

**A Comparative Analysis of Machine Learning models in the Classification Techniques
in the Prediction of Autism in Children**

**Ebierimunu ABULE
LCU/PG/002902**

**Being a MSc Viva Presentation to the Department of Computer Science, Faculty of Natural
& Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria**

**In Partial Fulfillment of the Requirements for the Award of Master of Science Degree (MSc)
in Computer Science**

Certification

This is to certify that this study titled “A Comparative Analysis of Machine Learning models in the Classification Techniques in the Prediction of Autism in Children” was conducted by Ebierimunu ABULE with Matric No. LCU/PG/002902 for the award of Master of Science Degree (MSc) in Computer Science in the Department of Computer Sciences, Faculty of Natural and Applied Sciences, Lead City University Ibadan, Oyo State, Nigeria under my supervision, and that this has not been previously submitted.

.....

Dr. Ayoade Akintayo

Supervisor

.....

Date

.....

Dr. V.B Oyekunle

Head of Department

.....

Date

Dedication

This thesis is dedicated to Almighty God.

Lead City University Ibadan DO NOT COPY

Acknowledgement

The Researcher wish to acknowledge and express my sincere gratitude to the Management of Lead City University Ibadan, Oyo State for their support, aligned academic facilities, and ambient environment provided to facilitate the achievement of this programme.

I also in no small measure appreciated the contribution and effort of my able supervisor Dr. Ayoade Akintayo for his unrelenting support and constructive criticisms in ensuring that this project is actualized. Thank you, sir, for your fatherly love. May Almighty God continue to bless you and may he rescue you in point of need.

I cannot help but appreciate the effort and support of the HOD Dr. Sakpere Wilson. I equally appreciate the other members of lecturers in the department of computer science department, Lead City University, Ibadan for their impacts in my academic journey.

Also, deepest appreciation goes to my parents Mr. and Mrs. Lawrence Abule. To my guardian Mr and Mrs. Capt. Wilson Theophilus, to my siblings Engr K. A. Abule, Opuoyere L. Abule, Pharm Ayibanua Theophilus and host of others for their support and friendship role displayed. My God Almighty direct our part and show us the way to success.

Although, all the aforementioned Institutions and persons mentioned helped in one way or the other, meanwhile all errors in the research report if found are solely mind.

Abstract

Autism Spectrum Disorder (ASD) is a mild cognitive impairment known to affect a person's language, communication, thought process and social behaviour. Some recent advances in machine learning made it possible to predict ASD using behavioural, demographic, and medical data, but the choice of optimal algorithms and feature selection techniques remains an open research question. This study used the ASD for children dataset obtained from the University of California, Irvine repository to investigate the performance of the selected machine learning models through a comparative analysis. Eight nonparametric models were compared in total, with four models SVM, KNN, GNB, and MLP being base learners, and the other four ensemble methods RF, XGBoost, Bootstrap Aggregating (Bagging), and Stacking. Hyperparameters of the models were autotuned, and GridSearchCV was employed to choose the model with the best combination of parameters. All base models performed brilliantly, with SVM being the best and most precise in this experiment with a result of 1, amongst all classifiers even before hyperparameter tuning. After hyperparameter tuning, SVM and MLP ranked highest with all metrics returning scores of 100%. KNN returned 96.2% accuracy and recall and GNB trailed at 95.26% for both metrics for the ensemble methods, XGB and Bagging with SVM performed best at 100% for all metrics. RF produced an accuracy and recall of 98.1% while Bagging with KNN and GNB had both metrics at 95.2%. The classification performance of Bagging with MLP and the stacking classifier were equivalent across all metrics though, producing accuracy and recall of 99.5%. Based on the findings, it is recommended that thorough feature selection should be conducted to eliminate features that could lead to overfitting, such as "Q_Chat_10" in this study cross-validation technique was used to assess the generalizability of models to unseen data.

Keywords: Accuracy, Classifier, Defaults, Financial, Models, Predicting, Validation

Word Count: 300

Table of Contents

Content	Page
Title Page	i
Certification	ii
Dedication	iii
Acknowledgement	iv
Abstract	v
List of Tables	ix
List of Figures	x
List of Acronyms	xi
Chapter One: Introduction	
1.1. Background to the Study	1
1.2. Statement of the Problem	6
1.3. Aim and Objectives of the Study	7
1.4. Research Questions	7
1.5. Significance of the of Study	8
1.6. Scope of the Study	8
1.7. Limitations of the Study	9
1.8. Operational Definition of Terms	9
Endnotes	

Chapter Two: Literature Review

2.1.	Conceptual Review	13
2.1.1.	Autism Spectrum Disorder (ASD)	13
2.1.2.	Autism SPECTRUM Disorder (ASD) in Children and Children	15
2.1.3.	Detection of Autism Spectrum Disorder (ASD)	16
2.1.4.	Machine Learning	18
2.1.4.1	Types of Machine Learning	19
2.1.5	Classification of ASD Using Machine Learning	28
2.2.	Theoretical Framework	28
2.2.1	Support Vector Machine (SVM)	28
2.2.1.1	Hyperparameters of SVM	30
2.2.2	K-Nearest Neighbors (KNN)	31
2.2.2.1	Hyperparameters of KNN	33
2.2.3	Gaussian Naive Bayes (GNB)	33
2.2.3.1	Hyperparameters of a GNB Model	35
2.2.4	Rain Forest	36
2.2.4.1	Hyperparameters of an RF model	37
2.2.5	Multilayer Perceptron (MLP)	39
2.2.5.1	Hyperparameters of a Multilayer Perceptron	41
2.2.6	Ensemble Techniques	41
2.2.7	Model Evaluation	45
2.2.7.1	Cross Validation	45
2.2.7.2	Accuracy	47
2.2.7.3	Balanced Accuracy	48

2.2.7.4 Precision	49
2.2.7.5 Recall (Sensitivity)	50
2.2.7.6 Specificity	50
2.2.7.7 F1 Score	51
2.2.7.8 Receiver Operating Characteristic (ROC) Curve	51
2.2.7.9 Precision–Recall (PR) Curve	52
2.3 Review of Empirical Studies	52
2.4 Summary and Gap in Literature Reviewed	74

Endnotes

Chapter Three: Methodology

3.1. Research Approach	92
3.2 Requirements Specification	92
3.3. Research Design	93
3.3.1 Data Collection	93
3.3.2 Data Preprocessing	95
3.3.3 Dataset Splitting	96
3.3.4 Evaluating the Performance of the Models	96
3.3.5 Cross Validation	100
3.3.6 Model Retraining	100
3.3.7 Model Evaluation	101

Endnotes

Chapter Four: Results and Discussion of Findings

4.1 Result on Data Preprocessing	103
----------------------------------	-----

4.1.1	Data Importation and Correlational Analysis	103
4.1.2	Data Partitioning	105
4.2	Cross Validation	105
4.3	Application of Classification Algorithm	106
4.4	Classification Results	106
4.5	Discussion of Findings	113

Endnotes

Chapter Five: Conclusion

5.1.	Summary of Findings	118
5.2.	Conclusion	119
5.3	Recommendations	121
5.4	Contribution to Knowledge	122
5.5	Suggested Area of Further Studies	123

Bibliography	125
---------------------	-----

Appendices	147
-------------------	-----

Bio-data

148	
The University Compliance Certification	145

List of Tables

Table	Title	Page
--------------	--------------	-------------

2.1	A Confusion Matrix Layout	46
4.1	Model Runtime for Hyperparameter Tuning	113

Lead City University Ibadan DO NOT COPY

List of Figures

Figure	Title	Page
--------	-------	------

3.1	Design Framework	92
3.2	Conceptual Framework	93
4.1	First Correlation Heat Map	104
4.2	Second Correlation Heat Map	104
4.3	SVM Confusion Matrix	106
4.4	KNN Confusion Matrix	106
4.5	GNB Confusion Matrix	108
4.6	MLP Confusion Matrix	108
4.7	RF Confusion Matrix	109
4.8	XGB Confusion Matrix	127
4.9	Bagging with SVM	110
4.10	Bagging with KNN	110
4.11	Bagging with GNB	111
4.12	Bagging with MLP	111
4.13	Stacking Classifier	111
4.14	Box Plot Comparing the Stacking Classifier to Base Learners	112
4.15	Optimal Model Parameters after Tuning of Hyperparameters	115
4.16	Evaluation Metrics of Each Model	116
4.17	Grouped Bar Chart Comparing Evaluation Metrics of Classifiers	116

List of Acronyms

Abbreviation	Meaning
ABIDE	Autism Brain Imaging Data Exchange
ADOS	Autism Diagnostic Observation Scale

AI	Artificial Intelligence
ASD	Autism Spectrum Disorder
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
DL	Deep Learning
EEG	Electroencephalogram
fMRI	Functional Magnetic Resonance Imaging
GNB	Gaussian Naive Bayes
KNN	K-Nearest Neighbours
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NHS	National Health Service
RBF	Radial Basis Function
RF	Random Forest
ROC	Receiver Operator Characteristic
SVM	Support Vector Machine
UCI	University of California, Irvine
XGBoost	Extreme Gradient Boosting