**OPTIMIZATION OF SCHIZOPHRENIA DIAGNOSIS AND MANAGEMENT USING DATA MINING ALGORITHMS**

**BY**

**OLALEKAN RAHMAN ABIODUN**

**050409028**

**MSC PROJECT (yet defended)**

**BIOMEDICAL ENGINEERING**

**UNIVERSITY OF LAGOS NIGERIA**

**MARCH 2019**

**DEDICATION**

This project is dedicated to the Almighty for His total support.

**ACKNOWLEDGEMENT**

First, I would like to acknowledge the unfaltering support from my project advisor, Dr. Ephraimo Nwoye. I am grateful to have had him as a mentor through this experience. He has been supportive in every aspect of the development of this project. I would also want to thank all the lecturers in the Department of Biomedical Engineering of the University of Lagos for their dedication to the success of the programme. My gratitude also goes to the Departments of Psychiatry, Health Records Lagos University Teaching Hospital, Idi-Araba, Lagos, Nigeria, for their immense contributions and support towards this achievement.

Thank you to my classmates and friends in the department and beyond. And finally, I would like to appreciate the support of my family members towards the completion of the project.

Thank you all!

**CERTIFICATION**

This is to certify that this project titled **“OPTIMIZATION OF SCHIZOPHRENIA DIAGNOSIS AND MANAGEMENT USING DATA MINING ALGORITHMS”** will be carried out under my supervision.

………………………………………….

Dr. E.O. Nwoye

Department of Biomedical Engineering

School of Postgraduate Studies

University of Lagos

Nigeria.

**TABLE OF CONTENT**

Dedication ii

Acknowledgement iii

Certification iv

List of Figures vii

List of Tables viii

Abstract ix

Chapter 1: 1

* 1. Introduction 1
  2. Schizophrenia Diagnostic Process 5
  3. Schizophrenia Diagnostic Criteria 6
  4. Management Procedures of Schizophrenia 6

1.4.1.. Treatment Based on Medication 7

1.4.2. Psychosocial Intervention 7

1.4.3. Service Level Intervention 8

1.5. Data Mining Techniques 8

1.6. Statement of Problem 9

1.7 Aims and Objectives 10

1.8. Hypotheses 11

1.9. Scope 11

1.10. Ethical Approval 11

1.11. Novelty 12

Chapter 2: Literature Review 13

Chapter 3: 19

3.0 Materials and Methodology 19

3.1. Data Acquisition 19

3.2. DSM-V Schizophrenia Diagnostic Criteria 22

3.3. WEKA Tools 22

3.4. J48 Decision Tree Classifier 23

3.4.1. Theory behind Decision Making of J48 classifier 24

3.4.1 .1 The Splitting Criterion 24

3.4.1.2 Stopping Criterion 25

3.4.1.3 Decision Tree Pruning 26

3.5 Zero Classifier 27

3.6 Naïve Base Classifier 27

3.7 Pearson’s Correlation 27

3.8. Methodology 30

3.8.1 Flow Chart Showing Methodology 30

3.8.2. Data preparation, exploration and visualization 31

3.8.3. Feature Selection 33

3.8.4 Class Distribution in dataset 35

3.9. Model Development 35

3.10. Performance Statistical Parameters 38

3.10.1 Confusion Matrix 38

3.10.2. Sensitivity 39

3.10.3. Specificity 39

3.10.4. Diagnostic Odd Ratio 39

3.10.5 ROC Curves 40

3.10.6 Area Under ROC Curve 41

Chapter 4: 42

4.1. Results and Discussions 42

4.2 Validation of Model 44

4.3. Discussion 50

4.4. Proposed Optimal Schizophrenia Treatment Regime 51

4.4.1. Rational for Selecting Drugs in Algorithms 56

Chapter 5: 57

5.1. Conclusion 58

* 1. Limitations of the CS J48 Model
  2. 5.3 Recommendations 58

Appendix 1: Sequence of Codes for the Project Work in Jupiter Notebook 60

Appendix 2: CS J48 Model 68

Appendix 3: Cost-sensitive J48 Model tree 69

Appendix 4: Naïve Bayes Model 70

Appendix 5: J48 Model 74

Appendix 6 : Guide To Use the Model by Any Psychiatrist for

Schizophrenia Diagnosis 75

References 79

**List of Figures**

1. Figure 1: Features of Schizophrenia 5
2. Figure 2: Shows WEKA GUI Chooser 23
3. Figure 3: shows a Decision Tree being pruned 28
4. Figure 4: shows Flow Chat of the Methodology 30
5. Figure 5: shows heatmap showing missing value positions in white stripes 32
6. Figure 6: shows Pearson’s Correlation values between each pair of Attributes 34
7. Figure 7: shows the stack bar charts of distributions in Class 35
8. Figure 8: shows cost matrix optimizes prediction performance of J48 Model 37
9. Figure 9: shows ROC curve with perfect discrimination 40
10. Figure 10 :Bar Chart comparing accuracy of each model 45
11. Figure 11: shows Bar Chart comparing Sensitivity of each model 46
12. Figure 12 : shows Bar Chart showing Specificities of Models 46
13. Figure 13 : Bar Chart comparing ROC Area of each model 47
14. Figure 14 : Bar Chart comparing DOR of each model 47
15. Figure 15 : ROC curve of CS J48 model 48
16. Figure 16 : 16 shows ROC curve of J48 model 48
17. Figure 17: shows ROC curve of Naïve Bayes model 49
18. Figure 18 :shows Cost-sensitive J48 Model tree 49
19. Figure 19: shows EUFEST Algorithm for Optimal Treatment ofSchizophrenia 54
20. Figure 20: shows Adapted Algorithm for Optimal Diagnosis and Management ofSchizophrenia 55
21. Figure 21 :shows Cost-sensitive J48 Model tree 69
22. Figure 22: shows WEKA Explorer Environment 76
23. Figure 23 shows WEKA Object Pane 77
24. Figure 24 shows WEKA Classifier Errors Pane 78

**List of Tables**

1. Table 1 shows Representation of features in WEKA API and Description 19
2. Table 2 shows DSM-5 Criteria for Schizophrenia Diagnosis 22
3. Table3: shows Dataset Description 31
4. Tabl**e** 4 shows some section of dataset after replacing missing values33
5. Table 5 shows a typical Confusion matrix for a classifier 38
6. Table 6 Confusion Matrix of pruned J48 Model 42
7. Table 7 Confusion Matrix of cost-sensitive J48 pruned (CS J48) Model 43
8. Table 8 Confusion Matrix of ZeroR Model 43
9. Table 9 Confusion Matrix of Naïve Bayes Model 43
10. Table 10 shows Accuracies of Models 44
11. Table 11 shows Sensitivities of Models 44
12. Table 12 shows Specificities of Models 44
13. Table 13 shows DORs of Models 45
14. Table 14 shows ROC areas 45
15. Table 15 shows Feature Vector Sample 75

**ABSTRACT**

Schizophrenia is a severe debilitating cognitive disorder with a heterogeneous path and symptom profile characterized by positive and negative symptoms. The disease affects about 70 million of global population with no imminent cure in sight. Early diagnosis of Schizophrenia has been an effective tool towards managing the disease. This study aims at optimizing diagnosis and management of the Schizophrenia via automation using data mining techniques to analyze psychometric profiles of patients collected by psychiatrists. Machine Learning Classification algorithms have been reported to be reasonably accurate in medical diagnosis of diseases. For this study medical records were collected from the Record Office of Department of Psychiatry, Lagos University Teaching Hospital, Lagos, Nigeria. The unstructured paper records were converted on Ms.Excel and loaded on WEKA API via Python-WEKA wrapper on Jupiter Notebook for pre-processing and model development. The process was also repeated on WEKA Explorer GUI for easy use by psychiatrists. “ReplaceMissingValue” filter was implemented on the dataset to replace missing values in the data with modal value of corresponding column, and the process was followed by building of cost-effective J48 model for predicting Schizophrenia using the obtained dataset. Algorithm training was executed with 10-fold Cross-validation. K-fold Cross-validation partitions the dataset into k folds and selects k-1 folds of labeled instances for algorithm training and one fold for testing. The process is repeated until when each of the k folds has been used for testing. This process reduces errors due to effect of sampling bias in dataset. It also assists in effective algorithm training when data is not large enough. A cost was also awarded against the model’s False Negative prediction to increase model’s recall. Performance of cost-sensitive J48 was compared with those of Baseline algorithm, ZeroR, and conditional probability-based Naïve Bayes on the basis of Sensitivity, Selectivity and Diagnostic Odd Ratio, Receiver Operating Curve (ROC) and area under ROC (AUC). This was done to ensure that the model is reliable. The cost-sensitive J48 model developed performed reasonably well in predicting Schizophrenia at an accuracy of 78 per cent. This research ultimately presents a quicker and more efficient tool for diagnosing Schizophrenia; it also presents an optimized Workflow that could assist psychiatrists in scheming timely management plan for treatment at reduced cost.