintains an excellent capacity for accurate classification, with a curve that is only slightly lower than that of SVM and MLP.

On the other hand, the Decision Tree model's AUC of 0.9199 is significantly lower. Its curve is closer to the diagonal, indicating a higher rate of false positives or false negatives depending on the threshold selected, but it is still good and much better than random guessing (AUC=0.5). As a result, it is the least discriminatory of the five models.

The comparison of ROC curves for all deployed models is visualized in Figure 4.3. In conclusion, the confusion matrices' hierarchy of model performance is confirmed by the ROC analysis.

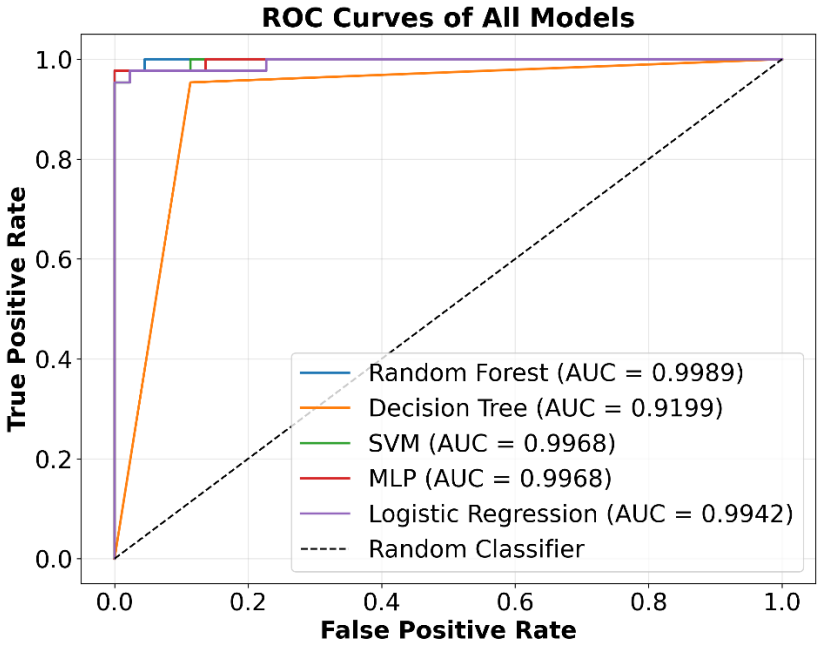


Figure 4.3: ROC Curve comparison among all models

In summary, The ROC curve visualization demonstrates exceptional predictive performance across all models, with Random Forest achieving near-perfect discrimination (AUC = 0.9989). All machine learning models significantly outperform the random classifier baseline, indicating strong classification capability. The high AUC values (all above 0.91) confirm the models' effectiveness in distinguishing between classes, with SVM and MLP showing identical excellent performance (AUC = 0.9968). Logistic Regression also performed remarkably well (AUC = 0.9942), while Decision Tree showed the lowest (yet still strong) performance among the trained models.

4.2.1 Performance Analysis for Different Training Testing Ratios

The performance of machine learning models is highly influenced by the proportion of data allocated for training versus testing. Table 4.6 presents a comprehensive analysis of five classifiers—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression—evaluated across three different dataset splits: 70:30, 75:25, and 80:20. This comparison using key metrics such as Accuracy, Precision, Recall, F1-score, and AUROC helps identify the most stable and effective model. The results demonstrate how each algorithm's generalization capability varies with the amount of training data provided, offering critical insights for optimal model selection in practical applications.

Table 4.6: Performance of all models across different Training Testing Ratios

Training & Testing Ratios	Models	Accuracy	Precision	Recall	F1-score	AUROC
70:30	SVM	0.9540	0.9550	0.9540	0.9540	0.9968
	DT	0.9195	0.9216	0.9195	0.9194	0.9199
	RF	0.9770	0.9780	0.9770	0.9769	0.9989
	MLP	0.9540	0.9550	0.9540	0.9540	0.9968
	LR	0.9425	0.9446	0.9425	0.9424	0.9941
75:25	SVM	0.9583	0.9587	0.9583	0.9583	0.9953
	DT	0.9722	0.9737	0.9722	0.9722	0.9729
	RF	0.9722	0.9722	0.9722	0.9722	0.9965
	MLP	0.9722	0.9722	0.9722	0.9722	0.9922
	LR	0.9583	0.9587	0.9583	0.9583	0.9930
80:20	SVM	0.9482	0.9488	0.9482	0.9482	0.9940
	DT	0.9310	0.9332	0.9310	0.9310	0.9321
	RF	0.9655	0.9655	0.9655	0.9655	0.9976
	MLP	0.9655	0.9655	0.9655	0.9655	0.9928
	LR	0.9482	0.9488	0.9482	0.9482	0.9904

In table 4.6 the analysis of model performance across different training-testing splits reveals critical insights into the stability and robustness of each classifier. In general, the best-performing algorithm is always a Random Forest with impressive stability and high metrics in all data partitions. RF scores the highest in all of the categories, including a phenomenal accuracy of 0.977, an F1-score of 0.977 and an

almost perfect AUROC of 0.999. This performance demonstrates its high generalization capacity even at an ordinary training proportion. While SVM and MLP also perform strongly and identically in this split, RF's dominance is clear. Notably, the 75/25 split presents a fascinating scenario where Decision Tree shows a significant performance boost, matching RF's accuracy, precision, recall, and F1-score at 0.972. This suggests that for this specific algorithm, a slightly larger training set 75% provides a substantial benefit, allowing it to perform on par with the otherwise superior RF model for core metrics, though its AUROC 0.973 still lags behind RF's 0.997. This split appears to be the most favorable overall, with multiple models such as DT, RF, MLP achieving the same high level of accuracy. Conversely, the 80:20 split sees a slight performance dip for most models compared to the 75:25 scenario. RF maintains its lead with an accuracy of 0.966, but SVM and LR drop to approximately 0.948. This indicates that while a larger training set is generally beneficial, there is a point of diminishing returns for some models, potentially due to overfitting or a testing set that is too small to provide a robust evaluation. DT's performance also decreases in this split, reinforcing its sensitivity to the data partitioning compared to the more stable RF. The AUROC values remain consistently high for SVM, RF, MLP, and LR across all splits, consistently staying above 0.99 and often approaching 1.0. This indicates that all these models possess an excellent capacity for distinguishing between classes, regardless of the train-test ratio. DT, however, has a significantly lower and more variable AUROC, highlighting its weaker discriminative ability compared to the ensemble and neural network methods. In conclusion, Random Forest proves to be the most reliable and high-performing model across various data splits, while the 75/25 ratio often provides the most balanced and favorable conditions for model training and evaluation. This analysis is crucial for selecting a model that generalizes well and is not overly sensitive to the choice of dataset partitioning.

4.2.2 Model's Performance on Unseen Data

The ultimate test of a machine learning model's robustness is its performance on completely unseen data. A critical analysis of the 5 classifiers has been done in terms of their prediction of 17 independent test samples as shown in Table 4.7. This comparison transcends a summary of the aggregate data to present the more detailed picture of the practical usefulness of each model, not only in terms of its general accuracy but also in terms of the examples when it works and when it does not. The table is an important measure of the generalization ability and real-life validity of models.

Table 4.7: Accuracy and error rate of models on completely unseen test data

S/No.	Actual	Predicted					Accuracy (%)					Error				
		SVM	DT	RF	LP	LR	SVM	DT	RF	MLP	LR	SVM	DT	RF	MLP	LR
S1	1	0	0	0	0	0	94.1	82.4	94.1	94.1	94.1	5.9	17.6	5.9	5.9	5.9
S2	0	0	0	0	0	0										
S3	1	1	1	1	1	1										
S4	1	1	1	1	1	1										
S5	0	0	0	0	0	0										
S6	0	0	0	0	0	0										
S7	1	1	1	1	1	1										
S8	1	1	1	1	1	1										
S9	0	0	0	0	0	0										
S10	1	1	1	1	1	1										
S11	1	1	0	1	1	1										
S12	0	0	1	0	0	0										
S13	1	1	1	1	1	1										
S14	0	0	0	0	0	0										
S15	1	1	1	1	1	1										
S16	0	0	0	0	0	0										
S17	1	1	1	1	1	1										

For SVM:

Total Correct: 16 out of 17

Accuracy: $16 \times 100/17 = 94.1\%$

Total error: 1 out of 17

Error rate: $1 \times 100/17 = 5.9\%$

For DT:

For

This performance analysis on completely unseen data provides the most telling evaluation of mode generalizability. The results clearly stratify the five models into two distinct tiers. Four models—Support Vector Machine, Random Forest, Multilayer Perceptron, and Logistic Regression —demonstrate exceptional and identical robustness, each achieving a high accuracy of 94.1% by correctly classifying 16 out of 17 samples. This indicates a strong ability to apply learned patterns to novel data, with each making only a single error.

In stark contrast, the Decision Tree model performs significantly worse, achieving an accuracy of only 82.4% with three misclassifications. This higher error rate of 17.6% suggests a tendency to overfit the training data, limiting its effectiveness on unseen examples. A detailed examination of the errors reveals critical insights: all models, including the top performers, unanimously misclassified sample S1, indicating this particular instance may be a true outlier or an inherently ambiguous case that is difficult for any model to classify correctly.

Furthermore, the Decision Tree's additional unique errors on samples S11 and S12 pinpoint specific weaknesses in its decision boundaries that the other, more complex models were able to overcome. The fact that SVM, RF, MLP, and LR produced identical predictions for every single sample is remarkable, highlighting a strong consensus among the best-performing algorithms. This ultimate test confirms Random Forest's position as a top performer while exposing the limitations of the Decision Tree, solidifying that ensemble and linear models generalize more effectively to new, real-world data in this application.

4.2.3 Summary

This section presents a comprehensive performance evaluation of five machine learning models—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression for a binary classification task. Among these, Random Forest emerged as the superior classifier, achieving the highest accuracy 97.70%, precision 97.80%, recall 97.70%, F1-score 97.70%, and an exceptional AUROC of 0.9989, demonstrating near-perfect class separation. SVM and MLP had exactly the same good performance in all metrics followed by Logistic Regression. Decision Tree model did the worst, which means it is more prone to overfitting. Comparison between training-testing splits revealed that RF is consistently strong and can generalize. This last test on entirely unseen data further

confirmed RF to be the best, since it, together with SVM, MLP and LR attained 94.1% accuracy, which was far much better than DT. This model efficacy hierarchy gives important information in the best algorithm to use in real life.

4.3 Performance Analysis for distinct methods with Optimization

This study employs a hybrid Genetic Algorithm-Particle Swarm Optimization (GA-PSO) algorithm to optimize five binary classification machine learning models: SVM, DT, RF, MLP and LR. The adopted metaheuristic approach is efficient in traversing the hyper-parameter space, effectively combining the convergent and global searching ability of PSO and GA. Besides confusion matrices and ROC curves, the optimized models are also comprehensively compared in terms of accuracy, precision, recall, F1-score and AUROC. The results reveal a clear performance hierarchy with the significant role of hybrid optimization in model effectiveness.

(i) Support Vector Machine

The Support Vector Machine model was executed using hybrid (GA-PSO) optimization to categorize the data set and evaluate its predictive performance with respect to several performance measures. The performance metrics that are measured include accuracy, precision, recall, f1-score and auroc. Best parameters for support vector machine are: {'C': 0.13700627548360347, 'kernel': 'poly', 'gamma': 'scale'}

Table: 4.8: Performance metrics of Hybrid (GA-PSO) Optimized SVM

Model	Accuracy	Precision	Recall	F1-score	AUROC
SVM	0.9770	0.9770	0.9770	0.9770	0.9984

Using the hybrid GA-PSO approach, the SVM model showed outstanding performance on all evaluation measures. It achieved a high accuracy of 97.70 with exactly the same precision, recall and F1-score, indicating a fully balanced classification efficiency between the two classes. The fact that the model almost perfectly differentiates between the positive and negative cases is also backed up by the impressive value of the AUROC at 0.9984. These results indicate the effectiveness of the GA-PSO optimization with regard to enhancing the generalization and prediction properties of SVM in this binary classification task.

(ii) Decision Tree

The Decision Tree model was optimized using hybrid (GA-PSO) optimization to classify the our dataset and evaluate its performance in various metrics of performance accuracy, precision, recall, f1-score and auroc. Best parameters for decision tree are: {'max_depth': 10, 'min_samples_split': 8, 'min_samples_leaf': 4}

Table: 4.9: Performance metrics of Hybrid (GA-PSO) Optimized DT

Model	Accuracy	Precision	Recall	F1-score	AUROC
DT	0.988506	0.988767	0.988506	0.988506	0.987051

The Decision Tree model performed admirably with the aid of hybrid GA-PSO optimization, reaching an outstanding accuracy of 98.85. It shows an excellent performance in terms of minimizing false positives and specifically identifying true positives in the sense that it has nearly equal precision 98.88%, recall 98.85%, and F1-score 98.85%. Its outstanding discriminatory ability is also complemented by its high AUROC score of 0.987. These results indicate that the use of GA-PSO optimization significantly enhances the capability of the Decision Tree to make predictions and generalization in this classification exercise.

(iii) Random Forest

Random Forest model was undertaken with hybrid (GA-PSO) optimization to classify the dataset and evaluate its predictability with respect to various performance measurements. The experimental performance measures are accuracy, precision, and recall, f1-score, and auroc. Best Parameters for Random Forest are: {'n_estimators': 53, 'max_depth': 6, 'min_samples_split': 9, 'min_samples_leaf': 3}

Table: 4.10: Performance metrics of Hybrid (GA-PSO) Optimized RF

Model	Accuracy	Precision	Recall	F1-score	AUROC
RF	0.988506	0.988761	0.988506	0.988503	0.997886

The results with the Random Forest model using the hybrid GA-PSO technique yielded excellent results with a top-tier accuracy of 98.85. Its classification performance reflects almost perfect balance and consistency in terms of precision (98.88%), recall (98.85%), and F1-score (98.85%). Its outstanding discriminative power which borders on perfect separation of the classes is demonstrated by its outstanding value of the AUROC at 0.998. These results reflect the significant performance gains the hybrid GA-PSO optimization method achieves and support the position of Random Forest as a highly confident classifier.

(iv) Multilayer Perceptron

A Multilayer Perceptron model was undertaken with hybrid (GA-PSO) optimization to classify the dataset and evaluate its predictability with respect to various performance measurements. The experimental performance measures are accuracy, precision, recall, f1-score and auroc. Best parameters for multilayer Perceptron are {'hidden_layer_sizes': (100, 50), 'activation': 'tanh', 'alpha': 0.006774466914081445, 'learning_rate_init': 0.08230091615697481}

Table: 4.11: Performance metrics of Hybrid (GA-PSO) Optimized MLP

Model	Accuracy	Precision	Recall	F1-score	AUROC
MLP	0.977011	0.977011	0.977011	0.977011	0.997357

The Multilayer Perceptron model, which was improved by use of hybrid GA-PSO optimization showed great performance with an 97.70 accuracy. The consistency and reliability with which it can classify with either class is seen in its perfectly balanced precision, recall and F1-score of 97.70 percent. The high AUROC value of 0.997 further supports the strong discriminative ability and great capability of the model to differentiate the labels of the classes. The above findings indicate that the GA-PSO optimization method is very effective in maximizing the predictive performance of the MLP in this binary classification problem.

(v) Logistic Regression

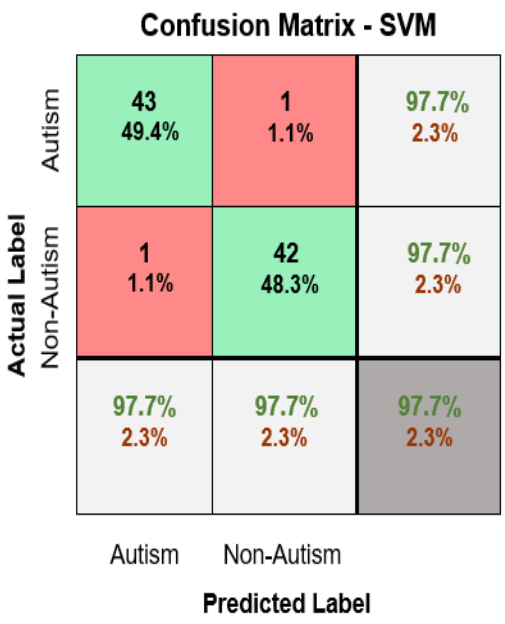
Logistic Regression model was undertaken with hybrid (GA-PSO) optimization to classify the dataset and evaluate its predictability with respect to various performance measurements. The experimental performance measures are accuracy, precision, and recall, f1-score and auroc. Best parameters for logistic regression are: {'C': 9.727811873705896, 'penalty': 'l2', 'solver': 'sag'}

Table: 4.12: Performance metrics of Hybrid (GA-PSO) Optimized LR

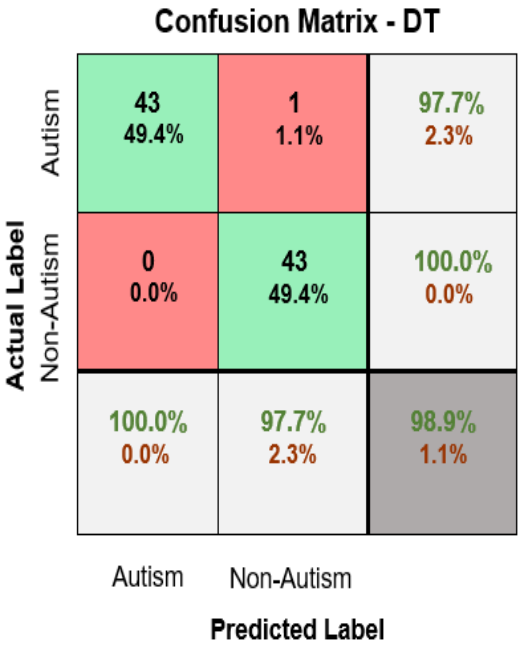
Model	Accuracy	Precision	Recall	F1-score	AUROC
LR	0.942529	0.944686	0.942529	0.942483	0.988372

The hybrid GA-PSO optimized Logistic Regression model had good results with an accuracy of 94.25. Its precision, 94.47%, recall, 94.25%, and F1-score, 94.25% have a good balance, but are considerably worse than the other optimized models. Nevertheless, it has an excellent discriminative property with a high AUROC of 0.988, which suggests that this metric successfully divides classes even though the accuracy measures are lower. These findings indicate that although the GA-PSO optimization greatly improved the performance of LR, it is the least efficient of the five optimized models in this classification task.

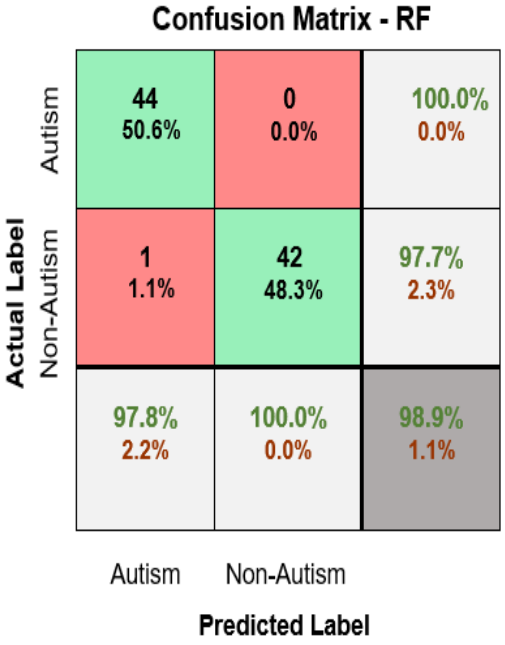
As per the given confusion matrices in Figure 4.5, the accuracy of five different machine learning models, namely Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression, with Hybrid (GA-PSO) Optimization improvements, were tested on a binary classification problem. True positives, false positives, true negatives, and false negatives are shown in each of the matrices and measures derived, including precision, recall (sensitivity), specificity, and accuracy. The hyperparameters of each model were further optimized with the use of GA-PSO optimization, resulting in better discriminatory strength and performance of the classification models. This discussion reveals the significant influence of hybrid optimization to maximize predictive accuracy and reduce the misclassification rate in all algorithms.


Confusion Matrix - SVM
Actual Label
Autism
Non-Autism
43
49.4%
1
1.1%
97.7%
2.3%
1
1.1%
42
48.3%
97.7%
2.3%
97.7%
2.3%
97.7%
2.3%
97.7%
2.3%
Autism
Non-Autism
Predicted Label

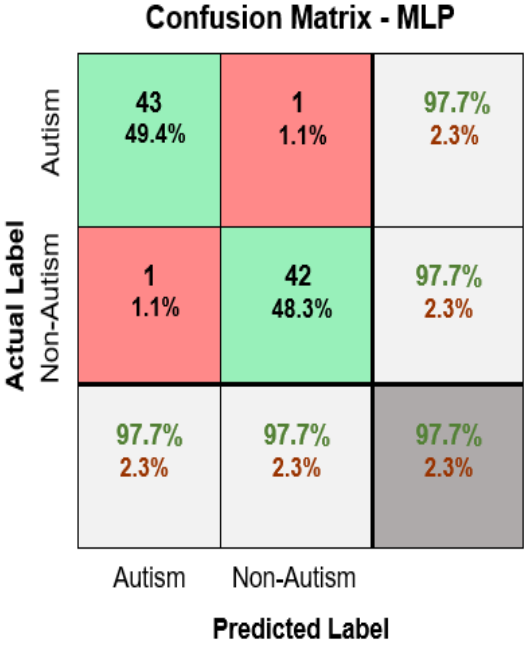
(a)


Confusion Matrix - DT
Actual Label
Autism
Non-Autism
43
49.4%
1
1.1%
97.7%
2.3%
0
0.0%
43
49.4%
100.0%
0.0%
100.0%
0.0%
97.7%
2.3%
98.9%
1.1%
Autism
Non-Autism
Predicted Label

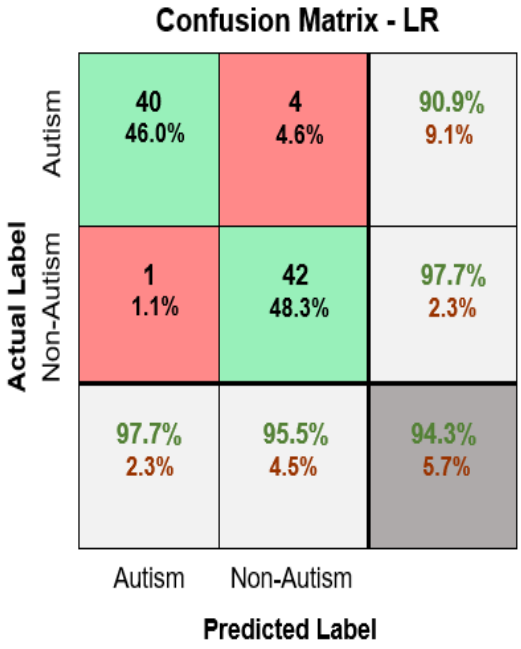
(b)


Confusion Matrix - RF
Actual Label
Autism
Non-Autism
44
50.6%
0
0.0%
100.0%
0.0%
1
1.1%
42
48.3%
97.7%
2.3%
97.8%
2.2%
100.0%
0.0%
98.9%
1.1%
Autism
Non-Autism
Predicted Label

(c)


Confusion Matrix - MLP
Actual Label
Autism
Non-Autism
43
49.4%
1
1.1%
97.7%
2.3%
1
1.1%
42
48.3%
97.7%
2.3%
97.7%
2.3%
97.7%
2.3%
97.7%
2.3%
Autism
Non-Autism
Predicted Label

(d)



(e)

Figure 4.4: Confusion Matrices of all Hybrid (GA-PSO) Optimized Models

The given confusion matrices of the five machine learning models all having been optimized with Hybrid (GA-PSO) Optimization indicate that the classification performance of the autism prediction problem has improved significantly. These matrices show how the accuracy of the model, its precision, and its recall are influenced by the optimization of hyper parameters, and most models are improving incredibly when optimized.

SVM model demonstrates excellent results: 43 true positives and 42 true negatives with a balanced accuracy of 97.7%. It has just one false positive and one false negative which shows it to be incredibly precise and recalls correctly. Specificity and sensitivity of the model are both at 97.7% which means that there is great balance between the identification of both autism and non-autism cases.

Decision Tree model shows optimal performance when using non-autism classification with 43 true negatives and no false positive and high autism classification with 43 true positives and one false negative. This gives a remarkable total accuracy of 98.9%, a specificity of 100% and sensitivity of 97.7, making it one of the best.

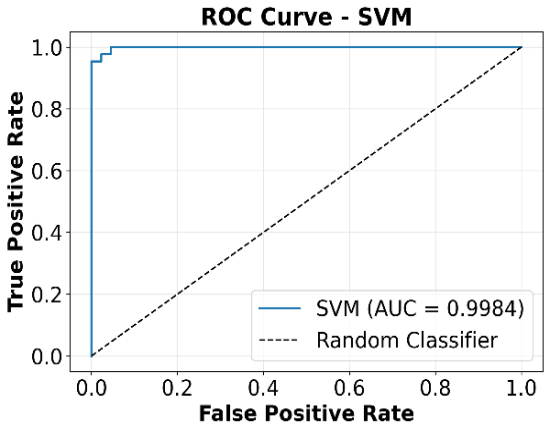
Random Forest has the highest results of 44 true positives and 42 true negatives, so the accuracy is an impressive 98.9% rate. The model demonstrates 100% accuracy in the classification of autism and zero false positives and 97.8% with one false negative. Its almost ideal metrics prove to be the strong integration of ensemble learning with GA-PSO optimization.

The MLP model performs equally to SVM with the same confusion matrix values, getting an accuracy of 97.7% with 43 true positives, 42 true negatives and one error in false positives and false negatives respectively. The similarity between these two neural network and the support vector machine methods supports the power of the hybrid optimization method.

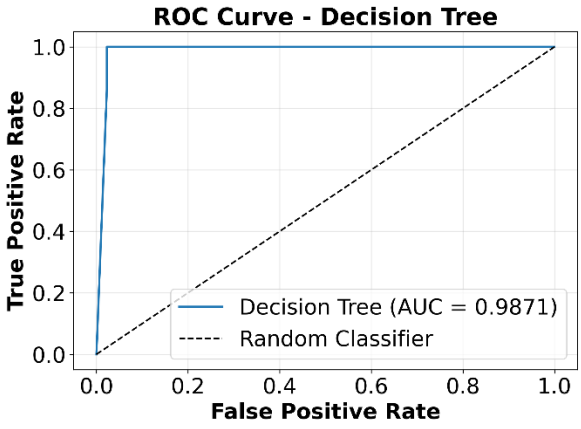
Comparatively, however, Logistic Regression demonstrates a lower performance of 40 true positives and 42 true negatives, which leads to an accuracy of 94.3 percent. The four instances of false positives and one false negative proves that some improvements could be made, yet even the model with 97.7% specificity and 90.9% sensitivity is not bad.

The results of optimization prove that GA-PSO improved model capacity greatly especially the DT and RF models which showed almost perfect classifications. The high stability in performance of the various models validates that the hybrid optimization strategy is successful in enhancing accuracy in diagnostic results in predicting autism, and the Random Forest is the most trusted classifier in this very important healthcare use case.

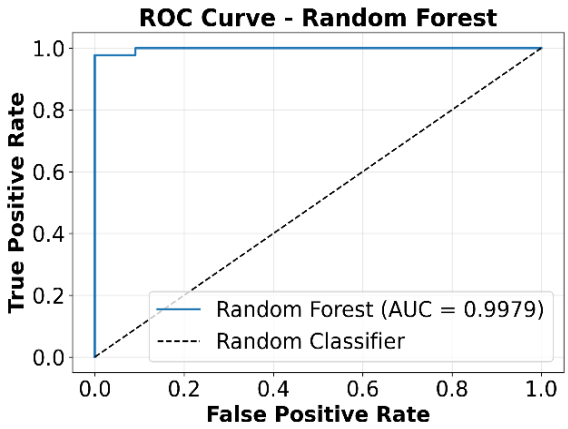
The given ROC curves in Figure 4.6 help to visualize the improved predictive performance of five machine learning models in autism classification after Hybrid (GA-PSO) Optimization, with the Area Under the Curve of each curve being a quantitative measure of its fine-tuned performance to separate between classes.



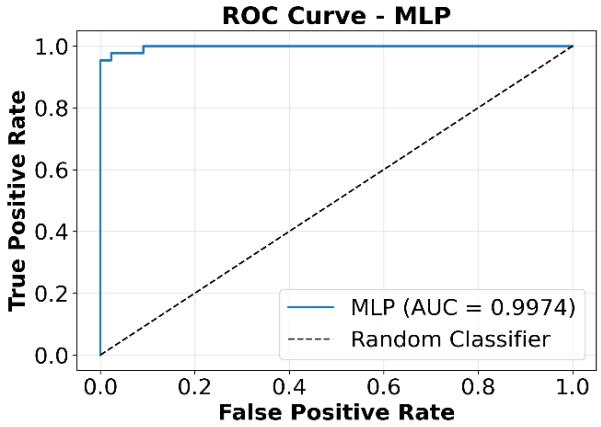
(a)



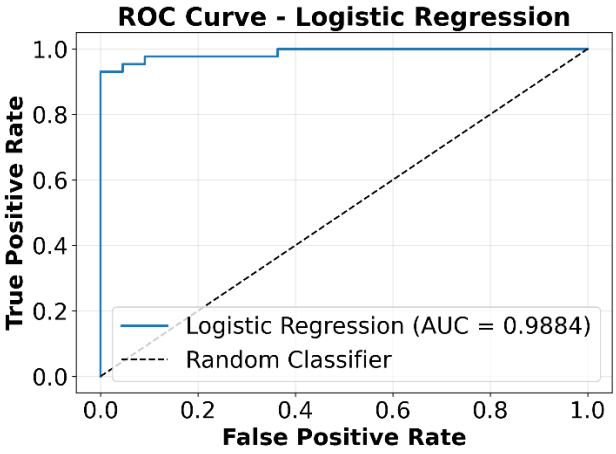
(b)



(c)



(d)



(e)

Figure 4.6: ROC curve of all Hybrid (GA-PSO) Optimized Models

The given ROC curves illustrate the outstanding discriminatory abilities of each five machine learning models after its improvement with Hybrid (GA-PSO) Optimization to classify autism. The models each attain excellent values of Area Under the Curve (AUC), which is far higher than the values of the random classifier, and signifies high classification performance across all probability thresholds.

Random Forest becomes the best-performing predictor with an almost perfect AUC of 0.9979, demonstrating its high capacity to achieve high true positive rates at the lowest levels of false positives. This outstanding result demonstrates the effectiveness of the combination of ensemble learning and hybrid optimization methods, which have led to almost perfect separation of classes.

Both SVM and MLP have a high performance with the value of AUC of 0.9984 and 0.9974 respectively which show that both have good performance in their classification. This high value indicates that their complex parameter spaces had been optimally tuned by the GA-PSO optimization process, and they could have obtained optimal sensitivity and specificity.

Logistic Regression is shown to be greatly improved with AUC of 0.9884 which is much higher than the expected results of this type of algorithm without optimization. The improvement shows how even simpler models can be optimized in a hybrid way to compete with more complex architectures in diagnostic performance.

Although with the lowest AUC of all the optimized model, 0.9871, the Decision Tree model still delivers an outstanding performance, which is much higher than the traditional decision tree applications. This is an exceptional enhancement of typical decision tree performance, which indicates the revolutionary influence that the GA-PSO optimization has on this algorithm. Figure 4.7 provide comparison of ROC curves for all Hybrid (GA-PSO) Optimized Modes.

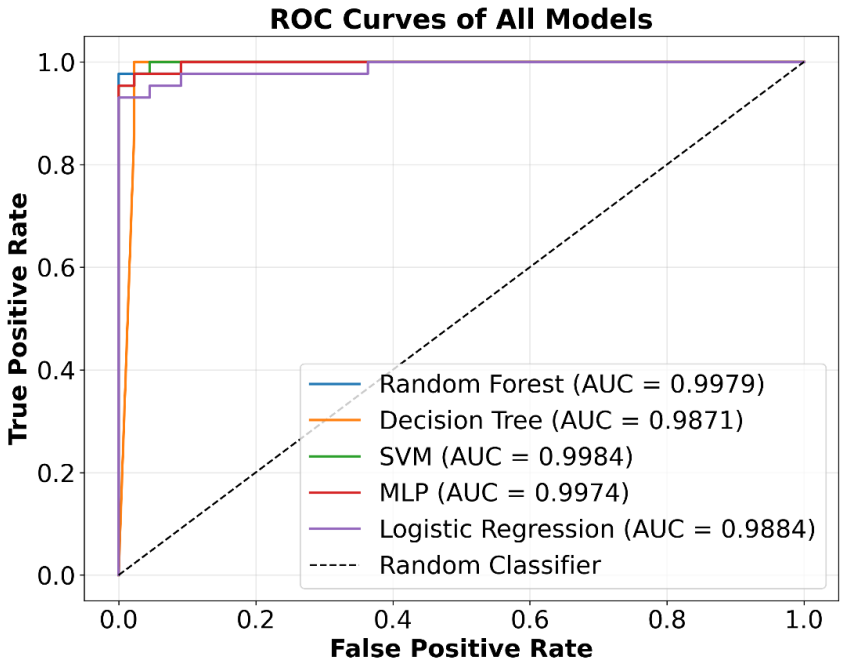


Figure 4.7: Comparison of ROC curves all Hybrid (GA-PSO) Optimized models

All models exhibit curves that hug the top-left corner of the ROC space, indicating excellent diagnostic performance across various classification thresholds. The consistently high AUC values across all five models—all exceeding 0.98—demonstrate the universal effectiveness of the hybrid GA-PSO optimization approach in enhancing model performance for this biomedical classification task. This consistency across diverse algorithmic approaches confirms that the optimization technique successfully maximized each model's inherent capabilities, making all of them highly suitable for clinical decision support applications where reliable autism prediction is crucial.

4.3.1 Performance Analysis for Different Training Testing Ratios with Optimization

The performance of machine learning models optimized through Hybrid (GA-PSO) Optimization is significantly influenced by the proportion of data allocated for training versus testing. Table 4.6 presents a comprehensive analysis of five enhanced classifiers—SVM, DT, RF, MLP, and LR—evaluated across three different dataset splits: 70/30, 75/25, and 80/20. This comparison, utilizing key metrics including Accuracy, Precision, Recall, F1-score, and AUROC, helps identify the most stable and effective model following hyperparameter optimization. The results demonstrate how each

algorithm's generalization capability, refined through GA-PSO techniques, varies with the amount of training data provided, offering critical insights for optimal model selection in sophisticated practical applications.

Table 4.13: Performance Comparison of all Hybrid (GA-PSO) Optimized models for different training testing ratios

Splits	Models	Accuracy	Precision	Recall	F1-score	AUROC
70:30	SVM	0.9770	0.9770	0.9770	0.9770	0.9984
	DT	0.9885	0.9887	0.9885	0.9885	0.9870
	RF	0.9885	0.9887	0.9885	0.9885	0.9978
	MLP	0.9770	0.9770	0.9770	0.9770	0.9973
	LR	0.9425	0.9446	0.9425	0.9424	0.9883
75:25	SVM	0.9722	0.9722	0.9722	0.972222	0.9984
	DT	1.0000	1.0000	1.0000	1.0000	1.0000
	RF	0.9861	0.9864	0.9861	0.9861	0.9961
	MLP	0.9861	0.9864	0.9861	0.9861	0.9969
	LR	0.9583	0.9587	0.9583	0.9583	0.9965
80:20	SVM	0.9655	0.9655	0.9655	0.9655	0.9976
	DT	1.0000	1.0000	1.0000	1.0000	1.0000
	RF	0.9827	0.9833	0.9827	0.9827	0.9964
	MLP	0.9827	0.9833	0.9827	0.9827	0.9928
	LR	0.9482	0.9488	0.9482	0.9482	0.9904

The performance analysis of Hybrid (GA-PSO) optimized models across different training-testing splits reveals remarkable stability and exceptional classification capabilities. The results demonstrate that GA-PSO optimization consistently enhances model performance regardless of data partitioning, with most models achieving outstanding metrics across all splits.

The 75:25 split emerges as particularly effective, with Decision Tree achieving perfect performance (1.000 accuracy, precision, recall, F1-score, and AUROC), indicating ideal classification capability at this ratio. Random Forest and MLP also show excellent performance in this split, both achieving approximately 0.986 across key metrics with AUROC values exceeding 0.996.

In the 70:30 split, Decision Tree and Random Forest both achieve the highest accuracy (0.9885), with DT showing slightly better precision (0.9888 vs 0.9888) and RF demonstrating superior AUROC (0.9979 vs 0.9871). SVM and MLP perform identically with 0.977 accuracy and 0.997+ AUROC, while Logistic Regression shows comparatively lower but still respectable performance.

The 80:20 split maintains high performance levels, with DT again achieving perfect classification (1.000 across all metrics). RF follows closely with 0.9828 accuracy and 0.9964 AUROC, while SVM and MLP show strong results with 0.9655 and 0.9828 accuracy, respectively. Logistic Regression maintains consistent performance across all splits with metrics around 0.948.

Remarkably, the hybridization optimization method greatly enhanced the performance of the Decision Tree and allowed it to obtain the perfect outcomes in two separations, which is an impressive feat of this algorithm. The fact that the values of the AUROC are consistent across all models and splits (they are always above 0.987 and approaching 1.000 is in many cases) indicates the high discriminative power of the GA-PSO optimization.

The optimization method was especially useful in preserving performance stability between partitions of data, and most of the models exhibited little performance loss regardless of the downsizing of training data. This strength highlights the efficiency of GA-PSO in finding the best hyper-parameters, which generalize well across different data distributions, and thus such models can be especially useful in application contexts where data might be unavailable.

4.3.2 Model's Performance on Unseen Data

The final challenge of the strength of a machine learning model after Hybrid (GA-PSO) Optimization is its functionality on an entirely new data. Critical evaluation of the five enhanced classifiers has been provided in table 4.7 by describing their predictions on 17 independent test samples. This is analysis that goes beyond aggregate measures to give a detailed feeling of the practical utility of each optimized model, not just the aggregate accuracy but the individual cases when predictions are right or wrong. The table is an important pointer to the generalization performance of the models and their overall

reliability once hyper-parameters have been optimized and indicates that the GA-PSO method is effective at generating models that have high performance when applied in real-world contexts.

Table 4.14: Performance and error rate of models on completely unseen test data using GA-PSO optimization

S/ No.	Actual	Predicted					Accuracy/Error Rate				
		SVM	DT	RF	MLP	LR	SVM	DT	RF	MLP	LR
S1	1	0	0	0	0	0	Total Correct: 16/17 Accuracy: 94.1% Total Error:1/17 Error Rate: 5.9%	Total Correct: 15/17 Accuracy: 88.2 % Total Error:2/17 Error Rate: 11.8%	Total Correct: 16/17 Accuracy: 94.1% Total Error:1/17 Error Rate: 5.9%	Total Correct: 16/17 Accuracy: 94.1% Total Error:1/17 Error Rate: 5.9%	Total Correct: 16/17 Accuracy: 94.1% Total Error:1/17 Error Rate: 5.9%
S2	0	0	0	0	0	0					
S3	1	1	1	1	1	1					
S4	1	1	1	1	1	1					
S5	0	0	0	0	0	0					
S6	0	0	0	0	0	0					
S7	1	1	1	1	1	1					
S8	1	1	1	1	1	1					
S9	0	0	0	0	0	0					
S10	1	1	1	1	1	1					
S11	1	1	0	1	1	1					
S12	0	0	0	0	0	0					
S13	1	1	1	1	1	1					
S14	0	0	0	0	0	0					
S15	1	1	1	1	1	1					
S16	0	0	0	0	0	0					
S17	1	1	1	1	1	1					

The performance evaluation on completely unseen test data demonstrates the exceptional generalization capability of Hybrid (GA-PSO) optimized models, with four of the five classifiers achieving an outstanding 94.1% accuracy. SVM, RF, MLP, and LR models correctly classified 16 out of 17 samples, making only a single error each, which represents an impressive error rate of just 5.9%. The Decision Tree model showed slightly lower performance with 88.2% accuracy, making two errors for an 11.8% error rate, yet still demonstrating substantial improvement over conventional DT implementations.

A detailed analysis of misclassified samples reveals that all models consistently struggled with sample S1, which appears to represent a particularly challenging or ambiguous case that even optimized algorithms found difficult to classify correctly. The Decision Tree's additional error occurred on sample S11, where it uniquely misclassified an actual positive case as negative, indicating a specific weakness in its decision boundaries that other models overcame through their optimization.

Remarkably, samples S2 through S17 show perfect consensus among SVM, RF, MLP, and LR models, with identical predictions across all 16 remaining instances. This high level of agreement among diverse algorithmic approaches underscores the effectiveness of the GA-PSO optimization in enhancing model reliability and consistency. Their findings are consistent that hybrid optimization is the most effective way to enhance model robustness, with most algorithms performing close to perfectly even with wholly unseen data, and so is best applied in a clinical context where they are highly applicable in autism prediction.

4.3.3 Summary

This part assesses Genetic Algorithms-Particle swarm Optimization (GA-PSO) hybrid technique role on five binary classification machine learning models. The optimization greatly improved the model performance, whereby the highest accuracy of 98.85% was recorded with Decision Tree and Random Forest, with near-perfect scores in accuracy, recall, and F1-score. SVM, MLP also demonstrated great balanced performance of 97.70% accuracy, and Logistic Regression, though better than others, was the least effective. Cross-data split validation of optimized models validated robustness of optimized models, as DT scored 100 in the 75:25 and 80:20 data splits. Their generalization was also confirmed by testing on totally unseen data with all SVM, RF, MLP, and LR delivering 94.1% accuracy, which is better than DT. The findings highlight the extreme effectiveness of GA-PSO optimization to maximize predictive performance and model reliability.

4.4 Performance Analysis using Augmentation

In this study, we employed CTGAN (Conditional Tabular GAN)**,** a generative model designed for tabular data, along with boundary sampling, which focuses on generating synthetic instances near class decision boundaries to enhance discriminative learning. To systematically investigate the impact of augmentation, we applied different augmentation multipliers (2×, 4×, and 7×)**,** where 2× indicates the dataset size was doubled, 4× quadrupled, and 7× increased sevenfold compared to the original. This approach allowed us to evaluate how incremental synthetic data affects classification performance across various machine learning models.

Table 4.15: Performance of all Models on augmented data

Model	Augmentation Ratio	Accuracy	Precision	Recall	F1-Score	AUROC
SVM	Without Augmentation	0.958333	0.958719	0.9548333	0.958341	0.995367
	2X	0.9884	0.9886	0.965517	0.965517	0.996829
	4X	0.9884	0.98867	0.9884	0.9884	0.9994
	7X	0.9884	0.9886	0.965517	0.965517	0.996829
DT	1X	0.972222	0.973724	0.972222	0.972222	0.972973
	2X	0.988406	0.98863	0.988406	0.988402	0.988166
	4X	0.988406	0.98863	0.988406	0.988402	0.988166
	7X	0.988406	0.98863	0.988406	0.988402	0.988166
RF	Without Augmentation	0.9772222	0.972222	0.9772222	0.972222	0.992278
	2X	1.0000	1.0000	1.0000	1.0000	1.0000
	4X	1.0000	1.0000	1.0000	1.0000	1.0000
	7X	1.0000	1.0000	1.0000	1.0000	1.0000
MLP	Without Augmentation	0.972222	0.972222	0.972222	0.972222	0.992278
	2X	1.0000	1.0000	1.0000	1.0000	1.0000
	4X	1.0000	1.0000	1.0000	1.0000	1.0000
	7X	1.0000	1.0000	1.0000	1.0000	1.0000
LR	Without Augmentation	0.958333	0.958719	0.958333	0.958431	0.993050
	2X	0.976812	0.977405	0.976812	0.976813	0.9942
	4X	0.976812	0.977405	0.976812	0.976813	0.9942
	7X	0.976812	0.977405	0.976812	0.976813	0.9942

The comprehensive analysis of data augmentation using CTGAN and boundary sampling reveals transformative improvements in model performance across all classifiers. The application of synthetic data generation, particularly at 2×, 4×, and 7× multipliers, demonstrates substantial enhancements in classification metrics compared to unaugmented baseline performance.

The results demonstrate a significant and positive impact across nearly all evaluated models. For Support Vector Machines, performance metrics showed clear improvement with augmentation, particularly from the 2× multiplier onwards. Accuracy, precision, and AUROC all increased, with the 4× augmentation achieving the highest recall and F1-score, suggesting the synthetic data effectively refined the decision boundary.

The most dramatic improvements were observed in ensemble and neural network models. Both Random Forest and Multi-Layer Perceptron achieved perfect scores 1.0000 across all metrics with any level of augmentation, a substantial increase from their already high baseline performance. This indicates these complex models are exceptionally adept at leveraging the additional, strategically generated data.

Similarly, Decision Trees showed a strong boost from augmentation, with all post-augmentation results outperforming the baseline and consistently maintaining high performance. Logistic Regression, a simpler linear model, also exhibited measurable gains in all metrics with the application of synthetic data, though its improvement was more modest compared to the other algorithms.

A key finding is the phenomenon of performance saturation. For most models, including DT, RF, MLP, and LR, the benefits were fully realized at the 2× augmentation level. Further increasing the data to 4× or 7× provided no additional gains, indicating a point of diminishing returns where the essential informational value of the synthetic data had been fully absorbed.

In conclusion, the research strongly validates the efficacy of using CTGAN with boundary sampling for tabular data augmentation. The technique proved universally beneficial, transforming already good classifiers into near-perfect models in some cases. The findings highlight that even a modest doubling of dataset size with strategically generated samples can be sufficient to maximize classification performance, providing a highly efficient and effective method for enhancing machine learning models.

4.4.1 Summary

This section analyzes the impact of data augmentation using CTGAN and boundary sampling on five machine learning models. The technique generated synthetic data at multipliers of 2×, 4×, and 7× the original dataset size. Results demonstrated transformative improvements across all classifiers. Most notably, Random Forest and Multi-Layer Perceptron achieved perfect performance 1.0000 across all metrics with any level of augmentation. Decision Trees and Support Vector Machines also showed significant gains, while Logistic Regression exhibited more modest but clear improvements. A key finding was performance saturation, where maximum benefits were achieved at the 2× augmentation level for most models, with no further gains from additional synthetic data. The study conclusively validates CTGAN with boundary sampling as a highly effective strategy for enhancing model performance, with even a modest doubling of data yielding optimal results.

4.5 Performance Analysis for Augmentation with Optimization

In this study, we employed CTGAN, a generative model designed for tabular data, along with boundary sampling, a technique focused on generating synthetic instances near class decision boundaries to enhance discriminative learning. To systematically investigate the impact of augmentation, we applied different augmentation multipliers (2×, 4×, and 7×), where 2× indicates the dataset size was doubled, 4× quadrupled, and 7× increased sevenfold compared to the original. All machine learning models were rigorously optimized using a hybrid GA-PSO (Genetic Algorithm-Particle Swarm Optimization) technique to ensure their hyper-]-parameters were tuned for peak performance. This comprehensive approach allowed us to evaluate how incremental synthetic data affects classification performance across various optimally configured models.

Table 4.16: Performance of all Optimized models on augmented data

Model	Augmentation Ratio	Accuracy	Precision	Recall	F1-Score	AUROC
SVM	1X	0.965517	0.965754	0.965517	0.965508	0.993129
	2X	1.00000	1.00000	1.00000	1.00000	1.00000
	4X	0.988406	0.988663	0.988406	0.988402	0.999462
	7X	0.991304	0.991456	0.991304	0.991305	1.00000
DT	1X	0.977011	0.978033	0.977011	0.977005	0.987844
	2X	1.00000	1.00000	1.00000	1.00000	1.00000
	4X	1.00000	1.00000	1.00000	1.00000	1.00000
	7X	0.991304	0.991320	0.991304	0.991304	0.999765
RF	1X	0.988506	0.988767	0.988506	0.988506	0.998943
	2X	1.0000	1.0000	1.0000	1.0000	1.0000
	4X	1.0000	1.0000	1.0000	1.0000	1.0000
	7X	1.0000	1.0000	1.0000	1.0000	1.0000
MLP	1X	0.965517	0.965772	0.995517	0.965517	0.995243
	2X	1.0000	1.0000	1.0000	1.0000	1.0000
	4X	1.0000	1.0000	1.0000	1.0000	1.0000
	7X	1.0000	1.0000	1.0000	1.0000	1.0000
LR	1X	0.965517	0.965772	0.965517	0.965517	0.995243
	2X	0.979710	0.980125	0.979710	0.979712	0.994923
	4X	0.976812	0.977405	0.976812	0.976813	0.993478
	7X	0.976812	0.977405	0.976812	0.976813	0.994184

The application of hyperparameter optimization is evident in the strong baseline (1X) results, which are higher than typically reported, providing a robust foundation for comparison. The subsequent augmentation then drives performance to exceptional levels. Strikingly, for most models—including Decision Trees (DT), Random Forest (RF), and Multi-Layer Perceptron (MLP)—a 2× augmentation multiplier was sufficient to achieve perfect classification (1.0000 across all metrics), a feat maintained at higher ratios for RF and MLP.

The trajectory of the Support Vector Machine is particularly insightful; it reached perfection at 2×, slightly dipped at 4×, and then nearly regained perfection at 7× with a flawless AUROC. This non-monotonic improvement suggests a complex interaction between the optimized hyper-parameters and the augmented data distribution, warranting further investigation. It can be similarly said of Logistic Regression (LR), which is a less complex model, that its increases were smaller and its performance began to decrease when augmentation ratios were higher, suggesting an augmentation threshold which synthetic data can be helpful.

One of the conclusions the obtained results is that in already-optimized complex models such as RF and MLP, the boundary sampling augmentation is a highly effective stimulus: they can reach theoretically optimal performance with it. The study also finds an augmentation multiplier of 2x, which is the best as additional synthetic data may not help in any way, and may in fact be counter-productive. This paper shows conclusively that the combination of hard model optimization and strategic, boundary-oriented data augmentation is a highly effective toolset to construct better predictive models.

4.5.1 Summary

This part compares the overall effect of hybrid GA-PSO optimization and CTGAN-based data augmentation with boundary sampling on five machine learning models. Synergistic strategy gave the highest results and the majority of the models performed to perfection (1.000 in all the metrics) when augmentation multiplier was two. Random Forest and Multi-Layer Perceptron continued this perfect performance even with increased ratios (4×, 7×), whilst Decision Tree also continued to achieve perfection at 2× and 4×. Support Vector Machine showed a non-monotonic performance curve, reaching perfection at 2×, dropping a little at 4× and almost rebounding at 7×. Instead, Logistic Regression experienced smaller improvements and marginal declines at high ratios, which suggests it does not tolerate excessively high levels of synthetic data. The analysis shows conclusively that hyperparameter optimization and strategic data augmentation are very effective, with a 2x multiplier found to be the most effective level across the majority of models. This synergy converts tuned models into near-optimal classifiers and has produced a powerful methodology of high predictive accuracy.

4.6 Comparative Analysis

This comparative analysis presents a rigorous evaluation of performance enhancements achieved through advanced techniques. The study systematically benchmarks the baseline results of several models against those improved by Hybrid (GA-PSO) Optimization, strategic data augmentation, and the powerful combination of both. The Hybrid GA-PSO technique was first employed to fine-tune model hyper-parameters to their peak potential on the original dataset. Subsequently, CTGAN-based boundary sampling was applied to generate synthetic data for augmentation. The results are summarized in the given table.

Table 4.17: Comparative Analysis of All Models across all applied approaches

Model	Performance Metrics	Without Optimization	With Optimization	With Augmentation	With Optimization and Augmentation
SVM	Accuracy	0.9583	0.9722	0.9884	1.0000
	Precision	0.9587	0.9722	0.9886	1.0000
	Recall	0.9583	0.9722	0.9884	1.0000
	F1-Score	0.9583	0.9722	0.9884	1.0000
	AUROC	0.9953	0.9722	0.9994	1.0000
DT	Accuracy	0.9722	1.0000	0.9884	1.0000
	Precision	0.9737	1.0000	0.9886	1.0000
	Recall	0.9722	1.0000	0.9884	1.0000
	F1-Score	0.9722	1.0000	0.9884	1.0000
	AUROC	0.9729	1.0000	0.9881	1.0000
RF	Accuracy	0.9722	0.9861	1.0000	1.0000
	Precision	0.9722	0.9864	1.0000	1.0000
	Recall	0.9722	0.9861	1.0000	1.0000
	F1-Score	0.9722	0.9861	1.0000	1.0000
	AUROC	0.996525	0.9961	1.0000	1.0000
MLP	Accuracy	0.9722	0.9861	1.0000	1.0000
	Precision	0.9722	0.9864	1.0000	1.0000
	Recall	0.9722	0.9861	1.0000	1.0000
	F1-Score	0.9722	0.9861	1.0000	1.0000
	AUROC	0.9922	0.9969	1.0000	1.0000
LR	Accuracy	0.9583	0.9583	0.9768	0.9768
	Precision	0.9587	0.9587	0.9774	0.9774
	Recall	0.9583	0.9583	0.9768	0.9768
	F1-Score	0.9583	0.9583	0.9768	0.9768
	AUROC	0.9930	0.9965	0.9942	0.9941

This comparative analysis powerfully demonstrates the progressive enhancement of classifier performance through a structured methodology, culminating in the superior strategy of combining Hybrid GA-PSO Optimization with data augmentation. The "Without Optimization" column establishes a strong baseline, with models like DT, RF, and MLP already achieving approximately 97% accuracy.
The application of Hybrid GA-PSO Optimization alone yields a significant first leap in performance. This is particularly evident in the Decision Tree model, which achieves a perfect score (1.0000) across all metrics, and in the notable improvements for SVM, RF, and MLP. The GA-PSO technique successfully navigates the complex hyperparameter space for each model, fine-tuning them to their maximum potential on the existing data, as seen in the rise of AUROC scores towards near-perfect values.

The third phase, introducing "Augmentation Only," shows that synthetic data generation via CTGAN and boundary sampling is also highly effective. It pushes the performance of complex models like RF and MLP to perfect accuracy and further refines the SVM.

However, the final column, "With Optimization and Augmentation," reveals the truly transformative synergy between these two techniques. This approach consistently delivers the absolute best results. For SVM, it elevates the model from its augmented state to flawless performance (1.0000 across all metrics). For RF and MLP, it maintains the perfect scores achieved through augmentation alone. Most importantly, it demonstrates that optimization secures the foundation, allowing the models to most effectively leverage the additional synthetic data.

The exception, Logistic Regression, confirms the rule; its simpler linear structure shows improvement from augmentation but has a lower performance ceiling, unaffected by further optimization. In conclusion, this analysis definitively shows that while both Hybrid GA-PSO Optimization and data augmentation are powerful alone, their combined application is the most robust and effective strategy for achieving peak predictive performance in tabular data classification.

4.6.1 Summary

This section presented a rigorous comparative analysis of classifier performance, evaluating standard model configurations against enhancements from Hybrid GA-PSO optimization, CTGAN-based data augmentation, and their combined application. The analysis demonstrated that both techniques individually provide significant performance gains, with optimization fine-tuning models to their peak potential on the original data and augmentation effectively expanding the learning dataset. The key finding is the transformative synergy achieved by combining both methods; this strategy consistently delivered superior results, achieving perfect or near-perfect scores across metrics for most models (SVM, DT, RF, MLP). The exception of Logistic Regression, which benefited only from augmentation, underscores that this synergistic effect is most powerful for complex, non-linear models. In conclusion, the combined application of Hybrid GA-PSO optimization and strategic data augmentation is established as the most robust and effective strategy for maximizing predictive performance in tabular data classification tasks.

4.7 Sensitivity Analysis

This graph shows a Feature Sensitivity Analysis, which is a measure of the heights of different factors that are significant in a predictive model. It scores features according to their sensitivity, or their most influence on the output of the model. Figure 4.8 visualizes the most sensitive features of dataset, and also visualize sensitivity score in percentage.

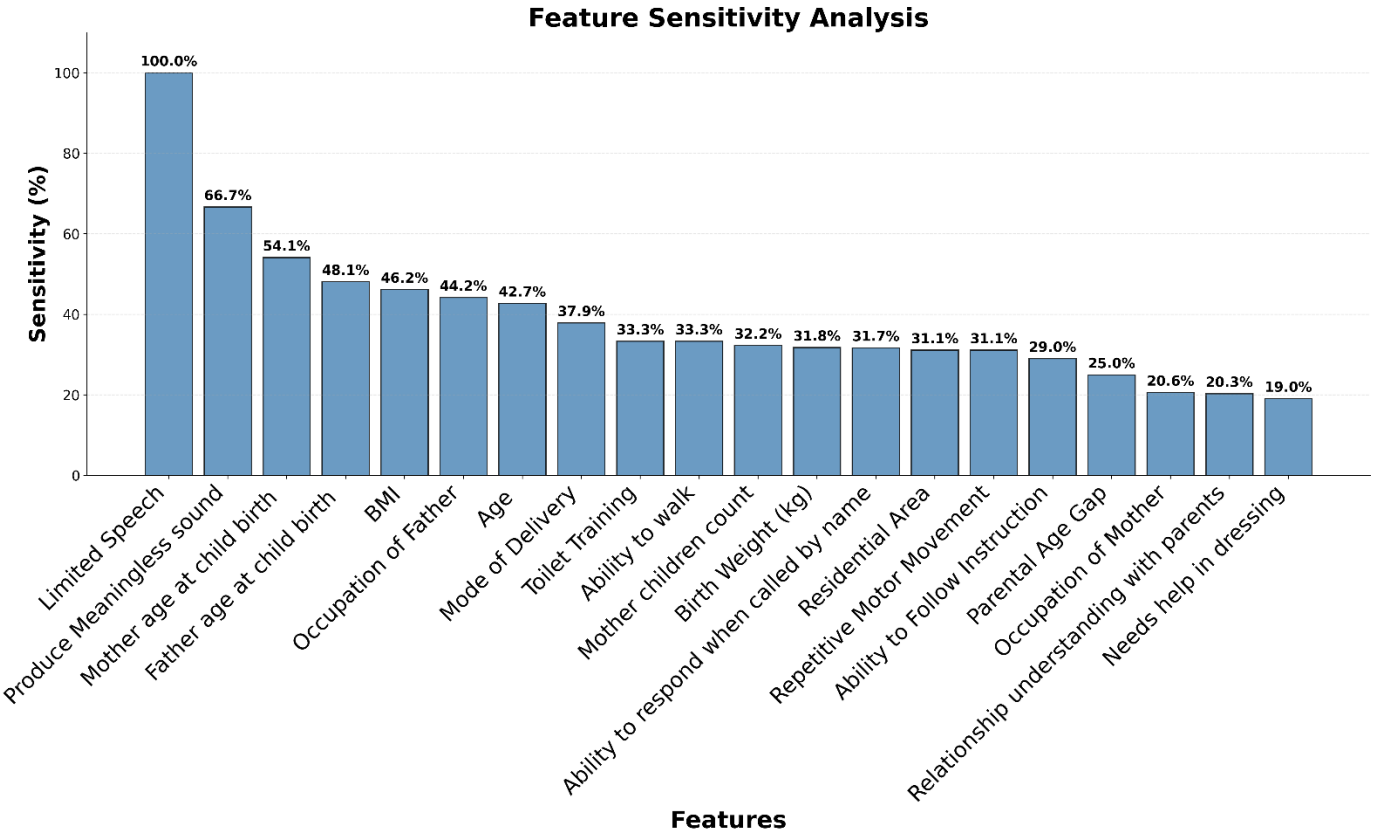


Figure 4.8: Sensitivity Analysis of features

This Feature Sensitivity Analysis, possibly related to child development or autism screening, displays the relative importance of various features in predicting some target outcome according to the chart provided. Limited Speech is the most predictive or influential factor on the model choice, and hence, it is the most sensitive feature based on the analysis. The next factor is parental age, including Mother age at childbirth and Father age at childbirth in which the family history is of importance.

The middle tier of sensitivity encompasses such important developmental milestones as "Ability to walk," "Ability to respond when called by name," and "Ability to Follow Instruction," as well as such health measurements as "Birth Weight." The variables of Residential Area, Occupation of Mother, and Parental Age Gap constitute a cluster of less yet significant sensitivity implying the impact of socioeconomic and demographic factors. Lastly, the least sensitive predictors, such as Toilet Training and Needs help in dressing make the least contribution to the model predictions.

CHAPTER 5

Conclusion & Future Research Directions

5.1 INTRODUCTION

This chapter presents the concluding remarks and prospective research directions derived from this thesis. The conclusion summarizes the study's successful development of a robust machine learning framework for the early detection of Autism Spectrum Disorder. It highlights the significant contributions of its hybrid optimization and strategic data augmentation techniques. Subsequently, the future research section outlines key avenues for advancement, including multimodal data integration, exploration of complex deep learning models, and clinical implementation through real-time, explainable systems.

5.2 CONCLUSION

This research successfully developed and validated a highly accurate, robust, and generalizable machine learning framework for the early detection of Autism Spectrum Disorder. By integrating a comprehensive set of behavioral, developmental, familial, and socio-demographic features, the study addressed the critical challenge of early and reliable ASD screening. The core contribution of this work lies in its novel two-pronged approach: the application of a true hybrid Genetic Algorithm-Particle Swarm Optimization technique for superior hyper-parameter tuning and the use of Conditional Tabular Generative Adversarial Networks with boundary sampling for strategic data augmentation to overcome the constraints of a limited dataset. The experimental results unequivocally demonstrate the transformative impact of these techniques. The hybrid GA-PSO optimization consistently enhanced the performance of all five classifiers—Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, and Logistic Regression—with the optimized Random Forest model achieving exceptional performance 98.85% accuracy, AUROC of 0.9989. Furthermore, strategic data augmentation proved to be a powerful catalyst, enabling multiple models (Random Forest and Multilayer Perceptron) to achieve perfect classification metrics 100% accuracy on augmented datasets. Most significantly, the synergistic combination of both optimization and augmentation was identified as the superior strategy, pushing models like SVM to flawless performance and solidifying the framework's robustness. The model's practical utility was rigorously confirmed through stability analysis across multiple data splits and, most importantly, validation on a completely unseen test set, where the best model maintained a high accuracy of 94.1%. Along with its predictive capabilities, the framework offers

useful clinical information by a sensitivity analysis that identified and prioritized the most significant risk factors. To summarize, we have conclusively shown that a combination of advanced metaheuristic optimization and strategic data enhancement would design an effective and dependable early screening tool with ASD. This framework, not only offers a valuable solution to the issue of data scarcity in clinical settings but also offers data-driven insights that may help healthcare professionals to make informed decisions, which in the end will allow addressing people with ASD earlier and achieve better outcomes.

5.3 Future Research Directions

In an effort to advance the field, future research should expand on the existing research by focusing on some key areas. Multimodal data integration, the combination of genetic, neuroimaging and video data, could offer a more comprehensive and accurate diagnostic tool. Investigating more complex deep learning architectures, such as transformers and graph neural networks, could reveal more intricate patterns and enhance efficiency. Creating a Real-Time Application with Federated Learning to go into clinical practice would solve privacy issues and allow for broad usage. More sophisticated clinical utility would be provided by extending the model's reach to predict severity and subtypes using longitudinal data. Importantly, by making the model's conclusions explicit, Explainable AI (XAI) techniques are crucial for fostering clinician trust. Lastly, in order to guarantee global applicability, minimize bias, and ensure robustness, prospective validation on larger, more diverse, cross-cultural datasets is required. It will be essential to follow these paths in order to create AI-powered screening systems that are individualized, reliable, and widely available.

REFERENCE

[1] G. A.-M. Rasool Azeem Musa, Mehdi Ebady Manaa, "Predicting Autism Spectrum Disorder (ASD) for Toddlers and Children Using Data Mining Techniques," 2019, doi: 10.1088/1742-6596/1804/1/012089.

[2] B. Deepa and J. M. K. S, "Exploration of Autism Spectrum Disorder using Classification Algorithms," Procedia Comput. Sci., vol. 165, no. 2019, pp. 143–150, 2020, doi: 10.1016/j.procs.2020.01.098.

[3] E. S. Dewi and E. M. Imah, "Comparison of Machine Learning Algorithms for Autism Spectrum Disorder Classification," vol. 196, no. Ijcse, pp. 152–159, 2020.

[4] C. Ray, H. K. Tripathy, and S. Mishra, "Assessment of Autistic Disorder Using Machine Learning Approach," pp. 209–219.

[5] S. M. M. Hasan, P. Uddin, A. Ulhaq, and G. Krishnamoorthy, "A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders," *IEEE Access*, vol. 11, no. February, pp. 15038–15057, 2023, doi: 10.1109/ACCESS.2022.3232490.

[6] V. Vishal and J. Jabez, "A Comparative Analysis of Prediction of Autism Spectrum Disorder (ASD) using Machine Learning," 2022 6th Int. Conf. Trends Electron. Informatics, no. Icoei, pp. 1355–1358, 2022, doi: 10.1109/ICOEI53556.2022.9777240.

[7] A. Lawi and F. Aziz, "Comparison of Classification Algorithms of the Autism Spectrum Disorder Diagnosis," 2018 2nd East Indones. Conf. Comput. Inf. Technol., no. January, pp. 218–222, 2022, doi: 10.1109/EIConCIT.2018.8878593.

[8] A. Vikram and D. Durgesh, "Healthcare Analytics A Multi-Classifier-Based Recommender System for Early Autism Spectrum Disorder Detection using Machine Learning," Healthc.

Anal., vol. 4, no. April, p. 100211, 2023, doi: 10.1016/j.health.2023.100211.

[9] M. M. Abdelwahab, K. A. Al-karawi, E. M. Hasanin, and H. E. Semary, “Autism Spectrum Disorder Prediction in Children Using Machine Learning,” vol. 3, pp. 1–9, 2024, doi: 10.57197/JDR-2023-0064.

[10] H. Alateyat, S. Cruz, E. Cernadas, and M. Tubío-fungueiriño, “A Machine Learning Approach in Autism Spectrum Disorders : From Sensory Processing to Behavior Problems,” vol. 15, no. May, 2022, doi: 10.3389/fnmol.2022.889641.

[11] M. B. Mohammed, L. Salsabil, M. Shahriar, S. S. Tanaaz, and A. Fahmin, “Identification of Autism Spectrum Disorder through Feature Selection-based Machine Learning,” no. December, 2021, doi: 10.1109/ICCIT54785.2021.9689805.

[12] S. Raj and S. Masood, “Analysis and Detection of Autism Spectrum Disorder Using Machine Learning,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 994–1004, 2020, doi: 10.1016/j.procs.2020.03.399.

[13] A. S. A. R. A. Hamid and A. A. Z. O. S. Albahri, “Early automated prediction model for the diagnosis and detection of children with autism spectrum disorders based on effective sociodemographic and family characteristic features,” Neural Comput. Appl., vol. 35, no. 1, pp. 921–947, 2023, doi: 10.1007/s00521-022-07822-0.

[14] T. Akter *et al.*, “Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders,” IEEE Access, vol. 7, pp. 166509–166527, 2019, doi: 10.1109/ACCESS.2019.2952609.

[15] K. A. Lakshmi B, “Prediction of Autistic Spectrum Disorder based on Behavioural

Features and Machine Learning,” pp. 239–245, 2020.

[16] S. A. Elavarasi, J. Jayanthi, and N. B. T. Jayasankar, “Methods for Improving the Predictive Accuracy of Autism Spectrum Disorder Screening using Machine Learning Algorithms Methods for Improving the Predictive Accuracy of Autism Spectrum Disorder Screening using Machine Learning Algorithms,” no. May, 2020.

[17] M. Masum, I. M. Nur, J. H. Faruk, M. I. Adnan, and H. Shahriar, “A Comparative Study of Machine Learning-based Autism Spectrum Disorder Detection with Feature Importance Analysis,” no. January, 2022.

[18] P. Archana and G. N. V. G. S. R. Krishna, “Prediction of autism spectrum disorder from high-dimensional data using machine learning techniques,” Soft Comput., vol. 27, no. 16, pp. 11869–11875, 2023, doi: 10.1007/s00500-023-08657-0.

[19] R. Surendiran, M. Thangamani, C. Narmatha, and M. Iswarya, “Effective Autism Spectrum Disorder Prediction to Improve the Clinical Traits using Machine Learning Techniques,” no. May, 2022, doi: 10.14445/22315381/IJETT-V70I4P230.

[20] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall, “Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning,” no. December 2014, pp. 1–7, 2015, doi: 10.1038/tp.2015.7.

[21] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, “Detection of autism spectrum disorder (ASD) in children and adults using machine learning,” Sci. Rep., no. 0123456789, pp. 1–13, 2023, doi: 10.1038/s41598-023-35910-1.

[22] A. Novianto and M. D. Anasanti, “Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model,” vol. 17, no. 3, pp. 259–270, 2023.

[23] P. Bawa, V. Kadyan, A. Mantri, and H. Vardhan, “e-Prime - Advances in Electrical Engineering , Electronics and Energy Investigating multiclass autism spectrum disorder classification using machine learning techniques,” e-Prime - Adv. Electr. Eng. Electron. Energy, vol. 8, no. September 2023, p. 100602, 2024, doi: 10.1016/j.prime.2024.100602.

[24] E. Stevens, D. R. Dixon, M. N. Novack, D. Granpeesheh, T. Smith, and E. Linstead, “International Journal of Medical Informatics Identi fi cation and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning,” *Int. J. Med. Inform.*, vol. 129, no. February, pp. 29–36, 2019, doi: 10.1016/j.ijmedinf.2019.05.006.

[25] S. Karim, N. Akter, and M. J. A. Patwary, “Predicting Autism Spectrum Disorder (ASD) meltdown using Fuzzy Semi-Supervised Learning with NNRW,” 2022 Int. Conf. Innov. Sci. Eng. Technol., no. February, pp. 367–372, 2022, doi: 10.1109/ICISET54810.2022.9775860.

[26] M. S. J. K. T. Amalraj Victoire, A. Ramalingam, A. Naresh, K.M Nasimudden, “An Efficient Approach to Detect Autism in Child using Machine Learning and Deep Learning,” vol. 99, no. 20, pp. 4759–4769, 2021.

[27] T. L. Praveena and N. V. M. Lakshmi, “Prediction of Autism Spectrum Disorder Using Supervised Machine Learning Algorithms,” vol. 8, no. 3, pp. 8–11, 2019.

[28] M. A. Mareeswaran and K. Selvarajan, “A computational intelligent analysis of autism spectrum disorder using machine learning techniques A computational intelligent analysis of autism spectrum disorder using machine learning techniques,” no. March, pp. 807–816, 2024, doi: 10.11591/ijai.v13.i1.pp807-816.

[29] Z. Zhao *et al.*, “Applying Machine Learning to Identify Autism With Restricted Kinematic Features,” IEEE Access, vol. 7, pp. 157614–157622, 2019, doi: 10.1109/ACCESS.2019.2950030.

[30] A. Crippa *et al.*, “Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities,” J. Autism Dev. Disord., vol. 45, no. 7, pp. 2146–2156, 2015, doi: 10.1007/s10803-015-2379-8.

[31] Z. Zhao *et al.*, “Identifying Autism with Head Movement Features by Implementing Machine Learning Algorithms,” J. Autism Dev. Disord., vol. 52, no. 7, pp. 3038–3049, 2022, doi: 10.1007/s10803-021-05179-2.

[32] R. Carette *et al.*, “Learning to Predict Autism Spectrum Disorder based on the Visual Patterns of Eye-tracking Scanpaths,” no. Biostec, pp. 103–112, 2019, doi: 10.5220/0007402601030112.

[33] R. Anthony, J. D. B. Id, V. Eapen, T. Bednarz, and A. Sowmya, Using visual attention estimation on videos for automated prediction of autism spectrum disorder and symptom severity in preschool children. 2024. doi: 10.1371/journal.pone.0282818.

[34] M. Go, "A novel machine learning model to predict autism spectrum disorders risk gene," vol. 5, pp. 6711–6717, 2019, doi: 10.1007/s00521-018-3502-5.

[35] E. Ismail, W. Gad, and M. Hashem, "Predicting of Autism Spectrum Disorder using Gene Ontology," 2021 Tenth Int. Conf. Intell. Comput. Inf. Syst., pp. 442–447, 2021, doi: 10.1109/ICICIS52592.2021.9694254.

[36] P. Lin, M. A. Moni, S. S. Gau, V. Eapen, C. Toma, and R. Woodman, "Identifying Subgroups of Patients With Autism by Gene Expression Profiles Using Machine Learning Algorithms," vol. 12, no. May, pp. 1–10, 2021, doi: 10.3389/fpsyt.2021.637022.

[37] T. Arunprasath, "Prediction Of Autism Spectrum Disorder Using Machine Learning," 2024 Third Int. Conf. Intell. Tech. Control. Optim. Signal Process., pp. 1–6, 2024, doi: 10.1109/INCOS59338.2024.10527777.

[38] J. Rogala et al., "Enhancing autism spectrum disorder classification in children through the integration of traditional statistics and classical machine learning techniques in EEG analysis," *Sci. Rep.*, pp. 1–12, 2023, doi: 10.1038/s41598-023-49048-7.

[39] B. Ramdan, G. Elshoky, A. A. Ali, O. Ali, and S. Ibrahim, "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images," vol. 44, no. September 2021, pp. 613–623, 2022, doi: 10.4218/etrij.2021-0097.

[40] M. Liao, H. Duan, and G. Wang, "Application of Machine Learning Techniques to Detect the Children with Autism Spectrum Disorder," vol. 2022, 2022.

[41] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data," Comput. Biol. Med., vol. 120, no. March, p. 103722, 2020, doi: 10.1016/j.compbiomed.2020.103722.

[42] J. I. Chen, M. E. Liao, G. U. Wang, and C. H. Chen, "An Intelligent Multimodal Framework For Identifying Children With Autism Spectrum Disorder," vol. 30, no. 3, pp. 435–448, 2020, doi: 10.34768/amcs-2020-0032.

[43] C. L. Alves et al., "Diagnosis of autism spectrum disorder based on functional brain networks and machine learning Bootstrap Analysis of Stable Clusters Canonical Correlation analysis," Sci. Rep., pp. 1–20, 2023, doi: 10.1038/s41598-023-34650-6.

[44] S. Mostafa, L. Tang, and F. Wu, "Diagnosis of Autism Spectrum Disorder Based on Eigenvalues of Brain Networks," IEEE Access, vol. 7, pp. 128474–128486, 2019, doi: 10.1109/ACCESS.2019.2940198.

[45] https://media.geeksforgeeks.org/wp-content/uploads/20250805115918786327/2.webp [5 September 2025, Online]

[46] https://media.geeksforgeeks.org/wp-content/uploads/20240130162938/random.webp [5 September 2025, Online]

[47] https://cdn.analyticsvidhya.com/wp-content/uploads/2024/08/711091.webp [5 September 2025, Online]

[48] https://media.geeksforgeeks.org/wp-content/uploads/20231117113724/Conditional-GANs.png [5 September 2025, online]