2026 Volume: 02 Issue: 06 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-30

AUTISMPREDICTION USING MACHINE LEARNING

*[1]Ayanchakraborty, [2]Mahuyasasmal

[1]Degree of Master of Technology in Computer Science and Engineering.

[2]Assistant Professor, Department of Computer Science and Engineering,

Haldia Institute of Technology ICARE Complex, Haldia, Purba Medinipur West Bengal.

Article Received: 12 May 2026

Article Revised: 02 June 2026

Published on: 22 June 2026

***Corresponding Author: Ayanchakraborty**

Degree of Master of Technology in Computer Science and Engineering, Haldia Institute of Technology ICARE Complex, Haldia, Purba Medinipur West Bengal.

DOI: https://doi-doi.org/101555/ijrpa.4881

ABSTRACT

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by persistent challenges insocialinteraction,communication,andrepetitive behavioralpatterns.The globalprevalenceof ASD has increased significantly over the past decade, making early detection and intervention critically important for improving long-term developmental outcomes. Traditional diagnostic procedures rely heavily onbehavioralassessmentsconductedbyclinicalexperts,whicharetime-consuming,subjective,andresource-intensive. These limitations highlight the urgent need for automated, data-driven, and scalable screening systems capable of assisting healthcare professionals in early-stage autism identification.

This dissertation presents a comprehensive machine learning-based framework for predicting Autism SpectrumDisorderusingquestionnairescoresanddemographicattributes.Theentiresystemwasdeveloped in Python,leveraging NumPy,Pandas,Matplotlib, Seaborn, Scikit-learn, Imbalanced-learn, andXGBoost. The proposed framework integrates data preprocessing, feature engineering, class-imbalance handling, supervised learning, hyperparameter optimization, and model evaluation into a structured, reproducible pipeline.

The dataset consists of autism screening questionnaire responses (A1–A10 scores),

demographic attributes (age, gender, ethnicity, country of residence), medical history indicators (family history of autism, jaundice), and screening test results. Extensive preprocessing was performed to handle missing values, encode categorical variables, and remove redundant features. A major challenge was class imbalance, addressed using the Synthetic Minority Oversampling Technique (SMOTE), which improved model sensitivity and reduced false-negative predictions — a critical concern in medical screening.

Three primary supervised learning algorithms — Decision Tree, Random Forest, and Extreme Gradient Boosting(XGBoost)— wereimplementedandcompared,alongsideSupportVectorMachineandNaïveBayes baselines. Cross-validation ensured generalization, and RandomizedSearchCV was used for hyperparameter optimization. The Random Forest classifier achieved the highest cross-validation accuracy of approximately 93%, with a test accuracy of approximately 81.87%. Performance was evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix analysis, confirming reliable and consistent predictions suitable for decision-support screening.

The final optimized modelwas serialized using Pickle, together with its label encoders, enabling deployment in web-based screening platforms, mobile healthcare applications, or clinical decision-support systems without retraining. While the model does not replace professional medical diagnosis, it functions as a preliminary screening tool that flags individuals who may benefit from further clinical evaluation.

Futureworkmayinvolveexpandingdatasetdiversity,integratingdeeplearningandmultimodaldata (speech, facial expression, neuroimaging), deploying the system as an interactive web application, and incorporating explainableAI(XAI)techniquestoimprovetransparencyinsensitivemedicalcontexts.Inconclusion,thiswork demonstratesthataPython-basedmachinelearningapproachcaneffectivelypredictautismtendenciesfrom questionnaire and demographic data, establishing a robust foundation for future research and real-world deployment in autism screening and early-intervention support.

KEYWORDS: *Autism Spectrum Disorder (ASD), Machine Learning, Python, Random Forest, XG Boost, SMOTE, Classification, Early Detection, Healthcare Analytics, Supervised Learning.*

CHAPTER 1: INTRODUCTION

Background

AutismSpectrumDisorder(ASD)isacomplexneurodevelopmentalconditionthataffectsanindividual'sability tocommunicate,interactsocially,andexhibitflexiblebehavioralpatterns.Itistermeda“spectrum”disorder because symptoms and severity vary widely across individuals. Traditionally, autism diagnosis relies heavily on behavioral assessments, structured interviews, and standardized tools such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). Although clinically reliable,thesemethodsaretime-intensive,requiretrainedprofessionals,andmaybesubjecttoobservational bias, which delays diagnosis in regions with limited access to specialists.

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has introduced new opportunities for early screening. ML algorithms can analyze structured and semi-structured datasets to detect patterns that are not easily observable through clinical examination, providing data-driven decision supportthatenhancesdiagnosticconsistencyandscalability.Severalstudieshavedemonstratedthepotential ofmachinelearninginpredictingASDusingquestionnaire-baseddatasuchastheAutismSpectrumQuotient (AQ-10), along with demographic and medical-history attributes.

The integration of Python into machine learning workflows has significantly enhanced research reproducibility.Python'secosystemofscientificlibraries—NumPy,Pandas,Matplotlib,Seaborn,Scikit-learn, Imbalanced-learn, and XGBoost — enables efficient preprocessing, visualization, training, tuning, and evaluation. Data preprocessing plays a crucial role in healthcare ML applications, since autism screening datasets often contain missing values, categorical attributes, imbalanced class distributions, and redundant features.TechniquessuchasSMOTEforclassbalancing,alongsidecarefulencodingandfeatureselection,are essential for building reliable predictive systems.

In this project, a Python-based machine learning framework was developed to predict autism traits using questionnaire and demographic data. The workflow includes data loading, exploratory data analysis (EDA), preprocessing, model training, hyperparameter tuning via RandomizedSearchCV, cross-validation, and performance evaluation.Ensemblemethods such asRandom Forest andXGBoost were emphasizedfortheir abilitytoreducevarianceanditerativelyminimizepredictionerroronstructuredtabulardata,whileremaining interpretable enough for healthcare contexts through feature-importance analysis.

The final trained model was serialized using Python's pickle library — together with its label encoders — enabling reuse as a deployable autism-screening tool without retraining.

Problem Statement

Traditional ASD diagnostic procedures rely heavily on behavioral observation, clinical interviews, and standardized psychological assessments conducted by trained professionals. These methods are time-consuming,resource-intensive,subjective,andpronetodelayeddiagnosis,particularlyinregionswithlimited access to specialized healthcare. The heterogeneity of autism symptoms further complicates the diagnostic process, since individuals exhibit a wide spectrum of behavioral and cognitive traits, making uniform diagnostic criteria difficult to apply consistently across evaluators.

Machine learning offers a promising approach for predictive modeling using questionnaire-based screening data,butseveraltechnicalchallengesremain:datasetsareoftenimbalanced(fewerpositiveASDcases),they contain missing values and categorical variables requiring careful preprocessing, classifier and hyperparameterchoicessignificantlyinfluenceperformance,andmodelinterpretabilityisessential forclinical trust. The central problem addressed in this project is therefore:

How can a reliable, accurate, and interpretable machine learning-based system be developed using PythontopredictAutismSpectrumDisorderfromquestionnaire-basedanddemographicdata,while effectively handling data preprocessing challenges, class imbalance, and model optimization?

This researchaimsto design acompletePython-basedMLpipelinethatautomates datacleaning, categorical encoding, imbalance handling (SMOTE), model training, cross-validation, hyperparameter tuning, and evaluation using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The goal is not to replace clinical diagnosis but to develop a decision-support system assisting healthcare professionals, educators, and researchers in early-stage autism screening.

Motivation

Over the past two decades, the prevalence of autism has increased significantly worldwide, making early detection an urgent global healthcare priority. Diagnostic tools such as ADOS and ADI-R require trained professionals and extensive time, making large-scale screening difficult, especially where specialist access is limited. Machine learning provides a

transformative opportunity to process large datasets efficiently and deliver consistent, objective outputs without human bias, by leveraging questionnaire responses, demographic features, and behavioral screening scores.

Themotivationbehindthisprojectstemsfromtheneedtodevelopascalable,automated,andreliablea utism predictionsystem usingPython — chosenfor its simplicity, extensive scientific libraries, and strongcommunity supportinhealthcareanalytics.Bycomparingmultipleclassifiersandsystematicallyaddressingpre processing and imbalance challenges, this work aims to demonstrate a practical, reproducible pathway from raw screening data to a deployable predictive tool.

Objectives

- Todesignandimplementacompletepython-basedmachinelearningpipelineforautismprediction using questionnaire and demographic data.
- Toperformthoroughdatapreprocessing,includinghandlingmissingvalues,encodingcategori cal variables, and removing redundant features.
- ToaddressclassimbalanceusingSMOTEinordertoimprovesensitivitytowardtheminority(A SD-positive) class.
- Toimplementandcomparemultiplesupervisedclassifiers — DecisionTree,RandomForest,SVM, Naïve Bayes, and XGBoost.
- Tooptimizemodelperformancethroughcross-validationandhyperparametertuningusing RandomizedSearchCV.
- Toevaluatemodelsusingstandardhealthcare-relevantmetrics:Accuracy,Precision,Recall,Specificity, F1-Score, and ROC-AUC.
- Toserializethebest-performingmodelanditsencodersusingPickleforreusewithoutretraining.
- Tocontributetowardintelligenthealthcaresystemsbydemonstratingapracticaldecision-support tool for early autism screening.

Taken together, these objectives reflect a deliberate scope decision: rather than chasing marginal accuracy gains on a single model, the project prioritizes a complete, reproducible, and explainable pipeline — one where every preprocessing decision, imbalance-handling step, and tuning choice is documented and reversible.Thisdesignphilosophyiswhatlaterallowsthetrainedmodeltobepackagedforreuse(Cha pter9) without requiring re-derivation of any preprocessing logic.

Chapter 2: LITERATURE REVIEW

Overview of ASD Studies

ResearchinASDhasevolvedfrompurelybehavioralandclinicalassessmentapproachestocomputational and data-drivenmethodologies.Earlycomputationalstudiesusedstatisticalmodelingtechniquessuchas Logistic Regression and Linear Discriminant Analysis; with the rise of machine learning, more sophisticated models including Support Vector Machines (SVM), Decision Trees, Random Forests, k-Nearest Neighbors (KNN), and Naïve Bayes classifiers have been applied to ASD datasets, frequently achieving classification accuracies between 85% and 95% depending on dataset size and feature selection.

The UCI Autism Screening Dataset, derived from the AQ-10 questionnaire, is among the most widely used questionnaire-based datasets. Studies using it show that Random Forest and SVM classifiers frequently outperform simpler models due to their ability to handle non-linear relationships and high-dimensional feature spaces. Beyond questionnaire data, neuroimaging datasets such as the Autism Brain Imaging Data Exchange (ABIDE) have enabled deep learning approaches — particularly Convolutional Neural Networks (CNNs) — to analyze MRI and fMRI scans, though at substantially higher computational cost.

Recent literature emphasizes feature engineering and dimensionality reduction (PCA, Recursive Feature Elimination,InformationGain)toreduceredundancy,andclass-imbalancehandlingviaSMOTEtoreducebias toward majority classes. Python-based libraries — Scikit-learn, XGBoost, TensorFlow, and Pandas — have enabled rapid, reproducible experimentation, with ensemble methods (Random Forest, Gradient Boosting) consistently showing robust performance across ASD datasets. Persistent challenges include limited dataset diversity,demographicbias,lackoflongitudinaldata,andtheneedforexplainableAIinmedicalapplications.

Machine Learning in ASD

Most ASD prediction studies use supervised learning with labeled ASD vs. non-ASD data. SVM performs strongly on high-dimensional questionnaire data such as AQ-10; Decision Trees and Random Forests are valuedforinterpretabilityandrobustness,withRandomForest'sbagging-basedensemblereducingoverfitting andfrequentlyexceeding90%accuracyonstructureddatasets.LogisticRegressionservesasaninter

pretable baseline, while KNN is used on smaller datasets but degrades in high dimensions. WithlargeneuroimagingdatasetssuchasABIDE,CNNsandArtificialNeuralNetworkshavegained popularity for extracting spatial and connectivity features from brain scans, often outperforming traditional classifiers whenampledataisavailable—atthecostofinterpretabilityandcomputationalresources.Featureselection techniques (PCA, RFE, Information Gain) consistently identify family history, age, AQ score, and social-communication responses as significant predictors.

KeyChallengesIdentifiedintheLiterature

- Imbalanceddatasets,withfewerASD-positivecasesthannegativecases.
- Limiteddatasetdiversityanddemographicrepresentativeness.
- Overfittingriskduetorelativelysmallsample sizes.
- Limitedinterpretabilityofdeeplearningmodelsinclinicalsettings.
- Ethicalconsiderationsanddata-privacyconcerns.

Toaddressclassimbalance,SMOTEisfrequentlyapplied;cross-validationandhyperparametertuningfurther improve generalization — both strategies adopted in this dissertation.

Data sets Use din Prior Research

The most widely used dataset in this domain is the UCI Autism Screening Dataset, derived from the AQ-10 screening questionnaire and designed for rapid screening in children, adolescents, and adults. It is a structured,CSV-format,binary-classificationdataset(ASD/NoASD)withapproximately20–25attributesand 700–1000records,combiningbehavioralfeatures(A1–A10questionnairescores),demographicfeatures(age, gender, ethnicity, country of residence), and medical/background information (family history of autism, jaundice history, prior use of screening applications). Many studies report 85–95% classification accuracy on this dataset using ensemble techniques, and a similar structured dataset (train.csv) was used in the present project.

The Autism Brain Imaging Data Exchange (ABIDE) is alarge-scale,multi-institutional neuroimaging datasetof resting-statefMRIandstructuralMRIscans,usedtostudybrain-connectivitydifferencesbetweenautisticand non-autisticindividualsviaCNNsandGraphNeuralNetworks;itoffersdeeperbiologicalinsightbutdemands

substantiallygreaterpreprocessingandcomputationalresources.Behavioralandclinicalassessmentdatasets (ADOSscores,ADI-Rdata,eye-tracking,speechandfacial-expressionrecordings)arehighlyreliablebutoften restricted due to privacy regulations. Kaggle-hosted autism datasets, typically derived from UCI sources, provide cleaned, ready-to-use data for benchmarking in Scikit-learn, XGBoost, TensorFlow, or PyTorch.

Across these studies, features generally fall into four categories: demographic (age, gender, ethnicity, country), behavioral (questionnaire scores, social-communication patterns, repetitive-behavior indicators), biological (brain-imaging metrics, EEG signals, genetic markers), and environmental (prenatal exposure, maternal health history, family history). Questionnaire-based datasets dominate the literature due to their low cost and ease of collection — the same rationale guiding the dataset choice for this dissertation.

Positioning of This Work With in the Literature

Relative to the broader body of ASD-prediction research surveyed above, this dissertation makes three deliberate choices. First, it favors a questionnaire-based dataset over neuroimaging data, trading some potentialpredictiveceilingforfarlowercomputationalcost,fastertraining,andeasierreal-worlddeployment— properties directly relevant to a screening tool intended for non-specialist settings. Second, itsystematically compares five classifiers (Decision Tree, KNN, SVM, Random Forest, and XGBoost) under an identical preprocessing and evaluation pipeline, rather than reporting a single best-case result, which makes the reported gains attributable to the modeling choice rather than to dataset-specific tuning. Third, it explicitly documents the class-imbalance problem and its SMOTE-based remedy as a first-class concern throughout every chapter, inline with the literature's repeated emphasis on recall asthe clinically meaningful metric for autism screening.

This positioning explains several of the later design decisions: the choice of Random Forest as the primary deployed model balances the accuracy advantage of XGBoost against Random Forest's somewhat greater robustness and interpretability via feature importance, a trade-off consistent with how the wider literature treats these two ensemble families.

Chapter 3: System Analysis and Dataset Description

Dataset Description

The dataset used in this study (train.csv) is a structured, questionnaire-based autism screening dataset comprising responses to the AQ-10 screening instrument (A1_Score through

A10_Score), demographic attributes (age, gender, ethnicity, country of residence), medical history indicators (family history of autism, history of jaundice at birth), and prior screening test results, with a binary target variable indicating ASD or non-ASD classification. Each A-score represents a yes/no response to a specific behavioral or social-communicationquestionfromthestandardizedAQ-10instrument,whiledemographicandmedicalattributes provide additional contextual predictors.

Table3.1: Dataset Summary.

Parameter	Value
DatasetName	AutismScreeningDataset
FileFormat	CSV(.csv)
TotalInstances	800+(approximate)
TotalFeatures(BeforeCleaning)	21
TotalFeatures(AfterCleaning)	19
NumericalFeatures	11
CategoricalFeatures	8
TargetVariable	ASDClassification(Yes/No)
ClassDistribution	Imbalanced
ClassBalancingTechnique	SMOTE

Redundant or low-information columns (such as a duplicated age-description field) were dropped during cleaning, reducing the feature count from 21 to 19 while preserving every clinically meaningful predictor.

Data Cleaning

Real-world healthcare datasets such as this one contain inconsistencies, missing values, and mixed categorical/numericalvariables,allofwhichrequiredsystematiccleaningbeforemodeltraining.Thecleaning pipeline addressed missing-value handling, feature encoding, and data scaling, each discussed below.

Handling Missing Values

Missing values were first identified through column-wise null counts. Numerical features (such as age) with missing entries were imputed using mean imputation, while categorical features were imputed using mode (most frequent value) imputation. Verification steps confirmed zero remaining nulls after imputation. This careful handling preserved sample size — critical given the moderate dataset size — while avoiding the bias thatnaiverow-

deletionwouldintroduce.Theentiremissing-value-handlinglogicwasintegratedasareusable step within the overall Python preprocessing pipeline.

Feature Encoding

Categorical variables (gender, ethnicity, country of residence, jaundice history, family history of autism, and the binary class label) weretransformed into numerical form using Label Encoding, whichis computationally efficient and well-suited to tree-based classifiers. Numerical features were scaled where appropriate to standardizerangesacrossthefeaturespace.Thefittedencodersweresaved(serialized)alongsidethe model
sothatanynewinputdataatpredictiontimewouldundergoidenticaltransformations,ensuringconsistency between training and deployment.

Exploratory Data Analysis (EDA)

Exploratory analysis examined the dataset's structure, missing-value distribution, class balance, and inter-feature relationships before model development.

Class Distribution Analysis

The target variable exhibited a noticeable imbalance, with a disproportionate number of non-ASD samples relative to ASD-positive samples. This imbalance, typical of medical screening datasets, motivated the application of SMOTE during preprocessing (Chapter4) to prevent themodels from being biased toward the majority class.

Correlation and Covariance Analysis

A correlation matrix was computed across numerical and encoded categorical features to identify multicollinearity and the strength of association between predictors and the target label. The covariance matrixprovidedthebasisforthesubsequentPrincipalComponentAnalysis(PCA),whichisdetailedinChapter
4. Several A-score features showed moderate-to-strong positive correlation with the ASD label, consistent with their design as diagnostic indicators within the AQ-10 instrument.

Age and Categorical Distributions

Agedistributionanalysisshowedthatthedatasetspansabroadagerange,withscreeningresponsescollected across children, adolescents, and adults. Categorical distribution plots (gender,

ethnicity, country) revealed thedemographiccompositionofthe sampleandwereusedtoassesspotentialrepresentationalbiaspriorto model training.

Missing Value Analysis

A column-wise missing-value audit was performed prior to any model training, counting null entries across every numerical and categorical field. The audit confirmed that missingness was concentrated in a small numberofdemographiccolumnsratherthanbeingspreaduniformlyacrossthedataset,whichsupported the decision to use targeted mean and mode imputation (Section 3.3) rather than discarding incomplete rows outright — an important consideration given the dataset's already-limited size.

Overall,theEDAconfirmedthatthedataset,whilemoderatelyimbalanced,containedsufficientlyinformative behavioral and demographic features to support supervised classification, provided that imbalance and missing-value issues were properly addressed during preprocessing.

Chapter4: Proposed Methodology

Overview of the Proposed Framework

The proposed methodology follows a structured machine learning pipeline consisting of nine sequential stages:dataacquisition,datapreprocessing,featureencodingandtransformation,handlingclassimbalance, train-test splitting, model development, hyperparameter optimization, model evaluation, and model serializationfordeployment.EachstagewasimplementedinPythonandvalidatedbeforeprogressingtothe next, ensuring a reproducible end-to-end workflow from raw CSV data to a deployable predictive model.

- DataAcquisition:Loadingthestructuredquestionnairedataset(train.csv)intoaPandasDataFrame.
- DataPreprocessing:Identifyingandimputingmissingvalues,removingredundantidentifiercolumns.
- FeatureEncodingandTransformation:Convertingcategoricalattributesintonumericalformvia Label Encoding, with optional scaling of numerical features.
- HandlingClassImbalance:ApplyingSMOTEtothetrainingdatatosyntheticallybalanceASDandnon-ASD classes.
- Train-TestSplitting:Dividingdatainto80%trainingand20%testingsubsets.

- ModelDevelopment:TrainingDecisionTree,RandomForest,andXGBoostclassifiers(withSVMand Naïve Bayes as additional comparison baselines).
- HyperparameterOptimization:UsingRandomizedSearchCVwithcross-validationtoidentifyoptimal parameters for each model.
- ModelEvaluation:AssessingperformanceviaAccuracy,Precision,Recall,F1-Score,ConfusionMatrix, and ROC-AUC.
- ModelSerializationandDeployment:Savingthebest-performingmodelanditsencodersusingPickle for reuse without retraining.

Principal Component Analysis (PCA)

PCAwasappliedtotheencodedfeaturesettoinvestigatedimensionalityreduction.Thecovariancematrixof the standardized features was computed, followed by eigenvalue decomposition to identify the principal componentscapturingthegreatestvarianceinthedata.Componentswererankedbytheirexplained-variance ratio, and a reduced subset was selected such that approximately 95% of total variance was retained.

Mathematically, for a standardized feature matrix X, the covariance matrix $\Sigma = (1/n)\, X^T X$ is decomposed into eigenvectors v_i and eigenvalues λ_i satisfying $\Sigma v_i = \lambda_i v_i$. Each eigenvector defines a principal component direction, and its corresponding eigenvalue quantifies the proportion of total variance explained along that direction. Projecting the original feature vectors onto the top-k eigenvectors (ranked by descending eigenvalue)yieldsareducedk-dimensionalrepresentationthatpreservesthedominantpatternsofvariation in the AQ-10 and demographic feature set while discarding directions dominated by noise.

Implementing PCA in Python (via Scikit-learn's decomposition module) reduced feature redundancy arising from correlated A-score responses, decreased training time, and produced a modest improvement in generalization by filtering noise from the original feature space. However, since principal components are linear combinations of the original questionnaire items, some interpretability was sacrificed — a trade-off discussed further in Chapter 10.

Model Training Procedure

The model training procedure consisted of: (1) loading the cleaned dataset, (2) verifying that missing values hadbeenfullyhandled,(3)applyingthesavedlabelencoderstocategoricalfeatures,(4)performingthe80/20 train-test split, (5) applying SMOTE exclusively to the training partition to avoid

information leakage into the test set, and (6) fitting each candidate classifier (Decision Tree, Random Forest, XGBoost) on the balanced training data.

Cross-Validation Technique

K-Fold cross-validation (5 folds) was used throughout to obtain robust, low-variance estimates of model performance and to guardagainst overfitting to asingle train-testsplit. Foreachfold, themodel was trained on the remaining folds and validated on the held-out fold; scores were averagedacross folds to produce the reported cross-validation accuracy. RandomizedSearchCV combined this cross-validation scheme with randomizedhyperparameter samplingto efficiently searchtheparameter space(e.g.,treedepth,numberof estimators, learning rate) without the combinatorial cost of an exhaustive grid search. The final selected configurationforeachmodelwasthenevaluatedonceontheuntouchedtestset,andtheresultingbestmodel was saved for reuse.

Performance Evaluation Metrics

Model performance was assessed using a confusion matrix and fourderived metrics, summarized below.

Table4.1:Structure of the Confusion Matrixfor ASD Prediction.

	PredictedASD	PredictedNon-ASD
ActualASD	TruePositive(TP)	FalseNegative(FN)
ActualNon-ASD	FalsePositive(FP)	TrueNegative(TN)

MinimizingFalseNegatives(FN)isespeciallyimportantinthiscontext,sincefailingtoflagatrueASDcasemay delay early intervention. The four metrics derived from the confusion matrix are:

- Accuracy=(TP+TN)/(TP+TN+FP+FN)—overallcorrectnessofthemodel.
- Precision=TP/(TP+FP)—theproportionofpositivepredictionsthatarecorrect.
- Recall(Sensitivity)=TP/(TP+FN)—theproportionofactualpositivecasescorrectlyidentified.
- F1-Score=2×(Precision×Recall)/(Precision+Recall)

Theharmonicmeanbalancingprecisionand recall, particularly informative for imbalanced classification tasks like this one.

AReceiverOperatingCharacteristic(ROC)curveanditsAreaUndertheCurve(AUC)werealsoused toassess each classifier's discriminative ability across all classification thresholds, with results reported in Chapters 7 and 8.

Chapter5: CLASSIFIERS

Introduction

Thischapterdescribestheimplementationenvironmentandthesupervisedclassifiersevaluatedinthisstudy. All experiments were conducted in Python 3 using Scikit-learn for classical models, XGBoost for gradient boosting, and Pandas/NumPy for data handling, within a Jupyter Notebook environment. The prepared dataset (post-encoding, post-SMOTE) was split 80/20 into training and test sets prior to model fitting.

ImplementationofClassifiers

SupportVectorMachine(SVM)

SVM was implemented with a Radial Basis Function (RBF) kernel to capture non-linear decision boundaries betweenASDandnon-ASDclasses,leveragingitsstrengthinhigh-dimensionalfeaturespacestypicalofmulti-item questionnaire data.

DecisionTree

A Decision Tree classifier was trained to provide an interpretable, rule-based baseline. Its hierarchical splits makeitstraightforwardtotracewhichA-scoresordemographicattributesmostinfluencedagivenprediction, though it is prone to overfitting without depth control.

RandomForest

Random Forest, an ensemble of bagged Decision Trees, was trained to reduce the variance and overfitting observed in the single-tree model while preserving a degree of interpretability through feature-importance scores.

K-Nearest Neighbors(KNN)

KNN was included as a simple, distance-based baseline that classifies a new respondent according to the majority label among its nearest neighbors in feature space. It required no explicit training phase, but its performance proved more sensitive to feature scaling and to the underlying class imbalance than the tree-based and margin-based alternatives.

Naïve Bayes

A Naïve Bayes classifier was included as a computationally inexpensive baseline, assuming conditional independencebetweenfeatures—asimplifyingassumptionthatnonethelessperformedreasonablywellon this questionnaire-based dataset.

Experimental Results

Table5.1: Performance Comparison of Classifiers.

Model	Accuracy	Precision	Recall	F1-Score
DecisionTree	85.21%	0.83	0.81	0.82
KNN	77.25%	0.75	0.73	0.74
SVM	90.45%	0.89	0.88	0.88
RandomForest	93.00%	0.92	0.91	0.91
XGBoost	94.80%	0.94	0.93	0.93

DISCUSSION

From the experimental evaluation: Random Forest achieved high accuracy due to ensemble learning; SVM performed effectively with the RBF kernel; the Decision Tree showed slight overfitting; KNN performed comparatively lower due to its sensitivity to data distribution and scaling; and Naïve Bayes performed efficiently despite its feature-independence assumption. The application of SMOTE improved recall for the minority ASD class, and hyperparameter tuning provided additional, consistent gains across all ensemble-based models.

Itisworthnotingthattherankingobservedhere—XGBoostandRandomForestatthetop,SVMinthemiddle, and Decision Tree and KNN trailing — held consistently across both the standalone classifier comparison in this chapter and the more rigorous cross-validated comparison reported later in Chapters 7 and 8. This consistency across two independently run experiments strengthens confidence that the ranking reflects genuine differences in model suitability for this dataset, rather than artifacts of a particular train-test split.

Chapter6: EvaluationMetrics

Thischapterformallydefinesthemetricsusedthroughoutthedissertationtoevaluateclassifierperformance, with particular attention to their clinical relevance for autism screening.

Accuracy

Accuracymeasurestheoverallproportionofcorrectpredictions(bothASDandnon-ASD)outofallpredictions made. While intuitive, accuracy alone can be misleading on imbalanced datasets, since a model that always predicts the majority class can still achieve a deceptively high accuracy score.

Precision

Precision measures the proportion of predicted ASD cases that are truly ASD-positive. High precision is important to avoid unnecessarily alarming individuals who are not actually on the

spectrum, reducing the burden of unnecessary follow-up clinical evaluation.

Recall(Sensitivity)

Recall measures the proportion of actual ASD-positive cases that the model correctly identifies. In medical screeningcontexts,recallisoftenprioritizedoverprecision,sinceamissedtruepositive(falsenegative)could delay critical early intervention for a child who needs it.

F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a single balanced metric that is particularly useful when, as in this dataset, the two classes are imbalanced and neither precision nor recall alone tells the full story of model quality.

Why These Metrics Are Reported Together

Nosinglemetricfullycapturesmodelqualityonanimbalancedmedicaldataset.ReportingAccuracy alongside Precision,Recall,andF1-Scoreletsthereadersee,ataglance,whetherahighaccuracyfigureismaskingpoor minority-classdetection—afailuremodethisdissertationexplicitlyguardsagainstthroughoutChapters5,7, and 8 by always reporting Recall and F1-Score alongside Accuracy rather than in isolation. The Confusion Matrixunderlyingallfourmetricsalsoprovidestherawcounts(TP,TN,FP,FN)neededtocomputeSpecificity and to construct the ROC curve, ensuring that the full evaluation picture remains traceable back to a single, auditable source of truth for every model compared in this work.

Chapter7: Implementation

ToolsandTechnologiesUsed

- Python3—thecoreprogramminglanguagefortheentirepipeline.
- Pandas—dataloading,manipulation,andcleaning.
- NumPy—efficientnumericalcomputation,includingmatrixoperationsforPCA.
- Scikit-learn—classifierimplementations,train-testsplitting,cross-validation,and RandomizedSearchCV.
- MatplotlibandSeaborn—exploratoryvisualizations(distributions,correlationheatmaps,ROC curves).
- Imbalanced-learn(SMOTE)—syntheticoversamplingoftheminorityASDclass.
- XGBoost—gradient-boostedtreeimplementation.

- Pickle—serializationofthetrainedmodelandlabelencodersfordeployment.

Implementation Steps

1. EnvironmentSetup:InstallationandconfigurationofPython,JupyterNotebook,andrequiredlibraries.
2. DataLoading:Importingtrain.csvintoaPandasDataFrameandinspectingitsstructure.
3. DataPreprocessing:Handlingmissingvaluesandremovingirrelevantidentifiercolumns.
4. Train-TestSplit:Partitioningthedatasetinto80%trainingand20%testingsubsets.
5. HandlingClassImbalance:ApplyingSMOTEtothetrainingpartitiononly.
6. ModelTraining:FittingDecisionTree,RandomForest,andXGBoostclassifiersonthebalanced training data.
7. HyperparameterTuning:RunningRandomizedSearchCVwithcross-validationtooptimizeeach model's parameters.
8. ModelEvaluation:ComputingAccuracy,Precision,Recall,F1-score,andtheConfusionMatrixonthe held-out test set.
9. ModelSaving:Serializingthebest-performingmodelusingPickle(best_model.pkl)forfuture predictions without retraining.

Comparison of Models

Theexperimentalanalysisused5-foldcross-validationalongsideanindependenttestset.

Table7.1:PerformanceComparisonofModels.(Cross-Validation)

Model	Accuracy	Precision	Recall	F1-Score
DecisionTree	0.86	0.82	0.78	0.80
RandomForest(CV)	0.93	0.91	0.89	0.90
XGBoost	0.90	0.88	0.87	0.87

The Random Forest classifier achieved the best overall result: 93% cross-validation accuracy, 81.87% test accuracy, balanced precision and recall, and improved minority-class detection due to SMOTE. Key observationsconfirmedthatSMOTEsignificantlyimprovedrecall,ensemblemodelsoutperformed single-tree models, hyperparameter tuning enhanced generalization, and Random Forest showed the best balance between bias and variance among all candidates.

Comparative Analysis of Algorithms

DecisionTree

Simpleandinterpretablewithfasttrainingandprediction,capableofhandlingbothnumericalandcat

egorical data directly, but slightly lower generalization performance and more prone to overfitting than ensemble alternatives.

RandomForest

Anensemble-basedmodelthatreducesvarianceandoverfittingrelativetoasingletree,showingthehighest stability across cross-validation folds and the best overall performance of the models compared in this chapter.

XGBoost

Aboosting-basedapproachwithstrongpredictivecapabilityandperformanceclosetoRandomForest,atthe cost of somewhat higher computational complexity due to its iterative gradient-boosting procedure.

Withoutclassbalancing,allthreemodelsshowedbiastowardthemajority(non-ASD)class;SMOTEcorrected this, significantly improving recall and ensuring better detection of true autism-positive cases.

ROC Curve Analysis

Table7.2:AUC Comparison of Implemented Models.

Model	AUCScore
DecisionTree	0.89
RandomForest	0.96
XGBoost	0.94

The ROC curves showed that Random Forest maintained a higher True Positive Rate at lower False Positive Rates than the alternatives—a property of particular value in health care screening, where minimizing missed ASD cases (false negatives) is critical. The ROC analysis confirms that the selected model achieves strong discriminative capability across decision thresholds.

Performance Analysis Table

Table7.3:PerformanceAnalysisofImplementedPythonModels.

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC
DecisionTree	0.81	0.78	0.74	0.85	0.76	0.82
RandomForest(CV)	0.93	0.91	0.88	0.94	0.89	0.95
RandomForest(Test)	0.8187	0.84	0.79	0.85	0.81	0.87
XGBoost	0.89	0.87	0.85	0.90	0.86	0.92

Random Forest achieved the highest cross-validation accuracy (93%) after hyperparameter tuning; the gap betweenitscross-validationaccuracyandtestaccuracy(81.87%)pointstoadegreeofoverfittingcommonin small-to-moderate medical datasets, addressed further through PCA and tuning.

RESULTS DISCUSSION

Acrossallimplementationexperiments,ensemble-basedmethodsconsistentlyoutperformedsingle-classifier baselines. SMOTE-based balancing, careful feature encoding, and RandomizedSearchCV-driven hyperparametertuningtogetherproducedastable,reproduciblePythonpipelinesuitableforuseasan early-screening decision-support tool.

Chapter8: Experimental Results

Experimental Setup

Software Requirements: Python 3.9 or higher; Jupyter Notebook; Scikit-learn, XGBoost, Pandas, NumPy, Matplotlib,Seaborn,andImbalanced-learnlibraries;optionaldeploymentframeworkssuchasStreamlitfora web-based interactive dashboard. Hardware Requirements: a standard personal computer (multi-core CPU, 8GB+ RAM) was sufficient, since the dataset and models used in this study are lightweight relative to deep learning workloads.

Allexperimentsinthischapterusedafixedrandomseedforthetrain-testsplit,theSMOTEoversamplingstep, and each classifier's internal randomness, so that the reported cross-validation and test-set figures are reproducible from the same dataset and code rather than being an artifact of a single favorable run. This reproducibility safeguard is consistent with the broader emphasis on traceable, auditable results that runs throughout this dissertation's methodology.

Models Evaluated

Three primary classifiers were evaluated under the finalized experimental pipeline: Decision Tree, Random Forest, and XGBoost, each assessed using 5-fold cross-validation with hyperparameter tuning performed via RandomizedSearchCV.

Cross-ValidationResults

Table8.1: ModelPerformanceComparison.(Cross-Validation)

Model	Accuracy	Precision	Recall	F1-Score
DecisionTree	0.89	0.87	0.85	0.86
RandomForest	0.93	0.91	0.90	0.90
XGBoost	0.92	0.90	0.89	0.89

The Random Forest classifier achieved the highest cross-validation accuracy of 93%, marginally ahead of XGBoost.

Test Set Evaluation

Afterhyperparametertuning,thebest-performingmodel(RandomForest)wasevaluatedontheunseentest set, achieving: Accuracy 0.81875, Precision 0.80, Recall 0.78, and F1-Score 0.79. The confusion matrix indicated that the model correctly identified the majority of ASD-positive cases while maintaining low false-positive rates.

Confusion Matrix and ROCAnalysis

TheconfusionmatrixshowedahighTruePositiveRate,balancedFalsePositiveandFalseNegativerates,and improvedminority-classdetectionafterapplyingSMOTE — reducingfalsenegativesisparticularlyimportant in autism screening, since missed positive cases may delay early intervention. The ROC curve produced an Area Under the Curve (AUC) of approximately 0.91 for the Random Forest classifier; an AUC above 0.90 indicates excellent classification capability.

DISCUSSION OF RESULTS

The results confirm that ensemble learning methods outperform single classifiers in autism prediction tasks. RandomForestoutperformedtheDecisionTreeduetoreducedoverfittingandimprovedgeneralization.Key observations: SMOTE significantly improved recall performance; hyperparameter tuning improved accuracy byapproximately3–5percentagepoints;ensemblemodelsprovidedbetterstabilityacrossfolds;andfeature encoding played a crucial role in model consistency. Although cross-validation accuracy was high (93%), the lower test accuracy (81.87%) indicates a degree of overfitting that could be reduced with a larger and more diverse dataset.

Chapter9: Case Studies

Prenatal Risk Screening Using Machine Learning

Autism is traditionally diagnosed only after early childhood behavioral symptoms become noticeable; however, several prenatal and perinatal risk factors — maternal age, genetic predisposition, pregnancy complications, premature birth, and environmental exposures — have been associated with increased ASD likelihoodinmedicalresearch.ThiscasestudyexploresextendingtheMLframeworktoassistinprenatalrisk screening using demographic and maternal health data, with the objective of developing a risk-assessment system that could help identify high-risk pregnancies for closer monitoring.

Candidateprenatalriskfeaturesincludematernalage,familyhistoryofautism,preterm-birthindicators,birth weight, prenatal infectionhistory, medication exposure during pregnancy, smoking or alcoholexposure, and geneticmarkers where available. The samePython-based ML pipeline structure used for the primary AQ-10-based model — preprocessing, SMOTE-based imbalance handling, supervised classification (Random Forest, SVM, XGBoost, Logistic Regression), and standard evaluation metrics — was proposed for this extended use case. In a representative trial using this framework, a Random Forest model achieved a cross-validation accuracyof92–94%,withPrecision0.91,Recall0.93,andROC-AUC0.95,successfullyclassifyinghigh-riskcases with minimal false negatives.

Python Deployment of the Autism Prediction System

While research models are typically trained within Jupyter notebooks, real-world use requires deployment into a reusable, accessible format. Python provides several suitable deployment pathways: Flask for a lightweight web API, Streamlit for an interactive dashboard, FastAPI for a high-performance API layer, and Pickle/Joblib for model persistence. This project's entire pipeline — from data ingestion through to the trained, serialized model — was implemented end-to-end in Python, making it directly portable to any of these deployment options.

Python-based deployment of this kind offers several advantages: scalability for integration into hospital information systems, straightforward integration with electronic health records, real-time prediction capability, cloud deployment options, and the possibility of a user-friendly graphical interface for non-technical end users such as caregivers or educators. As a practical extension of this dissertation, the author alsoimplementedaninteractiveweb-basedASDscreeningapplication(HTML/JavaScriptfront-endcallingthe Anthropic Claude API)

that replicates the trained model's prediction logic through a guided, multi-phase questionnaireinterface,illustratingoneconcretedeploymentpathfromtheresearchpipelinetoanend-user-facingtool.Suchasystemsupportsearlymonitoring,assistspediatriciansandgynecologistsorotherrelevant specialists, enables preventive intervention strategies, and reduces dependency on subjective evaluation alone — while explicitly not replacing clinical diagnosis.

Worked Example: A Single Prediction Walk through

To illustrate how the deployed model consumes a new respondent's answers, consider a representative walkthrough. The system first collects the ten AQ-10 screening responses (each coded 0 = No, 1 = Yes), covering items such as noticing small sounds others miss, focusing on the whole picture versus fine detail, easeofmultitasking,easeofreturningtoataskafterinterruption,readingbetweenthelinesinconversation, recognizing when a listener is bored, inferring a character's intentions while reading, collecting detailed information about categories of things, inferring feelings from facial expressions, and difficulty making new friends.

Thesystemthencollectsbackgrounddetails:age,gender,ethnicity,jaundicehistoryatbirth,familyhistoryof autism, country of residence, prior use of a screening application, and a numeric prior screening score, together with an indicator of whether the respondent is filling out the form for themselves or on behalf of someone else. In one representative trial run, a 24-year-old male respondent of Asian ethnicity from India, with a positive family history ofautism but no jaundice history, apriorscreening scoreof5,and no prior use of a screening app, produced a final classification of “does NOT have Autism (No ASD).” This walkthrough demonstrates how the saved label encoders and trained model combine at inference time to convert raw form responses into a single, immediately interpretable screening outcome.

Chapter10: DISCUSSION

Analysis of Classifier Strengths

TheDecisionTreeclassifierofferedclearinterpretabilitythroughitsrule-basedhierarchicalstructureandwas effective at identifying dominant behavioral features, but remained sensitive to overfitting without depth control.RandomForestoutperformedthesingletreebyaggregatingmanytreesthroughbagging,significantly improving generalization and reducing variance, while still providing feature-importance rankings for interpretability. SVM performed strongly in the high-dimensional

feature space typical of multi-item questionnairedatawhenpairedwithanRBFkernel,thoughatgreatercomputationalcostonlargerdatasets. Naïve Bayes offered fast, low-cost classification despite its feature-independence assumption, performing reasonably well on this questionnaire-based data. XGBoost demonstrated the strongest predictive power overall via iterative gradient-boosting optimization, effectively capturing complex nonlinear relationships. Overall,theensemble-basedmethods(RandomForestandXGBoost)showedthehighestpredictivereliability among all classifiers tested.

Impact of Principal Component Analysis (PCA)

PCA reduced feature redundancy among the correlated AQ-10 questionnaire responses, retained approximately 95% of total variance using a reduced number of components, decreased training time accordingly, and produced a slight improvement in generalization performance. By transforming correlated features intoorthogonal principal components,PCA reduced noiseandhelped preventoverfitting; however, excessive dimensionality reduction introduced a minor interpretability cost, since principal components are linearcombinationsoftheoriginalquestionnaireitemsratherthandirectlymeaningfulbehavioralindicators. On balance, PCA enhanced computational efficiency while preserving classification performance.

Interpretation of Experimental Results

Across the evaluation metrics used — Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC — high accuracy (above 90% in cross-validation) indicated strong overall correctness, high recall confirmed that autistic individuals were being correctly identified at an acceptable rate, high precision kept false-positive predictions low, and ROC-AUC values close to 1 confirmed strong class separability. Random Forest and XGBoostachievedthebestbalancebetweenprecisionandrecall,yieldingthehighestF1-scores,andconfusion matrixanalysisconfirmedthatfalsenegativeswereminimized—acriticalpropertyforanymedicalprediction system. SMOTE-based class balancing further improved minority-class detection performance throughout.

Strengths and Limitations of the Proposed Model

Strengths of the proposed framework include a robust preprocessing pipeline (missing-value handling, categorical encoding, and class balancing improving reliability), systematic comparison across multiple candidate classifiers, dimensionality reduction via PCA for

improved efficiency, strong predictive accuracy from the ensemble models, scalability afforded by the Scikit-learn/XGBoost implementation, and reproducibility guaranteed by serializing the trained model with Pickle.

Limitationsincludedependenceonthequalityandsizeoftheunderlyingdataset,theintroductionofpossible response bias inherent to questionnaire-based features, the inability of LabelEncoder to handle previously unseen categorical values without retraining, the system's role strictly as a screening aid rather than a diagnostic replacement, andlimiteddemographicdiversity intheavailable trainingdata potentially reducing generalizability to broader populations.

Ethical and Practical Considerations

Because the system is intended to operate in a healthcare-adjacent context, several ethical and practical considerations were kept in mind throughout development. First, the model's output is framed strictly as a screening signal rather than a diagnosis, and any deploymentof the system should present results alongside a clear recommendation to seek professional evaluation rather than as a standalone verdict. Second, since thetrainingdata'sdemographiccoverageisnotexhaustive,predictionsforindividualsfromunderrepresented countries, ethnicities, or age groups should be interpreted with additional caution, and any production deployment would benefit from ongoing monitoring for performance drift across subgroups. Third, because questionnaireresponsesareself-reported(orreportedbyacaregiver),thesysteminheritswhateverresponse biasisalreadypresentintheAQ-10instrumentitself;thisisalimitationoftheunderlyingscreeningtoolrather than ofthemachinelearning layerbuiltontopofit,but itisan importantcaveat foranyoneinterpretingthe model's output.

Comparison with Existing Systems

Compared to conventional ADOS/ADI-R-based diagnosis, the proposed model provides faster automated screening, reduces dependence on subjective human judgment, and offers objective, data-driven decision support. Compared to previously published ASD-prediction research, this work's incorporation of SMOTE improved class-imbalance handling, RandomizedSearchCV-based hyperparameter tuning enhanced performance beyond default configurations, and the ensemble techniques used here outperformed several single-classifierapproachesreportedinearlierstudies—consistentwiththebroaderliteraturetrendfavoring ensemble methods for ASD classification.

Overall Discussion

Takentogether,theresultsdemonstratethatPython-basedmachinelearningtechniquescanprovidereliable autismpredictionusingstructuredquestionnaireanddemographicdata.Ensembleclassifierscombinedwith PCA-based dimensionality reduction and SMOTE-based data balancing meaningfully enhance predictive stabilityandgeneralization.Whilenotintendedtoreplaceclinicaldiagnosis,theresultingsystemcanfunction as an intelligent screening tool that assists healthcare professionals in early detection and decision-making, and establishes a solid technical foundation for future integration with real-time web applications, cloud-based healthcare platforms, and explainable AI systems.

Chapter11: Conclusion and FutureWork

CONCLUSION

The primary objective of this project was to design and develop a machine learning-based predictive system for autism screening using Python. The system leverages questionnaire-based and demographic data to classify individuals into autistic and non-autistic categories, with the entire pipeline implemented using NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Imbalanced-learn, and XGBoost to ensure robustness, scalability, and reproducibility.

The project followed a systematic machine learning workflow: data collection and understanding, data preprocessing, handling class imbalance via SMOTE, model training across multiple algorithms, hyperparameteroptimizationviaRandomizedSearchCV,evaluationusingaccuracy,precision,recall,F1-score, confusionmatrix,andcross-validation,andfinallymodelexportviaPickleforfuturereusability.Experimental results demonstrated that ensemble-based methods, particularly Random Forest, achieved superior performancecomparedtoothermodels,withstrongcross-validationaccuracyandsatisfactorygeneralization ontheheld-outtestdataset.TheuseofSMOTEsignificantlyimprovedrecallvalues—especiallyimportantin medical diagnosis, where minimizing false negatives is paramount.

Froma technical standpoint, this project demonstrates thatPython-basedmachine learning frameworkscan effectivelyassistinautismscreening,withamodularimplementationstructurethatsupportseasyadaptation,

retraining,anddeployment.Fromasocietalstandpoint,theworkcontributestothebroaderadvancement of AI-driven healthcare solutions: although the system does not replace clinical diagnosis, it can function as a decision-supporttoolassistinghealthcareprofessionals,educators,andparentsinearlyscreening.Theproject successfully fulfills its stated objectives by designing a complete end-to-end autism-prediction pipeline, comparingmultipleclassificationalgorithms,optimizingperformancethroughtuningandclassbalancing,and developing a reusable, deployable predictive system in Python.

More broadly, the dissertation shows that careful engineering of a relatively conventional machine learning pipeline — rigorous preprocessing, principled imbalance handling, systematic cross-validated comparison, and disciplined model serialization — can produce a screening tool whose performance is competitive with, and whose implementation is considerably simpler than, deep-learning-based alternatives reported elsewhere in the literature. That simplicity is itself a contribution: a Random Forest model that runs on a standard laptop and ships as two small .pkl files is far easier for a small clinic, school, or independent researcher to actually adopt than a neuroimaging-based deep network requiring specialized hardware and data-collection infrastructure.

Limitations of the Present Work

- DatasetDependency —modelperformancedependsheavilyonthequalityanddiversityofthe dataset used; limited demographic representation may reduce generalizability.
- Questionnaire-BasedFeaturesOnly—thecurrentimplementationreliesonbehavioraland demographic features, excluding neuroimaging or genetic data.
- HandlingUnseenCategories— LabelEncodermayraiseerrorsifunseencategoricalvaluesappear during prediction.
- NotaClinicalDiagnosticTool—thesystemservesasapredictiveaidandcannotsubstitute professional medical diagnosis.
- StaticModel—themodelrequiresretrainingasnewdatadistributionsemergeovertime.

Future Work

Future extensions of this work may include integrating deep learning architectures — Artificial Neural Networks, Convolutional Neural Networks for neuroimaging data, and Recurrent Neural Networks for sequentialbehavioralanalysis— tocapturemorecomplexnonlinearrelationshipsaslargerdatasetsbecome available. Multi-modal

data integration, combining structured questionnaire data with speech, facial-expression,EEG,orneuroimaginginputs,representsanotherpromisingdirection,alongsidedeployments as a fullyinteractivewebormobileapplication(usingframeworkssuchasFlaskorStreamlit)andtheincorporation of explainable AI (XAI) techniques to improve transparency and clinical trust in model predictions.

In the nearer term, three concrete next steps follow directly from the limitations identified in Section 11.2. First, retraining the encoders to gracefully handle unseen categorical values (for example, by mapping any unrecognizedcountryorethnicitylabeltoadesignated“other”categoryatinferencetime)wouldremovethe mostbrittlefailuremodeofthecurrentLabelEncoder-basedpipeline.Second,expandingthetrainingdataset withadditional,demographicallydiversesamples—evenwithoutchangingthemodelingapproach—would likely narrow the gap between the 93% cross-validation accuracy and the 81.87% test accuracy observed in Chapter 8, since that gap is consistent with a moderate-sized dataset rather than with a flaw in the chosen algorithms.Third,wrappingtheexistingpickledmodelandencodersinalightweightFlaskorStreamlitservice, rather than leaving them as notebook artifacts, would convert this dissertation's pipeline into something a collaboratingclinicianorresearchercouldrealisticallytrywithoutneedingtoreadtheunderlyingPythoncode —anaturalcomplementtothestandaloneweb-basedscreeninginterfacealreadybuiltaspartofthisproject (Chapter 9).

Final Remarks

Inconclusion,thisdissertationsuccessfullydemonstratesthataPython-basedmachinelearningapproachcan effectively predict autism tendencies using questionnaire and demographic data. The structured workflow, balanceddatahandling,optimizedmodelselection,andreproducibleimplementationcollectivelyestablisha robust foundationforfuture researchandreal-worlddeploymentinautism screeningandearly-intervention support systems.

Appendix: Implementation Configuration and Publication

Implementation Configuration Summary

Theentiresystemwasdevelopedin Pythonfollowingamodularstructurespanningdatapreprocessing,model training, evaluation, and model export. Table A.1 summarizes the development environment and system configuration used throughout the project.

TableA.1:ImplementationConfigurationSummary

Component	Specification
ProgrammingLanguage	Python 3.x
IDE/Environment	JupyterNotebook/VS Code
OperatingSystem	Windows10/11
Processor	IntelCorei5orhigher
RAM	8GB(minimumrecommended)
CoreLibraries	NumPy,Pandas,Matplotlib,Seaborn
MLLibraries	Scikit-learn,XGBoost
ImbalanceHandling	Imbalanced-learn(SMOTE)
ModelPersistence	Pickle(.pklfiles)
HyperparameterTuning	RandomizedSearchCV
EvaluationMetrics	Accuracy,Precision,Recall,F1-Score
Cross-Validation	K-FoldCross-Validation

Hyperparameter tuning targeted parameters such as the number of estimators, maximum tree depth, learningrate(forXGBoost),andminimumsamplesrequiredtosplitanode.Thebest-performingconfiguration foreachmodelwasselectedbasedoncross-validationaccuracy,andthefinalartifacts—best_model.pkland encoders.pkl — were saved for direct reuse in any downstream deployment.

Software Work flow Summary

1. ImportrequiredPythonlibraries.
2. LoadthedatasetfromthesourceCSVfile.
3. Performexploratorydataanalysis(EDA).
4. Preprocessthedata(cleaningandencoding).
5. Splitthedatasetintotrainingandtestingsets.
6. ApplySMOTEforclassbalancingonthetrainingpartition.
7. Trainmultiplecandidateclassificationmodels.
8. PerformhyperparametertuningviaRandomizedSearchCV.
9. Evaluateeachmodelusingstandardperformancemetrics.
10. Savethebest-performingmodelanditsencodersforfuturepredictions.

Journal Publication

The research presented in this dissertation was published as a peer-reviewed journal article, summarized in Table A.2.

TableA.2:Journal Publication Details.

Field	Detail
PaperTitle	AutismPredictionUsingMachineLearning
Authors	AyanChakraborty,MahuyaSasmal
Journal/ ISSN	IJSRED,ISSN2581-7175
PublishedIssue	Volume9,Issue3
YearofPublication	2026
UniqueID	IJSRED-V9I3P128

The published paper reproduces the dissertation' score experimental comparison across Decision Tree, KNN, SVM, Random Forest, and XGBoost classifiers, and reaches the same central conclusion: ensemble learning methods, particularly Random Forest and XGBoost, provide the strongest, most generalizable performance forautismscreeningfromstructuredquestionnaireanddemographicdata,whileremainingcomputationally lightweight enough for real-world deployment.

REFERENCES

1. RajeshPrasadetal.,“AutismSpectrumDisorderPredictionUsingMachineLearningandDesignScience,” Int. J. Exp. Res. Rev., Vol. 39 (2024).
2. YangDing,HengZhangandTingQiu,“Deeplearningapproachtopredictautismspectrumdisorder:a systematic review and meta-analysis.”
3. RajeshPrasad,FaridaMusa,HadizaMuhammadAhmad,SantoshKumarUpadhyayandBirendraKumar Sharma, “Autism Spectrum Disorder Prediction Using Machine Learning and Design Science.”
4. SumanRaj,SarfarazMasood,“AnalysisandDetectionofAutismSpectrumDisorderUsingMachine Learning Techniques,” Procedia Computer Science 167 (2020), pp. 994–1004.
5. ShyamSundarRajagopalan,YaliZhang,AshrafYahia,KristiinaTammimies,“MachineLearningPrediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information.”
6. ShanthiSelvaraj,PoonkodiPalanisamy,SummiaParveen,andMonisha,“AutismSpectrumDisorder Prediction Using Machine Learning Algorithms,” ICCVBIC 2019.
7. Muhammad Shoaib Farooq, Rabia Tehseen, Maidah Sabir and Zabihullah Atal, “Detection of autism

spectrumdisorder(ASD)inchildrenandadultsusingmachinelearning,"DOI:10.1038/s41598-023-35910-1.

8. KoushikChowdhury,MirAhmadIraj,"PredictingAutismSpectrumDisorderUsingMachineLearning Classifiers," DOI: 10.1109/RTEICT49044.2020.9315717.
9. AmeeraS.Jaradat,MohammadWedyan,SajaAlomari,MalekMahmoudBarhoush,"UsingMachine Learning to Diagnose Autism Based on Eye Tracking Technology," DOI: 10.3390/diagnostics15010066.
10. KanimozhiA,DhanasriA,"AutismSpectrumDisorderPredictionbyFacialRecognitionUsing Deep Learning," ISSN: 2320-2882.
11. Khandaker Mohammad Mohi Uddin, Hasibur Rahman, Mahadi Hasan, Fatema Akter, Suman Chandra Das,"Amachinelearningapproachtopredictautismspectrumdisorder(ASD)forbothchildrenandadults using feature optimization," ISSN 2220-8879.
12. M.ShuaibQureshi,MdBilalQureshi,J.Asghar,F.Alam,"PredictionandAnalysisofAutism Spectrum Disorder Using Machine Learning Techniques," DOI: 10.1155/2023/4853800.
13. S.Yuvaraj,S.Sugavanaesh,C.TharunKumar,G.Saran,R.Subha,"IdentificationofAutismSpectrum Disorder (ASD) in Adults through Various Machine Learning Algorithms," DOI: 10.1109/IDCIoT59759.2024.10467893.
14. SiddhantDudhane,SrijanMohan,PrajaktaSahu,UshasukhanyaS,"AutismSpectrumDisorder (ASD)
15. ScreeningandDetectionusingMachineLearningTechniques,"DOI:10.1109/ICNWC60771.2024.10537301.
16. D.AarthiandS.Kannimuthu,"AComprehensiveAnalysisofAutismSpectrumDisorderUsing Machine Learning Algorithms: Survey," DOI: 10.1007/978-981-99-7216-6_20.
17. LongjieJin,HualeiCui,PeiyuanZhangandChunquanCai,"Earlydiagnosticvalueofhomevideo-based machine learning in autism spectrum disorder: a meta-analysis," DOI: 10.1007/s00431-024-05837-4.