

CONTENTS

CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF APPENDIX	xiv
STATEMENT	iii
PREFACE	iv
ABSTRACT	xv
Chapter 1 INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Scope	3
1.4 Research Objective	4
1.5 Research Benefits.....	4
1.6 Research Methodology	5
1.7 Structure of the Document	5
Chapter 2 LITERATURE REVIEW	7
Chapter 3 THEORETICAL BASIS	12
3.1 Alzheimer.....	12
3.2 Binary Classification.....	12
3.3 Decision Tree	12
3.3.1 Criterion	16
3.3.2 Maximum Depth	17
3.3.3 Minimum Samples Leaf.....	17
3.3.4 Max Leaf Nodes	17
3.3.5 Maximum Features.....	17
3.3.6 Pruning	18

3.4	Data Balancing Techniques	18
3.4.1	Oversampling	18
3.4.2	Undersampling/Downsampling.....	21
3.5	Hyperparameter Tuning	28
3.6	Evaluation Metric.....	28
3.6.1	Confusion Matrix	28
3.6.2	Precision	29
3.6.3	Recall.....	29
3.6.4	F1-Score	29
3.6.5	F2-Score	30
3.7	K-Fold Cross Validation	31
Chapter 4 RESEARCH METHODOLOGY		32
4.1	Research Description	32
4.2	Research Phases	33
4.2.1	Data Collection and Reading	33
4.2.2	Data Cleaning.....	34
4.2.3	Data Preprocessing.....	42
4.2.4	Model Development and Training	43
4.2.5	Model Evaluation	43
Chapter 5 IMPLEMENTATION		44
5.1	Data Collection	44
5.2	Data Cleaning & Exploratory	45
5.3	Data Preprocessing.....	48
5.4	Data Balancing.....	48
5.5	Model Development & Training.....	50
5.6	Model Evaluation.....	51
Chapter 6 RESULTS AND DISCUSSION		54
6.1	Data Cleaning and Preprocessing Results.....	54
6.2	Model Training and Tuning Results	55
Chapter 7 CONCLUSION AND FUTURE WORKS.....		c81



7.1 Conclusion	81
7.2 Future Works	81
REFERENCES.....	83
APPENDIX.....	89

LIST OF FIGURES

Figure 1.1 Dataset Distribution.....	3
Figure 3.1 Oversampling Illustration.....	18
Figure 3.2 SMOTE Illustration.....	20
Figure 3.3 Undersampling Illustration.....	21
Figure 3.4 Centroid-based Undersampling Illustration.....	22
Figure 3.5 ENN Illustration.....	23
Figure 3.6 Tomek Links Illustration.....	25
Figure 3.7 SMOTE+ENN Illustration.....	26
Figure 3.8 SMOTE Tomek Illustration.....	27
Figure 4.1 Research Phases.....	33
Figure 4.2 Original Dataset Distribution.....	34
Figure 4.3 Data Cleaning Steps.....	42
Figure 6.1 Visual Representation of Data Distribution Following Re- Classification: (a) before the re-classification and (b) after the re- classification.....	54
Figure 6.2 Training and Testing Class Distribution Before ROS.....	57
Figure 6.3 Training and Testing Class Distribution After ROS.....	57
Figure 6.4 Training and Testing Class Distribution Before SMOTE.....	60
Figure 6.5 Training and Testing Class Distribution After SMOTE.....	60
Figure 6.6 Training and Testing Class Distribution Before RUS.....	62
Figure 6.7 Training and Testing Class Distribution After RUS.....	62
Figure 6.8 Training and Testing Class Distribution Before ENN.....	64
Figure 6.9 Training and Testing Class Distribution After ENN.....	65
Figure 6.10 Training and Testing Class Distribution Before Tomek Links.....	67
Figure 6.11 Training and Testing Class Distribution After Tomek Links.....	67
Figure 6.12 Training and Testing Class Distribution Before Cluster Centroids...	69



Figure 6.13 Training and Testing Class Distribution After Cluster Centroids	70
Figure 6.14 Training and Testing Class Distribution Before SMOTEENN	72
Figure 6.15 Training and Testing Class Distribution After SMOTEENN	72
Figure 6.16 Training and Testing Class Distribution Before SMOTE-Tomek.....	74
Figure 6.17 Training and Testing Class Distribution After SMOTE-Tomek	75
Figure 6.18 Scatter Plot of Model Recall vs F2-Score	78

LIST OF TABLES

Table 2.1 Comparison of related works	8
Table 4.1 Some Examples of the Dataset.....	34
Table 4.2 Transformation of Origin Classes in the Dataset: Before and After Reclassification	35
Table 4.3 Irrelevant Columns and Their Respective Description	37
Table 4.4 Final Dataset Columns	38
Table 4.5 Final Dataset Description.....	39
Table 4.6 Experiment Scenario	43
Table 6.1 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the Decision Tree Baseline Model.....	56
Table 6.2 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the Random Oversampling Decision Tree.....	58
Table 6.3 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the SMOTE Decision Tree	61
Table 6.4 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the RUS Decision Tree	63
Table 6.5 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the ENN Decision Tree.....	65
Table 6.6 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the Tomek Links Decision Tree.....	68
Table 6.7 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the Cluster Centroids Decision Tree.....	70
Table 6.8 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the SMOTEENN Decision Tree	73
Table 6.9 Top 10 Recall with its Respective F2- Score and Top F2-Score with its Respective Recall of the SMOTETomek Decision Tree	75
Table 6.10 The Best Hyperparameter Combinations based on Recall.....	77



Table 6.11 Top 5 Confusion Matrix of Baseline Model based on Recall..... 79

Table 6.12 Top 5 Confusion Matrix of ENN Model based on Recall 79

LIST OF APPENDIX

Appendix 1 Baseline Model Full Results	89
Appendix 2 Random Oversampling Full Results	91
Appendix 3 SMOTE Full Results	93
Appendix 4 Random Undersampling Full Results	96
Appendix 5 ENN Full Results	98
Appendix 6 TomekLinks Full Results	101
Appendix 7 Cluster Centroids Full Results.....	103
Appendix 8 SMOTEENN Full Results	106
Appendix 9 SMOTETomek Model Full Results	108
Appendix 10 Data Distribution Before and After Data Balancing Techniques..	111
Appendix 11 Confusion Matrix for the Baseline Model.....	111
Appendix 12 Confusion Matrix for the ENN Model	123