Universiteit Leiden

Master Computer Science

Improving the estimation of age at onset for Huntington's Disease through the use of Machine Learning models with the Enroll-HD data.

Name:	Julio Cesar Marchiori Dias
Student ID:	s3095304
Date:	21/05/2024
Specialisation:	Bioinformatics
1st supervisor:	Dr. Eleni Mina
2nd supervisor:	Dr. Katherine Wolstencroft

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Abstract

Huntington's disease (HD) is a rare neurodegenerative disorder that is inherited in a dominant manner and caused by a prolonged CAG repeat in the huntingtin protein. It is characterised by motor, behavioural and cognitive abnormalities. Age-at-onset (AAO) in HD refers to the time when symptoms first appear in individuals carrying the CAG repeat mutation. HD is an incurable condition, which makes the determination of AAO crucial in identifying factors that may modify it, and in developing and evaluating therapies aimed at delaying its onset. The HD AAO is significantly correlated with the number of CAG repeats representing the most significant factor in estimating the AAO. Current models for AAO prediction utilise the length of the CAG repeat as the primary predictor variable. Results range between 47% to 72% of the variance considering the diversity in HD populations, indicating that there should be more factors influencing the onset than just the CAG repeat. Enroll-HD study is a global, longitudinal investigation of individuals affected by HD, with over 21,000 participants. It collects uniform clinical data (baseline and follow-up data) and biological samples from multiple study sites, in both manifest and premanifest stages of the disease. Although machine learning (ML) algorithms are powerful tools, they are not frequently utilised in rare disease research, mainly due to the scarcity of data required for a proper training of such models and subsequent estimations. In this study, we trained a series of complex ML models using a highly accurate subset (defined as patients enrolled as pre-manifest and subsequently manifesting the disease) from the Enroll-HD dataset, aiming to improve the current clinical AAO estimation method (known as the Langbehn formula). In addition to the CAG repeat, a group of extra variables were selected and incorporated into the training process, mostly related to lifestyle aspects. This project also assessed the efficacy of the trained models by submitting them to the remaining set of data from Enroll-HD (patients enrolled already as manifest). Results indicated that ML models outperformed the current used method. CatBoost ML model obtained a R^2 of 0.675 compared to 0.534 from Langbehn during test verification. In conclusion, the use of ML models in conjunction with Enroll-HD additional patient information facilitated the generation of more accurate predictions of HD AAO. Furthermore, the correlation between lifestyle aspects and different HD onsets was demonstrated.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Eleni Mina for her invaluable guidance and feedback. I also could not have undertaken this journey without my supervisor, Dr. Katherine Wolstencroft, who provided inspiring feedback and the opportunity to work with this wonderful project.

I am also grateful to my office mates in the BioSemantics group at LUMC for their feedback and moral support.

Finally, I would like to express my gratitude to my wife Leila for her unwavering support and patience. Her encouragement has been instrumental in keeping my spirits and motivation high throughout the process of life.

1 Introduction

1.1 Huntington's Disease (HD)

A rare disease is a medical condition that affects a small proportion of the population, typically fewer than 1 in 2,000 people, as defined by most health organizations [1]. Huntington's disease (HD) is a rare neurodegenerative disorder, with a prevalence of 4.88 cases per 100,000 inhabitants, with lower rates in Asia and higher rates in Europe, North America, and Australia [2]. HD is inherited in a dominant manner and it is characterized by uncontrolled movements (chorea), psychiatric and behavioral problems, and cognitive impairment [3, 4]. HD is caused by a mutation in the HTT gene, which translates into a mutated huntingtin protein [5]. Individuals with HD typically exhibit a cortico-striatal degeneration of white and gray matter, resulting in a selective loss of medium spiny neurons in the striatum and pyramidal neurons in the cortex [6]. The HTT mutation (mHTT) that causes HD involves a segment of DNA known as a CAG trinucleotide repeat, that occurs several times in a row. Normally, the CAG allele is repeated 10 to 35 times in a gene. In people with HD, the longest CAG segment is repeated 36 to more than 120 times. People with 36 to 39 CAG repeats may or may not develop the signs and symptoms of HD (reduced penetrance), while people with 40 or more repeats almost always do (fully penetrance) [7]. The length of the CAG trinucleotide expansion has a strong inverse relationship with the mean age of clinical onset [8]. When the CAG repeats are greater than 60, it is defined as juvenile Huntington's disease, which affects children and teenagers.

1.1.1 Onset

Age-at-onset (AAO) in HD refers to the time when symptoms first appear in individuals carrying the mutated prolonged CAG repeat. HD is an incurable condition, making the determination of AAO crucial in identifying factors that may modify it, and in developing and evaluating therapies aimed at delaying its onset [9]. The Langbehn formula is widely used by clinicians and HD researchers at the clinic for estimating the AAO [10]. This formula incorporates the CAG repeat length to estimate the AAO and accounts for between 47% to 72% of the variance in different HD populations. Residual variability in AAO can be attributed to either genetic and/or environmental factors [11]. These factors are the focus of current research to determine which specific environmental or genetic factors may influence the AAO.

1.2 Enroll-HD Platform

Enroll-HD is a worldwide longitudinal study of HD patients that aims to accelerate the development of therapies for HD by collecting uniform clinical data and biological samples to better understand the natural history of the disease [12]. This study collects baseline and follow-up data from multiple study sites, in the same way and using the same methods, from tens of thousands of pre-manifest (when participant is a CAG mutation carrier but has not yet shown HD symptoms) and manifest (when participant is already showing symptoms of the disease) patients. Enroll-HD was created by integrating two HD datasets: the Cooperative Huntington

Observational Research Trial (COHORT) in North America and Australia, and REGISTRY, an observational study of the European Huntington Disease Network (EHDN) [13]. Enroll-HD also extended research activities in Latin America and Asia. Enroll-HD collects clinical data from participants through annual visits. The assessments are performed by highly trained clinical personnel. Periodic Dataset Releases (PDS) offer information on Enroll-HD participants and are shared periodically. Each PDS comprises several files categorized into three groups: Participant-based (general study-independent information about the participant), Study-based (specific information about a participant within a study) and Visit-based (visit-dependent information for the study).

1.2.1 Data Structure

This project uses the Enroll-HD PDS-5 dataset (extraction date on October 31, 2020), which includes 21,116 participants. PDS-5 comprises 11 data files that are intrinsically connected through key variable components (as illustrated in Figure 1).

Participant-based data files contain information about the participant, including profile information, pharmacological and non-pharmacological therapies, and comorbid conditions and surgeries. Study-based data files store specific details about study participation, including participant status, study start and end dates, and whether the participant is enrolled in multiple studies. Finally, the visit-based data files contain information about the assessments and clinical data that were collected during each visit. [13]

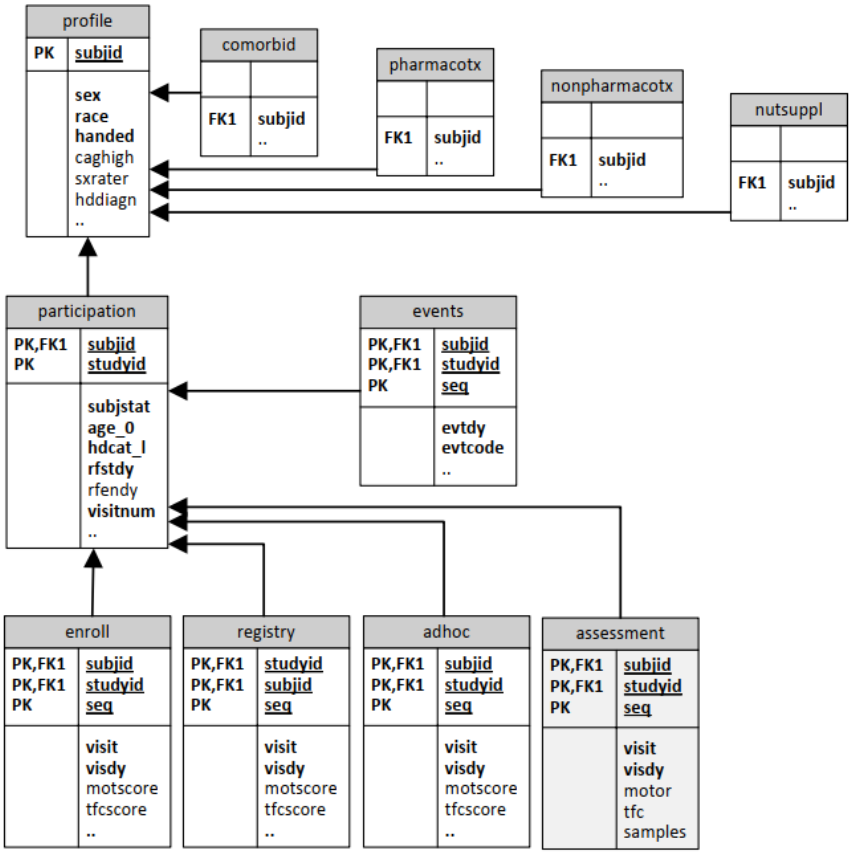

Figure 1: Diagram depicts the Enroll-HD entity relation, illustrating the relationship between the data file components and their key variables (primary keys [PK] and foreign keys [FK]) necessary for combining the data files. Diagram extracted from Enroll-HD Data Dictionary document [13].

1.2.2 Data Statistics

Enroll-HD PDS-5 documentation shares statistics about participants data [14]. Some of them characterized with respect to participant category, sociodemographic variables and clinical characteristics (as demonstrated in Figure 2).

An important variable called *hdcat* stores the participant category. It can take on the values 'manifest', 'pre-manifest', 'genotype negative', and 'family control'.

Annual study visits take place during the participant's routine clinical care visit. The baseline and annual study visits last between 45 minutes and a maximum of 2.5 hours [15]. A total of 55,975 annual study visits happened for the 21,116 participants over the years. Table 1 displays the distribution of annual study visits per maximum number of participants.

Visits	1	2	3	4	5	6	7	8	9
Participants	21116	14368	9498	6079	3261	1284	321	47	1

Table 1: Maximum participant counts for a specific number of Enroll-HD visits.

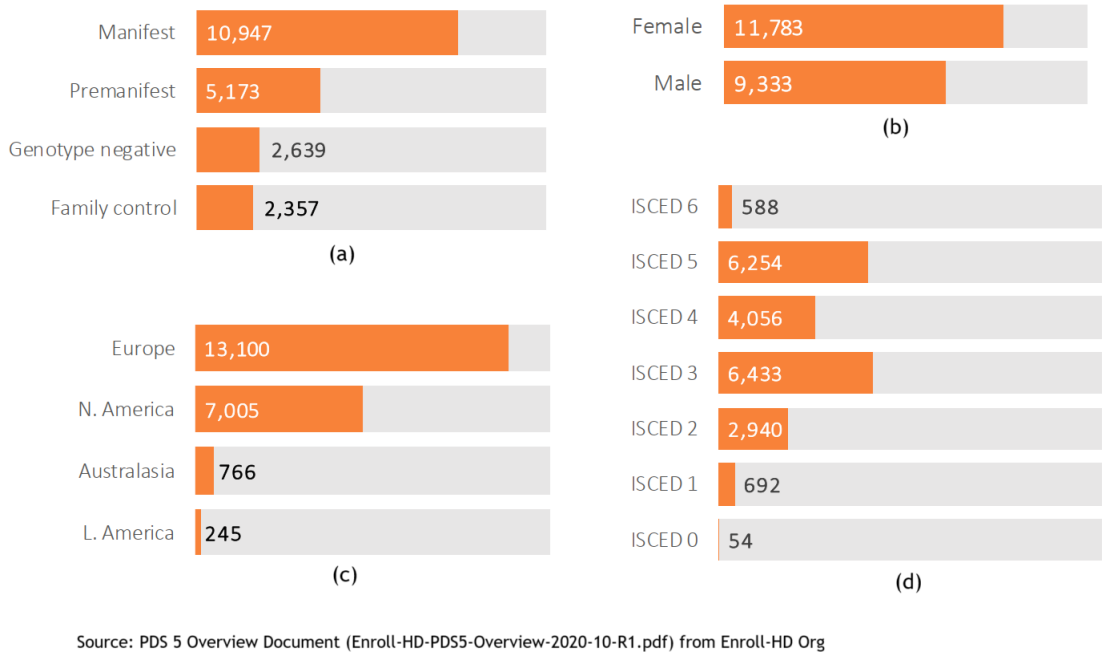

Figure 2: Figure shows samples of Enroll-HD PDS-5 statistics. Legend contains the parameter description and its variable data file name in parenthesis. **(a)** Participant category (hdcat) at baseline Enroll-HD visit. **(b)** Sex (sex). **(c)** Geographical region (region) **(d)** ISCED (isced) at baseline Enroll-HD visit. Figure extracted from Enroll-HD Data Support Document [14].

1.3 Research Objectives

The identification and diagnosis of rare diseases can be challenging due to the limited availability of data and expertise. Despite HD being a rare disease, the Enroll-HD dataset, which is sufficiently large, has enabled machine learning approaches to be used. Machine learning (ML) is a sub-field of Artificial Intelligence that utilises an algorithm, referred to as a model, to process and analyse data [16]. The data is used to train the ML model to make decisions or draw conclusions. After training, predictions can be made based on new data.

The main objective of this research is to improve the estimation of AAO for HD, which currently relies solely on CAG repeat length. Enroll-HD provides high-quality longitudinal data on HD patients, making it a valuable resource for ML algorithms.

1.4 Overview

Chapter 2 outlines the methods employed to address the research objectives, with relevant background information. The results of the experiment are shared in Chapter 3. Chapter 4 discusses all achievements and findings, proposes future works and concludes this research.

2 Methods

The major steps in ML for data analysis are as follows: data preparation, which encompasses data pre-processing, data selection, target definition and feature selection; algorithms or model proposition, which includes tasks such as feature scaling, split train/test datasets and method selection; model training, including hyper-parameter tuning; and results evaluation, with prediction and graph generation. These steps are illustrated in Figure 3 and were followed by this project. A comprehensive explanation of each step is provided in the subsequent sections.

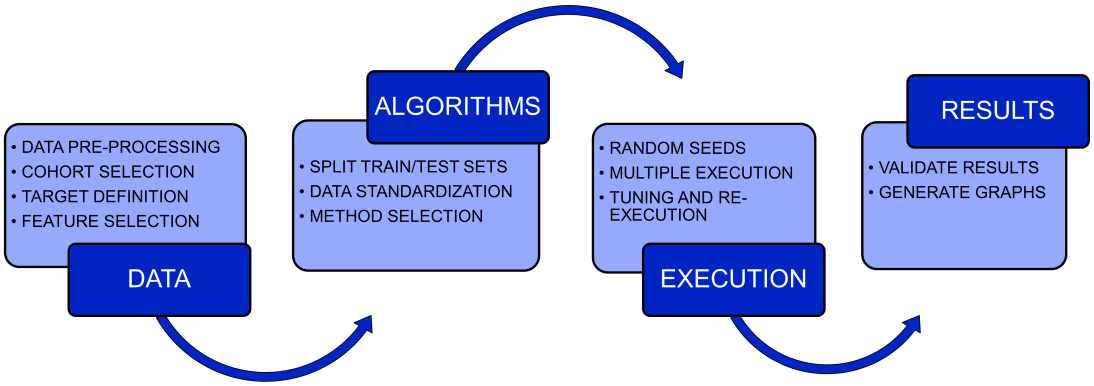

Figure 3: General method flowchart.

2.1 Software and Data Availability

The algorithms coded for this project were built in Python language (version 3.8.10), and used the processed and integrated HD Dataset called Enroll-HD. The Enroll-HD periodic dataset PDS5 was used. Access can be requested at Enroll-HD Data Access [17]. The Python Jupyter notebooks developed for this project can be reached at the cloud-based Git repository, GitHub [18].

Most of the code was developed using Jupyter Notebooks [19]. Jupyter is a project aimed at developing open-source software, open standards, and services for interactive computing across multiple programming languages. A Python library has been developed to support all the modules used in these notebooks. Table 2 presents the versions of the Python libraries used. Certain criteria and assumptions were used to guide the coding for this project. This includes the concept of randomness and the use of specific evaluation metrics.

Package	Version
keras	2.13.1
matplotlib	3.7.3
numpy	1.22.4
pandas	1.5.3
scipy	1.7.3
seaborn	0.13.0
sklearn	1.3.2
tensorflow	2.13.1

Table 2: List of Python libraries used into this project.

2.1.1 Randomness

The use of a fixed seed assists ML developers in achieving deterministic test execution and relieves them of the burden of dealing with randomness. However, it is unclear whether this is always the optimal approach or if there are alternative methods to address flakiness (when a test passes and fails non-deterministically for the same version of code) [20].

Although useful, this research acknowledges that obtaining an improved estimation of AAO cannot rely solely on fixed parameters. The use of fixed seeds was limited to certain scenarios, while experiment evaluations and results were based on the average values of multiple executions using random seeds. To ensure reproducibility of the results, a list of the seeds used was provided in the Jupyter notebooks.

2.1.2 Evaluation Metrics

The Mean Absolute Error (MAE) was the primary evaluation metric used to assess the accuracy of the forecasting model [21]. MAE is a statistical measure that calculates the average magnitude of the errors between predicted and actual values in the units of the response variable [22]. In this case, the response variable is years. When predicting the target, MAE accumulates absolute errors, so a lower measurement is preferred. The formula for MAE is:

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

The coefficient of determination, also called R-Squared (R^2), was used as a second evaluation metric. This metric was first introduced in 1921 [23]. In regression, R^2 is a statistical measure of how well the regression line fits the data. It represents the proportion of the dependent variable's variation that can be predicted by the independent variable(s). The coefficient of determination typically ranges from 0 to 1, with 1 indicating the best performance. Therefore, a higher measurement is desirable. R^2 can be expressed by the formula:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

During the training process, a third metric called Root-mean-square error (RMSE) was included. RMSE is a widely used measure for evaluating prediction quality [22]. It represents the standard deviation of the residuals, which are prediction errors. Residuals indicate the distance of data points from the regression line, while RMSE measures the spread of these residuals. In essence, it indicates the concentration of data around the line of best fit. RMSE formula can be represented as below:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

2.2 Data Preparation

By combining all Enroll-HD data files, it is possible to evaluate almost 200 variables per participant. For the purpose of using ML algorithms to analyse this data, it is essential that all variables are available for all observed participants. Unfortunately, this is not always the case, with PDS-5 having an estimated 47.15% of values being missing [24]. This can be explained by different factors, such as errors, inconsistencies, variables not being applicable to specific patients, or simply missing data. Thus, PDS-5 dataset required a pre-processing step prior to utilising ML algorithms.

2.2.1 Data Pre-Processing

To prepare the data for analysis by ML algorithms, various manipulation techniques were used. The BioSemantics team [25] at Leiden University Medical Center (LUMC) has been working with the Enroll-HD PDS-5 dataset for years. A pre-processing workflow was developed to manage this data. The workflow merges data files ('*enroll*', '*registry*', '*adhoc*', '*profile*' and '*participation*'), filters the necessary data for analysis, corrects inconsistent data, imputes missing values, and outputs the result. The workflow is represented in Figure 4.

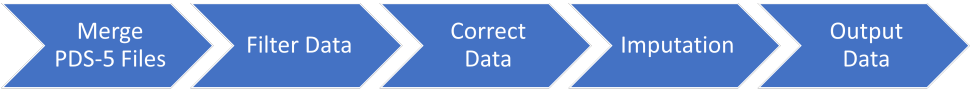

Figure 4: Flowchart representation of the workflow created by LUMC BioSemantics for the pre-processing of Enroll-HD dataset.

A valuable step in this workflow is the data aggregation that occurs when the results are output. The data aggregation reduces the number of features either by creating score variables based on assessments or by combining redundant information into fewer variables. To illustrate, rather than maintaining two distinct variables, namely '*momagesx*' (age when mother manifested HD) and '*dadagesx*' (age when father manifested HD), a single variable, designated '*parentagesx*' (either mother or father manifested HD), is employed to store the relevant information. In the event that both parents are afflicted with the disease, the variable '*parentagesx*' stores the AAO of the father. Another illustrative example is the creation of a variable to store whether the participant engages in drug abuse. A single new variable, named '*nmdrg*', is used to aggregate the abuse of different listed drugs, such as cocaine, marijuana, heroin, club drugs, amphetamines, Ritalin, hallucinogens, inhalants, opium, painkillers, barbiturates, tranquilizers, and others. Dealing with many variables can be complicated for any ML algorithm, so reducing the number of features used is extremely helpful for analysis.

The workflow also handles the transformation of string variables into numerical ones. After manipulating the variables, it may provide a name adjustment. Variables suffixed with '*_filled*' indicate that they have been transformed, while those suffixed with '*_impFill*' indicate an imputed filled feature. It is possible for variables to have both suffixes.

2.2.2 Data Selection (cohort)

Focusing on making AAO prediction, some filters were applied during the pre-processing stage (section 2.2.1). This research focused on adult patients AAO estimation, so juvenile HD cases were removed from the dataset. Cases where the participant was not an HD genetic carrier were also removed.

The Enroll-HD patient dataset comprises participants who were enrolled both before and after exhibiting symptoms of HD. The patient's category at the time of enrollment is stored in the 'hdcat_0' variable, while their current category is stored in 'hdcat_l'. A value of 2 indicates 'pre-manifest' (gene mutation carriers that do not exhibit any symptoms) and a value of 3 indicates 'manifest' (enroll-HD participants that exhibit symptoms of HD). To ensure greater precision and accuracy in determining the age at onset of HD, only those patients who were enrolled as pre-manifest and demonstrated symptoms during yearly visits (reclassified as 'manifest') were included in our analysis. Using a more selective cohort for training and validation purposes would increase the reliability of the ML model in determining the AAO. Patients enrolled as manifest were selected for ML model testing purposes. Figure 5 illustrates the step-by-step filtering process that was used to determine the cohort. Enroll-HD is a longitudinal dataset and may contain multiple registers for the same patient. To ensure objectivity and avoid bias, this project has selected only one entry per patient, specifically the first row when the patient became manifest.

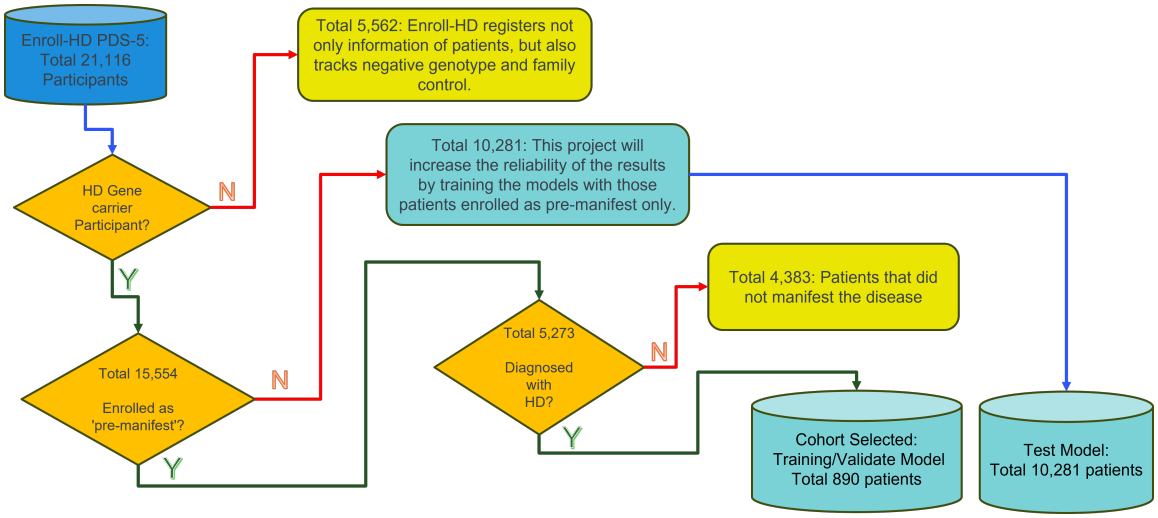

Figure 5: Cohort selection.

The length of the CAG extension is one of the best variables available for grouping and analysing patient characteristics. The graph in Figure 6 correlates the patients, either enrolled as pre-manifest or manifest, with the CAG extension length. Due to its limited scale visualization, a second graph, represented in Figure 7, illustrates only the curve of patients enrolled as pre-manifest (in blue). These graphs allow for the visualisation of the similarity between all curves.

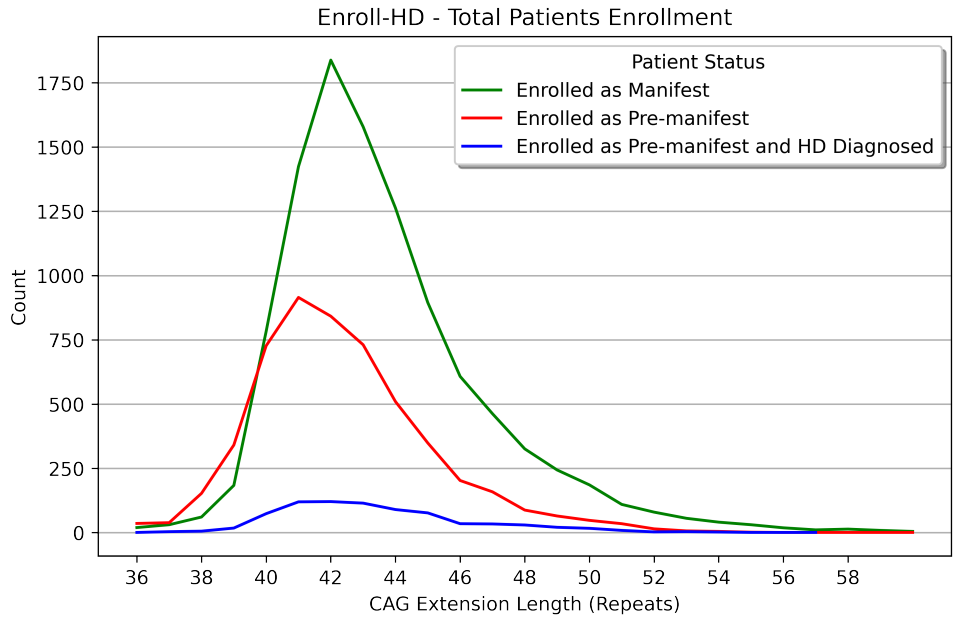

Figure 6: Graph shows the distribution of patients based on their CAG length extension.

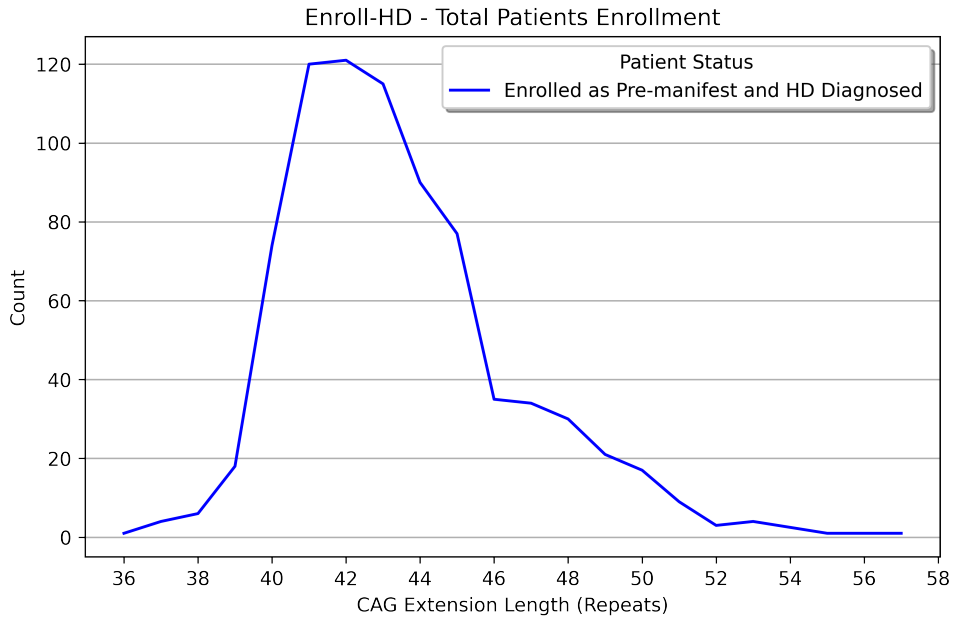

Figure 7: Graph is an amplification of the patients enrolled as pre-manifest and later diagnosed with HD.

2.2.3 Target Definition

HD typically causes a range of movement, cognitive, and psychiatric disorders, each with its own set of signs and symptoms. Enroll-HD provides information of each possible onset, which is grouped into motor ('*ccmtrage*'), cognitive ('*cccogage*'), and behavioural categories. The behavioural category has distinct onsets for aggressive behaviour ('*ccvabage*'), apathy ('*ccaptage*'), depression ('*ccdepage*'), irritability ('*ccirbage*'), perseverative obsessive behaviour ('*ccpobage*'), and psychosis ('*ccpsyage*'). The diagnosis of HD can be made based on any of these three different groups. The HD onset (in years) is then recorded as a variable called '*hddiagn*'. The variable '*hddiagn*' covered all possible onset scenarios and was therefore selected as the target for training the prediction algorithms for this project. To fill in any missing values for this variable, the cohort selection algorithm used the following content in order of priority: '*ccmtrage*' (motor onset), '*sxrater*' (rater's estimate of symptom onset), '*sxfam*' (age when symptoms were first noted by family) and '*sxsubj*' (age when symptoms were first noted by participant).

2.2.4 Feature Selection

Feature selection is the process of reducing input variables in a model by using only relevant data and eliminating noise [22, 26]. It programmatically selects relevant features for an ML model based on the problem being solved. Reducing input variables is desirable to both decrease computational costs and improve model performance [27]. Statistical-based feature selection methods assess the relationship between each input variable and the target variable. The input variables with the strongest relationship to the target variable are selected.

Prior to the feature selection process, a preliminary selection of variables from the Enroll-HD dataset was conducted. The objective was to minimize the potential for introducing noise and bias into the model [28]. The model was designed to accommodate a heterogeneous patient population, with the capacity to predict the onset of HD in those who are closer to being diagnosed with the disease, or even many years before it. The information extracted from Enroll-HD to train the model is the first entry that registers when the patient is manifesting the symptoms of HD. Consequently, all variables related to the assessment visits were removed. This was mainly because the assessment information measured when the participant is diagnosed with HD is already compromised (thereby introducing a degree of bias into the model). The pre-selection was performed by filtering information related to participants (labeled as 'patient profile'), which includes data such as sex, race, weight, and lifestyle factors (labeled as 'lifestyle aspect'), which includes variables such as caffeine use and drug abuse. The variables listed to be analyzed by the feature selection process are presented in Table 3.

Variable	**Category**	**Description**
caghigh	patient profile	Research larger CAG allele determined from DNA
caglow	patient profile	Research smaller CAG allele determined from DNA
sex	patient profile	Gender
race	patient profile	Ethnicity
region	patient profile	Continent
maristat	patient profile	Marital Status
isced	patient profile	ISCED education level
parenthd	patient profile	Father or mother affected
parentagesx	patient profile	Age at onset of symptoms in father or mother
emplnrsn	patient profile	Reason (if not employed)
height	patient profile	Height
weight	patient profile	Weight
handed	patient profile	Handedness
res	patient profile	Residence
jobclas	lifestyle aspect	Employment status
hxtobab	lifestyle aspect	Has the participant ever smoked
hxtobcpd	lifestyle aspect	Cigarettes per day
hxtobyos	lifestyle aspect	Years of Smoking
hxalcab	lifestyle aspect	Has the participant had alcohol problems in the past
alcunits	lifestyle aspect	Alcohol usage: Units per week
nmdgr	lifestyle aspect	Drug abuse
nmdgr	lifestyle aspect	Drug abuse frequency
cafab	lifestyle aspect	Current caffeine use
cafpd	lifestyle aspect	Does participant drink more than 3 cups of coffee, tea and cola drinks combined per day?

Table 3: Pre-selected Enroll-HD variables used to feed the feature selection algorithm. The names may suffer adjustments according to the pre-processing manipulation.

The algorithm for feature selection was composed using three feature selection techniques, all of which shared the same set of input data for training and testing. The data was randomly split by the usage of '*train_test_split*' python library module (available in Scikit Learn [29]), with 80% of the cohort dataset selected for training and 20% for validation. The approach employed to identify the variables to be utilised at the final feature selection stage involved the execution of multiple runs, with the addition of a single feature at a time. The model's performance was then evaluated as the feature set was incrementally increased.

2.2.4.1 SelectKBest - This feature selection technique is available in Scikit Learn Python library [30]. The feature selection on this method is performed by evaluating the relationship between each feature and the target variable. Statistical tests such as the chi-squared test, ANOVA F-test, or mutual information score are used by this method to score and rank the features based on their relationship with the output variable [31]. The K features with the highest scores are then selected to be included in the final feature subset. The higher the score, the more relevant the feature is considered. The SelectKBest object is instantiated with a scoring function that selects the feature selection method. For this project, the '*f_regression*' scoring function was chosen. This function performs univariate linear regression tests and returns F-statistic and p-values [32].

2.2.4.2 Mutual Information - Information Gain (IG) calculates the reduction in entropy or surprise resulting from transforming a dataset. It is a measure used in decision trees to determine the relevance of a feature and define the basic criteria for splitting a node [33]. IG can also be used for feature selection, by evaluating the gain of each variable in the context of the target variable. The Python library *mutual_info_regression*, from Scikit Learn [34], estimates mutual information for a continuous target variable using the principle of IG. It relies on non-parametric methods based on entropy estimation from k-nearest neighbors distances.

2.2.4.3 Lasso - This is a linear model from Scikit Learn [35] that estimates sparse coefficients. This method can be beneficial as it tends to prefer solutions with fewer non-zero coefficients. If two features are linearly correlated, their joint presence will increase the value of the cost function. Therefore, Lasso regression will attempt to reduce the coefficient of the less significant feature to 0 to select the best features [22].

2.3 Proposed Models

In order to select an appropriate model for training purposes, it is necessary to undertake a series of preparatory steps on the data. Firstly, the features must be rescaled. Next, the cohort must be divided into training and validation datasets.

2.3.1 Feature Scaling

Datasets frequently contain various types of data, with multiple dimensions. In Enroll-HD, features may have a broad range of values, including years, boolean variables, and categorical values (where the values simply represent a different classification). Using the original scale may give more weight to variables with a larger range. To address this issue, it is necessary to apply feature scaling to the independent variables. This project used standardization, also known as "Z-score normalization", as a feature scaling technique. Standardization rescales the values of each feature in the data to ensure that the mean and standard deviation are 0 and 1, respectively [22]. The standardization process was applied before the training process. The Scikit Learn library's *StandardScaler* [36] method was utilised. The standardization equation is shown below:

$$x' = \frac{x - \bar{x}}{\delta} = \frac{x - mean(x)}{standard\ deviation(x)}$$

2.3.2 Training and Validation datasets

Training data, also known as a training dataset, is the initial set of data used to train ML models. ML algorithms are trained using these datasets to enable them to make predictions. The validation dataset is a subset of the cohort dataset used to evaluate the trained model. Both the training and validation datasets must be diverse, representative, and unbiased. To adhere to these assumptions, we used the python library module ('*train_test_split*'), which prepared the training and validation datasets, for both feature and model selection. This module was also used during the training process. As a new or random seed was employed, a new set of test and training datasets was generated as a consequence.

2.3.3 Model Selection

This project evaluated two different model approaches on how to predict the HD AAO. One approach was based on the called classic ML models, as referenced in ref. [37]. The other approach was based on neural network algorithms. As neural networks represent a specific type of ML method, which will be explained in more details below, for the purposes of this discussion, ML models will be used to refer to those non-neural network models.

2.3.3.1 ML Models - Multiple ML models were tested, such as *LinearRegression*, *RandomForest*, *MLP*, *CatBoosting* and *XGBoost*, which were also evaluated by Ouwerkerk J. et al. (2023) [24]. Furthermore, additional models were included into this list, such as *ExtraTree*, *AdaBoosting* and *Bagging*. The models were submitted to multiple executions using their default parameters and the cohort dataset. A final and definitive evaluation was conducted utilising the three most efficacious ML models, in conjunction with the a neural network model. The models underwent into a minor hyper-parameter adjustment during the final evaluation. The results can be verified at section 3.2.

2.3.3.2 Neural Networks - Neural network (NN) is a type of ML algorithm that form the basis of deep learning (DL) models. NNs consist of layers of decision-making nodes, including an input layer, multiple decision-making layers, and an output layer [16, 22]. Each node is an artificial neuron that makes a computational decision based on its weight and threshold [22]. Although our cohort was relatively small for NNs, we decided to train a NN model and compare the results with the ML models. This project utilised a Feedforward Neural Network (FNN) model. FNNs were the first type of artificial neural network invented. Its architecture consists of three types of interconnected layers of neurons: the input layer, hidden layers, and the output layer [38]. The interconnection between layers is achieved through weights. Each neuron in one layer is connected to every neuron in the next layer, making this a fully connected network. The information moves in only one direction—forward—from the input nodes, through the hidden nodes, and to the output nodes [22], as illustrated in Figure 8. The TensorFlow/Keras [39] Python library was used to support this model.

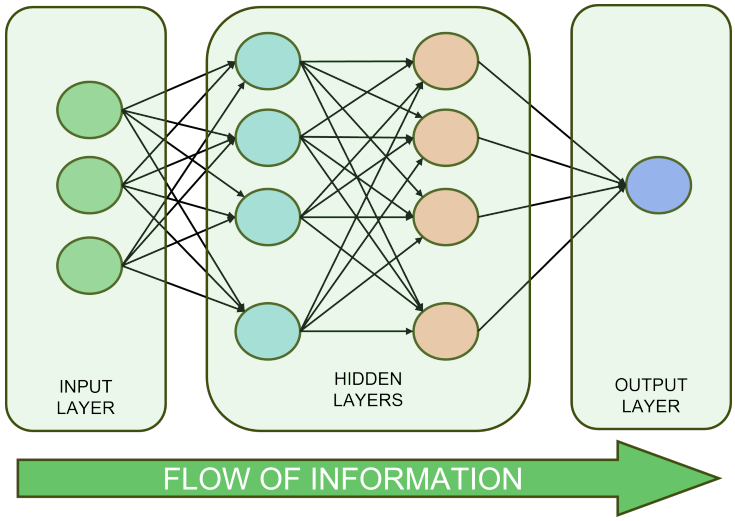

Figure 8: Typical FNN architecture.

2.4 Training the Models

In addition to the FNN model, the most effective ML model was selected for the training procedure. During the training execution, both models were optimised by adjusting their hyper-parameters.

2.4.1 Hyper-parameters Tuning

Hyper-parameters are configuration variables used in ML to control the learning process. They are set before the model training process begins, and adjusted as the results are verified. The objective of hyper-parameter tuning is to identify the values that result in optimal performance for a given task. Models may have numerous hyper-parameters, and determining the optimal combination of parameters can be approached as a search problem [40]. GridSearchCV is a

method from the Scikit Learn library that was used to search for the best set of parameters. This approach fits the model using all possible combinations after creating a grid of potential hyper-parameter values [41]. In this project, we used this method to tune the hyper-parameters of CatBoost. A comparable methodology was employed to optimise the parameters of the FeedForward Neural Network. However, in this instance, a distinct code (accessible via GitHub) was developed to enhance control and visibility over the validation process. Multiple runs were executed until a final set of parameters was determined.

2.4.1.1 CatBoost hyper-parameters - The tuning process for CatBoost involved manipulation of various parameters. The final configuration is presented in Table 4. All remaining parameters were set to their default values.

Parameter	**Description**	**Value**
Depth	Depth of the trees	3
Iterations	The maximum number of trees that can be built	1000
Learning Rate	Used for reducing the gradient step	0.01
L2 Leaf Reg	Coefficient at the L2 regularization term of the cost function	1
Seed random state	Used to make the behavior of the model deterministic	random

Table 4: CatBoost parameter tuning details.

2.4.1.2 FeedForward Neural Network hyper-parameters - The FNN architecture was designed with a single node on the output layer and two hidden layers in the intermediate structure. The final approach and its set parameters are listed in Table 5. Remaining parameters were set to their default values.

Parameter	Description	Value
Input Layer	Total neurons on input layer	128
Hidden Layers	Total neurons on first hidden layer	256
Hidden Layers	Total neurons on second hidden layer	128
Output Layers	Total neurons on output layer	1
Kernel Initializer	Initializer for the kernel weights matrix	normal
Optimizer	Optimizer algorithm	adam
Learning Rate	Step size taken by the optimizer during each iteration	0.005
Epsilon	A small constant for numerical stability. This epsilon is "epsilon hat" by Kingma and Ba [42]	0.001
Activation Function	Calculates the output of the node	ReLU
Batch Size	Number of samples per gradient update	125
Epochs	Number of times the entire training dataset is passed through the model	10
Seed	Used to make the behavior of the initializer deterministic	random
Loss function	Compute the quantity that a model should seek to minimize	MAE

Table 5: FNN Architecture and parameter details.

2.4.2 Training Execution

The models were trained in two distinct ranges. The first range, designated the penetrant range, encompassed patients exhibiting a CAG repeats extension between 40 and 60. The second range, encompassing the full spectrum of patients, included those with a CAG repeats ranging between 35 and 60.

The data was standardized during the training process. Once the model was trained, the target and estimated values were returned to their original scale, so the evaluation could be more realistic, being measured in years. Although the target information is presented as an integer variable in years, the estimated output has not been rounded, in order to allow for a more precise evaluation of models performance. For each execution range, 500 runs were performed using a random seed and the selected cohort. The parameters used were obtained from the hyper-parameter tuning step. A list of the seeds used was exported for reproducibility purposes. The performance of each run was measured using the mean absolute error (MAE), the root-mean-square error (RMSE) and the coefficient of determination (R^2). For the purpose of comparison and easier comprehension, the MAE and RMSE values were converted back to their original scale once the dataset was standardized. This allows for the error margin to be read in years.

The final evaluation is based on the average of all runs. After each run, the model's efficiency was evaluated by examining the test set estimation. This project used the term average to indicate the arithmetic and statistic mean calculated over the values, once all values were

used and no outliers observed. To establish a baseline for the experience and compare the performance with a well-used estimation method, we applied the Langbehn formula to the test dataset.

The combination of hyper-parameters tuning and execution processes is demonstrated in Figure 9.

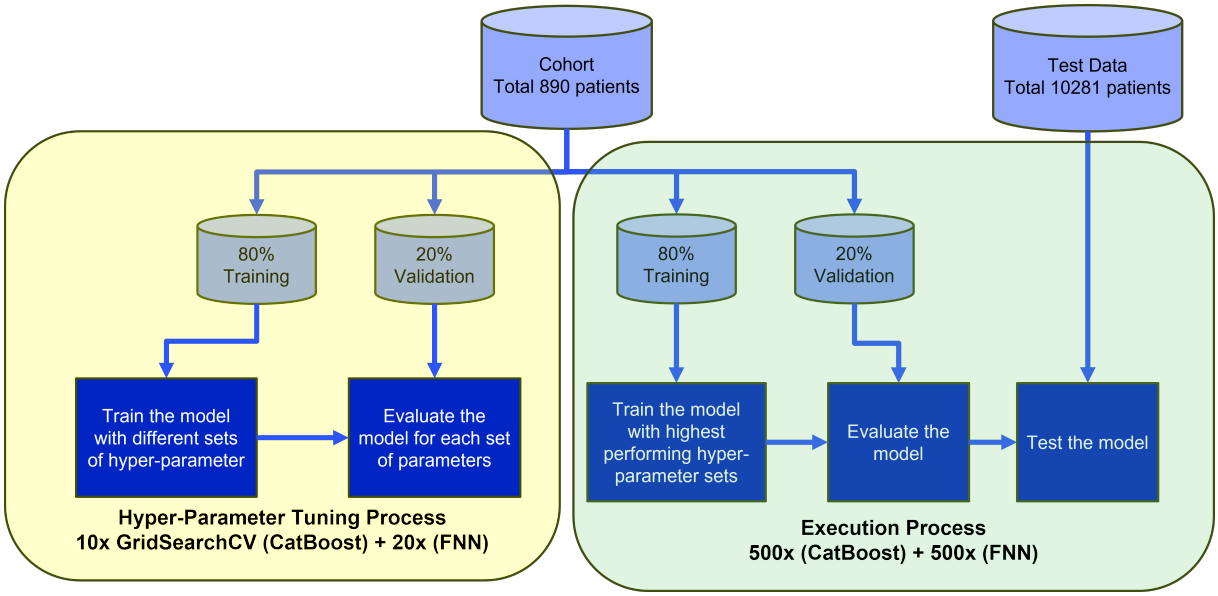

Figure 9: Figure illustrates dataset distribution for the hyper-parameter tuning and execution processes.

3 Results

This chapter presents the outcomes of the various processes implemented throughout the project. It is divided into four sections: the first section covers the selection of features; the second section shows the model selection; the third section presents the AAO estimations results; and the final section presents additional experiments and correlation analysis.

3.1 Selected Features

The feature selection algorithms evaluated and ranked the relative importance of the input variables (as previously listed in Table 3). It was found that '*caghigh*' had the highest impact on the target variable, as expected. Other variables, such as '*parentagesx*' also demonstrated significant impact. It is not coincidental that these two variables are proposed in the Ranen formula [49]. Table 6 demonstrates the percentage of importance of the 10 most relevant variables analyzed throughout the different feature selection methods, summing between 79.57% (Lasso) to 98.47% (Select KBest). Figure 10 demonstrates the level of importance and how it accumulates over the analyzed features.

Variable	**Select KBest**	**Mutual Information**	**Lasso**
caghigh	54.46	38.44	31.58
parentagesx_impFill	29.39	21.49	17.91
maristat_filled_impFill	6.72	7.67	6.63
emplnrsn_filled_impFill	2.41	8.58	6.69
hxtobyos_impFill	1.97	8.77	7.22
jobclas_filled_impFill	1.24	4.68	0.00
region_impFill	0.68	1.23	5.52
alcunits_filled_impFill	0.66	1.30	1.94
nmdrg_filled_impFill	0.50	1.14	0.00
cafab_filled_impFill	0.44	0.16	2.08
Total Accumulated	98.47	93.46	79.57

Table 6: Percentage of importance from different variables analyzed by feature selection methods.

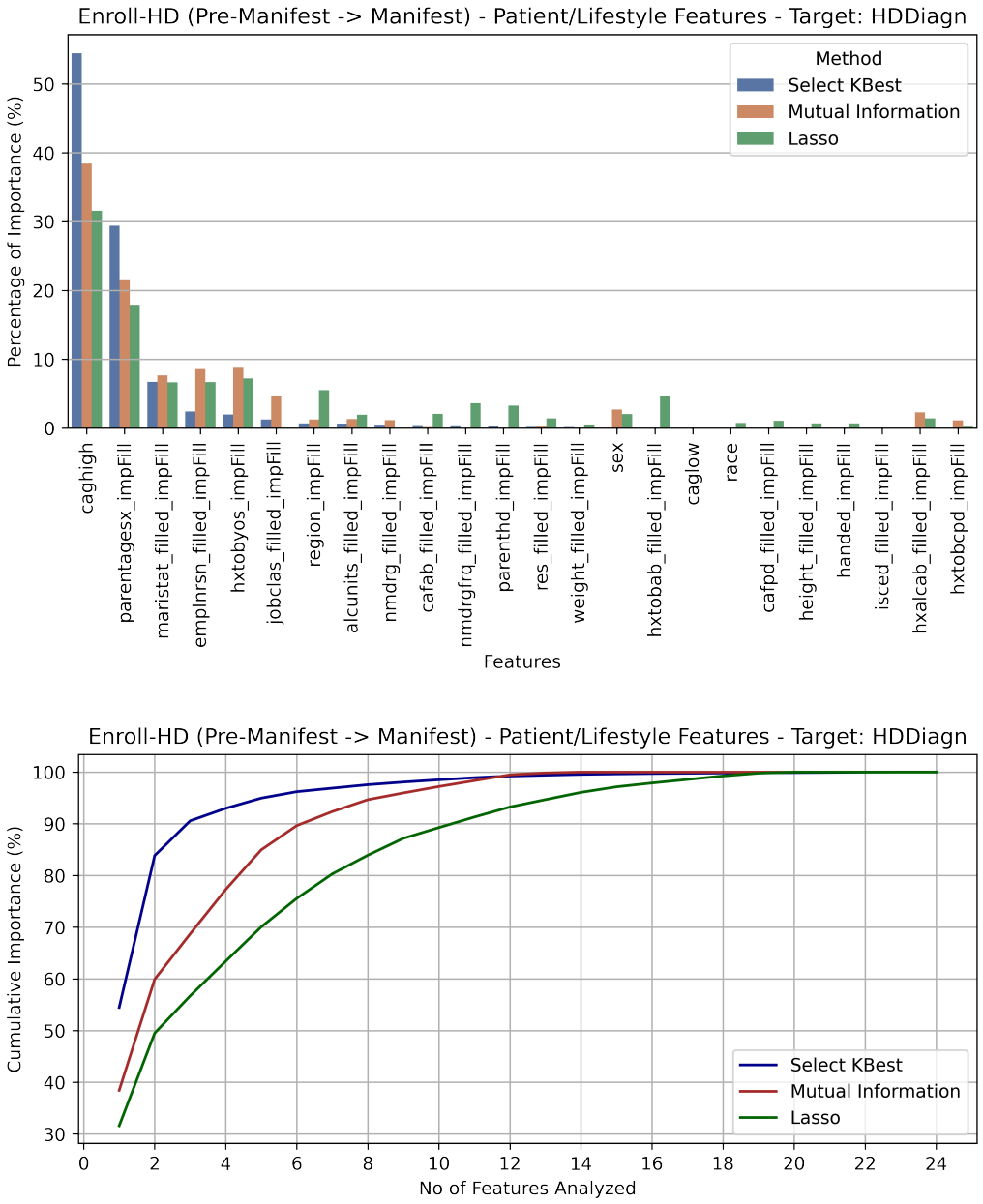

Figure 10: **On top**: Feature selection results comparing the three methods. SelectKBest provided the sort criteria used at this graph, from the most relevant to the less one. **On bottom**: Graphs illustrate the cumulative importance obtained by each feature selection method over feature analyses.

The feature selection process (section 2.2.4) tested a prediction evaluation by adding one feature at a time, and submitting them into a training and performance verification. The estimation response increased as more features were added, until the quality of the model began to degrade. At this point, the process ranked the variables for use in the training phase.

We evaluated multiple combinations of feature sets composed of the main relevant features, such as '*caghigh*', '*parentagesx_impFill*', '*hxtobyos_impFill*', '*emplnrsn_filled_impFill*', '*maristat_filled_impFill*', '*region_impFill*', '*nmdrgfrq_filled_impFill*' and '*hxtobab_filled_impFill*'. The performance achieved among these different feature sets did not demonstrate significant difference. Figure 11 illustrates how the error detected in estimating the AAO and how it differs slightly between different feature sets. Following the evaluation of the model's performance, the set of features selected for training the models were: '*caghigh*', '*parentagesx_impFill*', '*hxtobyos_impFill*', '*emplnrsn_filled_impFill*', '*maristat_filled_impFill*', '*region_impFill*' and '*nmdrgfrq_filled_impFill*'.

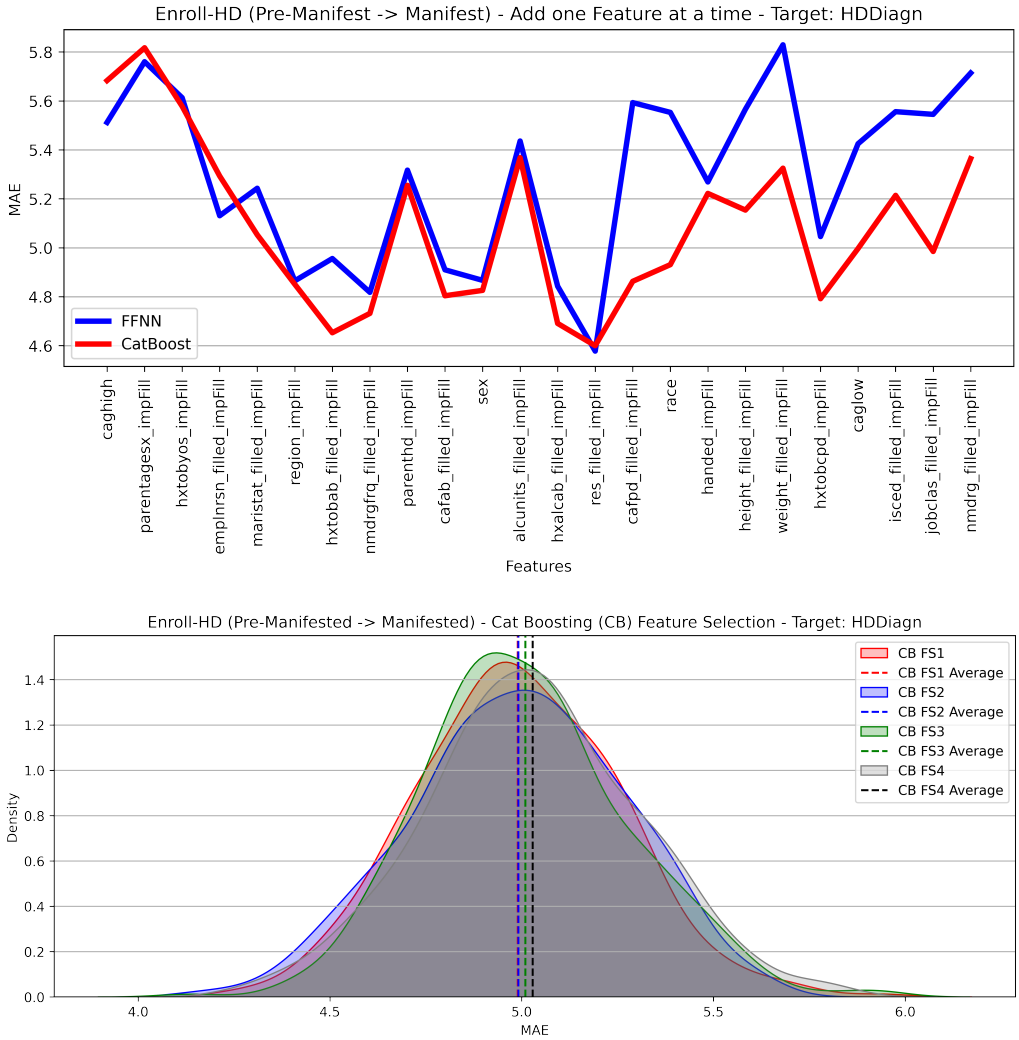

Figure 11: **On top**: Feedforward neural network and CatBoost ML model performance evaluated over error margin (MAE), when each variable is added to the feature set being trained. **On bottom**: Different features sets (from 1 to 4) training evaluation after multiple runs on CatBoost model.

3.2 Selected Models

The model selection process was conducted in two stages. In the initial stage, the models were evaluated using their default parameters and a fixed seed, over the penetrant range cohort. Table 7 shows the results.

Model	**MAE**	**RMSE**	R^2
Gradient Boosting	**5.091941**	**6.534478**	**0.645500**
Random Forest	**5.221921**	**6.759470**	**0.621520**
CatBoost	**5.277280**	**6.768308**	**0.620086**
Linear Regression	5.339975	6.821404	0.615203
Extra Tree	5.345446	6.877290	0.607164
Linear SVM	5.418970	6.882807	0.607589
FNN	5.526948	7.067763	0.585985
XGBoost	5.572010	7.181605	0.572636
Ada Boosting	5.616858	7.072828	0.584721
Bagging	5.691004	7.311850	0.557478
Langbehn	5.907501	7.833137	0.490806
MLP	7.058624	8.960238	0.334353
KNN	7.213200	8.959386	0.334674

Table 7: Model Selection.

The final stage of the model selection process involved the evaluation of the three most effective ML models. These were Gradient Boosting, Random Forest and CatBoost. Given the differences on methodologies employed, we have elected to include Ada Boosting and FNN models in this final round of evaluation. In order to ensure a better performance of the models, a hyper-parameter tuning process has been implemented. Figures 12 and 13 display the results from multiple runs using random seeds.

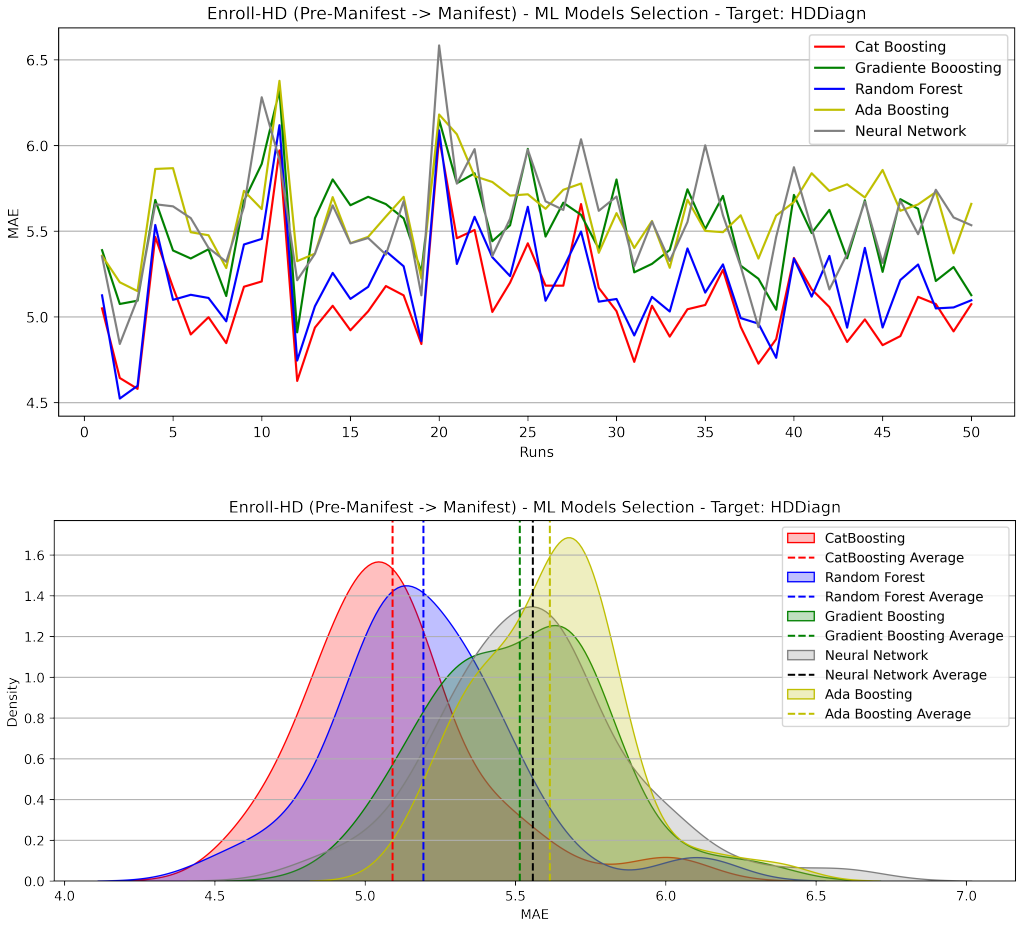

Figure 12: **On top**: Margin error analysis of different ML models after 50 runs. **On bottom**: Graphs illustrate the MAE density distribution over the 50 runs executed.

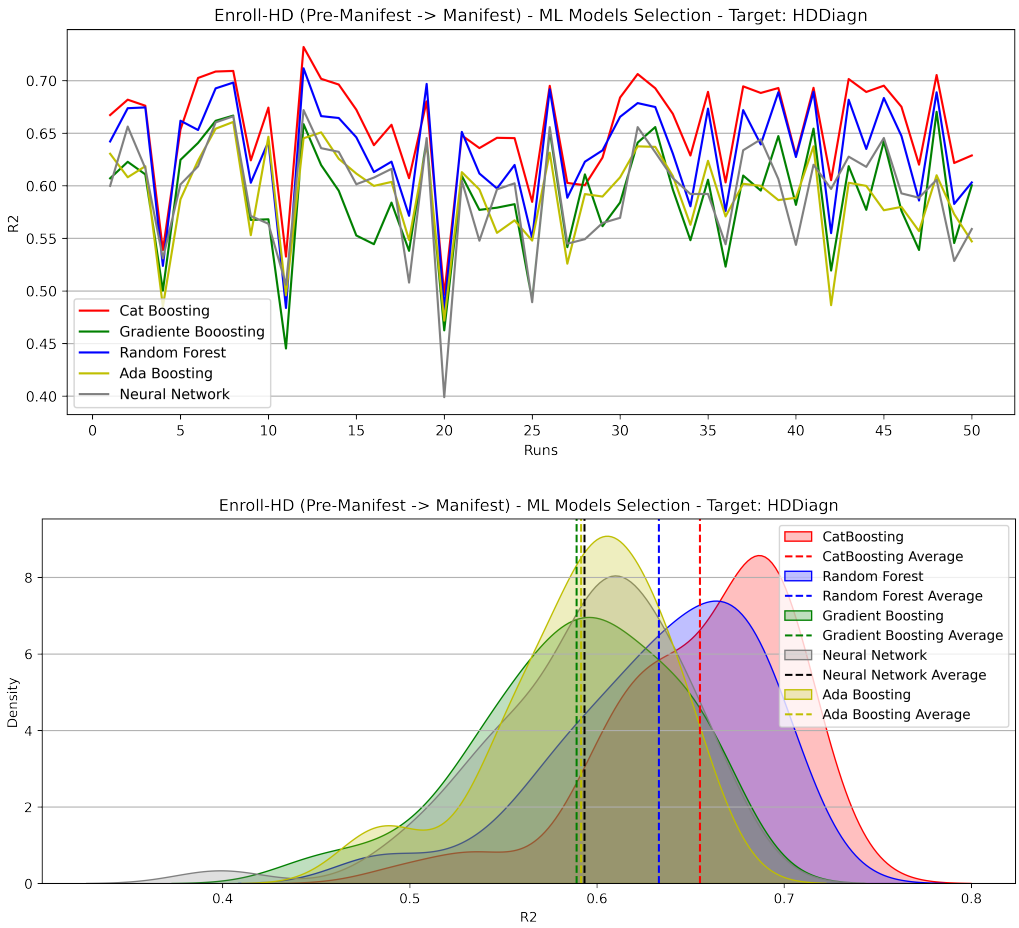

Figure 13: **On top**: R^2 analysis of different ML models after 50 runs. **On bottom**: Graphs illustrate the R^2 density distribution over the 50 runs executed.

CatBoost model was identified as the most effective ML model. The mean absolute error (MAE) results were significantly reduced on the CatBoost model, with higher efficiency observed in the R^2 evaluation. Besides CatBoost model, FNN was also selected for the AAO estimations, due to its neural network approach technique.

3.3 AAO Estimations

The AAO estimation results were divided into two categories: first one covering the 'penetrant range' patients (40 <= CAG length <= 60 repeats) and the second one for the 'full range' patients (35 <= CAG length <= 60 repeats). A comparative analysis was conducted between CatBoost ML (CatBoost), FeedForward Neural Network (FNN) models, and the state-of-the-art prediction formula (Langbehn Formula). This analysis was performed on both the training and test datasets. The Langbehn Formula was calculated exclusively using the test dataset, which served as the basis for comparison with the models during both the training and testing processes. Three different performance metrics were collected: MAE, RMSE and $R2$. The

presented results are an output of the multiple runs produced. It is important to note that even after performing a different set of 500 runs, the results differ only slightly. The following content represents the last execution, also shared on GitHub.

3.3.1 Penetrant Range Results

Penetrant range refers to those HD patients who are mutation carriers with CAG length greater than 40 repeats. The first metric evaluated was MAE. The initial graphical analysis demonstrates how this metric varies across each execution, both for CatBoost and FNN. This can be verified in Figure 14. The horizontal lines in figure represent the average of all the runs. During training, CatBoost had an average MAE of 5.017, while FNN had an average MAE of 5.242. Both models had lower MAE values than Langbehn, which had an average of 5.961. Testing returned an average MAE of 5.220 for CatBoost while 5.561 was measured for FNN. These values can be checked in Table 8.

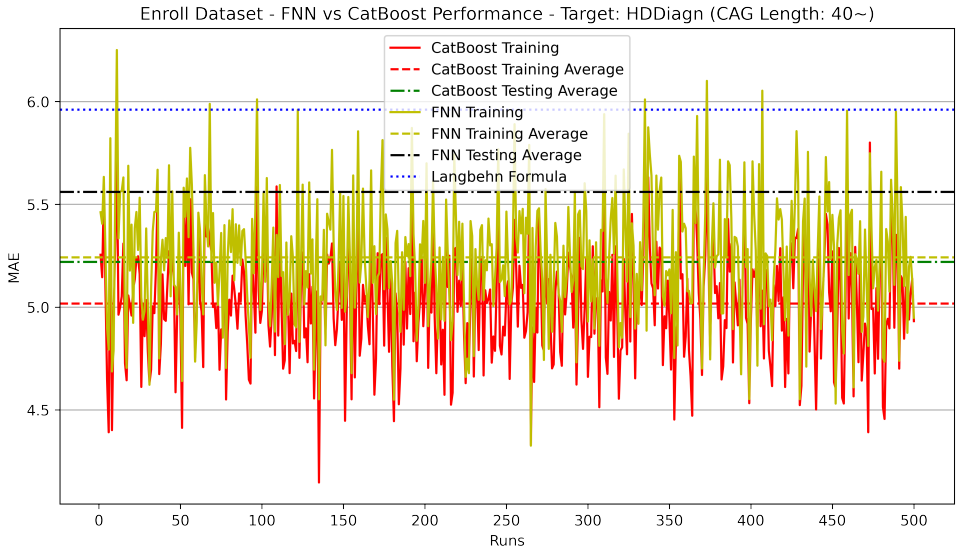

Figure 14: Penetrant range patients MAE measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and testing results.

Model	MAE (Training)	MAE (Test)
CatBoost	5.017453	5.220201
FNN	5.242455	5.560703
Langbehn*	5.960552	5.960552

Table 8: Penetrant Range - MAE average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

A kernel density estimate (KDE) plot of MAE training values was created to allow a better visualization. KDE is a technique used to display the distribution of observations in a dataset, similar to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions [44]. Figure 15 shows that the training curve shapes for both CatBoost and FNN are normally distributed, with the average vertical lines almost aligned with the center of the curves. Additional vertical lines were added representing the testing process.

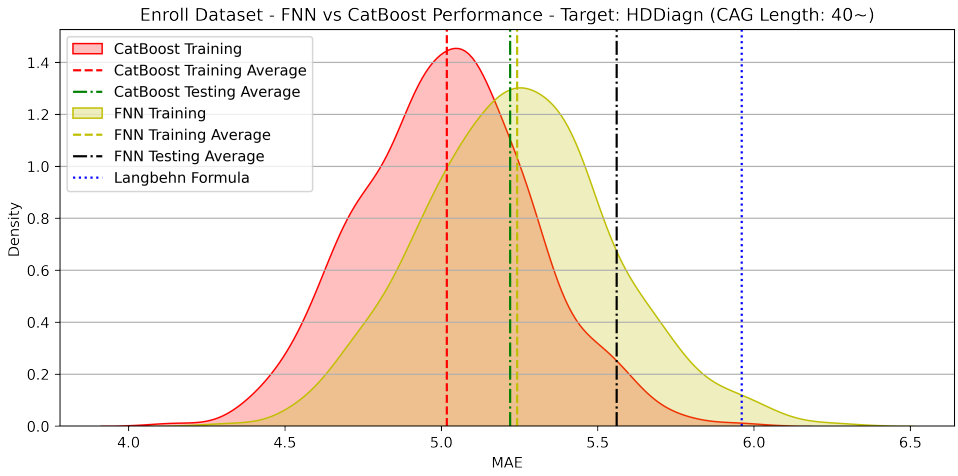

Figure 15: Graph shows the KDE well-distributed curves for MAE on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

The average, median, standard deviation and variance for the MAE values are presented, in Table 9 for training process and in Table 10 for testing process.

Model	Average	Median	Standard Deviation	Variance
CatBoost	5.017453	5.016913	0.272819	0.07443
FNN	5.242455	5.241287	0.301433	0.090862

Table 9: Penetrant Range - MAE Training Statistics.

Model	Average	Median	Standard Deviation	Variance
CatBoost	5.220201	5.219023	0.039583	0.001567
FNN	5.560703	5.556367	0.09596	0.009208

Table 10: Penetrant Range - MAE Testing Statistics.

A final MAE graph was generated, representing the same KDE distribution, with the addition of a testing process curve, as can be observed in Figure 16. The testing curve for density

distribution is very narrow, creating a clear distinction between the testing and training responses.

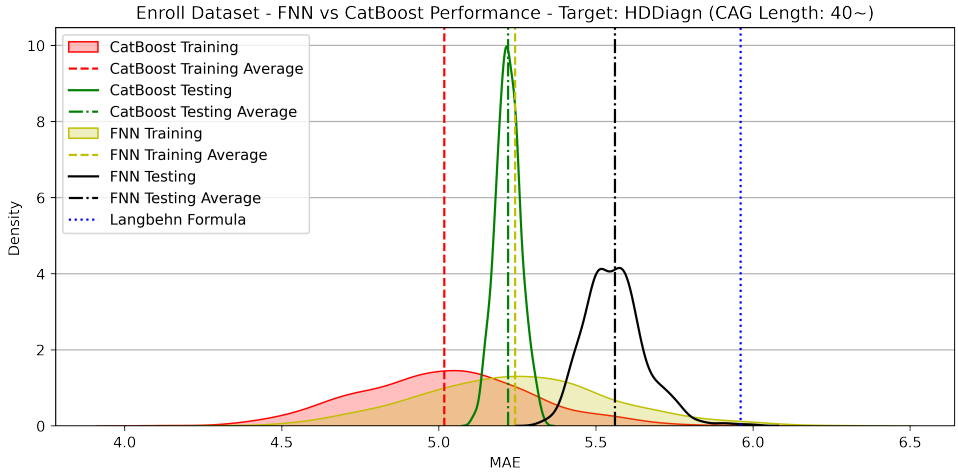

Figure 16: MAE KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

The second metric evaluated was RMSE, which, like MAE, measures the error margin. Therefore, the lower the value, the better. The results produced by RMSE are highly comparable to those produced by MAE, particularly in terms of data distribution. Table 11 compares the RMSE values between training and testing processes. Figure 17 illustrates the RMSE value measured for each training run. In order to facilitate a more comprehensive analysis, average horizontal lines were included, for both the training and testing measures.

Model	**RMSE (Training)**	**RMSE (Test)**
CatBoost	6.424858	6.778187
FNN	6.706516	7.195773
Langbehn*	7.692482	7.692482

Table 11: Penetrant Range - RMSE average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

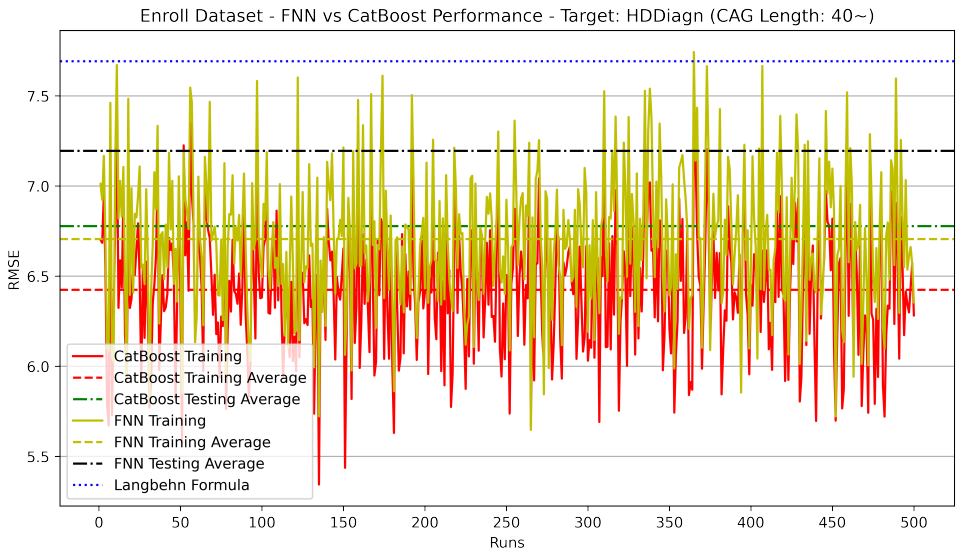

Figure 17: Penetrant range RMSE measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and testing results.

Tables 12 and 13 show complementary RMSE statistic measures. Figure 18 demonstrates the RMSE training density distribution, once again with well-balanced curves. As initially observed during MAE analysis, Figure 19 shows the narrow CatBoost testing curve, as an indication of how the model was well-adjusted to the test set.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	6.424858	6.422787	0.35412	0.125401
FNN	6.706516	6.698522	0.382076	0.145982

Table 12: Penetrant Range - RMSE Training Statistics.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	6.781384	6.779560	0.046852	0.002195
FNN	7.193796	7.188037	0.112857	0.012736

Table 13: Penetrant Range - RMSE Testing Statistics.

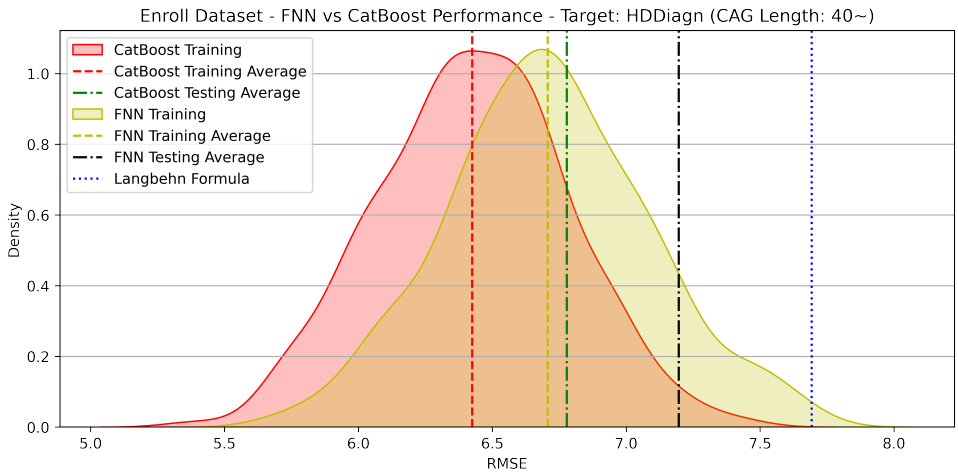

Figure 18: Graph shows the KDE well-distributed curves for RMSE on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

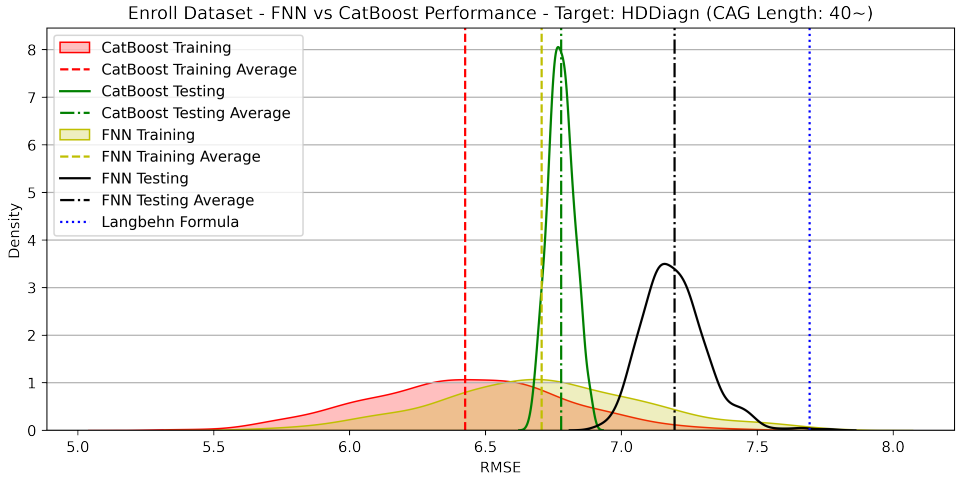

Figure 19: RMSE KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

Lastly, the coefficient of determination R^2 is analyzed. Unlike the previous two metrics, this measure is directly associated with the efficiency of the model. The maximum achievable value is 1, and in this case, a higher value indicates better performance. Table 14 shows the comparison between the models during training and testing processes.

Model	R^2 **(Training)**	R^2 **(Test)**
CatBoost	0.658454	0.669183
FNN	0.628005	0.627084
Langbehn*	0.573936	0.573936

Table 14: Penetrant Range - R^2 average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

The series of graphs generated for MAE and RMSE were employed in a similar manner for the calculation of R^2. Figure 20 shows the R^2 measures during the executions, while Figure 21 demonstrates the training KDE distribution. Besides the fact that higher performance for the models means higher R^2, both graphs have similar behavior compared to MAE and RMSE. The only difference lies in the reversed analysis.

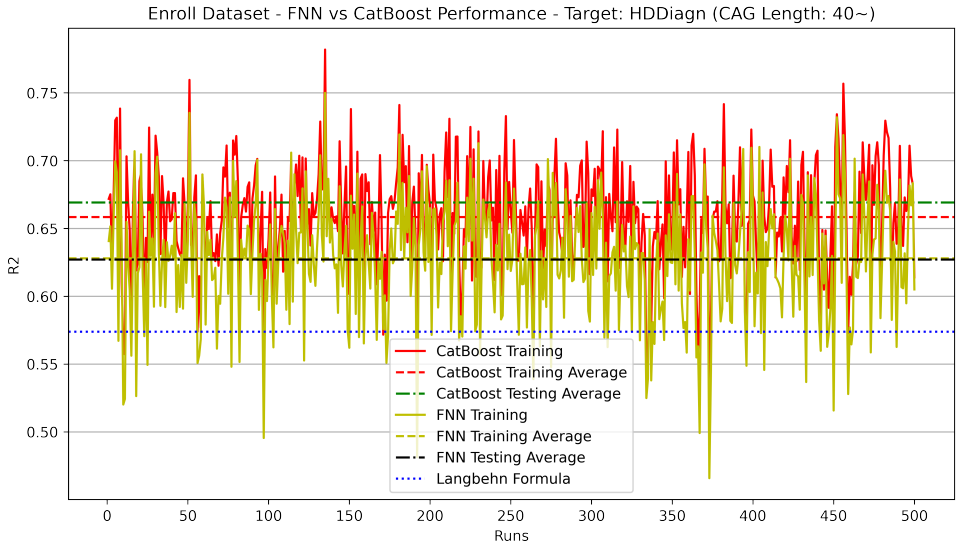

Figure 20: Penetrant range R^2 measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and testing results.

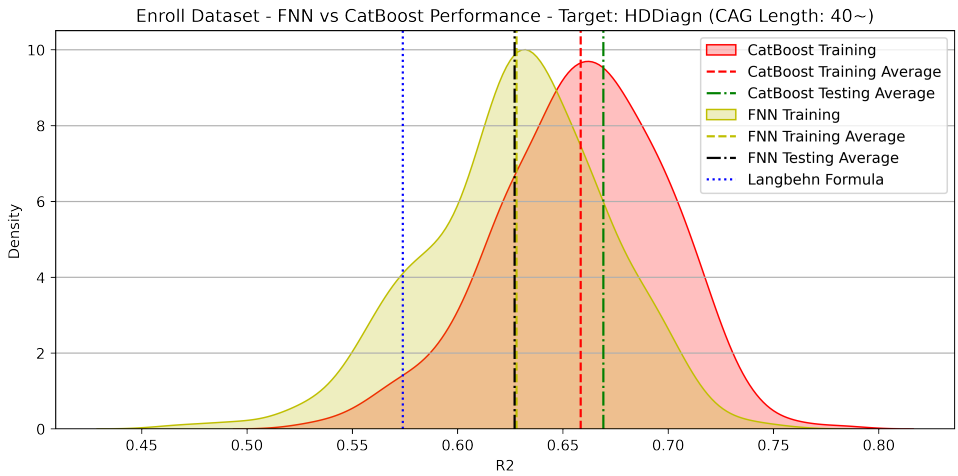

Figure 21: Graph shows the KDE well-distributed curves for R^2 on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

The average, median, standard deviation and variance for the R^2 values are presented in Tables 15 and 16. Figure 22 illustrates the KDE distribution, with the addition of the testing process curve.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	0.658454	0.659634	0.04006	0.001605
FNN	0.628005	0.630412	0.043034	0.001852

Table 15: Penetrant Range - R^2 Training Statistics.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	0.669183	0.669369	0.004494	0.000020
FNN	0.627084	0.628134	0.012221	0.000149

Table 16: Penetrant Range - R^2 Testing Statistics.

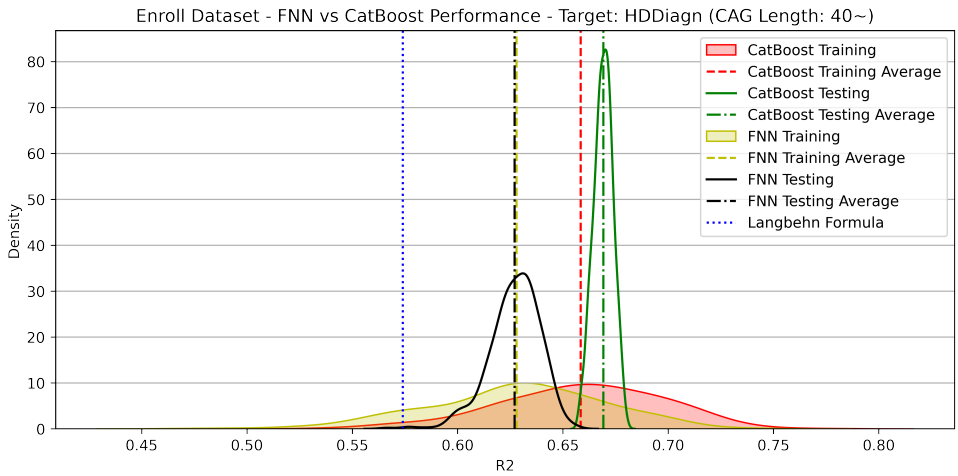

Figure 22: R^2 KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

3.3.2 Full Range Results

The Full Range model refers to the results of training across the entire range of CAG (length equals or greater than 35 repeats). The same analytical structure utilised previously for the penetrant range, was extended to cover the full range. The initial graphical analysis of the first metric evaluated, MAE, demonstrates how this metric varies across each execution, both for CatBoost and FNN. This can be seen in Figure 23. During the training phase, the CatBoost model had an average MAE of 5.126, while the FNN model demonstrated an average MAE of 5.308. Both models exhibited lower values than the Langbehn model, which reached an average MAE of 6.204. These values can be verified in Table 17. The test phase yielded an average MAE of 5.282 for CatBoost, 5.611 for FNN, and Langbehn exhibited an average MAE of 6.204.

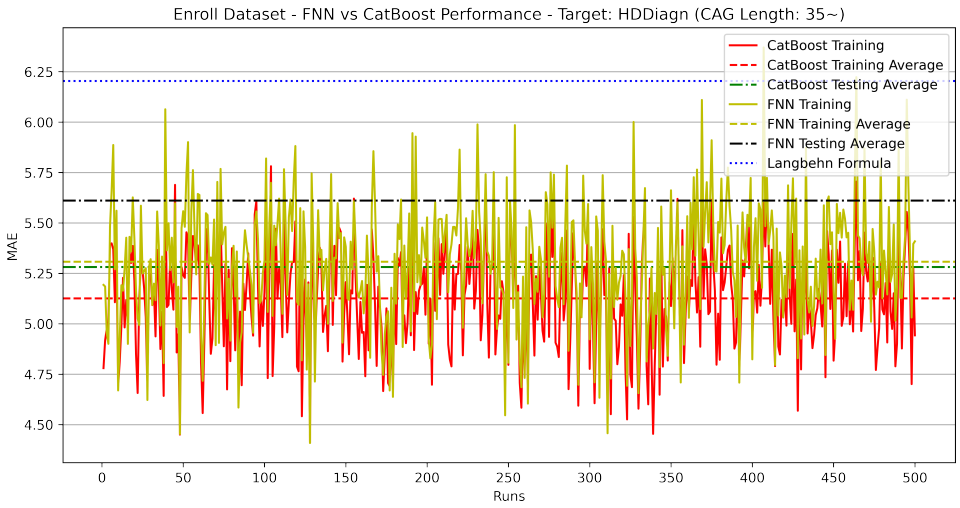

Figure 23: Full range patients MAE measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and test results.

Model	**MAE (Training)**	**MAE (Test)**
CatBoost	5.126210	5.281829
FNN	5.308015	5.611258
Langbehn*	6.204296	6.204296

Table 17: Full Range - MAE average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

The KDE plot of the MAE training values was subjected to analysis. Figure 24 illustrates that the curves for both CatBoost and FNN, with the average vertical lines almost aligned with the center of the curves, indicate a balanced distribution of measured values.

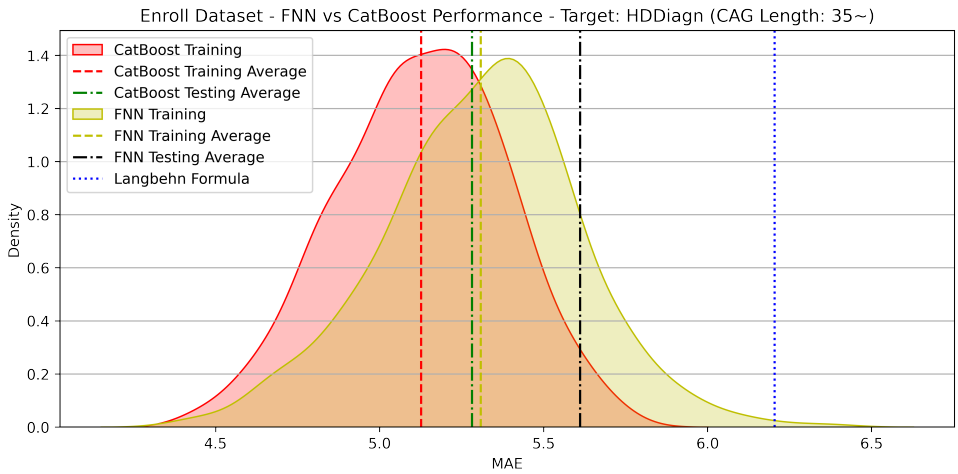

Figure 24: Graph shows the Full range KDE well-distributed curves for MAE on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

The average, median, standard deviation and variance for the MAE values are presented, in Table 18 for the training process and in Table 19 for the test process.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	5.12621	5.141751	0.256939	0.066017
FNN	5.308015	5.331178	0.299391	0.089635

Table 18: Full Range - MAE Training Statistics.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	5.281829	5.277104	0.038977	0.001519
FNN	5.611258	5.60588	0.080977	0.006557

Table 19: Full Range - MAE Testing Statistics.

A final MAE graph was generated, representing the same KDE distribution, with the addition of a test process curve, can be observed in Figure 25. The test curve for density distribution is very narrow, reinforcing the clear distinction between the test and training outcomes.

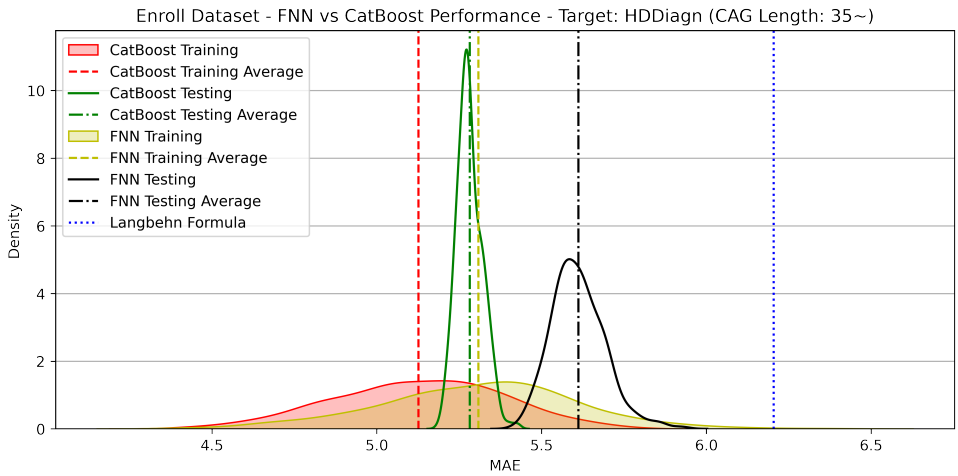

Figure 25: Full range MAE KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

The second metric evaluated was RMSE. Table 20 compares the RMSE values between training and testing processes. Figure 26 illustrates the RMSE value measured for each training run, for both CatBoost and FNN. In order to make it easier for analysis, average horizontal lines have been included for training and test measures.

Model	**RMSE (Training)**	**RMSE (Test)**
CatBoost	6.557404	6.850843
FNN	6.829016	7.278167
Langbehn*	8.199180	8.199180

Table 20: Full Range - RMSE average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

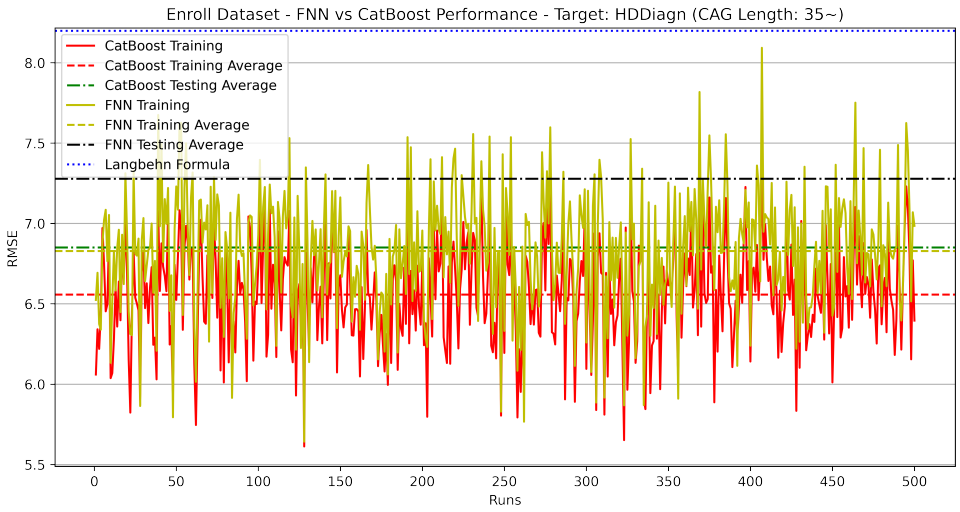

Figure 26: Full range RMSE measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and test results.

Tables 21 and 22 show complementary RMSE statistic measures. Figure 27 demonstrates the RMSE training density distribution, once again demonstrating well-balanced curves. As previously observed during MAE analysis, Figure 28 illustrates the narrow CatBoost test curve, indicating that the model was well-adjusted to the test set.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	6.557404	6.566822	0.32037	0.102637
FNN	6.829016	6.853157	0.370344	0.137155

Table 21: Full Range - RMSE Training Statistics.

Model	**Average**	**Median**	**Standard Deviation**	**Variance**
CatBoost	6.850843	6.84656	0.041282	0.001704
FNN	7.278167	7.267424	0.103467	0.010705

Table 22: Full Range - RMSE Testing Statistics.

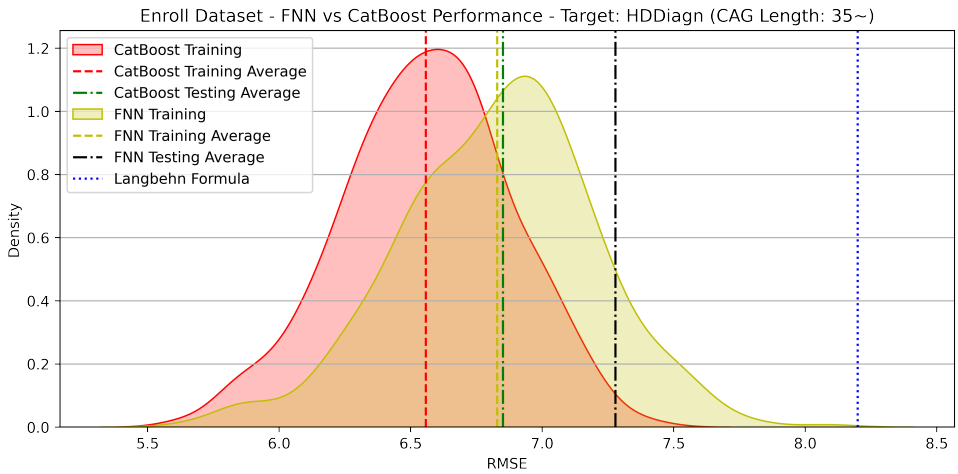

Figure 27: Graph shows the KDE well-distributed curves for RMSE on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

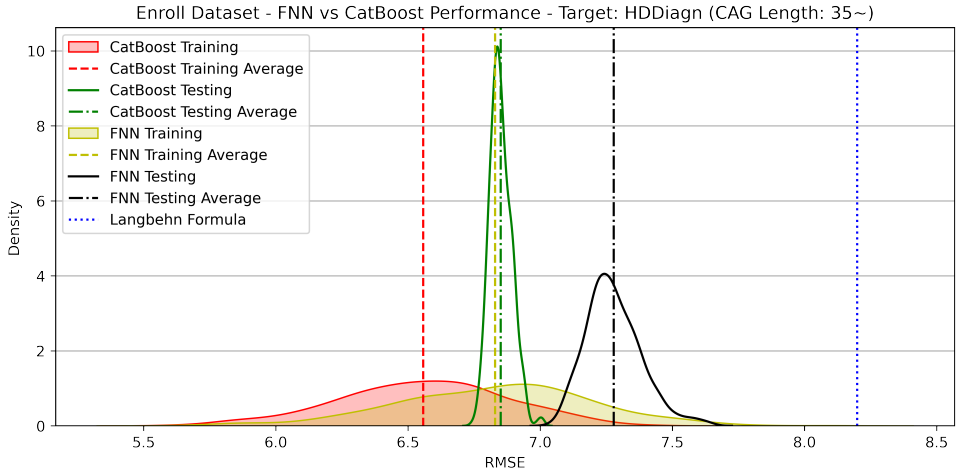

Figure 28: RMSE KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

Finally, the R^2 metric was provided for the full range results. Table 23 shows the comparison results between the models during training and testing processes.

Model	R^2 **(Training)**	R^2 **(Test)**
CatBoost	0.672105	0.67456
FNN	0.644561	0.632634
Langbehn*	0.533868	0.533868

Table 23: Full Range - R^2 average results. *Langbehn value is just a reference that was obtained by applying its formula over the test dataset.

Figure 29 illustrates the R^2 values observed during the execution of the models, while Figure 30 depicts the training KDE distribution.

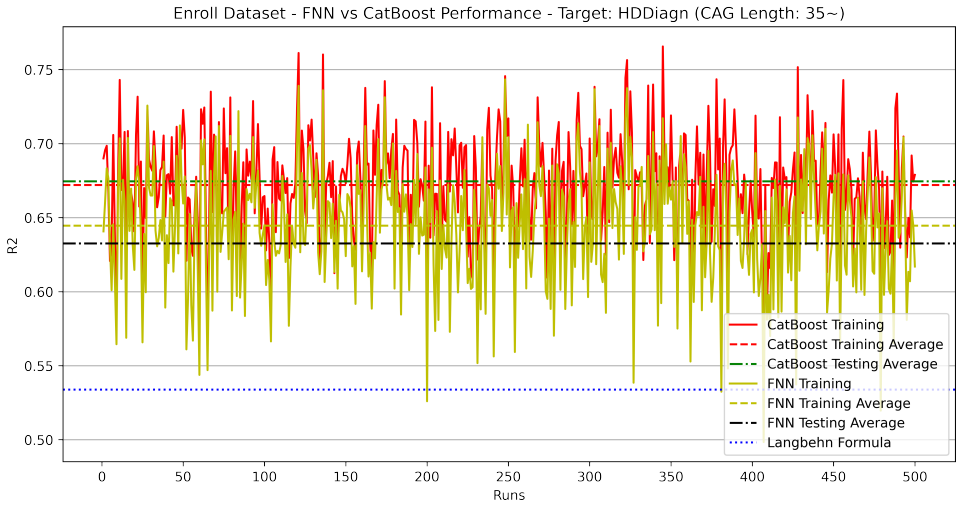

Figure 29: Full range R^2 measurements, obtained during the multiple executions for the models training process. Horizontal lines indicate average values for training and testing results.

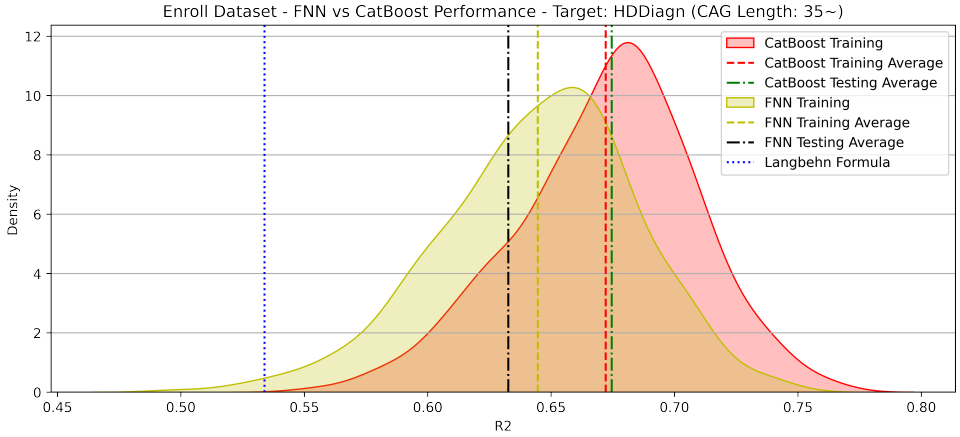

Figure 30: Graph shows the Full range KDE well-distributed curves for R^2 on both trained models. Vertical lines indicate the average values, including the baseline Langbehn.

Tables 24 and 25 show the average, median, standard deviation and variance obtained from R^2 measures. Figure 31 shows the R^2 full range results curve behavior, where the vertical line for CatBoost testing average is higher than the training one.

Model	Average	Median	Standard Deviation	Variance
CatBoost	0.672105	0.676800	0.036403	0.001325
FNN	0.644561	0.648622	0.039711	0.001577

Table 24: Full Range - R^2 Training Statistics.

Model	Average	Median	Standard Deviation	Variance
CatBoost	0.67456	0.674979	0.003928	0.000015
FNN	0.632634	0.633792	0.010492	0.000110

Table 25: Full Range - R^2 Testing Statistics.

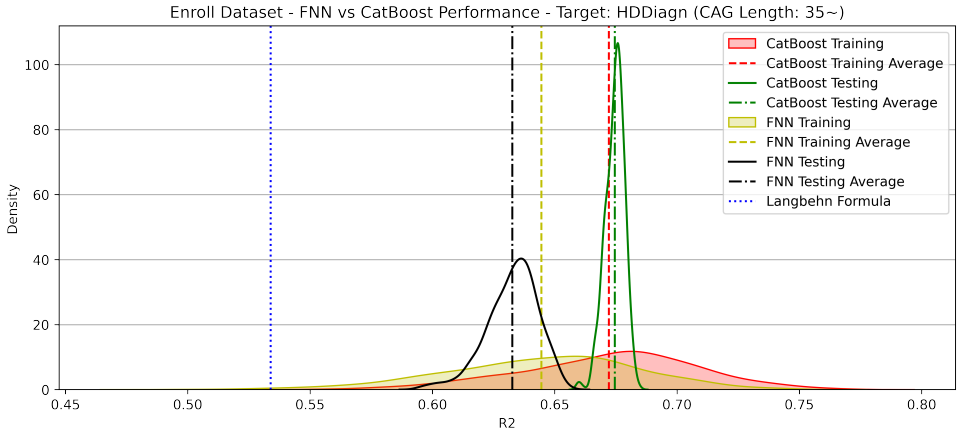

Figure 31: Full Range R^2 KDE plots for training and testing processes. Vertical lines indicate the average values, including the baseline Langbehn.

3.3.3 Summary

Table 26 presents a summary of the results obtained during the training and testing steps for both the CatBoost and FNN models. All the metrics employed during the processes are listed. Additionally, relevant statistical information is provided, including average, median, standard deviation (Std), and variance values.

Phase	**Metric**	**Statistics**	**Penetrant Range**		**Full Range**	
			CatBoost	FNN	CatBoost	FNN
Training	MAE	Average	**5.0175**	5.2425	**5.1262**	5.3080
		Median	5.0169	5.2413	5.1418	5.3312
		Std	0.2728	0.3014	0.2569	0.2994
		Variance	0.0744	0.0909	0.0660	0.0896
	RMSE	Average	**6.4249**	6.7065	**6.5574**	6.8290
		Median	6.4228	6.6985	6.5668	6.8532
		Std	0.3541	0.3821	0.3204	0.3703
		Variance	0.1254	0.1460	0.1026	0.1372
	R^2	Average	**0.6585**	0.6280	**0.6721**	0.6446
		Median	0.6596	0.6304	0.6768	0.6486
		Std	0.0401	0.0430	0.0364	0.0397
		Variance	0.0016	0.0019	0.0013	0.0016
Testing	MAE	Average	**5.2202**	5.5607	**5.2818**	5.6113
		Median	5.2190	5.5564	5.2771	5.6059
		Std	0.0396	0.0960	0.0390	0.0810
		Variance	0.0016	0.0092	0.0015	0.0066
	RMSE	Average	**6.7782**	7.1958	**6.8508**	7.2782
		Median	6.7764	7.1866	6.8466	7.2674
		Std	0.0460	0.1172	0.0413	0.1035
		Variance	0.0021	0.0137	0.0017	0.0107
	R^2	Average	**0.6692**	0.6271	**0.6746**	0.6326
		Median	0.6694	0.6281	0.6750	0.6338
		Std	0.0045	0.0122	0.0039	0.0105
		Variance	2.0e-05	0.0001	1.1e-05	0.0001

Table 26: Summary table with all the statistics results after training and testing the CatBoost and FNN models.

The baseline information for comparison is the Langbehn formula, which resulted in the data summarized in Table 27.

Range	**MAE**	**RMSE**	R^2
Penetrant Range	5.9605	7.6924	0.5739
Full Range	6.2042	8.1991	0.5338

Table 27: Summary of Langbehn formula results applied to the test dataset.

3.4 Additional Experiments

3.4.1 Training Different Target

The target variable used for the HD AAO estimation experiment was '*hddiagn*', which was described as the age at which the patient was diagnosed with HD. This variable exhibited varying degrees of correlation with different HD onsets. This is demonstrated in Figure 32. As illustrated, motor onset (represented by variable '*ccmtrage*') exhibited the strongest correlation (0.94) with '*hddiagn*'.

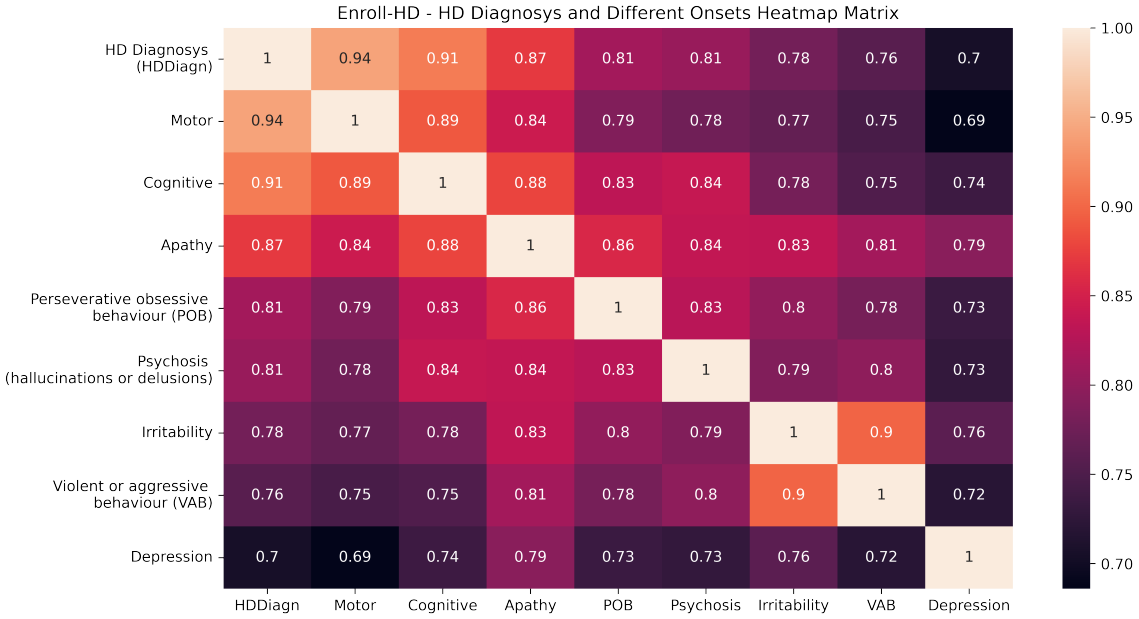

Figure 32: Heatmap correlating hddiagn and other different HD onsets.

In light of the observed correlation between '*hddiagn*' and '*ccmtrage*', an additional experiment was conducted. The same infrastructure (cohort, features, models and hyper-parameters) used for the '*hddiagn*' prediction task, was employed for a new training run targeting '*ccmtrage*' instead. Table 28 contains all the results obtained from the new experiment. Figure 33 shows the MAE results obtained when training the CatBoost and FNN models targeting the motor onset.

Phase	Metric	Model	Penetrant Range		Full Range	
			hddiagn	ccmtrage	hddiagn	ccmtrage
Training	MAE	CatBoost	5.0175	**4.7853**	5.1262	**4.8937**
		FNN	5.2425	4.9887	5.3080	5.0647
		Langbehn	5.9605	5.3479	6.2042	5.6555
	RMSE	CatBoost	6.4249	**6.1977**	6.5574	**6.3294**
		FNN	6.7065	6.4808	6.8290	6.5881
		Langbehn	7.6924	7.0246	8.1991	7.7208
	R^2	CatBoost	0.6585	**0.6749**	0.6721	**0.6846**
		FNN	0.6280	0.6445	0.6446	0.6584
		Langbehn	0.5739	0.6121	0.5338	0.5482
Testing	MAE	CatBoost	5.2232	**5.0913**	5.2792	**5.1503**
		FNN	5.5589	5.4806	5.6137	5.5259
		Langbehn	5.9605	5.3479	6.2042	5.6555
	RMSE	CatBoost	6.7813	**6.5841**	6.8481	**6.6483**
		FNN	7.1937	7.0787	7.2778	7.1432
		Langbehn	7.6924	7.0246	8.1991	7.7208
	R^2	CatBoost	**0.6688**	0.6592	**0.6748**	0.6650
		FNN	0.6272	0.6060	0.6326	0.6131
		Langbehn	0.5739	0.6121	0.5338	0.5482

Table 28: Summary table with all the statistics results after training and testing the CatBoost and FNN models for the '*ccmtrage*' target, compared to previous '*hddiagn*' runs.

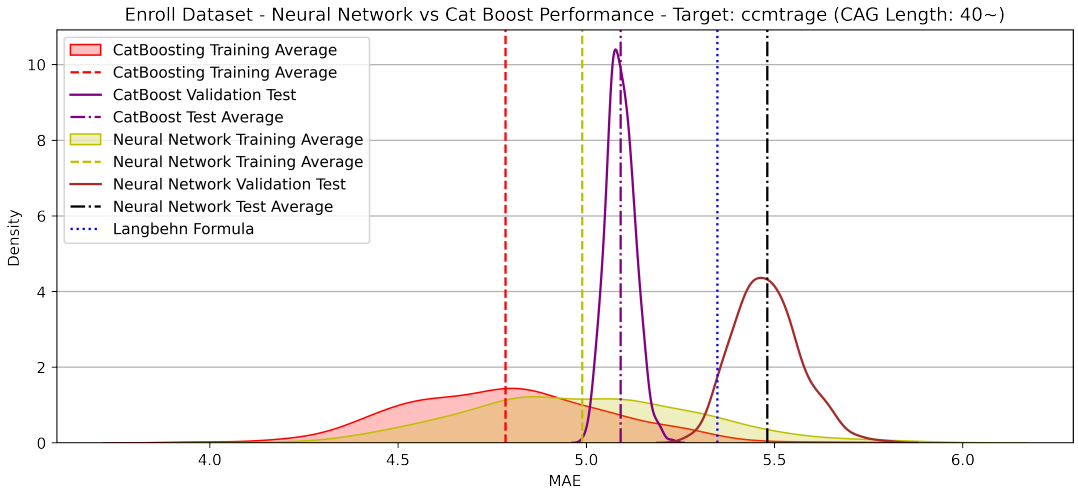

Figure 33: MAE KDE plots for training and testing processes for motor onset target ('*ccmtrage*'). Vertical lines indicate the average values, including the baseline Langbehn.

Additionally, this project produced statistical correlational analysis. The graphical results assisted in the comprehension of how specific environmental and lifestyle variables could contribute to the onset estimation, further discussed in chapter 4. Figure 34 shows the distribution of different HD onsets over the Enroll-HD dataset, and how each one of these onsets triggers the patients over different ages.

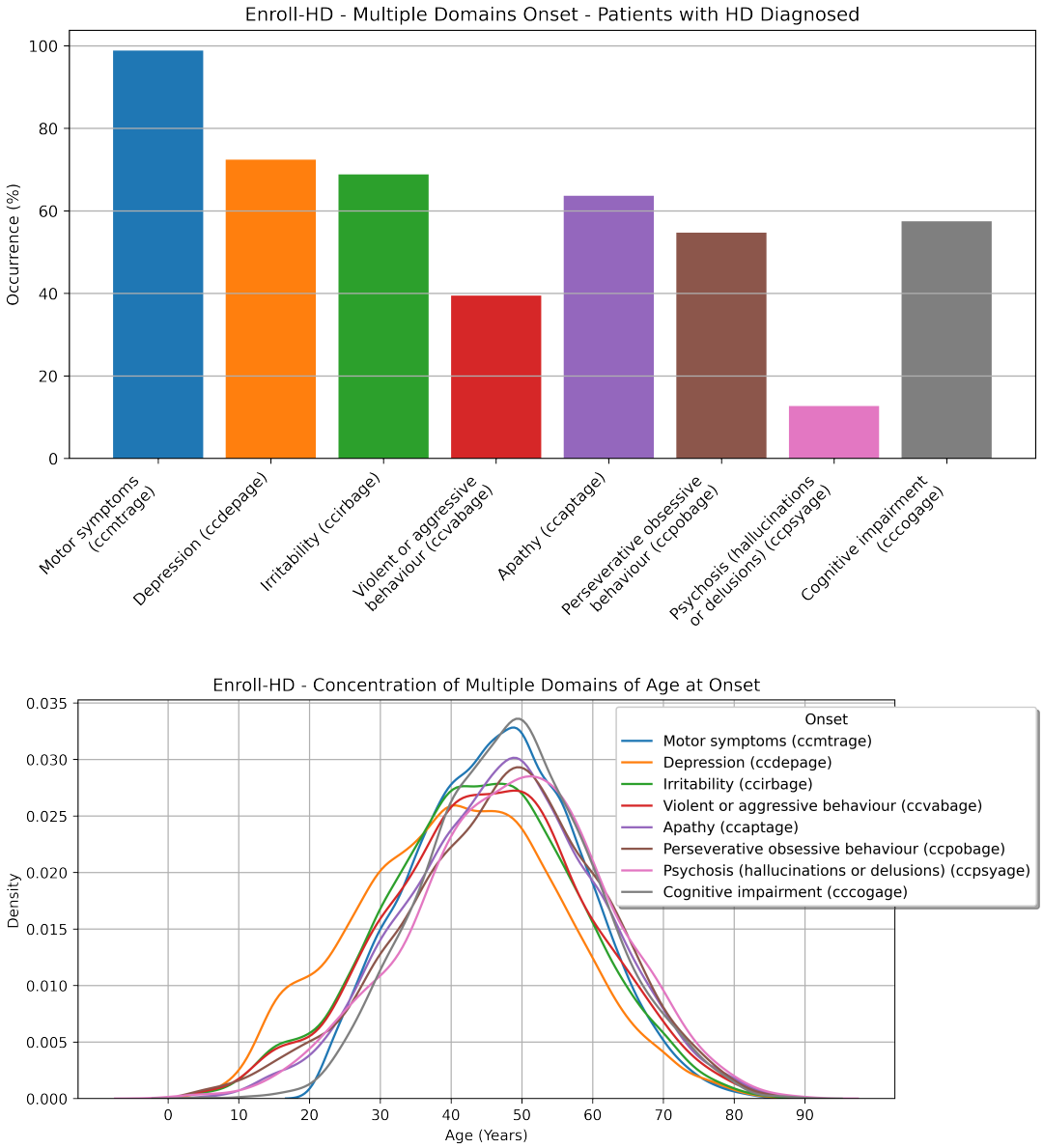

Figure 34: **On top**: Distribution of HD onsets. **On bottom**: KDE density graph showing the occurrences of HD onsets over range of ages.

Figure 35 demonstrates the onset variation (in years) over the different CAG repeats range. It can be observed that the range between 36 and 39 repeats exhibits a non-linear behaviour

that differs from the above 40 repeats.

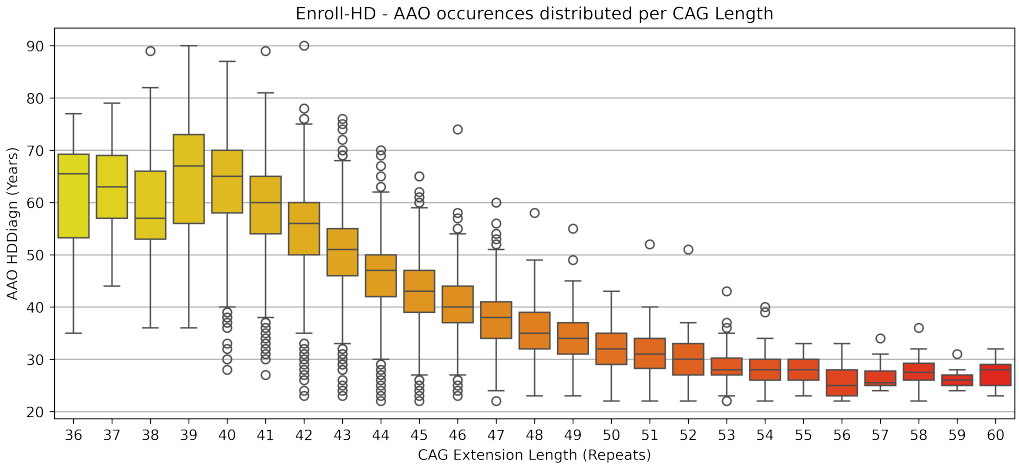

Figure 35: Age at onset variation over different CAG repeats.

Considering that the CAG repeats length can be used as the most efficient variable to group the HD patients together, different HD onsets can be visualized according to this variable. Figure 36 shows 4 different HD onsets correlated with the CAG repeats.

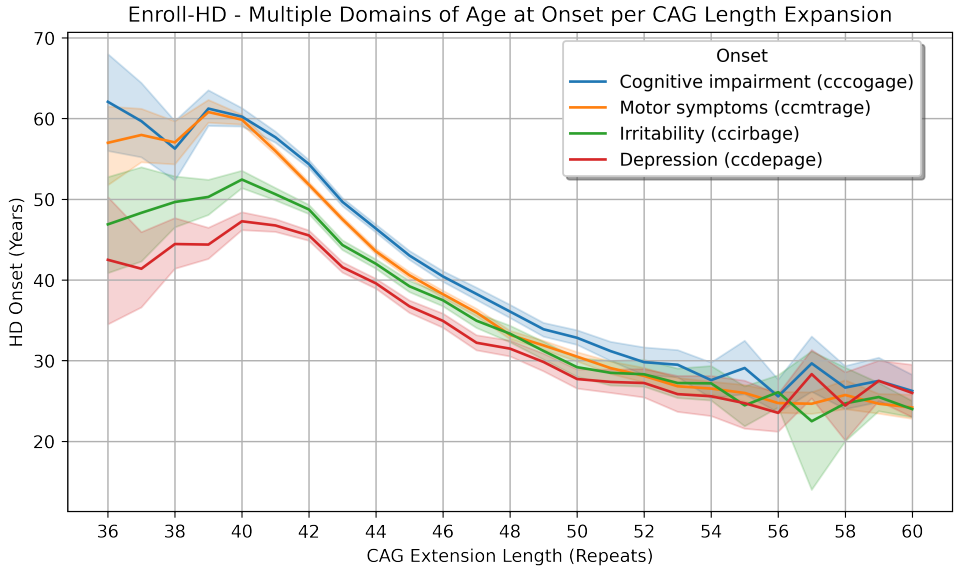

Figure 36: Graph shows the HD onset occurrences correlated with CAG repeats.

3.4.2 Correlational Analysis

A noteworthy range for the observation of CAG repeats is between 40 and 45. This range accounts for 76% of the total Enroll-HD cases, as it is illustrated in Figure 37.

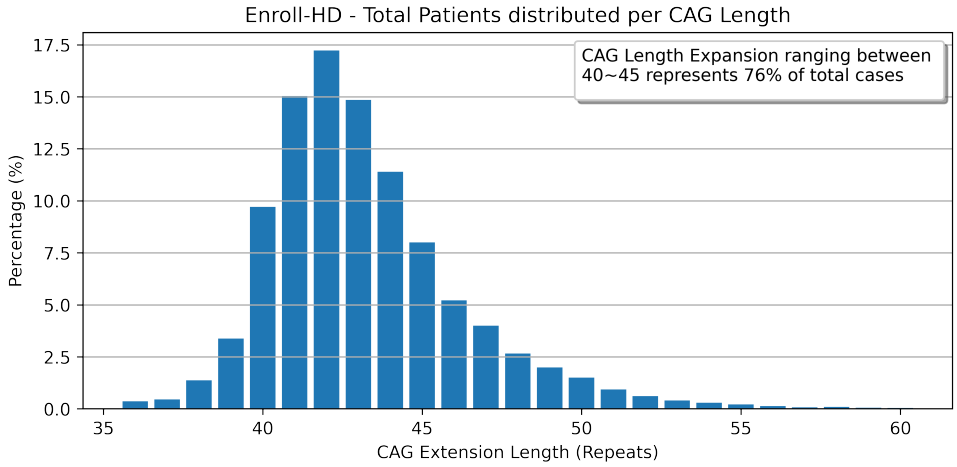

Figure 37: Graph shows the HD onset occurrences correlated with CAG repeats.

The range between 40 and 45 CAG repeats also represents the group of patients with a wider range of onset, displaying the majority of outlier cases (as previously demonstrated in Figure 35), sometimes with the onset ranging from 23 years up to 90 years old, as indicated for those patients with CAG length of 42 repeats. Table 29 shows the statistics of this range, where the onsets are based in years.

CAG Repeats	Onset avg	Onset min	Onset max
40	63.5	28.0	87.0
41	59.3	27.0	89.0
42	55.0	23.0	90.0
43	50.2	23.0	76.0
44	46.3	22.0	70.0
45	43.2	22.0	65.0

Table 29: Enroll-HD occurrence statistics. First column represents the length of CAG repeats for the carriers patients. Other columns display statistics (average, minimum and maximum values) of the HD onset.

A number of graphical correlations were developed by cross-referencing lifestyle variables with CAG repeat values, ranging between 40 and 45 repeats, with the objective of identifying different HD onset patterns. These correlations are presented in the following figures. Figure 38 contains the percentage of patients who self-identified as drug abusers and the age when

depression was triggered on them, stratified by CAG repeats. Figure 39 illustrates the number of HD patients diagnosed with depression, stratified by marital status, and the distribution of these statuses across the different age groups of onset, once again organized by CAG repeats.

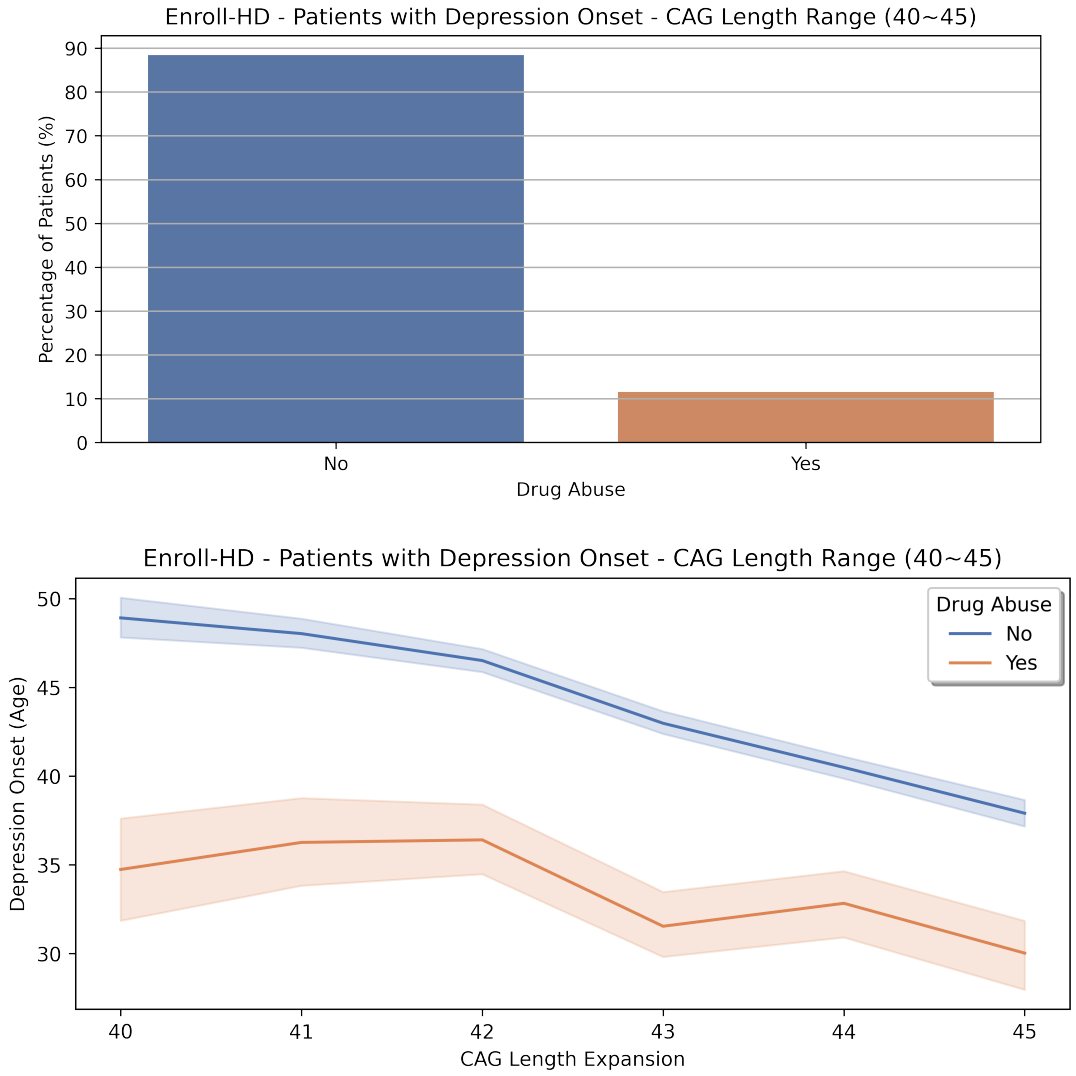

Figure 38: **On top**: Percentage distribution of patients diagnosed with HD depression self-identified as drug abusers. **On Bottom**: Correlation between the use of drugs and depression onset, stratified by CAG repeats.

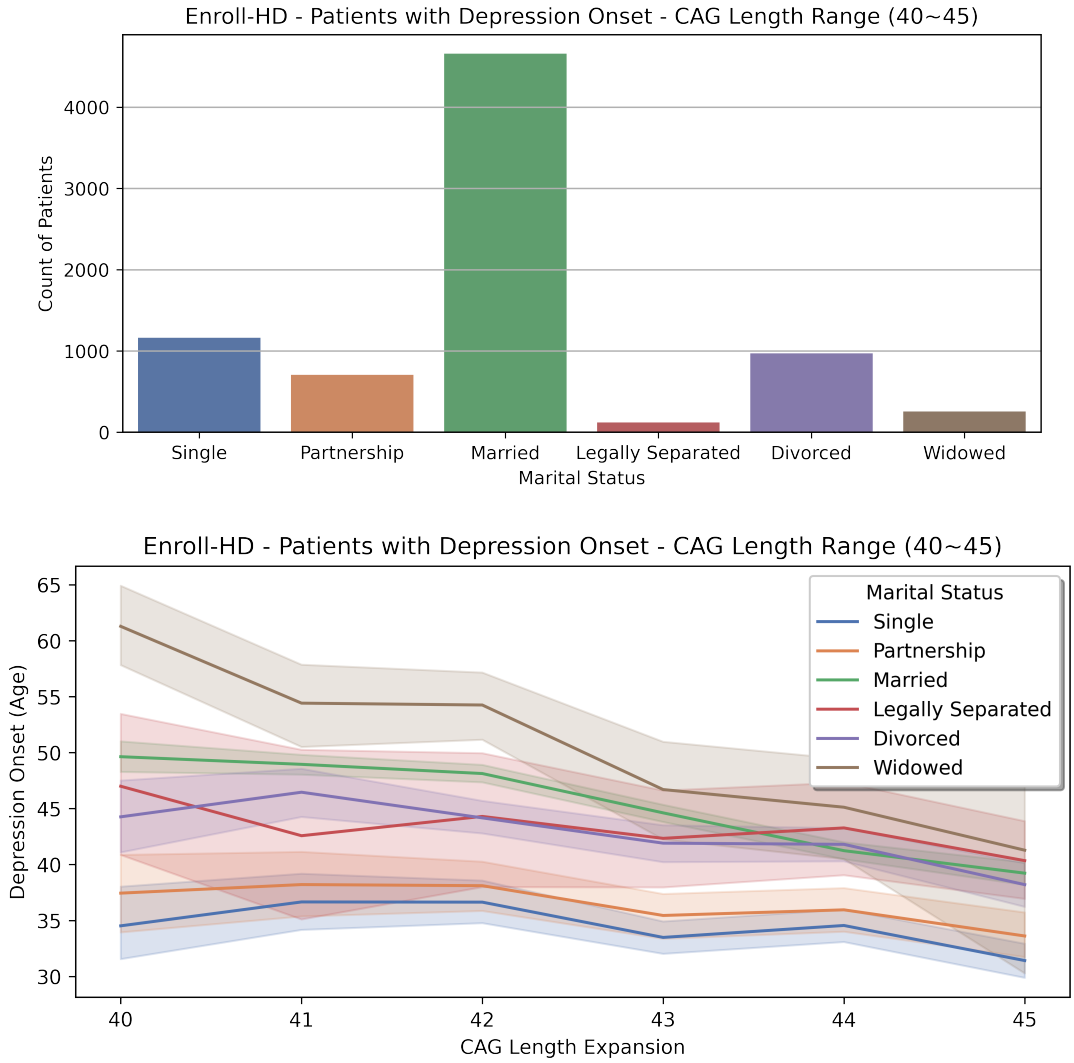

Figure 39: **On top**: Total number of Enroll-HD patients diagnosed with depression, organized by marital status. **On Bottom**: Average curves of depression onsets, distributed by marital status and grouped by CAG repeats.

4 Discussion and Conclusion

4.1 Achievements

In this project we aimed to improve the AAO estimation for HD gene carriers. Although ML algorithms are powerful tools, they are not frequently utilised in rare disease research due to various factors. For instance, these algorithms require access to a large number of samples to effectively detect patterns, which can be scarce in the case of rare diseases [45]. Moreover, even when a sufficient number of samples is available, biological data may contain several features and characteristics that can either introduce noise to the analysis or cause issues when dealing with high-dimensional data, commonly referred to as the 'Curse of Dimensionality' [22], which requires a precise feature selection effort. The processes we created for feature and model selection helped finding the optimal combination of algorithm hyper-parameters and set of variables, thus enhancing the accuracy of the AAO estimation. A cohort extracted from the Enroll-HD dataset was the dataset used to train two different models: FNN and CatBoost. The results demonstrated that both CatBoost and FNN outperformed the current formula that is being used for AAO estimation, the Langbehn formula. The small variation between the different runs of the algorithms indicates the stability of the models. The proximity between the average and median indicates the balance of the results. The reduced variance observed corroborates the assertion that the number of executions was sufficient, and that the models did not diverge significantly from the median across the runs. The low standard deviation observed in the testing measures indicate that the models were well-adjusted to the test dataset, as reported in Table 26. To illustrate, in the penetrant range CatBoost analysis, a training MAE average of 5.02 was registered with a standard deviation of 0.27, while for the testing process the MAE average was 5.22 with a significantly reduced standard deviation of 0.04. Nevertheless, FNN did not demonstrate the same level of performance as CatBoost. As is typical of neural networks, it requires a substantial quantity of data in order to achieve a level of understanding and performance that is superior to that of traditional ML algorithms. Table 30, extracted from section 3.3.3, illustrates the average reduced MAE error margin and efficiency achieved by both models in comparison to the Langbehn formula.

Phase	**Metric**	**Penetrant Range**			**Full Range**		
		CatBoost	FNN	Langbehn	CatBoost	FNN	Langbehn
Training	MAE	**5.0175**	5.2425	5.9605*	**5.1262**	5.3080	6.2042*
	R^2	**0.6585**	0.6280	0.5739*	**0.6721**	0.6446	0.5338*
Testing	MAE	**5.2202**	5.5607	5.9605	**5.2818**	5.6113	6.2042
	R^2	**0.6692**	0.6271	0.5739	**0.6746**	0.6326	0.5338

Table 30: Result table demonstrating the CatBoost and FNN performances in comparison with Langbehn formula. *Same values obtained from testing process were used for comparison purpose.

The findings indicated that the models exhibited a slight reduction in MAE performance when trained with the full range CAG repeat in comparison with the penetrant range. However, this reduction in performance was more pronounced for the Langbehn formula. To illustrate, CatBoost differed its MAE measures from 5.2202 to 5.2818 (1.18% loss) for the test dataset, FNN went up from 5.5607 to 5.6113 (0.91% loss) and Langbehn from 5.9605 to 6.2042 (4.09% loss). The R^2 measures worked differently across the models, which does not necessarily indicate a problem (as R^2 is not the optimal evaluation metric [46]), but it is nevertheless noteworthy. For example, in the case of CatBoost, the test R^2 performance exceeded the training one, for both penetrant and full ranges. Furthermore, Langbehn demonstrated losses when comparing the R^2 performance from penetrant to full range. In fact, the CAG repeat range between 36 and 39 repeats exhibits a non-linear behaviour that differs from the above 40 repeats, as previously demonstrated in Figure 35. The discrepancy for this specific range justifies the erroneous calculation from Langbehn when such a range is involved, as it relies on a linear regression calculation. In contrast, CatBoost and FNN employ distinct methodologies for handling the predictions, resulting in superior performance. In a previous research, Ouwerkerk J. et al. (2023) [24] successfully developed ML models using a similar version of the Enroll-HD dataset to improve HD AAO estimation and compared it to the well-used Langbehn formula. Ouwerkerk also used two different ranges to compare his results with the Langbehn formula, as it was initially proposed for CAG repeats between 41-56. In our project, the range definition differs slightly from Langbehn's and Ouwerkerk's previous works, providing a wider spectrum of analysis.

An additional achievement of this project was the utilisation of the models to validate the Enroll-HD dataset. The training process was limited to those patients that were enrolled as pre-manifest and became manifest, with the objective of selecting the most reliable data from Enroll-HD. In parallel, the remaining patients (i.e., those enrolled already as manifest) were used as a test dataset. The results indicated that training dataset was sufficiently representative for training the model effectively (especially for CatBoost), and that the Enroll-HD test dataset was accurate to a certain extent, once the trained models performed very well when submitted to the test dataset. This conclusion is evidenced by the Results (Chapter 3), as illustrated in Figures 25, 28, and 31. In addition to the reduction in standard deviation observed between the test and training datasets, these Figures illustrate the high efficiency achieved by CatBoost training, as well as its narrow testing curve.

4.2 Additional Findings

In section 3.4, we performed a new experiment using the motor onset '*ccmtrage*' instead of the '*hddiagn*' target variable. The results demonstrated an enhanced performance, especially for CatBoost model. As '*hddiagn*' can be defined as the age at which different forms of HD symptoms may manifest (motor, cognitive or behavioural), the implementation of individualized training estimation facilitated a more comprehensive understanding and accurate onset predictions. It is evident that there has been an improvement in performance, especially in error margin metrics. There were exceptions on performance enhancement, particularly in the

case of R^2 evaluation for CatBoost and FNN testing runs. The R^2 results presented do not fully reflect the gain obtained in terms of the error margin measure. This is why such experiments do not rely on a single performance metric, especially R^2, which may be misinterpreted as a result [46]. The observed improvement in motor onset model performance, even in the absence of any feature change or hyper-parameter adjustment, suggests that there is potential for further enhancement of onsets estimation, especially if each one of them is treated in an individualized form. This implies that not only hyper-parameters can be adjusted, but also the features selected to compose the training dataset. Indeed, it is acceptable that the variation of HD onsets may be influenced by multiple lifestyle aspects at different levels. As observed in Figure 36, the various onset curves exhibit a comparable shape, although they indicate the occurrence in different stages of life. Once again, it can be noted the non-linear behaviour for those HD carrier patients with CAG repeats ranging between 36 and 40.

The analysis of individual HD onsets, crossed with some lifestyle aspects, and grouped by such specific CAG repeats ranges, reveals some intriguing correlations, that inspire further investigation into the possible underlying causes and consequences. Some examples were demonstrated in Figures 38 and 39, where the correlation suggests how marital status or drug abuse may influence the HD depression onset, once again subject of further analysis.

4.3 Future Work

The achievements and findings provided by this project encourage further researches. To begin with, previous studies have shown that genetic factors can affect the HD AAO [47, 48]. This project demonstrated that variables associated with environmental and lifestyle factors can also contribute in the onset estimation. The enhanced performance achieved with ML models, compared to the Langbehn formula (that is currently being used by clinicians for HD AAO estimation), can be further augmented by training such models with a new and expanded cohort. The criteria used to select the training data for this project can be extended to include more participants, including patients who were enrolled as manifest in Enroll-HD, once the models demonstrated that the information entered by clinicians is as reliable as that from patients enrolled as pre-manifest. This cohort expansion would also contribute to an improvement in the NN models' performance. Furthermore, new approaches can be used to improve the HD AAO estimation, such as survival analysis. Survival analysis is a statistical method that examines the expected duration of time until a well-defined event occurs [49]. It means that, as the Enroll-HD is a longitudinal study, all the historical data related to the participants could be used and contribute to the AAO estimation, rather than only the data specific to when the onset happened.

Moreover, this project demonstrated that different types of HD onsets (e.g. motor onset) can be estimated in an individualized way. A significant proportion of HD studies are focused on predicting the motor onset. However, the use of different ML models, which shall receive independent training, and have their own set of selected features, can provide more accurate and improved behavioural and cognitive onset estimations. This approach can also assist in identi-

fying which variable factors have a stronger relationship with the onset being examined.

Furthermore, the HD AAO prediction models can be enhanced by incorporating additional features. In addition to the participants' personal information and lifestyle factors, data related to non-pharmaceutical, pharmaceutical, and nutritional information can be incorporated into the feature selection process.

Finally, all the processes can be updated with the latest Enroll-HD PDS release, which includes more years of clinical data, and a collection of prediction models can be constructed to provide a more robust and multi-technique tool, capable of achieving more accurate HD AAO estimations and being accessible to clinicians to guide their patients.

4.4 Conclusion

In conclusion, the results of this project demonstrated that the use of machine learning algorithms to estimate the age at onset of Huntington's disease, using a highly reliable cohort (defined as patients who were enrolled as pre-manifest before becoming manifest) extracted from the Enroll-HD dataset, is more accurate than the widely used Langbehn formula method. This was achieved by training a neural network (feedforward neural network) and an algorithm for gradient boosting on decision trees (CatBoost) using variables selected from patients' profiles and their lifestyle information. This demonstrated the importance of such aspects on the course of the disease. Furthermore, a reliability assessment was conducted on the remaining Enroll-HD data using the patients enrolled already as manifest as a test dataset. This demonstrated excellent performance with the trained machine learning models. Moreover, we have outlined future research directions based on our findings and achievements of this study, that can influence lifestyle decisions of individuals carrying the HD gene mutation.

References

[1] KNOWLEDGE ECOLOGY INTERNATIONAL - Selected Government Definitions of Orphan or Rare Diseases. URL: https://www.keionline.org/wp-content/uploads/KEI-Briefing-Note-2020-4-Defining-Rare-Diseases.pdf

[2] Medina A, Mahjoub Y, Shaver L, Pringsheim T. Prevalence and incidence of Huntington's disease: an updated systematic review and meta-analysis. Mov Disord. 2022;37(12):2327-2335. doi:10.1002/mds.29228

[3] McColgan P, Tabrizi SJ. Huntington's disease: a clinical review. Eur J Neurol. 2018 Jan;25(1):24-34. doi: 10.1111/ene.13413. Epub 2017 Sep 22. PMID: 28817209.

[4] De Souza RA, Leavitt BR. Neurobiology of Huntington's Disease. Curr Top Behav Neurosci. 2015;22:81-100. doi: 10.1007/7854_2014_353. PMID: 25205327.

[5] Saudou F, Humbert S. The Biology of Huntingtin. Neuron. 2016 Mar 2;89(5):910-26. doi: 10.1016/j.neuron.2016.02.003. PMID: 26938440.

[6] Waldvogel HJ, Kim EH, Tippett LJ, Vonsattel JP, Faull RL. The Neuropathology of Huntington's Disease. Curr Top Behav Neurosci. 2015;22:33-80. doi: 10.1007/7854_2014_354. PMID: 25300927.

[7] Bean L, Bayrak-Toydemir P. American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. Genet Med. 2014 Dec;16(12):e2. doi: 10.1038/gim.2014.146. Epub 2014 Oct 30. PMID: 25356969.

[8] Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, Frontali M, Folstein S, Ross C, Franz M, Abbott M, et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet. 1993 Aug;4(4):387-92. doi: 10.1038/ng0893-387. PMID: 8401587.

[9] Orth, M., Schwenke, C. Age-at-onset in huntington disease. PLoS Currents 3, 1258 (2011). doi: 10.1371/currents.RRN1258. PMID: 22453877.

[10] Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR; International Huntington's Disease Collaborative Group. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet. 2004 Apr;65(4):267-77. doi: 10.1111/j.1399-0004.2004.00241.x. Erratum in: Clin Genet. 2004 Jul;66(1):81. PMID: 15025718.

[11] N. Ahmad Aziz and Jorien M.M. van der Burg and Sarah J. Tabrizi and G. Bernhard Landwehrmeyer. Overlap between age-at-onset and disease-progression determinants in Huntington disease. 2018. doi: 10.1212/WNL.0000000000005690.

[12] Enroll-HD Clinical Research Platform. URL: https://enroll-hd.org/

[13] Enroll-HD Clinical Research Platform - PDS6 Data Dictionary Document. URL: https://enroll-hd.org/enrollhd_documents/2023-10-R2/ENROLL-HD_DataDictionaryPDS6R2_v20230915.pdf

[14] Enroll-HD Docs - https://enroll-hd.org/for-researchers/data-support-documentation/

[15] Enroll-HD Clinical Research Platform - Protocol Document. URL: https://enroll-hd.org/enrollhd_documents/Enroll-HD-Protocol-1.0.pdf

[16] Judith, H. and Daniel, K. (2018). Machine Learning For Dummies. IBM Limited Edition. ISBN: 978-1-119-45494-6 (ebk)

[17] Enroll-HD Data Access Request. URL: https://enroll-hd.org/for-researchers/access-data/

[18] GitHub - Julio Marchiori.
URL: https://github.com/JulioMarchiori/hd_aao_estimation_improvement_pub

[19] Jupyter Notebooks. URL: https://jupyter.org/

[20] S. Dutta, A. Arunachalam and S. Misailovic, "To Seed or Not to Seed? An Empirical Analysis of Usage of Seeds for Testing in Machine Learning Projects," 2022 IEEE Conference on Software Testing, Verification and Validation (ICST), Valencia, Spain, 2022, pp. 151-161, doi: 10.1109/ICST53961.2022.00026.

[21] Willmott, Cort J.; Matsuura, Kenji, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". 2005. doi:10.3354/cr030079

[22] Aurélien Géron. Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed., O'Reilly, 2019. ISBN: 9781492032649

[23] Wright, S. (1921). "Correlation and causation". Journal of Agriculture Research. 20 (7): 557-585

[24] Ouwerkerk J, Feleus S, van der Zwaan KF, et al. Machine learning in Huntington's disease: exploring the Enroll-HD dataset for prognosis and driving capability prediction. Orphanet J Rare Dis. 2023;18(1):218. Published 2023 Jul 27. doi:10.1186/s13023-023-02785-4

[25] LUMC Leiden University Medical Center - Department of Human Genetics - BioSemantics Team - https://www.lumc.nl/en/afdelingen/human-genetics/biosemantics/

[26] Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health. 2020;8(1):e000262. Published 2020 Feb 16. doi:10.1136/fmch-2019-000262

[27] Jason Brownlee. "How to Choose a Feature Selection Method For Machine Learning". August 20, 2020. Data Preparation

[28] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. "Bias in machine learning software: why? how? what to do?". In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2021). Association for Computing Machinery, New York, NY, USA, 429–440. https://doi.org/10.1145/3468264.3468537

[29] Scikit Learn. URL: https://scikit-learn.org

[30] Scikit Learn - Feature Selection - SelectkBest. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[31] Kavya D. Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning. 2023. URL: https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48

[32] SciKit Learn - Feature Selection - sklearn.feature_selection.f_regression. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression

[33] Ross BC. Mutual information between discrete and continuous data sets. PLoS One. 2014;9(2):e87357. Published 2014 Feb 19. doi:10.1371/journal.pone.0087357

[34] Scikit Learn - Feature Selection - Mutual Information. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

[35] Scikit Learn - Feature Selection - Lasso. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso

[36] Scikit Learn - PreProcessing - Standard Scaler. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[37] Faouzi J, Colliot O. Classic Machine Learning Methods. 2023 Jul 23. In: Colliot O, editor. Machine Learning for Brain Disorders [Internet]. New York, NY: Humana; 2023. Chapter 2. Available from: https://www.ncbi.nlm.nih.gov/books/NBK597496/ doi: 10.1007/978-1-0716-3195-9_2

[38] Vikas Gupta. Understanding Feedforward Neural Networks. 2017.
URL: https://www.learnopencv.com/understanding-feedforward-neural-networks/

[39] TensorFlow - Keras. URL: https://www.tensorflow.org/versions/r2.13/api_docs/python/tf/keras

[40] Sayak Paul. "Hyperparameter Optimization in Machine Learning Models" Radar AI Edition. URL: https://www.datacamp.com/tutorial/parameter-optimization-machine-learning-models

[41] Scikit Learn - Model Selection - GridSearchCV. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[42] Diederik P. Kingma, Jimmy Ba, "Adam: A Method for Stochastic Optimization" Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. doi:10.48550/arXiv.1412.6980

[43] Ranen NG, Stine OC, Abbott MH, et al. Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. Am J Hum Genet. 1995;57(3):593-602.

[44] Seaborn - Kernel density estimate plot.
URL: https://seaborn.pydata.org/generated/seaborn.kdeplot.html

[45] Dhaenens CM, Burnouf S, Simonin C, Van Brussel E, Duhamel A, Defebvre L, Duru C, Vuillaume I, Cazeneuve C, Charles P, Maison P, Debruxelles S, Verny C, Gervais H, Azulay JP, Tranchant C, Bachoud-Levi AC, Dürr A, Buée L, Krystkowiak P, Sablonnière B, Blum D; Huntington

French Speaking Network. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. Neurobiol Dis. 2009 Sep;35(3):474-6. doi: 10.1016/j.nbd.2009.06.009. Epub 2009 Jul 8. PMID: 19591938.

[46] Ford, C. 2020. "Understanding Robust Standard Errors." UVA Library StatLab. https://library.virginia.edu/data/articles/understanding-robust-standard-errors/ (accessed February 1, 2023). Is R-squared Useless?.

[47] Dhaenens CM, Burnouf S, Simonin C, Van Brussel E, Duhamel A, Defebvre L, Duru C, Vuillaume I, Cazeneuve C, Charles P, Maison P, Debruxelles S, Verny C, Gervais H, Azulay JP, Tranchant C, Bachoud-Levi AC, Dürr A, Buée L, Krystkowiak P, Sablonnière B, Blum D; Huntington French Speaking Network. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. Neurobiol Dis. 2009 Sep;35(3):474-6. doi: 10.1016/j.nbd.2009.06.009. Epub 2009 Jul 8. PMID: 19591938.

[48] Jong-Min Lee, Vanessa C. Wheeler, Michael J. Chao, Jean Paul G. Vonsattel, Ricardo Mouro Pinto, Diane Lucente, Kawther Abu-Elneel, Eliana Marisa Ramos, Jayalakshmi Srinidhi Mysore, Tammy Gillis, Marcy E. MacDonald, James F. Gusella, Denise Harold, Timothy C. Stone, Valentina Escott-Price, Jun Han, Alexey Vedernikov, Peter Holmans, Lesley Jones, Seung Kwak, Mithra Mahmoudi, Michael Orth, G. Bernhard Landwehrmeyer, Jane S. Paulsen, E. Ray Dorsey, Ira Shoulson, Richard H. Myers, Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease, Cell, Volume 162, Issue 3, 2015, Pages 516-526, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2015.07.003.

[49] Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesth Analg. 2018 Sep;127(3):792-798. doi: 10.1213/ANE.0000000000003653. PMID: 30015653; PMCID: PMC6110618.