Exploring machine learning approaches for phenotype prediction of Huntington's disease

Caterina Fuses
Department of Biomedical Sciences
Creatio, Production and Validation
Center of Advanced Therapies
Faculty of Medicine and Health
Sciences, Universitat de Barcelona
Barcelona, Spain
cfuses@ub.edu

Josep M Canals
Department of Biomedical Sciences
Creatio, Production and Validation
Center of Advanced Therapies
Faculty of Medicine and Health
Sciences, Universitat de Barcelona
Barcelona, Spain
jmcanals@ub.edu

Jordi Abante
Department of Biomedical Sciences
Creatio, Production and Validation
Center of Advanced Therapies
Faculty of Medicine and Health
Sciences, Universitat de Barcelona
Barcelona, Spain
jordi.abante@ub.edu

ABSTRACT

Age of onset of Huntington's disease is currently being predicted by using the CAG trinucleotide expansion as the main predictor, but it does not explain the entire variability of the phenotype. The present study explores the potential of machine learning algorithms trained with a broader set of genetic data to improve the modeling of this remaining unexplained variance. The data used are single nucleotide polymorphism genotypes from the Enroll-HD dataset.

KEYWORDS

Huntington's disease, Genetic Modifiers, Machine Learning

ACM Reference Format:
Caterina Fuses, Josep M Canals, and Jordi Abante. 2024. Exploring machine learning approaches for phenotype prediction of Huntington's disease. In *Proceedings of Knowledge Discovery and Data Mining 2024 (KDD '24)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/

1 INTRODUCTION

Huntington's disease (HD) is a hereditary neurodegenerative disease whose first symptoms can appear at different points of a lifetime. Importantly, the age of disease onset correlates strongly with the length of the mutation related to the disease, a CAG trinucleotide expansion in the hungtingtin gene (*HTT*) [3, 4]. Genetic testing and clinical prediction for age of onset (AO) both rely on the length of this expansion. Nevertheless, this is not a perfect predictor, as the standard deviation of AO at a specific CAG repeat length is quite large, specially for short expansions (Fig. 1).

Expansion length accounts for 40-70% of the variability of AO, while the remaining variance shows a high degree of heritability [8, 11]. Large genetic studies have been done during the last two decades and are still ongoing in the search for genetic modifiers of AO, genes or genetic elements involved in the process of disease onset either by accelerating or delaying the emergence of motor symptoms. Amongst the proposed modifiers are *FAN1*, *MLH1*, and *MSH3* [7, 10].

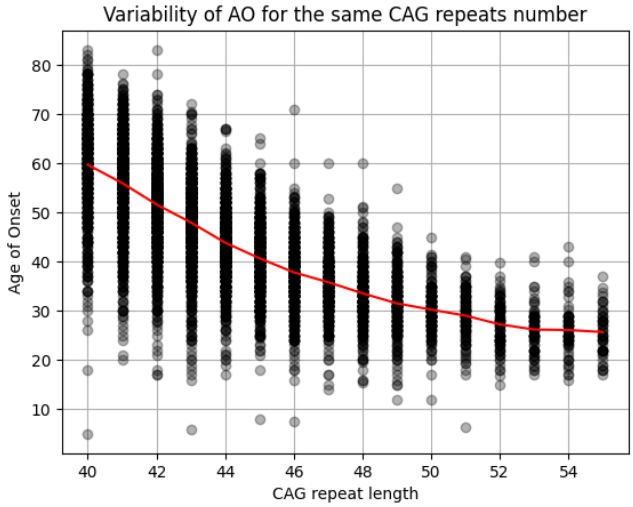

Figure 1: Inverse correlation of age of onset and CAG repeat length observed in the Enroll-HD data used in this project.

Phenotype prediction in HD is not only interesting from a clinical point of view for life planning addressed to mutation carriers. It can also provide valuable insight of the disease mechanisms that generate such phenotype, and apply this knowledge into the design of clinical trials, where the effect of putative modifiers of HD pathogenesis needs to be controlled for the effect of genetic background [5]. Knowing more about disease onset mechanisms can also reveal possible targets for treatment that could delay onset of symptoms [4], as we still lack treatments to prevent, delay or cure the disease.

Here, we present an exploratory analysis using machine learning (ML) techniques to test whether ML models are able to find significant genetic factors contributing to a higher degree of explainability for AO variability beyond the number of CAG repeats.

2 METHODS

2.1 Data preprocessing

The data used for this analysis comes from Enroll-HD, a worldwide observational study aimed towards the development of therapeutics for HD. The dataset contains whole-genome single nucleotide

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/

polymorphisms (SNPs) genotypes of 9064 patients, their CAG trinucleotide expansion length in number of CAG triplet repeats, and their AO.

Training algorithms with all available data would require an unfeasible amount of computing power. Hence, various filtering steps were taken (Fig. 2). First, a list of biological processes related to HD was assembled after a thorough literature review [13, 10, 9, 1, 5, 12, 4, 7], contained in Table 1. The genes related to each gene ontology (GO) term corresponding to the selected processes form the set of core genes of the analysis.

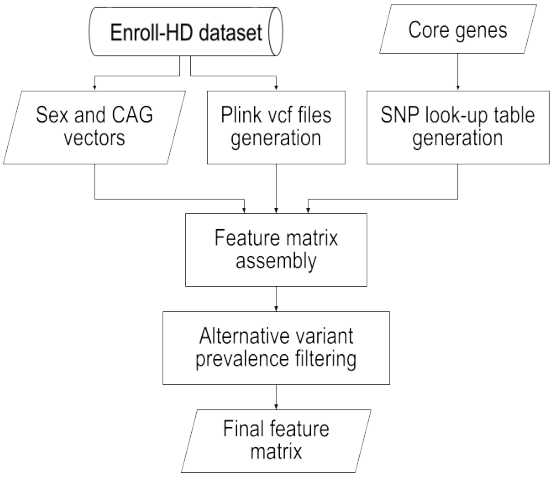

Figure 2: Data preprocessing flow chart.

A look-up table was retrieved from Ensembl's Biomart relating each SNP to the corresponding gene through its genomic coordinates. We then used this table to filter the original dataset, keeping only those SNPs corresponding to the core genes set. The filtering was done directly after generating VCF (Variant Call Format) files with *Plink2* using the raw Enroll-HD data. We also established a minimum alternative variant prevalence, discarding those SNPs in which more than 99% of the samples have the reference variant, as such SNPs would contribute minimally to the models. The resulting number of SNPs was 339,801. The data was numerically encoded following the standard correspondence (0: homozygous for reference variant, 1: a reference and an alternative variant, 2: both alternative variants). The final feature matric used is described in Table 2.

2.2 ML models

Given the high dimensionality of the data, we tested both regularized linear and non-linear methods. In particular, we considered Lasso and Elastic Net for the former class, and Random Forest and XGBoost for the latter. For XGBoost we considered the two approximated solutions (approximation and histogram) for faster computation. These models were benchmarked against an ordinary least-squares (OLS) with only sex and CAG as features, used as the baseline. In all cases both CAG and AO vectors where scaled, down to 0-1 with a MinMax scaler in the CAG case, and using a standard scaler in the case of AO.

Model training and evaluation was performed with Python, all models except XGBoost (which has its own package, *XGBoost Python Package*) were implemented using *Sklearn*. Importantly, all used functions accept the feature matrix loaded as a sparse matrix, which allows for faster computations while not requiring as much computational power.

Data was split into training and testing sets using the *Sklearn*'s function `train_test_split` from the module *model_selection* with a fixed random state number to ensure the samples used in each set were the same across models. The testing set size was set to the 30% of the total number of samples.

Hyperparameter tuning was performed with a grid search with 5 fold cross-validation using *Sklearn*'s function `GridSearch` from the *model_selection* module, over the training sets, using the default R^2 scoring as the metric to evaluate the trained model on the validating set in each fold. The final used hyperparameters in each model are presented in Table 3.

The performance of the models was first checked with the percentage of deviance explained by the predictions of the training samples to ensure the model was not overfitting. The testing set of samples was used to evaluate the model both graphically and numerically. Three different metrics were computed using the *Sklearn* module metrics: the coefficient of determination R^2, the mean squared error (MSE) and the mean absolute error (MAE). Graphically, we plotted the predictions over the actual values, alongside the residuals over the predictions. This graphical representation of predictions gives information about whether the error in predictions is bigger along a specific range of ages or it is homogeneous across all ages.

Results regarding which features contribute the most in each model where represented following the Manhattan plot format. For the Lasso and Elastic Net models, the feature importance was assessed through their coefficients values. For the rest of the methods, the Gini index was used for Random Forest, and the feature gain for the two XGBoost approaches. Features representing SNPs were related back to what gene and GO term they corresponded in order to extract biological conclusions from the models.

3 RESULTS

3.1 Prediction Accuracy

The metrics from each resulting model are shown in Table 4, were the baseline is also included for easier comparison. The best results in terms of metric values were obtained with XGBoost (with the histogram tree method), improving by 3.4% the baseline R^2. The approximate tree method is almost as good, while also taking less time to train (with a 5 fold cross-validation, the histogram method took 13.7 min and the approximate method took 8.3 min). Random Forest Regressor also achieved good results with a similar R^2 score, but it took 28 minutes to train, making it the slowest model. The linear approaches (Lasso and Elastic Net) did not improve the prediction with respect to the baseline.

Comparing the plots produced to graphically study the predictions we also see interesting differences between the two methods. Fig. 3 shows the predictions of the testing set of the best performing model of each method. Predictions of the least squares methods

Table 1: Included GO terms as core genes.

Process	GO term	Related Genes
Mismatch repair	GO:0006298	*FAN1, MLH1, MSH3, MLH3*
Synaptic transmission, Glutamatergic	GO:0035249	*GRIK2, GRIN2A, GRIN2B*
Omega peptidase activity	GO:0008242	
Cysteine-type endopeptidase activity	GO:0004197	
Proteasome-mediated ubiquitin-dependent protein catabolic process	GO:0043161	
Ubiquitin binding	GO:0043130	
Ubiquitin protein ligase binding	GO:0031625	
Protein deubiquitination	GO:0016579	*UCHL1*
Transcription regulator activity	GO:0140110	*TCERG1, TP53*
Neuron apoptotic process	GO:0051402	*DFFB, MAP3K5, MAP2K6*
Lipoprotein metabolic process	GO:0042157	*APOE*
Axonal transport	GO:0098930	*HAP1*
Folic acid metabolic process	GO:0046655	*MTHFR*
Energy reserve metabolic process	GO:0006112	*PPARGC1A*

Table 2: Feature matrix description.

Statistic		Value
Nº features		339886
Nº samples		9064
Distribution of labels		
	Male/Female	4417/4647
	CAG repeat length	40-55, mean 44, std. 3.06
	AO	5-83, mean 45.54, std. 11.58

Table 3: Hyperparameter values chosen by GridSearch.

Model	Parameter	Value
Lasso	alpha	0.01
	max_iter	1000
Elastic Net	alpha	0.01
	l1_ratio	0.9
	max_iter	10000
Random Forest	ccp_alpha	0.001
	max_depth	6
	n_estimators	30
XGBoost (approx)	reg_alpha	0.5
	max_depth	3
	n_estimators	10
XGBoost (hist)	reg_alpha	0.1
	max_depth	2
	n_estimators	20

have non symmetric residuals, underestimating AO at older ages and over estimating AO at younger ages. Tree based methods, on the contrary, have a more homogeneous distribution of residuals, although they fail to predict the youngest AOs.

Table 4: Model metrics and number of features used as regressors in each model, including the baseline OLS.

Model	R^2	MAE	MSE	Nº features
OLS	0.5572	0.5124	0.4462	2
Lasso	0.5478	0.5188	0.4556	867
Elastic Net	0.5395	0.5247	0.464	1164
Random Forest	0.5856	0.4886	0.4175	884
XGBoost (hist)	**0.5908**	**0.4844**	**0.4123**	33
XGBoost (approx)	0.5847	0.4904	0.4184	34

3.2 Feature Importance

We further inspected the trained models by looking at feature importance. As expected, all models use CAG as the most important regressor, but the regressors that follow differ between models. Interestingly, sex was not used in any model as a regressor, showing that AO variability was not found to be sex dependant. Throughout all models we can see SNPs which are used repeatedly, mainly rs144287831 (*MLH1*) and rs61997076 (*FAN1*), both known HD modifiers [2, 6]. Note that although the metrics of the linear methods did not point to a good regressor, the models acknowledge the effect of SNPs from experimentally tested HD modifiers. The advantage of the linear methods is that their explainability also shows if a certain feature delays or hastens disease onset, an information given by the sign of the feature's coefficient. *MYT1L* and *CDYL2* SNPs usage is also consistent between XGBoost models, and the highest contributing SNPs in the linear models also overlap strongly.

To provide further biological interpretability we can do GO analysis, which results in a list of GO terms of the genes of all the SNPs used in each model alongside the background proportions of the original input feature matrix. Fig. 4 represents this information through a segmented bar chart. Lasso, Elastic Net, and Random Forest models, with approximately a thousand features each, closely resemble the largest background proportions. In contrast, the XGBoost methods exhibit GO term proportions that deviate more from

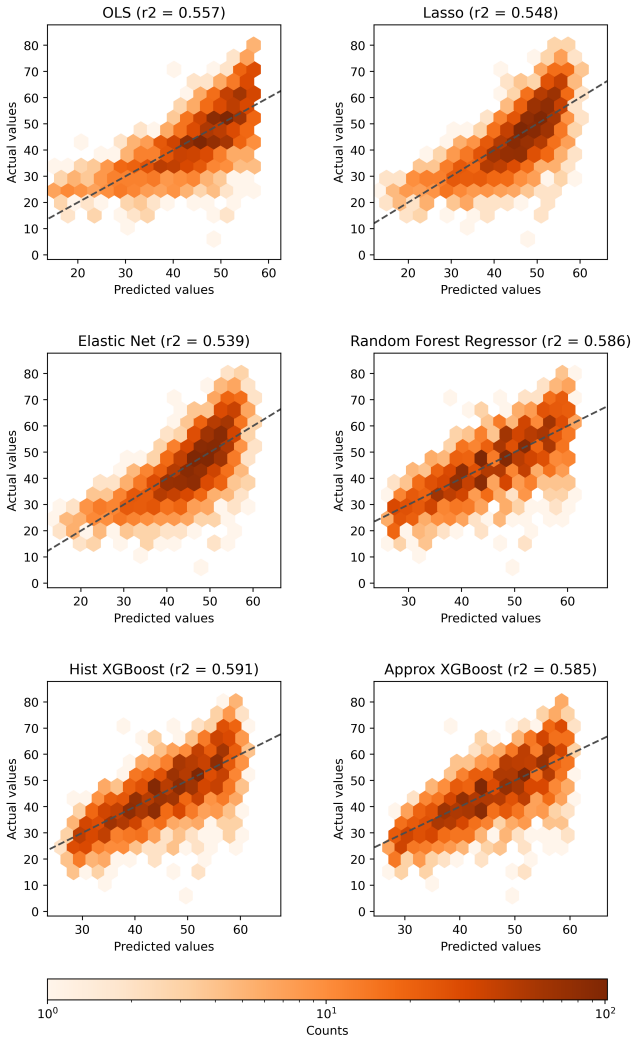

Figure 3: Actual vs. predicted outcomes with all trained models, including the baseline OLS.

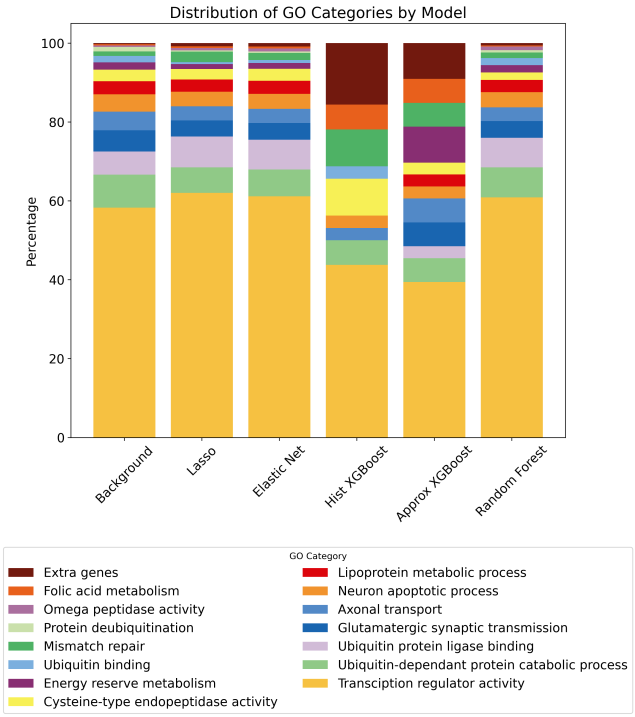

Figure 4: GO terms proportions of the features used in each model plus the background proportions.

the background and vary between them. Across all models, over 40% of the SNPs used as regressors belong to GO:0140110 (transcription regulator activity), followed by terms related to protein degradation and ubiquitin. The enrichment analysis (done with a Fisher's test) shows the best performing XGBoost model has three significantly enriched terms: GO:0006298 (mismatch repair), GO:0046655 (folic acid metabolism) and the *Extra genes* group. This last group is enriched by all models, primarily due to multiple *FAN1* SNPs. The term of transcription regulation is only significantly enriched in Lasso, Elastic Net and Approx XGBoost.

4 DISCUSSION

Findings about SNP contribution to the models should not be interpreted as being genomic positions directly related to the phenotype. We could be viewing a SNP in the same linkage disequilibrium block as the effect causing polymorphism. These blocks are loci (positions in the genome) which are non-randomly associated. These loci most probably fall inside the same gene, so the gene conclusions we can extract from this analysis do hold biological significance.

All models were able to find positive controls which have been previously described as HD modifiers such as *FAN1* and *MLH1*, but we only improved the baseline predictions with tree-based methods. This leads to the conclusion that the relationship between genetic information and AO is probably non-linear, as the models that perform best are the ones which can model non-linear relations, while the regularized linear relations failed in finding new information in the provided features to better explain AO variability. This directs the future research path following this project towards algorithms which can capture better this non-linearity. This, combined with the fact that HD modifiers are suspected to work at a protein interaction level [12], points towards a very promising continuation of this study by working with Graph Neural Networks, representing the core genes in a graph relating them through their protein-protein interactions, and encoding in each node the SNPs of each gene.

CODE AVAILABILITY

All code generated to produce the presented results is uploaded in the Github repository ML-HD (URL: https://github.com/cfuses/ML-HD.git). The original data is not available, so the data preprocessing scripts cannot be tested. To test the ML scripts and reproduce similar results, a toy example of 70k SNPs and 900 samples is available in the repository. This toy example was build by taking the SNPs of a set of literature-based HD modifiers candidates (*MLH1, MLH3, GRIK2,*

GRIN2A, GRIN2B, UCHL1, APOE, ASK1, MAP3K5, PPARGC1A) and the SNPs of the *HTT* gene, and randomly picking the rest to sum up to 70k.

ACKNOWLEDGEMENTS

This study was supported by grants from the Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación and European Regional Development Fund (ERDF A way of making Europe) (Red de Terapias Avanzadas, RD21/0017/0020); European Union NextGenerationEU/PRTR; Generalitat de Catalunya (2021 SGR 01094); "la Caixa" Foundation under the grant agreement LCF/PR/HR21-00622"; and Red Española de Supercomputación (RES) under project BCV-2024-2-0010.

Biosamples and data used in this work were generously provided by the participants in the Enroll-HD study and made available by CHDI Foundation, Inc. Enroll-HD is a clinical research platform and longitudinal observational study for Huntington's disease families intended to accelerate progress towards therapeutics; it is sponsored by CHDI Foundation, a nonprofit biomedical research organization exclusively dedicated to collaboratively developing therapeutics for HD. Enroll-HD would not be possible without the vital contribution of the research participants and their families.

REFERENCES

[1] Gillian P. Bates et al. 2015. Huntington disease. (Apr. 2015). DOI: 10.1038/nrdp.2015.5.

[2] Marc Ciosi et al. 2019. A genetic association study of glutamine-encoding dna sequence structures, somatic cag expansion, and dna repair gene variants, with huntington disease clinical outcomes. *EBioMedicine*, 48, (Oct. 2019), 568–580. DOI: 10.1016/j.ebiom.2019.09.020.

[3] M Duyao et al. 1993. Trinucleotide repeat length instability and age of onset in huntington's disease. (1993). http://www.nature.com/naturegenetics.

[4] Emilia M. Gatto, Natalia González Rojas, Gabriel Persi, José Luis Etcheverry, Martín Emiliano Cesarini, and Claudia Perandones. 2020. Huntington disease: advances in the understanding of its mechanisms. *Clinical Parkinsonism & Related Disorders*, 3, 100056. DOI: 10.1016/j.prdoa.2020.100056.

[5] James F Gusella and Marcy E MacDonald. 2009. Huntington's disease: the case for genetic modifiers. *Genome Medicine*, (Aug. 2009).

[6] James F. Gusella, Jong Min Lee, and Marcy E. Macdonald. 2021. Huntington's disease: nearly four decades of human molecular genetics. (Oct. 2021). DOI: 10.1093/hmg/ddab170.

[7] Jong Min Lee et al. 2015. Identification of genetic factors that modify clinical onset of huntington's disease. *Cell*, 162, (Aug. 2015), 516–526, 3, (Aug. 2015). DOI: 10.1016/j.cell.2015.07.003.

[8] Jian-Liang Li et al. 2003. A genome scan for modifiers of age at onset in huntington disease: the hd maps study. (2003).

[9] Branduff McAllister et al. 2022. Exome sequencing of individuals with huntington's disease implicates fan1 nuclease activity in slowing cag expansion and disease onset. *Nature Neuroscience*, 25, (Apr. 2022), 446–457, 4, (Apr. 2022). DOI: 10.1038/s41593-022-01033-5.

[10] Davina Moss et al. 2017. Identification of genetic variants associated with huntington's disease progression: a genome-wide association study. *The Lancet Neurology*, 16, 701–711, 9. DOI: 10.1016/S1474-4422(17)30161-8Ãŕ.

[11] Michael Orth and Carsten Schwenke. 2011. Age-at-onset in huntington disease. *PLoS Currents*, 3, (July 2011), RRN1258. DOI: 10.1371/currents.RRN1258.

[12] Ricardo Mouro Pinto et al. 2013. Mismatch repair genes mlh1 and mlh3 modify cag instability in huntington's disease mice: genome-wide and candidate approaches. *PLoS Genetics*, 9, (Oct. 2013), 10, (Oct. 2013). DOI: 10.1371/journal.pgen.1003930.

[13] Karen A. Sap, Karlijne W. Geijtenbeek, Sabine Schipper-Krom, Arzu Tugce Guler, and Eric A. Reits. 2023. Ubiquitin-modifying enzymes in huntington's disease. (Feb. 2023). DOI: 10.3389/fmolb.2023.1107323.

Received 27 May 2024