Huntington's Disease: Supervised Machine Learning Analysis of the Enroll-HD Features in Relation to BMI in the Manifest Stage

MSc Artificial Intelligence and Data Science

Osagie Elliot Aibangbee

202115576

August 2022

ABSTRACT

To determine the relationship between Huntington's Disease (HD) progression in a well-defined sample of persons with HD followed at over 155 sites by the Enroll-HD observational study, BMI was assessed in 866 adults, 546 manifest (with a confirmed diagnosis and manifestation of HD symptoms) and 320 control (not having the HD genetic mutation). They were followed for an estimated (mean ± SD) 5.4±0.6 years. Using the expansive Enroll-HD registry cohort, we investigated the relationship(s) between HD progression measures and body mass index (BMI) in patients with motor-manifest HD. The simple vector machine (SVM) and logistic regression models were trained to predict BMI classes at fifth follow-up based on selected variables which were analysed to determine their association with the target variable. After tuning both models, the SVM model gained 11% (51% - 62%) in accuracy, while the GLM gained 9% (45% - 54%). The SVM model showed superiority to the GLM model. Both models' precision/sensitivity and f1-score of the underweight class of fifth follow-up BMI was notably poorer than other classes. We found CCC (combined clinical characteristics) and combined feedself, novel derived variables, to be relevant predictors of BMI in addition to other established variables such as baseline BMI, CAP score, age of onset of impairment (motor, cognitive, etc.), etc. While the predictive influence of variables like cross-sectional chorea, baseline age, and gender appeared to be minimal. As more data become available, we are likely to unravel more causal relationships within the Enroll-HD variables.

INTRODUCTION & BACKGROUND

According to the Huntington Disease Society of America (HDSA), Huntington's disease (HD) is an incurable predominantly hereditary progressive neurodegenerative disorder that is caused by a defect in the huntingtin gene. The neuropathology of HD begins with the deterioration of the striatum, which coordinates many aspects of movement, mood, and memory resulting in functional decline and loss of independence (HDSA, 2022). Clinical presentation of HD is induced by an elongation of a cytosine-adenine-guanine (CAG) trinucleotide sequence in the huntingtin gene. People normally have around 20 CAG repeats, but HD carriers (premanifest and manifest) have 36 – 39 CAGs (reduced penetrance allele) or 40 or more CAGs (full penetrance allele) (HDSA, 2022; Lokhande, 2017; Langbehn, 2004). Medical diagnosis of HD is presently based on motor symptoms and changes in brain structure and physiology (Rizk-Jackson et al., 2011). Most people develop HD symptoms between ages 30 and 50, but it may occur in children and young adults (juvenile HD). The number of genetic CAG repeats, the most significant contributor to age of onset, is inversely proportional to average age of onset (Wright et al., 2019; Djousse et al., 2003; Wexler et al., 2004). Rosenblatt et al. (2012) found that the presence of shorter CAG repeats led to a slow decline in clinical progression in comparison to larger CAG repeats.

There has been links between HD and weight loss and the inability to eat despite cases of affected persons commonly having increased appetite and higher energy intake than non-affected controls (Trejo et al., 2004). Hamilton et al. (2004) found a weak but significant correlation between weight loss in HD-diagnosed adults and increasing chorea (abnormal involuntary movement disorder) severity, worse baseline motor performance, less severe baseline depressed mood, and poorer baseline independence rating. Using support vector machine (SVM) and linear regression models on neuroimaging data, Rizk-Jackson et al. (2011) observed a correlation between the estimated years to clinical onset (generated from genetic information and age) and established measures of disease progression. Landwehrmeyer et al. (2017) observed a correlation between the smaller allele and larger allele CAG repeats in healthy controls (genotype negative, family control, and community controls), and found smaller allele CAG lengths to be larger in manifest than controls. Ghazaleh et al. (2021) trained a random forest model to predict changes in clinical outcomes in manifest HD based on Enroll-HD variables and observed that cognitive impairment, age at diagnosis, tetrabenazine use, etc., were reliable predictors in addition to the established CAG repeat length and CAG-Age-Product (CAP) score (the product of CAG length and age). Khan et al. (2021) detected a pattern between severe weight loss and measures of cognitive, functional ability, and behavioural variables in premanifest HD.

BMI is the value derived by dividing a person’s weight (in kilograms) by their height squared (in meters). Healthcare providers use it as a screening tool in addition to other tools and tests to evaluate someone’s health status and risks. One of the most common factors associated with HD weight loss in patients is insufficient calorific intake (Trejo et al., 2004). However, Hamilton et al. (2004) concluded that weight loss after onset of symptom is not a consistent characteristic of HD.

Using machine learning (ML) predictive models on large datasets, as in those related to HD, allows finding hidden patterns that are often not revealed by clinical observations (Kelley & Ideker, 2005). In Myers & Montgomery (1997), generalized linear models (GLM) provide a unified framework to accommodate distributions of the exponential family (Gaussian/Normal, Poisson, Binomial, Gamma, etc.). According to Zuur et al. (2009), GLM comprise three steps: distribution of the outcome variable (random component), specification of predictor variables (systematic component), and the link function (linking the mean of the outcome variable to the predictors). Sub-types of GLM include

logistic regression, poisson regression, and survival analysis (Kabacoff, 2017). Logistic regression (or logit), a supervised ML algorithm, is traditionally used to estimate binary outcome based on one or more predictor variable(s) (Hilbe, 2011). For the model to train well, three assumptions are made: predictor variables are significantly related to the outcome variable, predictor variables are not correlated with one another, and model observations are uncorrelated (Hilbe, 2011).

Support Vector Machine (SVM) is a supervised ML algorithm that learns linear and nonlinear relationship(s) between the input variables and outcome variable using a decision boundary. It is popular for its high accuracy and resilience in dealing with highly dimensional data from diverse sources (Scholkopf et al., 2004). At the center of the decision boundary lies the separating hyperplane which can be a point, line, or plane depending on the level of dimension at which the data is separated (Cortes et al, 1995; Ben-hur & Weston, 2007). The SVM algorithm uses the kernel trick to transform otherwise inseparable data points onto higher dimensions at which they become linearly separable (Ben-hur & Weston, 2007; Noble, 2006). During training, the optimal hyperplane is determined by maximizing the separation margin of each class label while minimizing the magnitude of classification error (Noble, 2006). The trained model is then tested on a novel subset of data (i.e., test set) to determine the model's ability to accurately predict the class labels of the novel data (Rizk-Jackson et al., 2011).

The present investigation conducted the supervised learning analysis of HD in the manifest stage using a logistic regression and SVM classifier algorithms to better understand important Enroll-HD variables in the prediction of fifth follow-up BMI classes. The performance of each model was evaluated and compared based on accuracy, recall/sensitivity, precision, and f1-score. The insight gained from the performance of each model was then applied in analysing feature associations with each BMI class labels at fifth follow-up.

METHODOLOGY & EXPERIMENTAL SET UP

Hardware & Software Specification

Data cleaning, pre-processing, analysis, and machine learning experiment implementation were performed using Python v3.9.7 (van Rossum et al., 2006) on the Jupyter Notebook IDE (Kluyver et al., 2016). The python libraries used in the creation and evaluation of the supervised learning algorithms as administered to the selected datasets include: Imblearn (Mishra, 2017), SciPy (Virtanen et al., 2020), SciKit-Learn (Pedregosa et al., 2011). The hardware used for all experiments was a HP EliteBook Folio 1040 G3 laptop with an Intel Core i5-6300U 2.40GHz, 2496Mhz, 4 processors and 8GB of RAM.

APPROACH

This research was conducted with the aim of understanding the relationship between BMI and selected Enroll-HD variables in manifest and control (genotype negative and family controls) participants using ML algorithms. Two supervised ML algorithms (logistic regression and SVM) were implemented and trained on pre-selected features based on research and expert knowledge. The GLM and SVM classifiers were each trained to predict fifth follow-up BMI classes of manifest and control participants depending on a list of predictor variables (table 1). The outcome labels were integers 0 – 4, each representing unique BMI classes (table 2). Both model performances were compared in terms of classification metrics including accuracy, recall, precision, and f1-score.

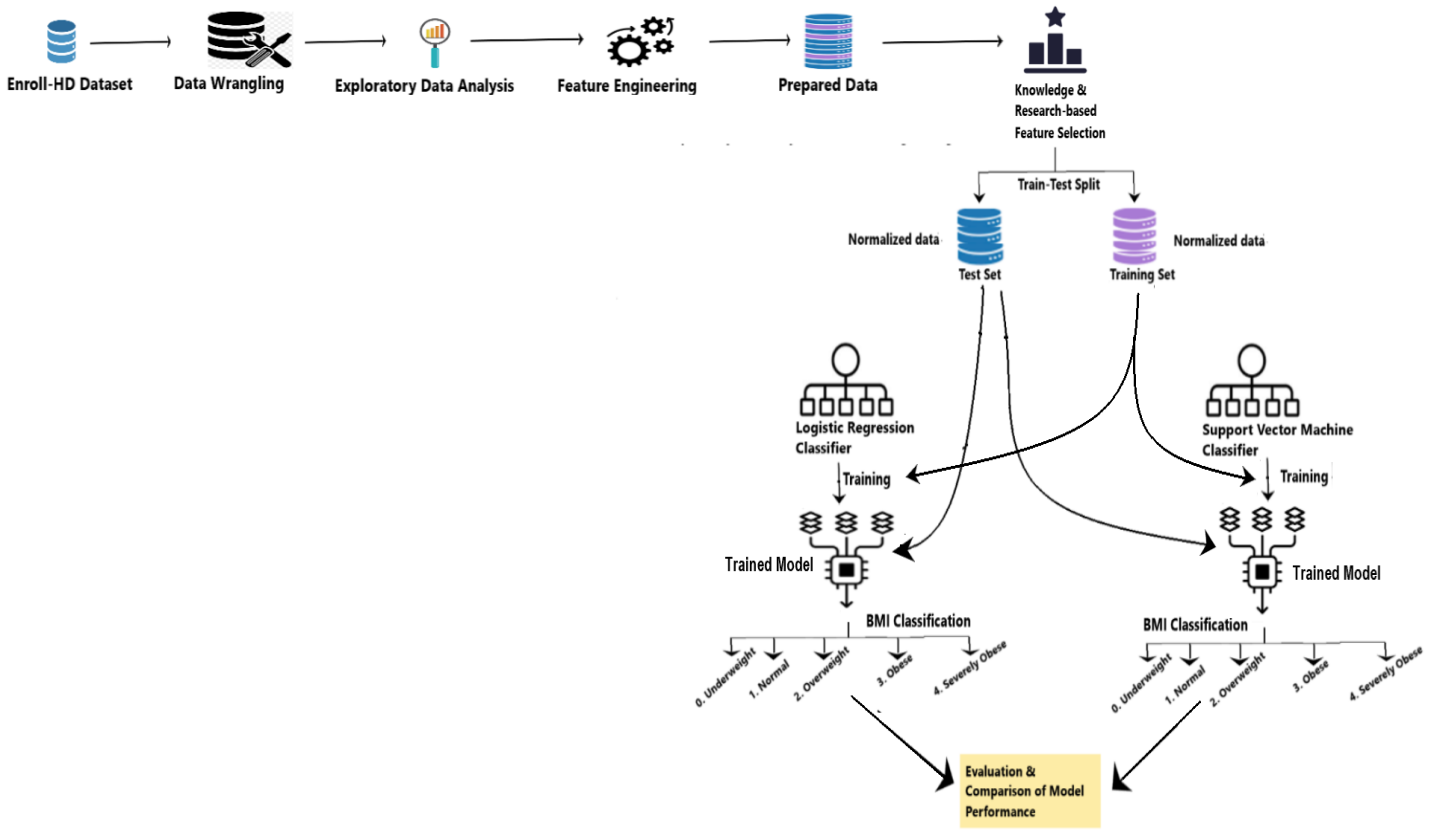

Fig 1. A flow-chart illustration of the research project. The dataset was read from csv files and then cleaned and formatted in preparation for EDA. Some relevant features were engineered from existing variables before pre-determined features were selected following expert knowledge and research. Train-test split was done using the ratio of 2:1 and minority classes were randomly oversampled to tackle the imbalanced distribution of the outcome variable. Uniformly distributed training set was then used to train both ML algorithms to predict BMI class labels (0 – 4) at fifth follow-up. Finally, both model performances were evaluated and compared in terms of accuracy, precision, recall, and f1-score.

Dataset

Enroll-HD is an international longitudinal observational study setup to simplify clinical research into HD. Participants at annual clinical visits provide a wide range of clinical, physiological, behavioural, etc., information which is strictly monitored for accuracy and quality. It is mandatory for contributing clinical sites to obtain and maintain local ethical approval. The derived periodic dataset (PDS) consists of data from 21,116 total participants (i.e., 5,173 premanifest, 10,947 manifest, 2,639 genotype negative, and 2,357 family control) (Enroll-HD, 2020). After excluding all premanifest participants, records from the first six visits (i.e., visit 1 - 6) of participants with six or more visits (988) were selected because it was observed that most of the participants had 3 – 5 follow-up records as observed by Khan et al. (2021). Before performing analysis and model development, we further excluded participants who, at baseline, were below age thirty, had received swallowing therapy or had taken high calorie nutritional supplement in the past. Samples were grouped into manifest (546 motor-manifest) and control (144 genotype negative & 176 family controls). Finally, 866 total participant records were used for analysis and model development.

A manifest HD participant is an individual with clinical characteristics diagnostic of HD. A family control participant does not possess the HD genetic mutation and is unrelated by blood but lives with/cares for someone with the genetic mutation (e.g., a spouse/partner). A genotype negative participant is defined as a blood relative of someone who has the HD genetic mutation but does not have the mutation themselves. Table 1 contains selected dataset variables. As this study aims to understand the relationship(s) between BMI and other influencing variables and compare performances of the GLM and SVM models, we included the following features as listed in table 1.

Table 1. Set of features used in training the ML models. The choice of features used was informed by expert advice and research. Detailed explanation of these variables is provided in the enroll-HD data dictionary (Enroll-HD, 2020).

Features		
Baseline Age; Sex; Baseline BMI; *Caghigh (*Larger allele CAG length); *Caglow (*Smaller allele CAG length); *CAP* (CAG-Age-Product) score: derived by multiplying larger allele CAG length and age. <u>Clinical Characteristics</u> *ccmtr:* medical history of motor impairment; *cccog*: medical history of cognitive impairment; *ccapt*: medical history of apathy; *ccdep*: medical history of depression; *CCC* (combined clinical characteristics): sum of clinical characteristics component measures. <u>Age of Onset</u> *ccmtrage:* age of onset of motor *impairment.* *cccogage:* age of onset of cognitive impairment. *ccaptage:* age of onset of apathy. *ccdepage:* age of onset of depression.	<u>Pharmacotherapeutic</u> *Tetrabenazine use; treated chorea.* <u>Nonpharmacotherapeutic</u> *Received swallowing therapy* <u>Nutritional</u> Took high calorie: medical history of receiving high calorie nutritional supplement as treatment for weight loss <u>Comorbid</u> *Had GI illness:* medical history of gastrointestinal illness <u>Choreic Measures</u> chorface; chorbol; chortrnk; chorrue; chorlue; chorrle; chorlle; *Chorea score*: sum of choreic component measures <u>Cognitive Scales</u> *sdmt1* (Symbol Digit Modality Test); *verfct5* (Verbal Fluency Test Category); *scnt1* (Stroop Color Naming Test); *swrt1* (Stroop Word Reading Test)	<u>Functional Capacity and Independence Scores</u> *indepscl* (subjects' independence scale); *fascore* (functional assessment); *tfcscore* (total functional capacity) <u>Psychiatric Evaluation</u> *exfscore* (executive function); *aptscore* (apathy); *depscore* (depression); *irascore* (irritability/aggression); *psyscore* (psychosis) *cognitive score*: sum of cognitive scale measures *behaviour score:* sum of psychiatric evaluation measures *motscore*: Motor impairment score; *Feedself;* *Independent feeding:* derived from indepscl

Data Wrangling and Domain Knowledge Representation

The original dataset contains periodic information of participants as unique records, resulting in multiple records per participant. Firstly, we extracted records of the listed features in table 1 and cleaned the dataset. Most missing values were replaced by the mean score per participant for quantitative features (e.g., height, etc.,) or the mode per participant for qualitative features (e.g., feedself, etc.,). However, missing BMI was calculated using weight and height variables. The remaining missing data were dropped from the dataset. Derived/engineered features based on domain knowledge and research are provided in tables 1 & 2.

Table 2. Domain knowledge and research-based categorization of numeric/quantitative features.

Derived Feature	Domain knowledge and research-based categories					
BMI class	BMI < 18.5	18.5 ≤ BMI < 25	25 ≤ BMI < 30	30 ≤ BMI < 40	BMI ≥ 40	
	0 (Underweight)	1 (Normal)	2 (Overweight)	3 (Obese)	4 (Severe obesity)	
Age bucket	Age<30	30≤age<40	40≤age<50	50≤age<60	60≤age<70	age≥70
	0	1	2	3	4	5
Larger allele CAG band	CAG < 27	27≤CAG<36	36≤CAG<40	CAG≥40		
	Normal	Intermediate	Reduced penetrance	Full penetrance		
Received swallowing therapy	cmtrt = 6	cmtrt != 6				
	1 (Yes)	0 (No)				
Took high calorie supplement	cmcat = 7	cmcat != 7				
	1 (Yes)	0 (No)				
Has had gastrointestinal illness	mhbodsy = 7	mhbodsys != 7				
	1 (Yes)	0 (No)				
Independent feeding	indepscl > 60	indepscl ≤ 60				
	1 (Yes)	0 (No)				
Treated chorea	cmindc__modify = Chorea	cmindc__modify != Chorea				
	1 (Yes)	0 (No)				
Tetrabenazine use	cmtrt__ing = Tetrabenazine	cmtrt__ing != Tetrabenazine				
	1 (Yes)	0 (No)				

In addressing our research questions, the continuous BMI variable was converted into five classes of BMI: underweight, normal, overweight, obese, severely obese (see table 2). The cleaned dataset had an uneven distribution of fifth follow-up BMI classes: 34 underweight (3.9%), 352 normal (40.6%), 256 overweight (29.6%), 185 obese (21.4%), and 39 severely obese (4.5%) participants. This challenge was resolved by randomly oversampling the minority class labels.

ML Algorithms

Logit and SVM models were implemented in python using sklearn library. Both models were trained to distinguish individual records into five BMI classes: underweight, normal, overweight, obese, and severely obese (see fig 1 & table 2). Imbalanced class distribution was handled only in the training set using the SMOTE (Synthetic Minority Oversampling Technique) algorithm implemented through the python library, imblearn (Chawla et al., 2002) by randomly oversampling the minority class labels resulting in a uniform class distribution of the training set (see Jupyter Notebook).

Feature Selection

Due to the high dimensionality (large number of features) of the dataset, it was deemed necessary to reduce dimensionality to improve model performance and optimise resources. For instance, Rizk-Jackson et al. (2011) observed a better performance in models (voxels) using relatively smaller number of features. This was done by selecting features based on domain knowledge and research. Some variables were dropped due to multicollinearity/autocorrelation (e.g., fascore, motscore, ccmtrage, etc) and zero-variance (received swallowing therapy, etc.,). 866 records were used in the development of both classification algorithms to predict BMI class labels at fifth follow-up. Multicollinearity check within predictor variables was performed using Pearson's correlation matrix (see Jupyter Notebook) and variables with an absolute correlation value above 0.70 with other variable(s) were dropped (e.g., combined motscore, ccmtrage, caghigh, and combined independent feeding).

Hyperparameter Optimization

The optimal state of hyperparameter values of both models were explored through sklearn's GridSearchCV function implemented with a five-fold cross validation using all 4 core processors (i.e., n_jobs=-1). The hyperparameter combination with the highest accuracy was retained as the best estimator and used to run predictions on the test set.

Model Evaluation

Before training each model, the dataset was randomly split into 66% training and 34% test sets. Model performance on test data was evaluated in terms of metrics such as precision, recall/sensitivity, f1-score, etc., (table 2) for each BMI class. Accuracy alone would not suffice due to the imbalanced distribution of the outcome variable after randomly splitting the dataset. Thus, recall/sensitivity, precision and f1-score were more relevant in evaluating the test data.

Precision is the models' ability to correctly predict a label, recall is its ability to capture the actual instances of a specific label, and f1-score is the weighted harmonic mean of precision and recall (Pedregosa et al., 2011). These are defined mathematically as follows:
Accuracy = (TP + TN) / (TP + TN + FP + FN);
Precision = TP / (TP + FP);
Sensitivity/Recall = TP / (TP + FN);
Specificity = TN / (TN + FP);
F1-score = (2 * Precision * Recall) / (Precision + Recall)
NOTE:
TN = True Negative; TP = True Positive.
FN = False Negative; FP = False Positive.

RESULTS

Participants

Data used for analysis and training of the GLM and SVM classifiers consisted of 866 participants who were 53.5±11.4 (mean ± SD) years old at baseline and had spent 5.4±0.6 years in the study. Manifest participants had at least 37 CAG repeats while control had a maximum of 35.

Table 3. Statistics of variables per HD status category (baseline vs. fifth follow-up)

Measure	Manifest (N = 546)	Control (N = 320)	Genotype Negative (N = 144)	Family Control (N = 176)
Baseline:				
Demographic data				
Age, years,				
Mean ± SD	53.3 ± 11.1	53.9 ± 11.9	50.4 ± 12.1	56.8 ± 11.0
Median (range)	53.5 (30; 82)	55.0 (30; 83)	52.0 (30; 83)	58.0 (30; 79)
Sex, males, n (%)	286 (52.4)	115 (35.9)	44 (30.6)	71 (40.3)
BMI, kg/m^2				
Mean ± SD	25.8 ± 5.2	29.4 ± 6.6	29.2 ± 7.1	29.5 ± 6.2
Median (range)	25.1 (15.8; 51.3)	28.0 (16.9; 57.1)	27.2 (16.9; 57.1)	28.3 (18.3; 52.8)
BMI Classes, n (%)				
Underweight	21 (3.8)	3 (0.9)	2 (1.4)	1 (0.6)
Normal	245 (44.9)	88 (27.5)	43 (29.9)	45 (25.6)
Overweight	183 (33.5)	100 (31.2)	41 (28.5)	59 (33.5)
Obese	91 (16.7)	107 (33.4)	47 (32.6)	60 (34.1)
Severely Obese	6 (1.1)	22 (6.9)	11 (7.6)	11 (6.2)
Comorbidities/therapies				
has had gastrointestinal illness, n (%)	206 (37.7)	98 (30.6)	48 (33.3)	50 (28.4)
HD Clinical Characteristics & Age of Onset				
motor impairment, n (mean ± SD)	543 (46.8±11.0)	182 (1.1±8.0)	7 (2.4±11.7)	175 (0±0)
cognitive impairment, n (mean ± SD)	356 (32.4±25.4)	1 (0.2±4.4)	1 (0.5±6.5)	0 (0±0)
apathy, n (mean ± SD)	381 (34.6±25.0)	15 (2.1±10.0)	15 (4.8±14.5)	0 (0±0)
depression, n (mean ± SD)	431 (35.1±21.5)	239 (8.4±18.1)	64 (18.6±23.3)	175 (0±0)
Research Genotyping				
Smaller allele CAG repeat length				
Mean ± SD	18.4 ± 3.3	16.8 ± 2.2	16.7 ± 2.5	16.9 ± 1.9
Median (range)	17 (9; 29)	17 (9; 29)	17 (9; 29)	17 (9; 22)
Larger allele CAG repeat length				
Mean ± SD	43.2 ± 2.9	19.7 ± 3.5	19.9 ± 3.7	19.6 ± 3.4
Median (range)	43 (37; 58)	19 (15; 35)	19 (15; 35)	19 (15; 35)
Larger allele CAG repeats classes, n (%)				
Normal	0 (0)	303 (94.7)	134 (93.1)	169 (96.0)
Intermediate	0 (0)	17 (5.3)	10 (6.9)	7 (4.0)
Reduced Penetrance	17 (3.1)	0 (0)	0 (0)	0 (0)
Full Penetrance	529 (96.9)	0 (0)	0 (0)	0 (0)

Fifth follow-up:				
Demographic Information				
BMI, kg/m^2				
Mean ± SD	25.4 ± 5.4	29.8 ± 6.9	29.8 ± 7.4	29.7 ± 6.4
Median (range)	24.5 (13.2; 51.4)	28.4 (17.5; 57.4)	27.8 (17.5; 57.4)	29.0 (17.5; 52.0)
BMI Classes, n (%)				
Underweight	31 (5.7)	3 (0.9)	2 (1.4)	1 (0.6)
Normal	268 (49.1)	84 (26.2)	37 (25.7)	47 (26.7)
Overweight	153 (28.0)	103 (32.2)	51 (35.4)	52 (29.5)
Obese	84 (15.4)	101 (31.6)	39 (27.1)	62 (35.2)
Severely Obese	10 (1.8)	29 (9.1)	15 (10.4)	14 (8.0)

Table 4. Cross-sectional BMI per larger allele CAG repeat classes

Measure	Normal (Control)	Intermediate (Control)	Reduced Penetrance (Manifest)	Full Penetrance (Manifest)
n	303	17	17	529
Baseline:				
BMI kg/m^2				
Mean ± SD	29.2 ± 6.3	31.6 ± 10.8	28.5 ± 6.9	25.7 ± 5.1
Median (range)	27.9 (16.9; 53.0)	28.3 (22.4; 57.1)	25.4 (18.4; 41.4)	25.1 (15.8; 51.3)
Fifth follow-up:				
BMI kg/m^2				
Mean ± SD	29.7 ± 6.8	30.2 ± 7.5	28.2 ± 7.2	25.3 ± 5.3
Median (range)	28.5 (17.5; 57.4)	27.9 (20.8; 45.3)	26.1 (16.6; 43.0)	24.5 (13.2; 51.4)

Genotyping

It was observed that the larger allele CAG length consisted of 94.5% normal and 5.5% intermediate for controls, and 96.8% full penetrance and 3.2% reduced penetrance for manifest (table 4).

The mean ± SD CAG size for the larger allele was higher in manifest versus controls (43.2 ± 2.9 vs. 19.8 ± 3.6; table 4). Despite having the same range and median, the CAG sizes for smaller allele were also higher in manifest than controls (mean ± SD: 18.4 ± 3.3 vs. 16.8 ± 2.2, respectively; $p<0.001$). A positive correlation was observed between smaller allele and larger allele CAG lengths in the control group (r=0.39; $p<0.001$), while no significant correlation was found in the manifest group.

BMI analysis of the classes of larger allele CAG length per participant category revealed that the difference between control and reduced penetrance mean baseline and fifth follow-up BMIs were insignificant. Although, full penetrance had significantly lower average BMI than reduced penetrance (baseline mean ± SD: 25.7 ± 5.0 vs. 28.5 ± 6.9; $p<0.05$, fifth follow-up mean ± SD: 25.2 ± 5.2 vs. 28.2 ± 7.2; $p<0.05$ respectively, table 5), there was no significant difference between the median BMI of reduced and full penetrance participants (baseline median difference: $p=0.38$; fifth visit median difference: $p=0.12$).

Clinical Assessments

As anticipated, the mean ± SD UHDRS motor impairment, total functional capacity, independence, behaviour, and chorea scores were significantly poorer in the manifest population than in control at baseline and fifth follow-up (table 3). Considering the UHDRS core assessment outcomes with respect to BMI classes, the underweight subgroup of the manifest population had higher motor impairment (mean ± SD: 41.0±19.4) than normal (31.3±16.2), overweight (29.6±18.1), obese (28.2 ± 14.9), and severely obese (19.2±12.7). A weak negative correlation was observed between baseline BMI and chorea ($r = -0.22$, $p<0.001$), and motor impairment ($r = -0.15$, $p<0.001$) scores in the manifest population. While the fifth follow-up BMI weakly correlated with chorea ($r = -0.15$, $p<0.001$), independence ($r = 0.14$, $p<0.05$), motor impairment scores ($r = -0.2$, $p<0.001$).

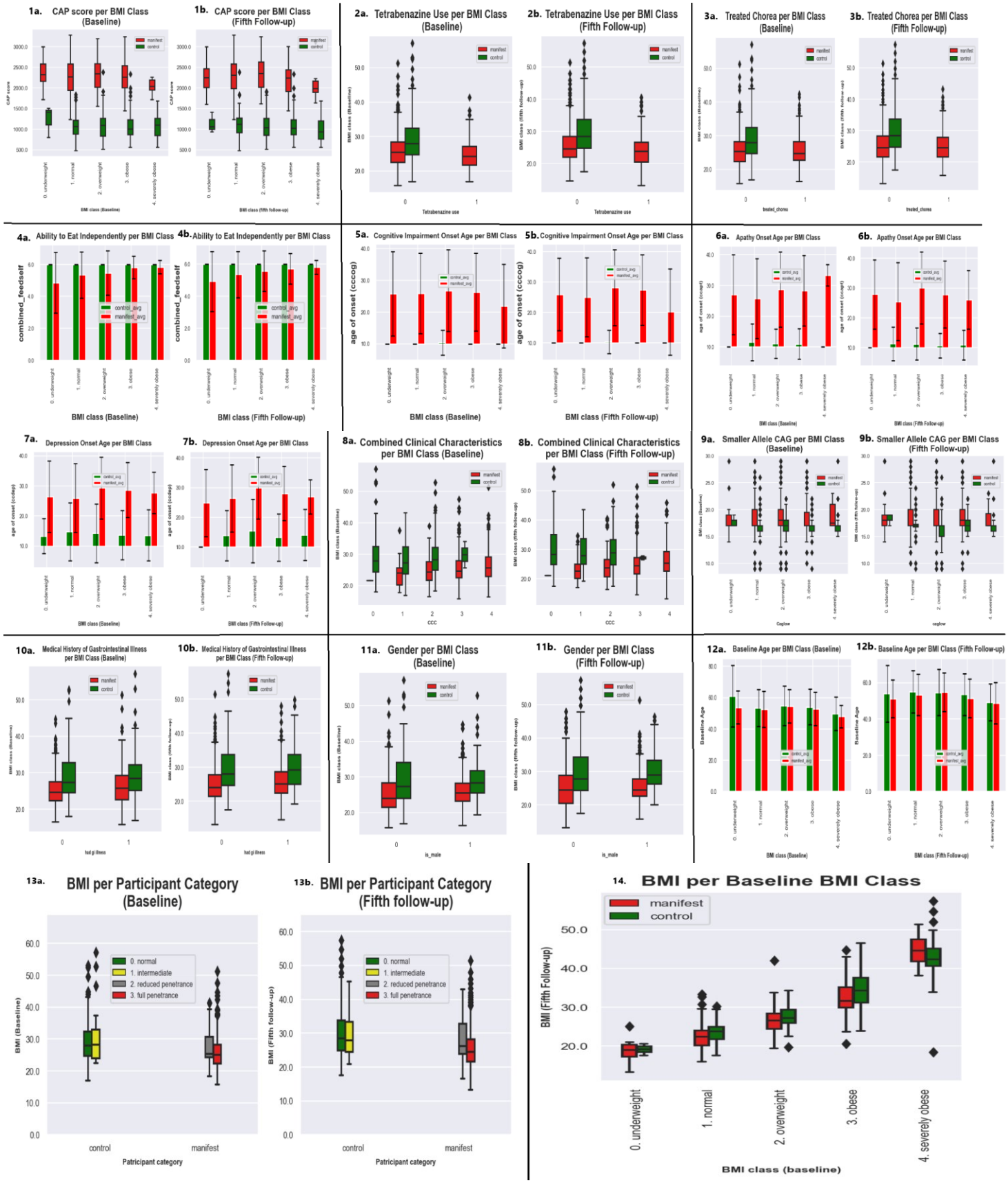

Fig 2. Cross-sectional association between BMI and selected Enroll-HD predictor features in manifest and control participants. These variables have a negative relationship with the outcome variable: CAP score (r=-0.5), tetrabenazine use (r=-0.24), treated chorea (r=-0.34), age of onset of cognitive impairment (r=-0.31), age of onset of apathy (r=-0.29), age of onset of depression (r=-0.24), CCC (r=-0.30), smaller allele CAG (r=-0.09), history of gastrointestinal illness (p=0.54), gender (r=-0.12), baseline age (r=-0.10). while ability to eat independently (r=0.23) has a positive correlation.

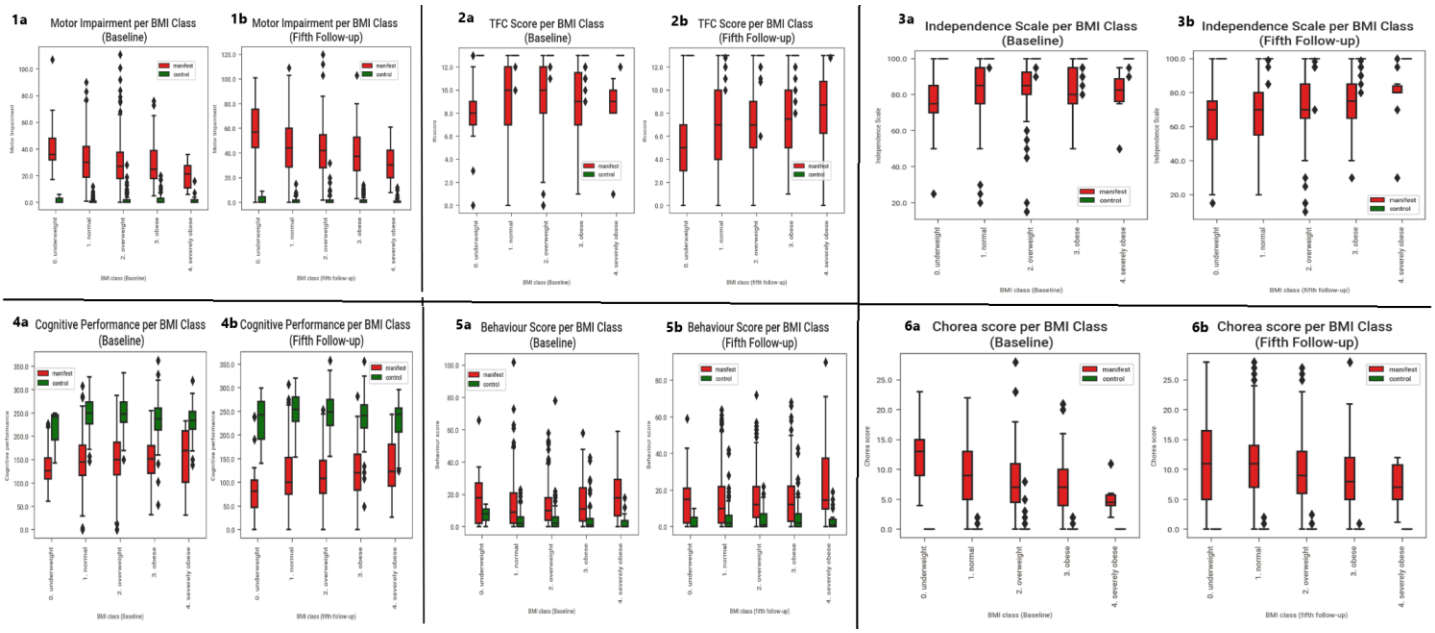

Fig 3. Cross-sectional relationship between BMI and UHDRS variable. The following UHDRS variables had a negative correlation with the target variable: motscore and chorea score, while cognitive_score, tfcscore, independence scale, and behaviour score had no correlation with the target variable.

Table 5. Architecture of tuned/optimized models

Hyperparameter	**Logistic Regression**	**SVM**
Epochs	350	N/A
Training Duration (Minutes)	0.03	0.004
C	7.5	750
Penalty/Regularization	L1 (Lasso)	N/A
Solver/Kernel	SAGA (Stochastic Average Gradient descent)	RBF (Radial Basis Function)
Tolerance	0.0001	0.012
Multiclass/Gamma	Multinomial	Auto
Random State	101	101
Break Ties	N/A	True

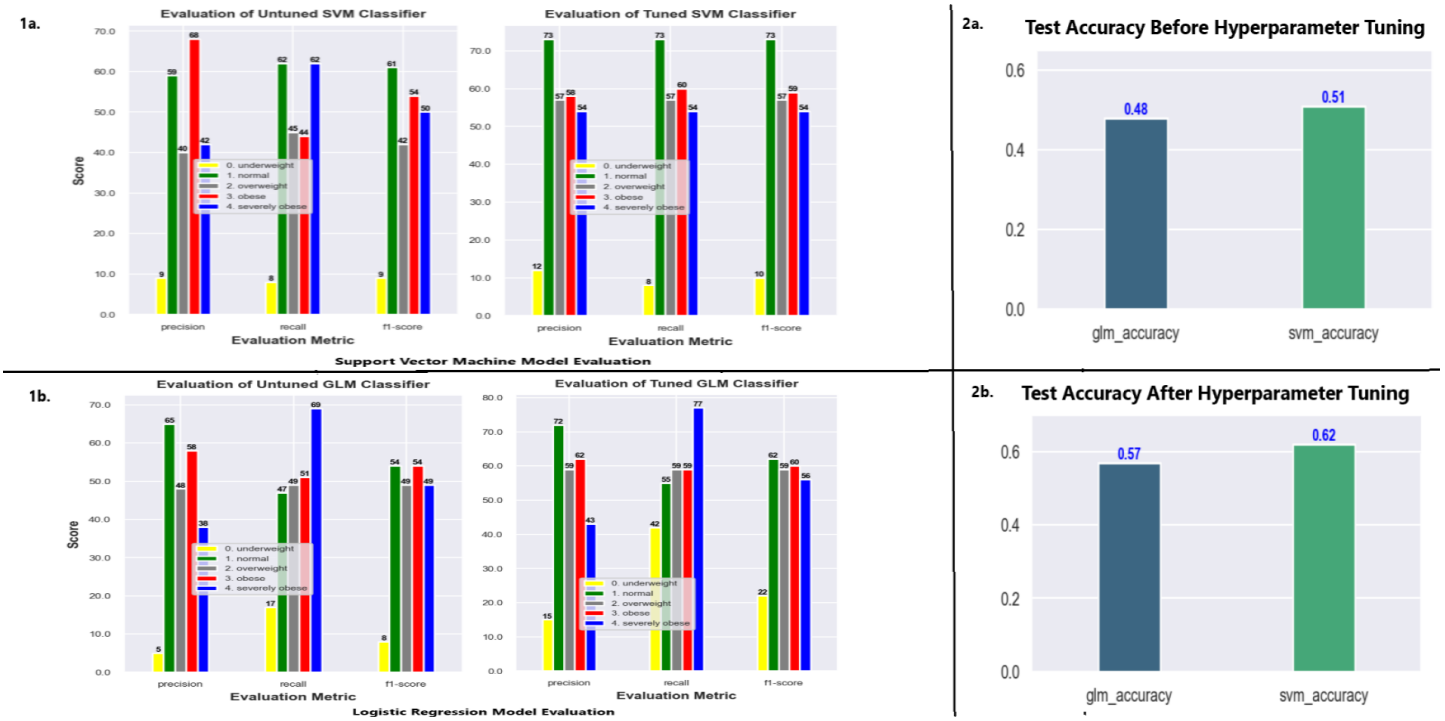

Fig 4. GLM & SVM model evaluation in terms of accuracy, precision, recall/sensitivity, and f1-score. SVM had slightly better accuracy than the GLM model before (2a) and after (2b) the tuning of hyperparameters. Hyperparameter tuning improved both model performance in terms of accuracy (SVM: 51% to 62%; GLM: 48% to 57%), recall, precision, and f1-score (SVM: 1a; GLM: 1b).

Taking the performance metrics of both algorithms into consideration (logit: 57% and SVM: 62% test accuracy), the SVM performed slightly better than the GLM.

DISCUSSION

In this study, the cross-sectional association between BMI and selected factors in manifest and control participants was investigated using logit and SVM classification models. The multi-dimensional associations between the predictor and outcome variables played a major role in both model performances (see fig 2 & 3). After training and tuning of hyperparameters, the logit model had 57% test accuracy at predicting the outcome, while the SVM classifier showed its superiority with a better accuracy of 62%. In terms of f1-score, both models performed relatively well at predicting each class labels except the underweight class which had the smallest number of instances in the test data (please see fig 4). Although, the tuned logit showed better precision, recall, and f1-score at predicting the underweight and overweight classes than the SVM.

Besides the clear association between baseline BMI and the outcome variable, this result supports (Ghazaleh et al., 2021; Langbehn et al., 2019; Tabrizi et al., 2013) who found that CAP score and CAG repeat length are among the most important factors for predicting HD progression. These were also two of the most important predictors of fifth follow-up BMI as manifest CAP scores were significantly higher than controls (p<0.001). Other top-ranked predictors such as cross-sectional motor impairment and behaviour_score (higher in manifest than controls, p<0.001) have a negative correlation with the outcome. While cognitive_score, tfcscore, and feedself (less in manifest than controls, p<0.001) each have positive correlation with the outcome variable. This indicates that having manifest HD causes increase in motor impairment and behavioural issues which in turn tends to result in decreased fifth follow-up BMI. CCC (higher in manifest than control, p<0.001) has a

similar association and influence on the outcome as tfcscore. However, the opposite is true for cognitive_score, tfcscore, and feedself. Interestingly, baseline age and cross-sectional chorea, and gender had no significant correlation with the outcome.

Investigation into the association between CAG repeat length and fifth follow-up BMI revealed that the CAG length of manifest (significantly less than control, $p<0.001$) has a weak negative correlation with outcome in manifest ($r = -0.1$, $p<0.05$) and no correlation in control (fig 6). It was observed that, at baseline, the mean BMI of reduced penetrant participant was significantly higher than that of fully penetrant participants and no significant difference was found between their fifth follow-up BMI. This may indicate that having a fully penetrant CAG repeat length causes a faster drop in BMI than having reduced penetrant CAG, which is consistent with the findings of Rosenblatt et al. (2012). However, the median values suggest that there is no significant difference across both periods. Hence, more data may help to clarify this observation. As in the work of Rizk-Jackson et al. (2011), age of onset of clinical impairments (cognitive, apathy, etc.,) were found to be relevant predictors of the BMI at fifth follow-up.

In the future, more data would help in training the machine learning algorithms more effectively without having to make too many copies of the data. Additionally, predicting a change in BMI or the magnitude of cross-sectional weight change may be more relevant in terms of helping manifest HD patients in better managing their health.

CONCLUSION

This research serves to further improve the understanding of factors responsible for the change in BMI as it affects manifest HD individuals. The result of this classification task confirms the superiority of the SVM over the GLM especially when dealing with a highly dimensional dataset with multiple classes as in this case, as both models benefitted equally from the tuning of their hyperparameters. This study did well to capture the multi-dimensional relationships between manifest BMI and HD progression as measured by selected UHDRS variables. In this research, we found CCC (combined clinical characteristics) and combined feedself, novel derived variables, to be relevant predictors of BMI in addition to other established variables such as baseline BMI, CAP score, age of onset of impairment (motor, cognitive, etc.), etc. While the predictive influence of variables like cross-sectional chorea, baseline age, and gender appeared to be minimal. The insights gained from this work is encouraging and can serve as a background for future investigations. As more data become available, we are likely to unravel more causal relationships within the Enroll-HD variables.

REFERENCES

Ben-hur, A. and Weston, J. (2007) A user's guide to support vector machines. *Methods in Molecular Biology.*

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002) SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.

Cortes, C., Vapnik, V. (1995) Support-vector networks. *Machine Learning* 20, 273–297.

Djousse, L., Knowlton, B., Hayden, M., Almqvist, E. W., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., et al (2003) Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *American Journal of Medical Genetics.* 119A, 279-282.

Enroll-HD (2020). Enroll-HD Periodic Dataset 5, electronic dataset, https://www.enroll-hd.org/, Unpublished Dataset.

Ghazaleh, N., Houghton, R., Palermo, G., Schobel, S. A., Wijeratne, P. A. and Long, J. D. (2021) Ranking the Predictive Power of Clinical and Biological Features Associated with Disease Progression in Huntington's Disease. *Frontiers in Neurology*, 814.

Hamilton, J. M., Wolfson, T., Peavy, G. M., Jacobson, M. W. and Corey-Bloom, J. (2004) Rate and correlates of weight change in Huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(2), 209-212.

HDSA (2022) *Overview of Huntington's Disease*. Available online: https://hdsa.org/what-is-hd/overview-of-huntingtons-disease/ [Accessed: 19/07/2022]

Hilbe, J. M. (2011) Logistic regression. *International encyclopedia of statistical science*, 1, pp.15-32.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science and engineering*, 9(3), 90–95.

Kabacoff, R. I. (2017) *Generalized Linear Models*. Available online: https://www.statmethods.net/advstats/glm.html [Accessed: 20/08/2022]

Kelley, R., and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* 23, 561-566.

Khan W, Alusi S, Tawfik H, Hussain A (2021) The relationship between non-motor features and weight-loss in the premanifest stage of Huntington's disease. *Plos one* 16(7): e0253817.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S. and Ivanov, P., (2016) Jupyter Notebooks-a publishing format for reproducible computational workflows. 87-90.

Landwehrmeyer, G. B., Fitzer-Attas, C. J., Giuliano, J. D., Gonçalves, N., Anderson, K. E., Cardoso, F., Ferreira, J. J., Mestre, T. A., Stout, J. C. and Sampaio, C. (2017) Data analytics from Enroll-HD, a global clinical research platform for Huntington's disease. *Movement disorders clinical practice*, 4(2), 212-224.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S., Hayden, M. R. and an International Huntington's Disease Collaborative Group (2004) A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical genetics*, 65(4), 267-277.

Langbehn, D. R., Stout, J. C., Gregory, S., Mills, J. A., Durr, A., Leavitt, B. R., et al. (2019) Association of CAG repeats with long-term progression in huntington disease. *JAMA Neurology* 76:1375–85.

Lokhande, S., (2017) Understanding Huntington's disease using machine learning approaches. *KGI Theses and Dissertations, 4*.

McKinney, W., 2010, June. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* 445(1), 51-56.

Mishra, S. (2017) Handling imbalanced data: SMOTE vs. random undersampling. *International Research Journal of Engineering and Technology*, 4(8), 317-320.

Myers, R. H. and Montgomery, D. C. (1997) A tutorial on generalized linear models. *Journal of Quality Technology*, 29(3), 274-291.

Noble, W.S., 2006. What is a support vector machine?. *Nature biotechnology*, 24(12),1565-1567.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.

Rizk-Jackson, A., Stoffers, D., Sheldon, S., Kuperman, J., Dale, A., Goldstein, J., Corey-Bloom, J., Poldrack, R. A. and Aron, A. R. (2011) Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *Neuroimage*, 56(2).788-796.

Rosenblatt, A., Kumar, B. V., Mo, A., Welsh, C.S., Margolis, R. L., and Ross, C. A. (2012) Age, CAG repeat length, and clinical progression in Huntington's disease. *Movement Disorders*. 27, 272-276.

Schölkopf, B., Tsuda, K. and Vert, J. P. (2004) Kernel methods in computational biology. *MIT press*.

Seabold, S., and Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with python. *In 9th Python in Science Conference*.

Tabrizi, S. J., Scahill, R. I., Owen, G., Durr, A., Leavitt, B. R., Roos, R. A., *et al.* (2013) Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurology* 12:637–49.

Trejo, A., Tarrats, R. M., Alonso, M. E., Boll, M. C., Ochoa, A. and Velásquez, L. (2004) Assessment of the nutrition status of patients with Huntington's disease. *Nutrition*, 20(2), 192-196.

Van Der Walt, S., Colbert, S. C. and Varoquaux, G. (2011) The NumPy array: a structure for efficient numerical computation. *Computing in science and engineering*, 13(2), 22-30.

van Rossum, G. and Drake, F. L. (2006) Python Reference Manual. Python Software Foundation. https://www.python.org/

Virtanen, P., Gommers R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.

Wexler, N. S. and US–Venezuela Collaborative Research Project* (2004) Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proceedings of the national academy of sciences*. 101, 3498-3503.

Wright, G. E., Collins, J. A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Bečanović, K., Drögemöller, B. I., Semaka, A., Nguyen, C. M. and Trost, B. (2019) Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *American Journal of Human Genetics*, 104(6),1116-1126.

Wright, R. E. (1995) Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 217–244). *American Psychological Association.*

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. and Smith, G. M. (2009) GLM and GAM for count data. *In Mixed effects models and extensions in ecology with R* (pp 209-243).