CLASSIFICATION OF HUNTINGTON'S DISEASE USING MACHINE LEARNING APPROACHES

JELIZA JUSTINE A/P SEBASTIN

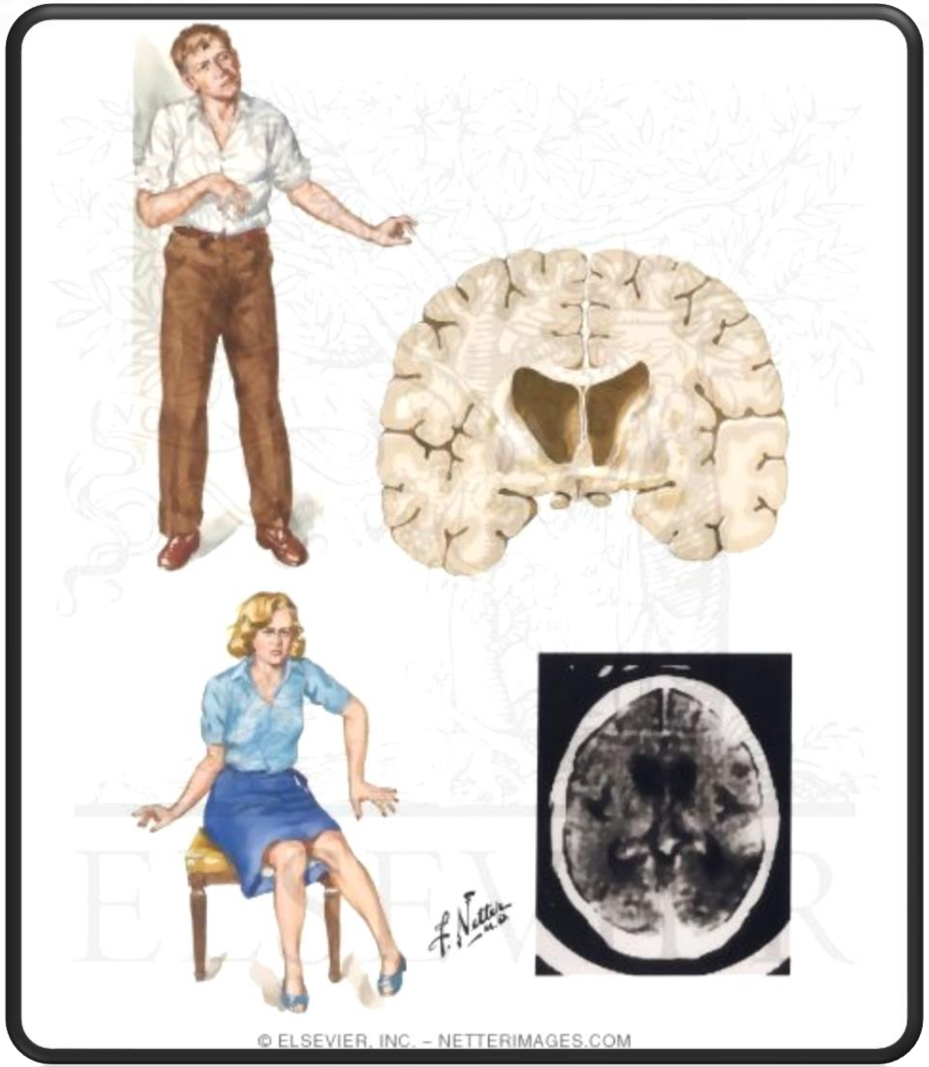
SUPERVISOR: DR. HASLINA BINTI HASHIM

Innovating Solutions

TABLE OF CONTENTS

- INTRODUCTION
- LITERATURE REVIEW
- METHODOLOGY
- RESEARCH DESIGN
- CONCLUSION

INTRODUCTION



© ELSEVIER, INC. – NETTERIMAGES.COM

- Huntington’s disease is a severe neurodegenerative disorder characterized by motor, cognitive, and psychotic symptoms.
- It is caused by a mutation in the HTT gene, leading to the production of a toxic mutant huntingtin protein.
- The disease results from an expansion of the CAG trinucleotide repeat in the huntingtin gene.
- Early detection and personalized treatments are essential for patient support and timely interventions.
- Conventional machine learning algorithms have moderate success in disease classification, so, this research will highlight on classification algorithms that show superior performance in classification task for Huntington’s disease.

Huntington’s disease is a fatal genetic disorder that destroys nerve cells in the brain, impairing movement, speech, and thinking.

The disease results from mutations in a gene on chromosome 4, leading to the buildup of harmful proteins in the brain.

Understanding the genetic basis of Huntington’s disease is crucial for developing targeted treatments and interventions.

There is a global health crisis due to the high incidence of Huntington’s disease, highlighting the need for better treatment options.

PROBLEM BACKGROUND

PROBLEM STATEMENT

There is an abundance of biological data, but precise and effective identification methods are needed.

Current diagnostic techniques lack sensitivity and specificity, leading to delays in intervention and suboptimal patient outcomes.

Reliable classification models are needed for early identification and classification of Huntington's disease.

Aim

The aim of this research is to develop and identify the most effective classification model for Huntington's disease by comparing advanced machine learning techniques, specifically bagged ensemble learning, generalized linear models, and decision trees.

RESEARCH OBJECTIVES

RO1 To investigate the methods that can be used to classify Huntington’s disease.

RO2 To implement bagged ensemble learning, generalized linear model and decision tree in the classification of Huntington’s disease.

RO3 To evaluate the performance of bagged ensemble learning, generalized linear model and decision tree in the classification of Huntington's disease.

RESEARCH SCOPES

01 Implementation of bagged ensemble learning algorithms, generalized linear models, and decision trees for building the classification models for Huntington's disease.

02 Evaluation and comparison of the performance of the developed models in terms of accuracy, precision, recall, and other relevant metrics.

03 Model training, validation, and optimization using cross-validation techniques for each classification method.

04 Comprehensive statistical analysis of model performance, including comparisons with baseline models, to identify the most effective classification method.

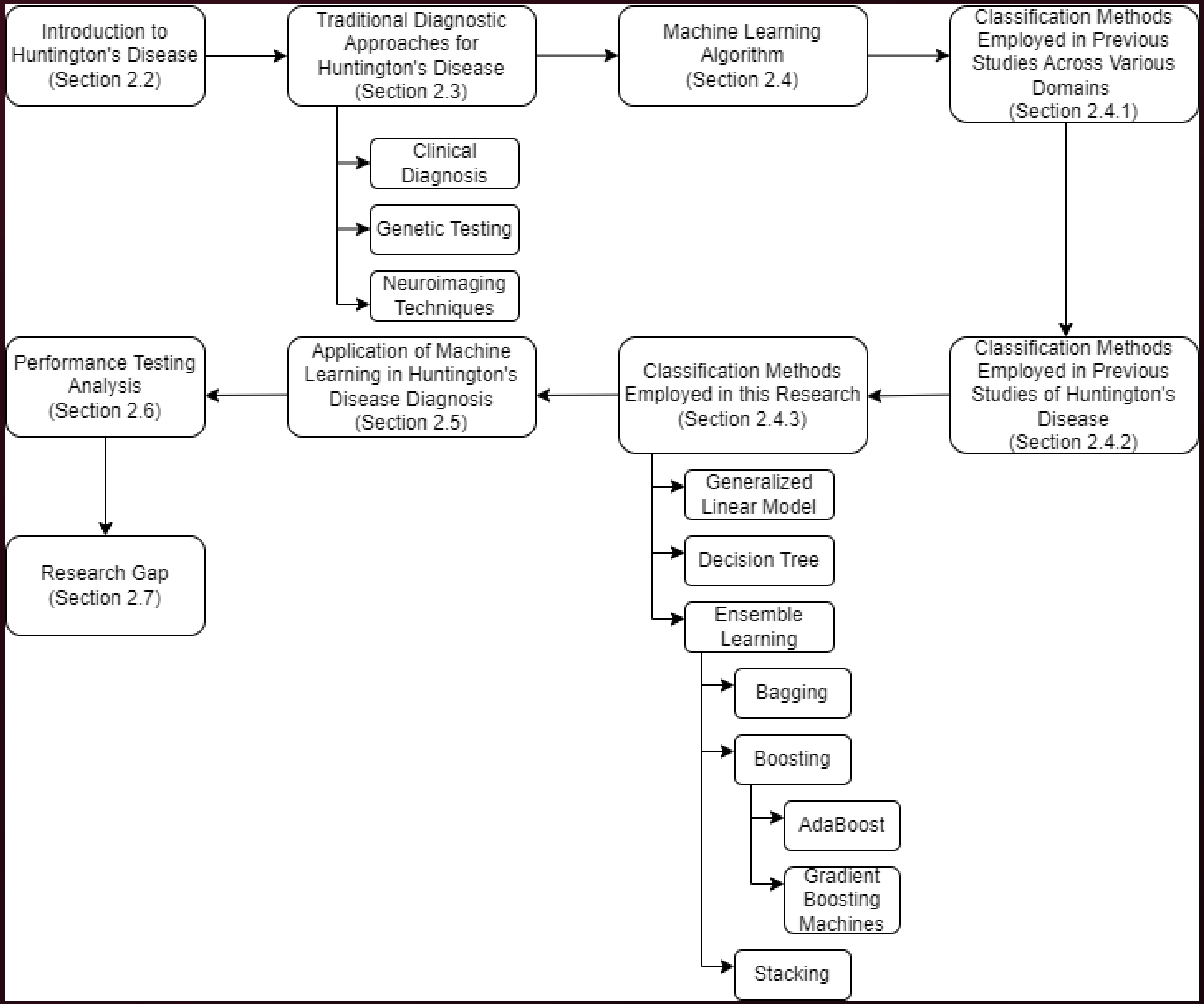
05 Feature selection and engineering to identify informative biomarkers and clinical variables associated with Huntington's disease, enhancing the classification accuracy of the models.

06 The dataset used in this study was obtained from the following source: TRACK-HD study, which is a multinational longitudinal observational study (Wiecki et al., 2016).

RESEARCH IMPORTANCE

- Addressing the need for precise and timely identification of individuals with Huntington’s disease.
- Utilizing advanced machine learning techniques is key to this research.
- Early detection allows for prompt intervention and individualized treatment plans.
- Improving patient outcomes and quality of life is a critical goal of this research.

CHAPTER 2: LITERATURE REVIEW



Introduction to Huntington's Disease (Section 2.2)

Traditional Diagnostic Approaches for Huntington's Disease (Section 2.3)

Clinical Diagnosis

Genetic Testing

Neuroimaging Techniques

Machine Learning Algorithm (Section 2.4)

Classification Methods Employed in Previous Studies Across Various Domains (Section 2.4.1)

Classification Methods Employed in Previous Studies of Huntington's Disease (Section 2.4.2)

Classification Methods Employed in this Research (Section 2.4.3)

Generalized Linear Model

Decision Tree

Ensemble Learning

Bagging

Boosting

AdaBoost

Gradient Boosting Machines

Stacking

Application of Machine Learning in Huntington's Disease Diagnosis (Section 2.5)

Performance Testing Analysis (Section 2.6)

Research Gap (Section 2.7)

Comparison Table of Various Domains Classification Methods

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Title	Method	Measurement	Performance
Classification of gene expression dataset for type 1 diabetes using machine learning methods (AlFefaai and AlRashid, 2023)	Support Vector Machine (SVM)	Accuracy	89.1%
Exploring the key factors related to the risk of heart disease by applying classification methods (You, 2023)	Random Forest	Accuracy	91%
Comparative Analysis of Machine Learning Methods for Breast Cancer Classification in Genetic Sequences (Kurian and Jyothi, 2022)	Gradient Boosting	Accuracy	95.82%
Machine Learning Algorithms for the diagnosis of Alzheimer's and Parkinson's Disease (Noella and Priyadarshini, 2020)	Bagged Ensemble Learning	Accuracy, Sensitivity, Specificity, Precision	Accuracy: 90.3% Sensitivity: 0.89 Specificity: 0.92 Precision: 0.87
Classification Based on Machine Learning Methods for Identification of Image Matching Achievements (Faroek, Umar, and Riadi, 2022)	Multilayer Perceptron (MLP)	Accuracy	87.5%

Anchor Paper

Comparison Table of Various Domains Classification Methods - continue

Three language political leaning text classification using natural language processing methods (Kosiv and Yokovyna, 2022)	Logistic Regression, Ensemble models combined TF-IDF vectorization, B-NBC meta-model, and base models including SVC, NuSVC, and LR	Macro F1-score	Logistic Regression – 0.7966, Ensemble models – 0.7966
A Novel Machine Learning Framework for Prediction of Early-Stage Thyroid Disease Using Classification Techniques (Gummadi and Reddy, 2022)	Random Forest	Accuracy	97.05%
Machine Learning in Bioinformatics: New Technique for DNA Sequencing Classification (Sarkar et al., 2022)	Naïve Bayes, Support Vector Machine (SVM)	Accuracy	93.16%
Deep Learning Classification Methods Applied to Tabular Cybersecurity Benchmarks (Noever and Noever, 2021)	MobileNetV2's convolutional neural network	Accuracy	97%
Evaluating the Efficiency of the Classifier Method When Analysing the Sales Data of Agricultural Products (Wang et al., 2022)	k-nearest neighbours	Accuracy	99%

Comparison Table of Huntington’s Disease Classification Methods

Title	Methods	Measurement	Performance
Identification of contributing genes of Huntington’s disease by machine learning (Cheng et al., 2020)	Generalized Linear Model	Accuracy, Precision, and Recall Metrics	Accuracy: 97.46±3.26%4 Precision: 95.96±5.14% Recall: 99.38±1.98%
A novel and proposed triad machine learning-based fra 's disease (D	Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB)	Accuracy, Sesitivity, Specificity	Accuracy of 85.45%, Sensitivity of 78.37%, Specificity of 76.55%
Ne untington Dis	Artificial Neural Networks (ANNs)	Forecasting Precision (Accuracy), Sensitivity, Specificity	The best results were obtained with a four-layer ANN, reaching a – - Forecasting precision of 0.92 when including control, pre-manifest, and manifest cases. - Sensitivity was 0.95, - Specificity was 0.804.
Optimizing Screening for Intrastriatal Interventions in Huntington's Disease Using Predictive Models (Barrett et al., 2024)	Logistic Regression	Area Under Curve (AUC)	AUC – 85.1%
Machine learning in Huntington’s disease:exploring the Enroll-HD dataset for prognosis and driving capability prediction (Ouwerkerk et al., 2023)	Light Gradient Boosting Machine (LGBM), and Recurrent Neural Networks (RNNs)	Predicting Age at Onset (AAO), Acuuracy of Assessing Driving Capability	1. Light Gradient Boosting Machine (LGBM) – AAO - improved prognosis by 9.2% 2. 1. Recurrent Neural Networks (RNNs) - Assessing Driving Capability - accuracy of 85.2%

Method Paper

Comparison Table of Huntington’s Disease Classification Methods - continue

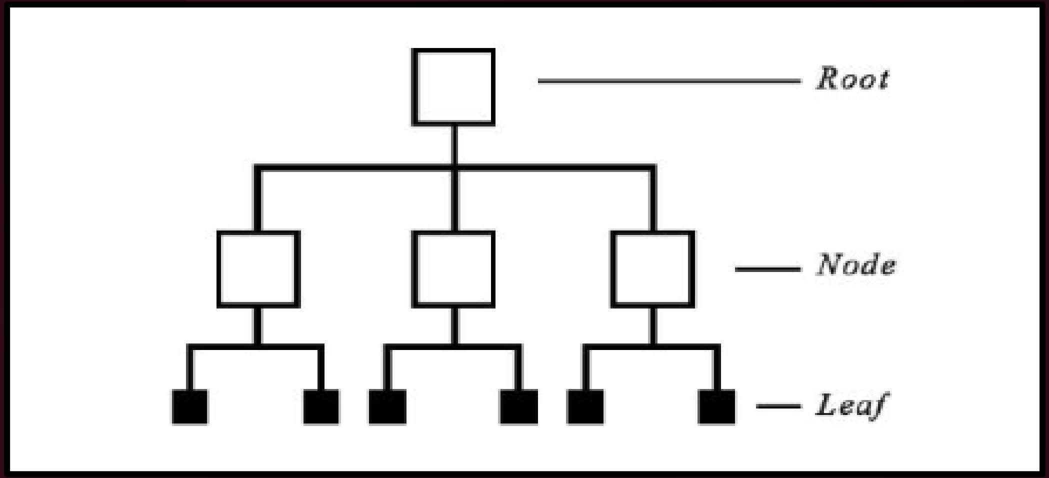
Method Paper

Predicting Severity of Huntington's Disease With Wearable Sensors (Scheid et al., 2022)	**Linear Discriminant Analysis**	**Accuracy, Sensitivity, Specificity**	**Accuracy - 96.4%** **sensitivity - 92.9%** **specificity - 100%**
assification for Disease Onset via Duality Methods (Woolnough et al., 2022)	Ball uncertainty robust SVM model (Ball-SVM)	Accuracy	95%
Exploring Huntington’s Disease Diagnosis via Artificial Intelligence Models: A Comprehensive Review (Ganesh et al., 2023)	Decision Tree	Accuracy	100%
Classification of Huntington's Disease Stage with Features Derived from Structural and Diffusion-Weighted Imaging (Lavrador et al., 2022)	Support Vector Machine (SVM)	Accuracy	85-95%
Using Machine Learning to identify microRNA biomarkers for predisposition to Juvenile Onset Huntington's Disease (Patel, Sheridan, and Shanley, 2022)	Random Forest	AUC	100%

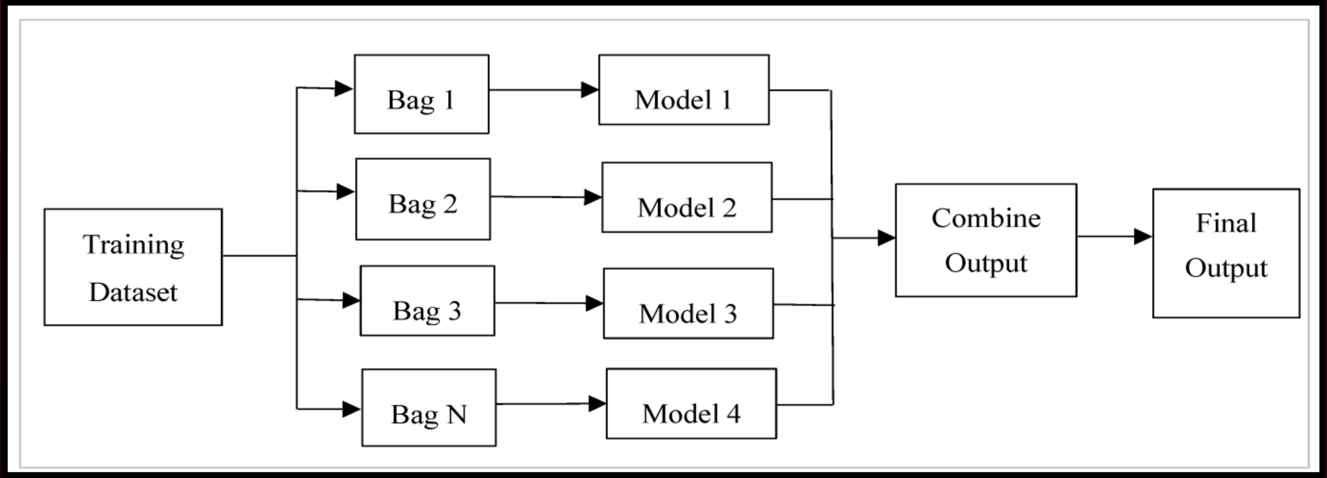
Classification Methods Employed in this Research

Machine Learning Algorithms	Advantages	Disadvantages
GLMs are parametric models that extend linear regression to accommodate response variables that follow different distributions. They are interpretable through their coefficients and require data preprocessing but are less flexible than decision trees.	• Flexible model suitable for analyzing various types of response variables and distributions. • Useful for understanding the relationship between response variables (e.g., Huntington's disease status) and predictors (e.g., gene expression profiles).	• Assumes linearity between predictors and the response variable, which may not always hold true in complex biological systems. • May not effectively capture non-linear relationships, potentially limiting its ability to model intricate interactions among genes in disease mechanisms.
A **decision tree** is a non-parametric model that uses a tree-like structure to make decisions based on feature values. It is highly interpretable but prone to overfitting and sensitive to small data changes.	• Simple to understand and intuitive. • Effective in both classification and regression tasks. • Achieved 100% accuracy in classifying gait signals in Huntington's disease patients, demonstrating robustness.	• Prone to overfitting, especially with small datasets or noisy data, leading to poorer generalization on unseen data.
Bagging (Bootstrap Aggregating) is an ensemble method that combines multiple decision trees to improve robustness and accuracy. It reduces overfitting and variance compared to a single decision tree by averaging the predictions of individual trees, though it is less interpretable than a single model.	• Reduces variance • Encourages diversity • Efficient for large datasets	• Neglects Interpretability • Potential Overfitting

DECISION TREE



BAGGED ENSEMBLE LEARNING



My Reseacrh

Dataset Paper

Paper name: A Computational Cognitive Biomarker for Early-Stage Huntington's Disease

Source: TRACK-HD study, which is a multinational longitudinal observational study (Wiecki et al., 2016).

Method Paper

-ML method in HD

Paper 1: Identification of contributing genes of Huntington's disease by machine learning (Cheng et al., 2020)
Method: Generalized linear model

Paper 2: Exploring Huntington's Disease Diagnosis via Artificial Intelligence Models: A Comprehensive Review (Ganesh et al., 2023)
Method: Decision tree

Anchor Paper

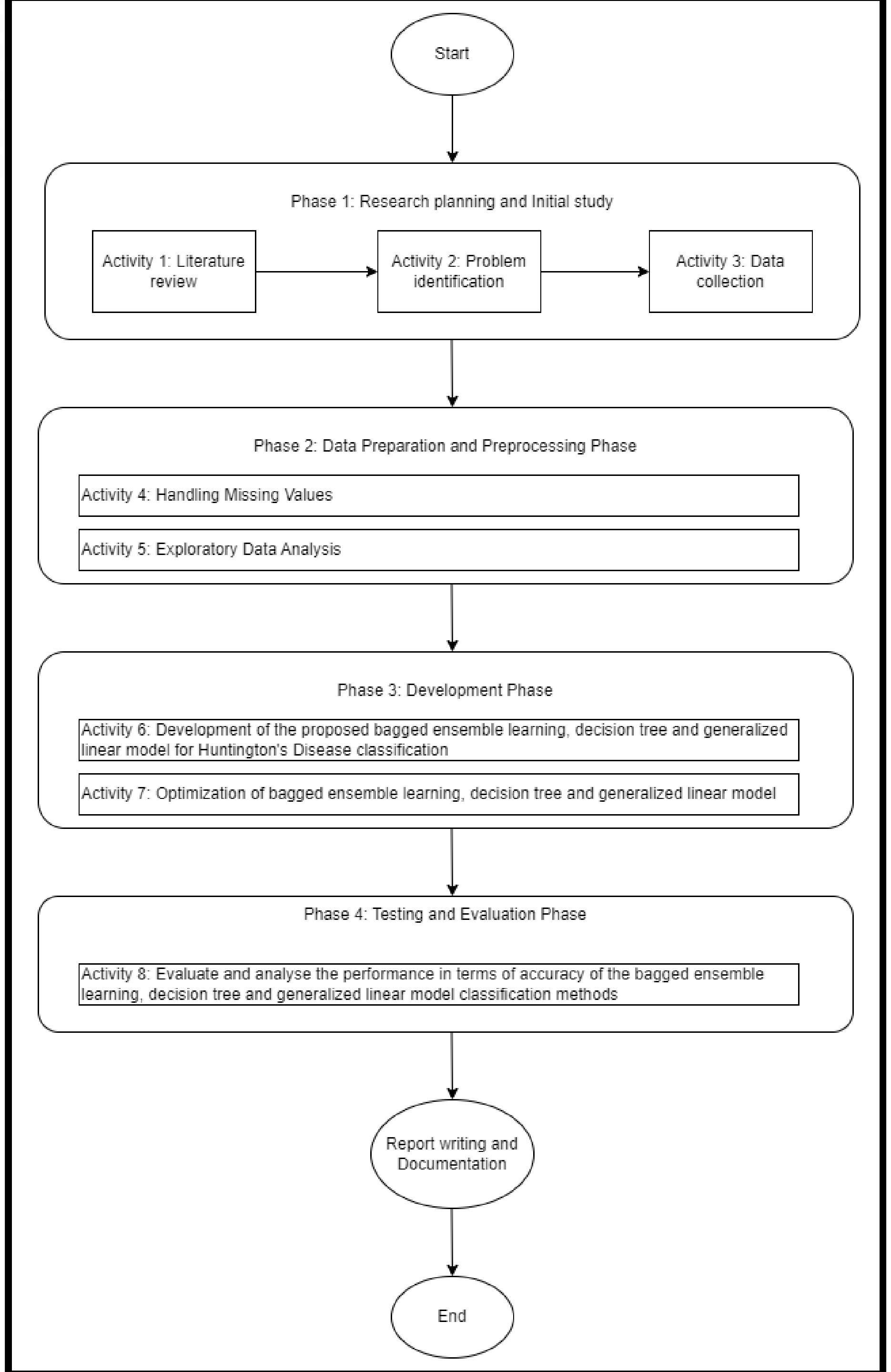
-ML method in other domain

Paper: Machine Learning Algorithms for the diagnosis of Alzheimer's and Parkinson's Disease (Noella and Priyadarshini, 2020)

Method: Bagged ensemble learning

CHAPTER 3: RESEARCH METHODOLOGY

RESEARCH FRAMEWORK



Start

Phase 1: Research planning and Initial study

Activity 1: Literature review

Activity 2: Problem identification

Activity 3: Data collection

Phase 2: Data Preparation and Preprocessing Phase

Activity 4: Handling Missing Values

Activity 5: Exploratory Data Analysis

Phase 3: Development Phase

Activity 6: Development of the proposed bagged ensemble learning, decision tree and generalized linear model for Huntington's Disease classification

Activity 7: Optimization of bagged ensemble learning, decision tree and generalized linear model

Phase 4: Testing and Evaluation Phase

Activity 8: Evaluate and analyse the performance in terms of accuracy of the bagged ensemble learning, decision tree and generalized linear model classification methods

Report writing and Documentation

End

RO1

To investigate the methods that can be used to classify Huntington's disease.

RO2

To implement bagged ensemble learning, generalized linear model and decision tree in the classification of Huntington's disease.

RO3

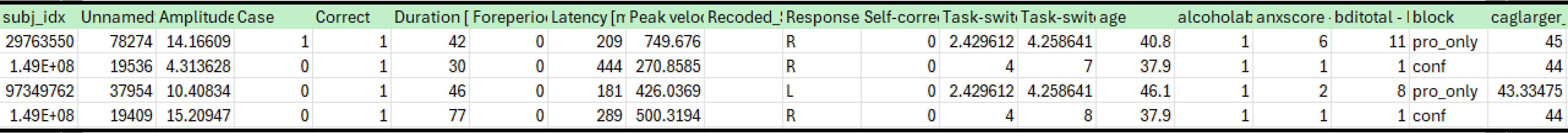
To evaluate the performance of bagged ensemble learning, generalized linear model and decision tree in the classification of Huntington's disease.

DATASET

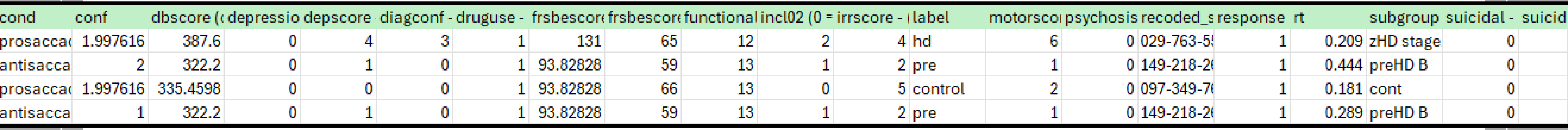
- Source: TRACK-HD study, which is a multinational longitudinal observational study (Wiecki et al., 2016)
- Link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148409

  https://figshare.com/articles/RT_accuracy_and_clinical_measures/2008407

  https://figshare.com/articles/Parameter_fits/2008404
- 90, 476 data samples, 41 attributes – but only 9, 048 data samples were used in this study



subj_idx	Unnamed	Amplitude	Case	Correct	Duration [	Foreperio	Latency [r	Peak velo	Recoded_	Response	Self-corre	Task-swit	Task-swit	age	alcoholab	anxscore -	bditotal -	block	caglarger_
29763550	78274	14.16609	1	1	42	0	209	749.676		R	0	2.429612	4.258641	40.8	1	6	11	pro_only	45
1.49E+08	19536	4.313628	0	1	30	0	444	270.8585		R	0	4	7	37.9	1	1	1	conf	44
97349762	37954	10.40834	0	1	46	0	181	426.0369		L	0	2.429612	4.258641	46.1	1	2	8	pro_only	43.33475
1.49E+08	19409	15.20947	0	1	77	0	289	500.3194		R	0	4	8	37.9	1	1	1	conf	44



cond	conf	dbscore (	depressio	depscore	diagconf -	druguse -	frsbescore	frsbescore	functional	incl02 (0 =	irrscore -	label	motorscor	psychosis	recoded_s	response	rt	subgroup	suicidal -	suicid
prosaccad	1.997616	387.6	0	4	3	1	131	65	12	2	4	hd	6	0	029-763-5	1	0.209	zHD stage	0	
antisacca	2	322.2	0	1	0	1	93.82828	59	13	1	2	pre	1	0	149-218-2	1	0.444	preHD B	0	
prosaccad	1.997616	335.4598	0	0	0	1	93.82828	66	13	0	5	control	2	0	097-349-7	1	0.181	cont	0	
antisacca	1	322.2	0	1	0	1	93.82828	59	13	1	2	pre	1	0	149-218-2	1	0.289	preHD B	0	

FEATURES

No.	Attributes	Description
1.	subj_idx	a unique identification number or index assigned to each individual participant in the study
2.	Unnamed: 1	unspecified data column
3.	Amplitude [deg]	the degree measurement of the amplitude, which could represent the magnitude or extent of a specific aspect related to the research study
4.	Case	the individuals or subjects participating in the study, categorized into different groups such as pre-manifest individuals with the Huntington's disease mutation, early symptomatic patients, and healthy controls
5.	Correct	the accuracy of performance during the antisaccade conflict task, particularly in terms of response inhibition and executive control
6.	Duration [ms]	the time measurements recorded in milliseconds during the performance of the antisaccade conflict task, specifically assessing reaction times and task completion durations
7.	Foreperiod [ms]	the duration of time in milliseconds between a warning signal and the presentation of the actual task stimulus during the antisaccade conflict task
8.	Latency [ms]	the time taken by participants to respond to the task stimuli during the antisaccade conflict task
9.	Peak velocity [deg/s]	the maximum speed of eye movements in degrees per second during the antisaccade conflict task, reflecting the efficiency of eye movement control
10.	Recoded_Subject	the individuals who participated in the study, including pre-manifest individuals carrying the HD mutation, early symptomatic HD patients, and healthy controls
11.	Response	the participants' reactions or actions during the antisaccade conflict task, indicating how they performed in terms of eye movements and cognitive control
12.	Self-corrected	instances where participants rectified their errors or made adjustments independently during the antisaccade task without external feedback
13.	Task-switch (2)	a variable representing participants' performance or behavior related to task-switching during the cognitive task
14.	Task-switch (3)	the cognitive process of shifting attention and mental resources between different tasks or activities
15.	age	the participants' chronological age at the time of the study

FEATURES

16.	**alcoholabuse - History of alcohol abuse (1 - never abused, 2 - ex-drug abuser, 3 - current abuse)**	**the variable indicating individuals' history of alcohol abuse**
17.	anxscore - ((HAD-SIS) Anxiety scale)	measuring anxiety levels in participants
18.	bditotal - Becks depression inventory (BDI-II)	a tool used to assess the severity of depression symptoms in participants
19.	block	a specific element or segment within the experimental design or task being discussed
20.	caglarger_value (CAG length)	a genetic characteristic associated with Huntington's disease
21.	cond	a specific condition or state that participants are subjected to during the study
22.	conf	a specific aspect or measure within the study
23.	dbscore (disease burden score)	a metric used to quantify the overall burden or severity of the disease within the study population
24.	depression	a psychological condition associated with feelings of sadness, hopelessness, and loss of interest in daily activities
25.	depscore - (HAD-SIS) Depression scale	the level of depressive symptoms individuals experience
26.	diagconf - Diagnostic confidence score (DCS)	the level of certainty or confidence in diagnostic assessments made during the study
27.	druguse - History of drug abuse (1 - never abused, 2 - ex-drug abuser, 3 - current abuse)	the history of drug abuse based on three categories: 1 - never abused, 2 - ex-drug abuser, 3 - current abuse
28.	frsbescore_f - Frontal behaviours (FrSBe) family rating	the family ratings of frontal behaviors assessed using the FrSBe scale
29.	frsbescore_s - Frontal behaviours (FrSBe) self-rating total score	the total score for self-rated frontal behaviors using the FrSBe scale
30.	functionalscore - Total functional capacity (TFC)	the Total Functional Capacity (TFC) assessment

FEATURES

31.	**incl02 (0 = control subject, 1 = premanifest gene carrier, 2 = early HD)**	**the classification of participants into three groups: 0 for control subjects, 1 for premanifest gene carriers, and 2 for early Huntington's Disease (HD) individulas**
32.	irrscore - (HAD-SIS) Irritability scale	the Irritability scale based on the HAD-SIS assessment
33.	label	classify or categorize data points into different groups.
34.	motorscore - Total motor score (TMS)	the Total Motor Score (TMS), used to assess motor function in participants
35.	psychosis	mental health conditions involving a loss of touch with reality and experiencing delusions or hallucinations
36.	recoded_subject	a transformed or coded version of subject identifiers for analysis purposes
37.	response	the reaction time and error rates recorded during the antisaccade conflict task, reflecting participants' cognitive processing speed and accuracy
38.	rt	the time taken by participants to respond to stimuli during the antisaccade task
39.	subgroup	a distinct subset of participants within the study, possibly categorized based on specific criteria or characteristics for analysis
40.	suicidal - Suicidal ideation	thoughts, plans, or desires related to suicide that the participants may experience
41.	suicide - Suicide attempts	actions taken by participants to harm themselves with the intent to die

PERFORMANCE MEASUREMENT

Confusion matrix is a tool used to evaluate the performance of classification models by presenting a summary of the model's predictions against the actual outcomes (Riehl, Neunteufel, and Hemberg, 2023)

Accuracy, Precision, Recall, Specificity will be calculated from the confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

1 Accuracy = (TP + TN) / (TP + TN + FP + FN)

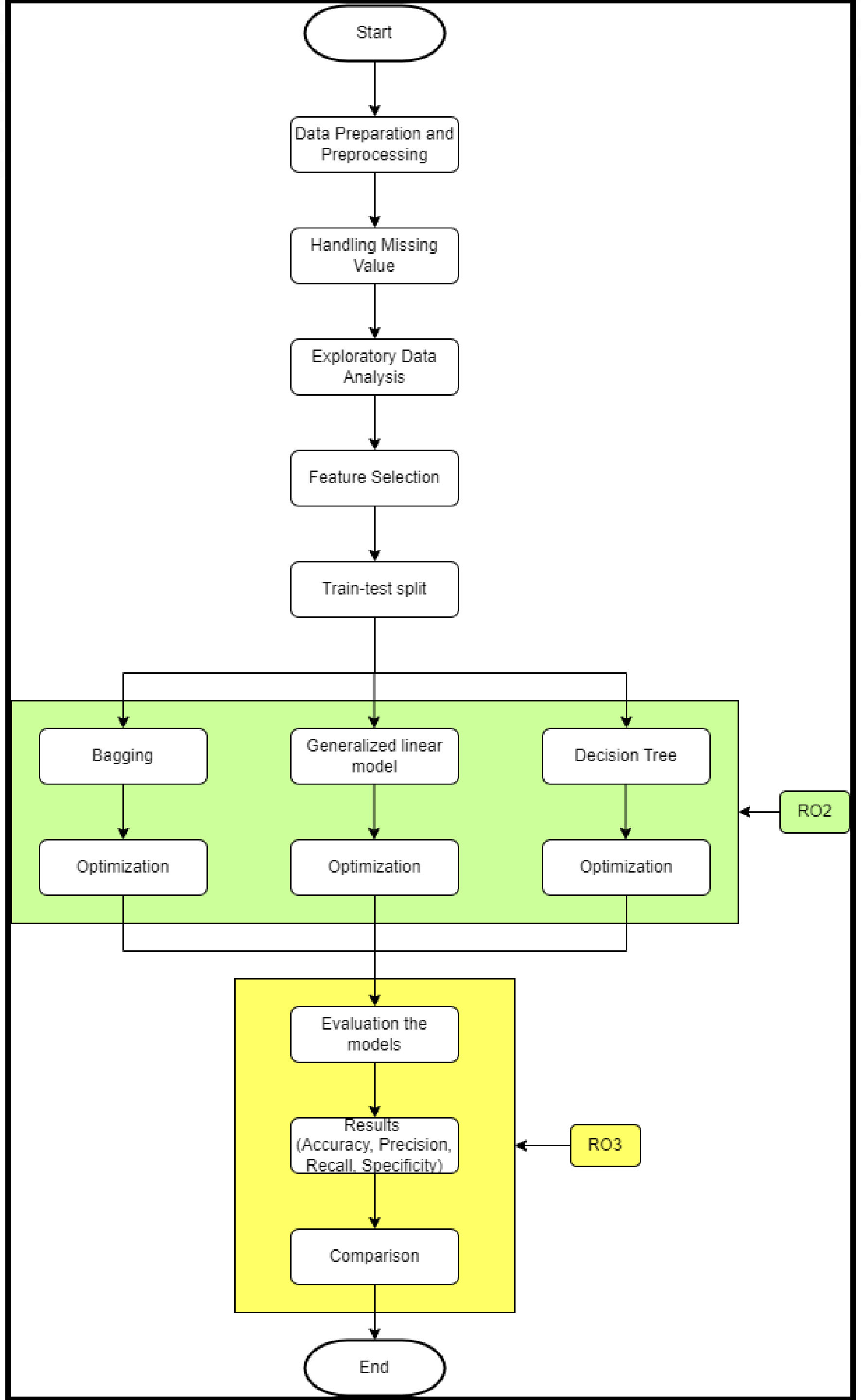
2 Precision = TP / (TP + FP)

3 Recall = TP / (TP + FN)

4 Specificity = TN / (TN + FP)

CHAPTER 4: RESEARCH DESIGN

EXPERIMENTAL METHODOLOGY FRAMEWORK



www.utm.my

Innovating Solutions

DATA PREPROCESSING

1. Handling Missing value

- ❑ Missing values in the dataset were identified and addressed.
- ❑ Various imputation methods were utilized according to the suitability for each variable.
- ❑ Missing numerical values were replaced with the mean or median, depending on their distribution.
- ❑ Categorical variables were imputed using the mode.
- ❑ Multiple imputation was used for some variables to improve dataset robustness and account for potential variability.

2. Exploratory Data Analysis

- ❑ The goal is to fully comprehend the primary features of the dataset using statistical summaries and visual aids.
- ❑ Techniques include correlation analysis, data visualization, and descriptive statistics.

HANDLING MISSING VALUE

- ➢ Missing values identified from the dataset

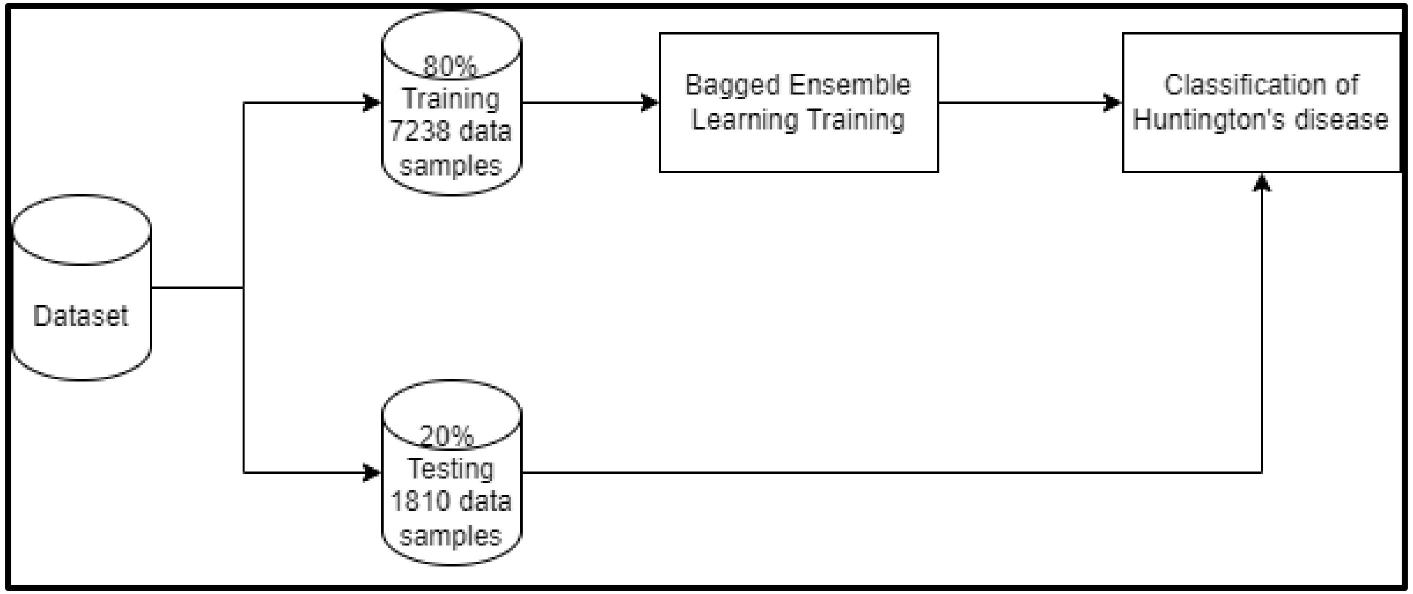
No.	Attributes	Missing Values
1.	Recoded_Subject	26
2.	Task-switch (2)	1497
3.	Task-switch (3)	1497
4.	age	26
5.	alcoholabuse – History of alcohol abuse (1 – never abused, 2 – ex-drug abuser, 3 – current abuse)	26
6.	anxscore – ((HAD-SIS) Anxiety scale)	303
7.	bditotal – Becks depression inventory (BDI-II)	276
8.	caglarger_value (CAG length)	2927
9.	conf	1497
10.	dbscore (disease burden score)	2927
11.	depression	26
12.	depscore – (HAD-SIS) Depression scale	303
13.	diagconf – Diagnostic confidence score (DCS)	26
14.	druguse – History of drug abuse (1 – never abused, 2 – ex-drug abuser, 3 – current abuse)	26
15.	frsbescore_f – Frontal behaviours (FrSBe) family rating	5484
16.	frsbescore_s – Frontal behaviours (FrSBe) self-rating total score	380
17.	functionalscore – Total functional capacity (TFC)	26
18.	incl02 (0 = control subject, 1 = premanifest gene carrier, 2 = early HD)	26
19.	irrscore – (HAD-SIS) Irritability scale	303
20.	label	26
21.	motorscore – Total motor score (TMS)	26
22.	psychosis	26
23.	recoded_subject	26
24.	subgroup	261
25.	suicidal – Suicidal ideation	26
26.	suicidal – Suicide attempts	26

EDA

- Overview of the dataset

Characteristics	Value
Number of variables	41
Number of observations	9048
Duplicate rows	0
Numeric variable type	21
Categorical variable type	18
Unsupported	1
Text	1

DATA DIVISION & MODEL DEVELOPMENT



Model	Method	Accuracy	Precision	Recall	F1-Score
Bagging	Bagging	98.95	98.87	98.77	98.82
Bagging	Random Forest	98.23	98.04	98.13	98.07
Bagging	Extra Trees	97.84	97.65	97.66	97.66

PRELIMINARY RESULTS

- **Techniques Used**:
  - **Random Forest Classifier**: Uses decision trees with random subsets of features.
  - **Extra Trees Classifier**: Increases robustness by dividing nodes at random.
  - **Bagging Classifier**: Uses bagged ensemble learning with random sampling (bootstrapping).
- **Bagging Ensemble Learning**:
  - Trains multiple model iterations on different subsets of training data.
  - Combines predictions from these models to improve overall performance.

CHAPTER 5: CONCLUSION

- Focus on decision trees, bagged ensemble learning, and generalized linear models (GLM).

- Compare these methods to determine effectiveness.

- Bagged ensemble learning likely provides better classification accuracy by reducing variance and preventing overfitting.

- GLM expected to produce reliable and understandable results, though potentially less accurate than the ensemble approach.

- Decision trees are easy to understand but may be less accurate and have more variance.

- The goal is to assist in choosing the best machine learning method for Huntington’s disease classification.

ACHIEVEMENT

- Significant progress made by investigating methods for classifying Huntington’s disease, completing the first research objective.
- Bagged ensemble learning was implemented and tested, achieving 98.95% accuracy.
- Next steps involve conducting feature selection to optimize input variables.
- Train all three models: bagged ensemble learning, GLM, and decision tree.
- Enable comprehensive evaluation of their performance.
- Aim to further improve classification techniques for Huntington’s disease.

RESEARCH CONSTRAINTS

01 Finding a comprehensive and high-quality dataset for Huntington's disease classification was difficult.

02 Understanding different machine learning models like decision trees, bagged ensemble learning, and GLM had a steep learning curve.

THANK YOU

univteknologimalaysia utm.my utmofficial