Supplementary Materials

A Machine Learning Approach for Early Diagnosis of Cognitive Impairment Using Population-Based Data

Supplementary Table 1. List of predictor variables and their representation

Category	Predictors	Representation
Social-demographic and genetic risk factor	Age, y	Continuous value
	Sex	Binary (0: female, 1: male)
	Race	Categorical variable
	Highest education attainment	
	Apolipoprotein ε4 (*APOE* ε4) carrier	Binary (0: no, 1: yes)
Vascular risk factors	Smoking	
	Body mass index (BMI)	Continuous value
	Blood pressure systolic, mmHg	
	Blood pressure diastolic, mmHg	
	Total cholesterol, mmol/L	
	Cholesterol high-density lipoprotein, mmol/L (HDL)	
	Cholesterol low-density lipoprotein, mmol/L (LDL)	
	Triglycerides, mmol/L	
	Glycated hemoglobin % (HbA1c)	
	Diabetes	Binary (0: no, 1: yes)
	Hyperlipidemia	
	Hypertension	
	Stroke history	
Neuroimaging markers	Presence of lacunes	
	Presence of cortical microinfarcts	
	Presence of cerebral microbleeds	
	Presence of infarct	
	Presence of intracranial stenosis	
	Total grey matter volume, ml	Continuous value
	Total white matter volume, ml	
	Hippocampus volume, ml	
	White Matter hyperintensities	
	Total intracranial volume, ml	
	Atrophy central R1	Binary (0: no to mild atrophy, 1: moderate to severe atrophy)
	Atrophy cortical R1	
	Atrophy medial temporal R1	

Supplementary Table 2. Grid search for hyperparameter

Algorithm	**Grid search inputs**	**Optimized parameters**
Logistic regression	Regularization parameter: [0.1,1], penalty: [L1, l2, elasticnet]	Regularization parameter: 1, penalty: l2
Support vector classifier	Regularization parameter: [0.01,0.1,1], kernel: [linear, radial, polynomial, sigmoid]	Regularization parameter: 1, kernel: radial-based
Gradient boosting	Learning rate: [0.01, 0.1, 0.5], maximum depth: [3,5,7], number of estimators: [10,30,50,70]	Learning rate: 0.1, maximum depth: 3, number of estimators: 30

Supplementary Table 3. Missing data for each variable

Variables	Count	Percent
Apolipoprotein ε4 (*APOE* ε4)	218	23.93
Body mass index	1	0.11
Total cholesterol, mmol/L	39	4.281
Cholesterol high-density lipoprotein, mmol/L	346	37.98
Cholesterol low-density lipoprotein, mmol/L	41	4.501
Triglycerides, mmol/L	32	3.513
Glycated hemoglobin % (HbA1c)	32	3.513
Total grey matter volume, ml	95	10.428
Total white matter volume, ml	95	10.428
Hippocampus volume, ml	86	9.44
White Matter hyperintensities	86	9.44
Total intracranial volume, ml	86	9.44
Presence of lacunes	85	9.33
Presence of cortical microinfarcts	90	9.879
Presence of cerebral microbleeds	89	9.769
Presence of infarct	85	9.33
Presence of intracranial stenosis	88	9.66

Supplementary Material 1. Description and mathematical expression of performance evaluation metrics

This section describes the performance measure used in past literature on ML-based prediction model for disease diagnosis. Performance metrics, including, Sensitivity, specificity, positive predictive value (PPV) and F1 score [https://doi.org/10.3390/healthcare1003054]. We defined participants with cognitive impairment as true positive (TP) otherwise as true-negative (TN) if participants are correctly predicted by the ML model. Participants were deemed as false positive (FP) or false negative (FN) if being wrongly predicted by the ML model.

Accuracy refers to the total correct predictions (TP+TN) out of the total number of samples. Accuracy is expressed in the mathematical formula as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + FP + TN + FN}$$

Accuracy can range between 0-1 and often expressed in percent with high percentage indicates better model performance.

Sensitivity also known as true positive rate or recall in the field of AI and ML. Sensitivity measures the model's ability to predict true positive among all participants with cognitive impairment in the sample. Sensitivity is expressed in the formula as follows:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + FN}$$

A ML model with high sensitivity will correctly predict most participants with cognitive impairment as positive cases (low false negative results).

Specificity also known as true negative rate. Specificity assesses the performance of the ML-based model to identify true negative case among all participants with no cognitive impairment in sample. Specificity is expressed in the formula as follows:

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{TN + FP}$$

A ML model with high specificity will correctly predict most participants with no cognitive impairment as negative cases (low false positive results).

Positive predictive value (PPV) also commonly known as precision in ML. PPV assesses the performance of the ML-based model in identifying true positive case among all cases predicted positive. PPV is expressed in the formula as follows:

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + FP}$$

Like PPV, **Negative predictive value (NPV)** assesses the performance of the ML-based model in identifying true negative case among all cases predicted negative. NPV is using the formula as follows:

$$NPV = \frac{\mathrm{TN}}{\mathrm{TN} + FN}$$

Lastly, **F1 score** is a harmonic metric that combines sensitivity (recall) and PPV (precision) based on a formula as follows:

$$F1\ score = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + Recall}$$

F1 score ranges between 0 and 1. Higher the recall and precision contribute to higher the F1 score. F1 score is suitable to assess ML model trained and tested on imbalance datasets.
Area under the **receiver operating characteristic curve (ROC)** was used to assess the models' discrimination over all classification thresholds (trade-off between the true positive rate and false positive rate). An ideal model is one that maximize the area under curve. A model with high AUC will have a ROC curve closer to the upper left corner of the plot.

Supplementary Material 2. Description on Shapley Additive Explanations (SHAP) explainer model

In view of the potential consequences of medical decisions, understanding the reasoning behind predictions is crucial [1]. Shapley Additive Explanations (SHAP) was applied to data generated by the ensemble ML model to understand how the algorithm make its prediction. This method has been previously described in [2-4] and was applied to studies on dementia and cognitive impairment.

In brief, SHAP is a post-hoc model-agnostic methods that originates from cooperative game theory. The SHAP algorithm compute the SHAP values to quantify how much each input features contribute to the predicted output. The SHAP values were used to identify and visualize important relationships and help users to understand how the ML models makes predictions in general.

REFERENCES

[1] Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, Xu S, Stub D, Smith K, Tacey M (2018) Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Med* **15**, e1002709.

[2] Hernandez M, Ramon-Julvez U, Ferraz F, Consortium wtA (2022) Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis. *PLoS One* **17**, e0264695.

[3] Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ (2020) Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci Rep* **10**, 20630.

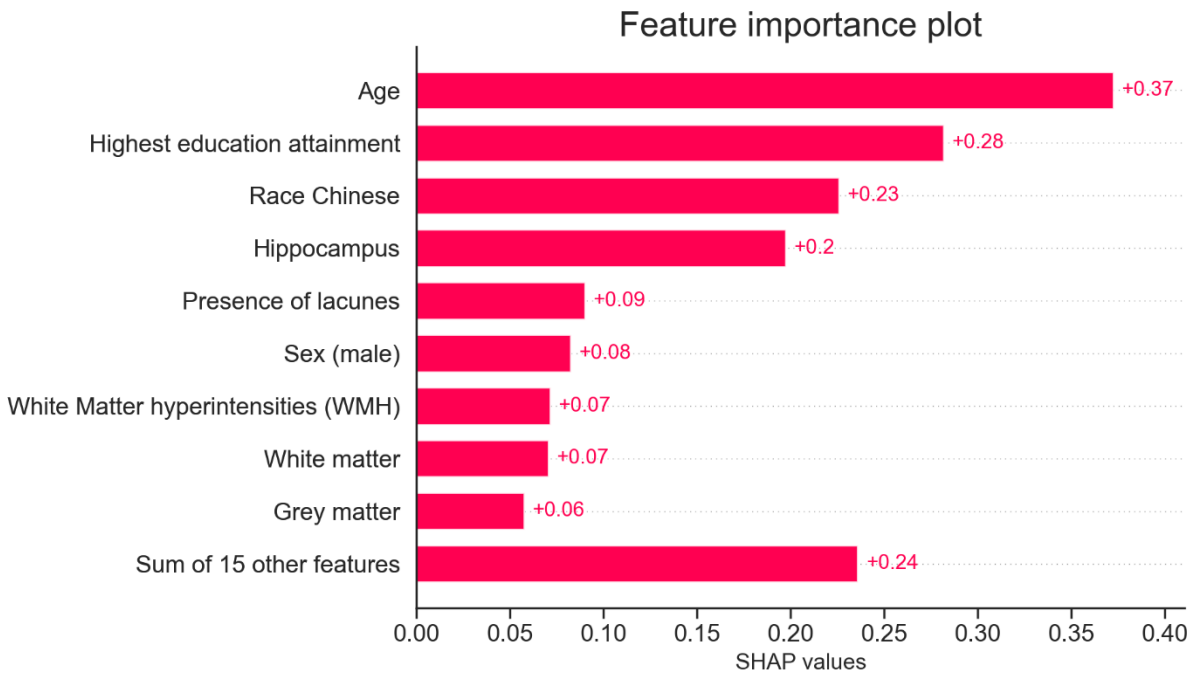
[4] Bloch L, Friedrich CM (2021) Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. *Alzheimers Res Ther* **13**, 155.

Supplementary Table 4. General characteristics comparing participants with complete and missing data

Risk factors	**Overall N= 911**	**Participants with complete data N=604**	**Participants with missing data N=307**	**p**
Age, y	69.8 ± 6.4	69.6 ± 6.3	70.1 ± 6.5	0.237
Sex (male)	452 (49.6)	303 (50.2)	149 (48.5)	0.693
Race	293 (32.2)	194 (32.1)	99 (32.2)	0.340
Chinese	293 (32.2)	194 (32.1)	99 (32.2)	
Indian	322 (35.3)	205 (33.9)	117 (38.1)	
Malay	296 (32.5)	205 (33.9)	91 (29.6)	
Highest education attainment	172 (18.9)	107 (17.7)	65 (21.2)	0.565
Nil	172 (18.9)	107 (17.7)	65 (21.2)	
Primary	380 (41.7)	258 (42.7)	122 (39.7)	
Secondary	255 (28.0)	172 (28.5)	83 (27.0)	
Tertiary	104 (11.4)	67 (11.1)	37 (12.1)	
Apolipoprotein ε4 (*APOE* ε4)	117 (16.9)	110 (18.2)	7 (7.9)	**0.023**
Vascular risk factors				
Smoking	257 (28.2)	162 (26.8)	95 (30.9)	0.219
Body mass index (BMI)	25.6 ± 4.6	25.8 ± 4.6	25.4 ± 4.6	0.200
Blood pressure systolic, mmHg	146.1 ± 19.1	146.5 ± 19.4	145.1 ± 18.4	0.296
Blood pressure diastolic, mmHg	77.1 ± 10.7	77.4 ± 10.6	76.5 ± 10.9	0.224
Total cholesterol, mmol/L	5.0 ± 1.1	5.0 ± 1.1	4.9 ± 1.2	0.293
Cholesterol high-density lipoprotein, mmol/L (HDL)	1.4 ± 0.4	1.4 ± 0.4	1.4 ± 0.4	0.447
Cholesterol low-density lipoprotein, mmol/L (LDL)	3.1 ± 0.9	3.1 ± 0.9	3.0 ± 1.0	0.332
Triglycerides, mmol/L	1.5 [1.0,2.1]	1.4 [1.0,2.1]	1.5 [1.1,2.1]	0.057
Glycated hemoglobin % (HbA1c)	6.0 ± 1.5	5.9 ± 1.4	6.1 ± 1.6	0.127
Diabetes	335 (36.8)	215 (35.6)	120 (39.1)	0.337
Hyperlipidemia	692 (76.0)	460 (76.2)	232 (75.6)	0.909
Hypertension	731 (80.2)	474 (78.5)	257 (83.7)	0.074
Stroke history	43 (4.7)	31 (5.1)	12 (3.9)	0.511
Neuroimaging markers				
Presence of lacunes	132 (14.5)	81 (13.4)	51 (16.6)	0.231
Presence of cortical microinfarcts	45 (4.9)	31 (5.1)	14 (4.6)	0.830
Presence of cerebral microbleeds	281 (30.8)	201 (33.3)	80 (26.1)	**0.031**
Presence of infarct	23 (2.5)	17 (2.8)	6 (2.0)	0.576
Presence of intracranial stenosis	104 (11.4)	72 (11.9)	32 (10.4)	0.575
Total grey matter volume, ml	515.9 ± 63.9	517.1 ± 59.6	512.4 ± 74.7	0.409
Total white matter volume, ml	352.6 ± 52.9	354.4 ± 52.5	347.4 ± 54.0	0.101
Hippocampus volume, ml	3.5 ± 0.4	3.5 ± 0.4	3.5 ± 0.4	0.201
White matter hyperintensities	1.5 [0.4,4.4]	1.5 [0.4,4.2]	1.5 [0.4,4.8]	0.943
Total intracranial volume, ml	1060.5 ± 111.2	1064.8 ± 109.8	1048.8 ± 114.3	0.072
Atrophy central R1	239 (26.2)	163 (27.0)	76 (24.8)	0.520
Atrophy cortical R1	383 (42.0)	268 (44.4)	115 (37.5)	0.054
Atrophy medial temporal R1	288 (31.6)	197 (32.6)	91 (29.6)	0.402

Continuous variables were expressed as a mean value (± SD), while categorical variables expressed as number (percentage %). Non normally distributed variables (triglyceride and white matter hyperintensities) were expressed as median [IQR]

Supplementary Figure 1. Mean SHAP plot



The mean SHAP plot aggregates the mean of the absolute SHAP values across all 911 participants. Predictors with large mean SHAP values have significant impact on the model's cognitive impairment predictions.