Supplementary Material for

Substitute or Supplement? The Role of Multimodal Digital Biomarkers in Mobile Cognitive Impairment Assessment Tools

Whani Kim[1†], Jin Sung Kim[2†], So Yoon Park[1], Hyun Jeong Ko[1], Byung Hun Yun[1,3], Yu Young Kim[1,3], Dong Han Kim[1], Ui Jun Kwon[1], Sang Kwon Lim[1], Bo Ri Kim[4], Jee Hang Lee[2,5*], Geon Ha Kim[6,*], Jin Woo Kim[1,3]

***Co-Correspondence:**

Geon Ha Kim (geonha@ewha.ac.kr)

Jee Hang Lee (jeehang@smu.ac.kr)

List of contents

Supplementary Figures

Supplementary Table

Table S5. Results of Ablation Study – Changes in Classification Performance When Recursively Eliminating Specific DBMs

Table S6. Three-Way ANOVA Results on AUC scores Variation by Task Combination

Table S7. Optimized Feature Sets Selected for Each Task Combination

I. Evaluation Metrics

Two independent performance metrics, (i) Receiver Operating Characteristic Area Under Curve, or ROC-AUC (Equation 5), and (ii) Accuracy (Equation 6), were used in the study. The ROC curve is a graph that graphically shows the relationship between the ratio of the True Positive Rate, TPR (Equation 2), and the ratio of the False Positive Rate, FPR (Equation 3) to the incorrectly identified case for HC. The ROC-AUC value is a value for the area below the ROC Curve, and the closer to 1, the better the model. We also used TPR, which is well known for its sensitivity, as well as True Negative Rate, TNR (Equation 4) which is well known for its specificity, as a performance metric. Additionally, to prevent data leakage, we filled the missing values in both the training and test data based on the mean values from the training data. We also considered applying SMOTE and Adasyn techniques, which are popular oversampling techniques as oversampling techniques, as the data is balanced (Chawla et al., 2022; He et al., 2008). However, it was not well applied in our data and was not adopted(Supplementary Table 1). This was likely due to the high dimensionality of our data, which made these techniques less effective(Jacqueline et al., 2021; Wensheng et al., 2022; Tharinda et al., 2023).

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$TNR = \frac{TN}{FP + TN} \tag{4}$$

$$ROC\text{-}AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where TP, FN, TN, and FP stand for true positives, false negatives, true negatives, and false positives, respectively.

II. SHAP value

The SHAP value for a feature *j* in a specific prediction is formulated as shown in Equation 1:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_S(X_{S \cup \{j\}}) - f_S(X_S)] \quad (1)$$

where ϕ_j is the SHAP value for feature j, S is a subset of all the features F excluding feature j, $|S|$ and $|F|$ are the cardinality subsets of sets S and F, f_s is the prediction function constrained to features in set S, and $X_{S \cup \{j\}}$ and X_s are feature sets with and without including feature j, respectively. The SHAP value represents the average marginal contribution of feature *j* across all possible feature combinations in the prediction model, which directly indicates the relative importance of this feature.

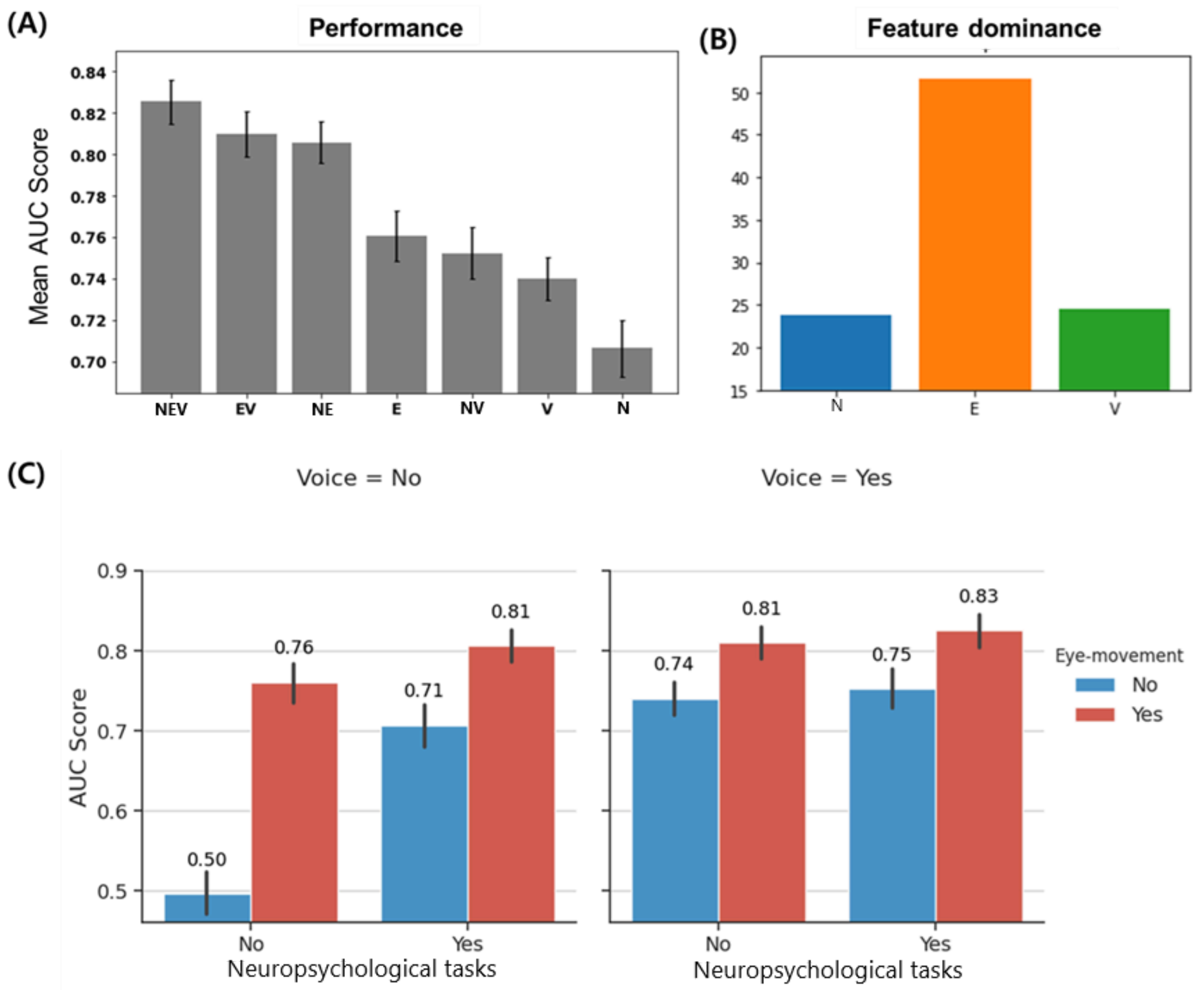

Figure S1. Result. **(A)** Mean AUC score for models that were trained with multiple combinations of DBMs. The X-axis shows the machine learning model with various DBM combinations, while the Y-axis indicates the Mean AUC score of each model. **(B)** Feature dominance of individual DBMs. The X-axis displays the individual DBMs, and the Y-axis shows the feature dominance that ranges from 0 to 100. **(C)** Results from the factorial design experiment among three DBMs. The X-axis indicates the presence of a neuropsychological task, and the Y-axis shows the mean AUC score. N refers to a neuropsychological task, E refers to eye movement, and V refers to voice assessment.

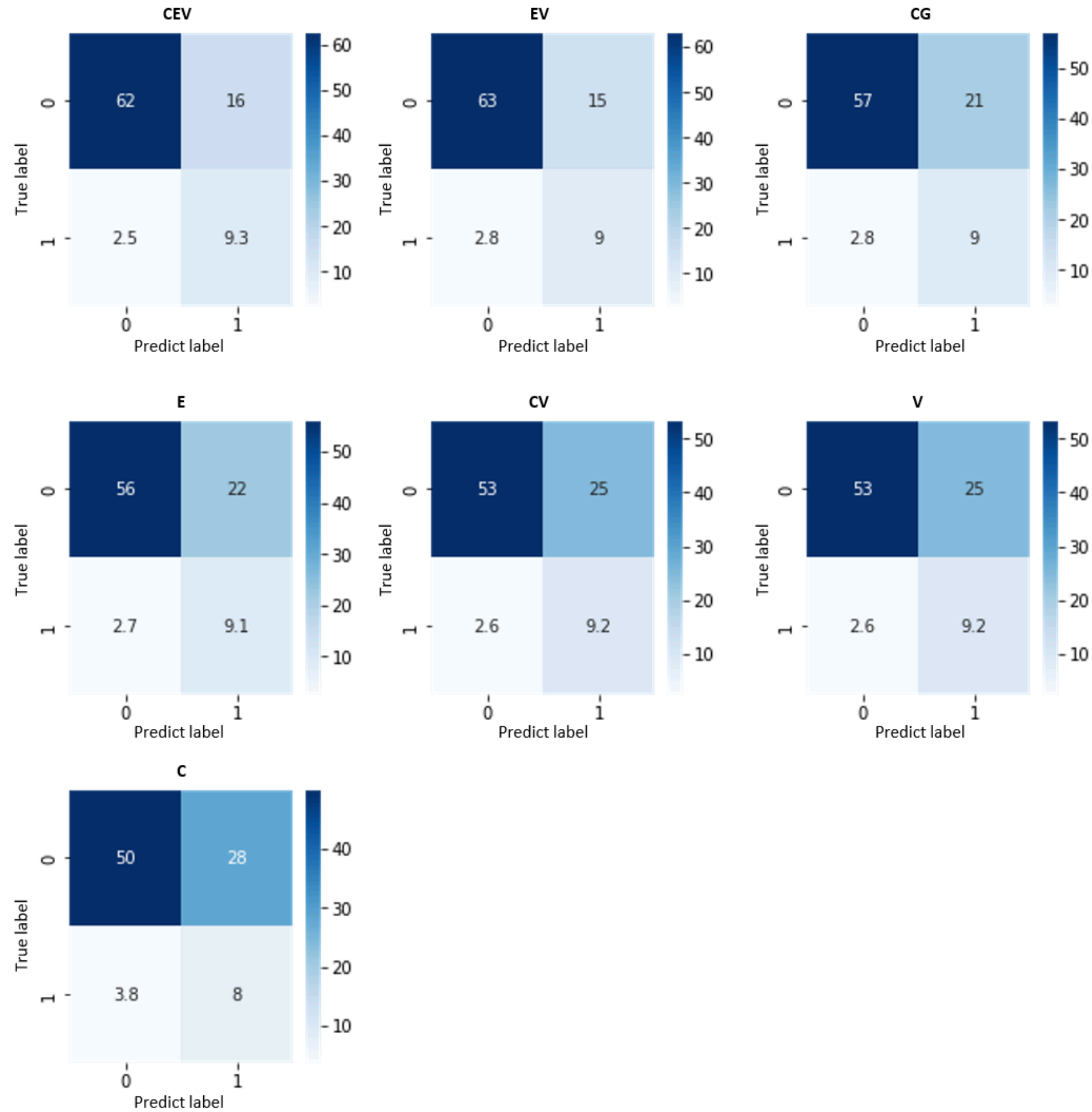

Figure S2. Confusion Matrices (2x2x2) for Various DBM Combinations

Table S1. Applications for Early Screening of Cognitive Impairment (CI): Types of Digital Biomarkers and Sensor Technologies

Digital Tool	Sensor	DBMs	Reference
Digital neuro signature (DNS)	Camera	Eye-movements	Meier et al., 2021 Meier et al., 2020
	Touchscreen	Motor(Speed/Accuracy)	
	Accel/gyro	Gait	
	Device use	Augmented Reality(AR)	
Linkt Health	Microphone	Speech/Voice	Staffaroni et al., 2020
	Accel/gyro	Gait	
Rey auditory verbal learning test (RAVLT)	Microphone	Speech/Voice	Banks et al., 2024
	Touchscreen	Motor(Speed/Accuracy)	
Cogstate brief battery (CBB)	Touchscreen	Motor(Speed/Accuracy)	White et al., 2023
TalkBank CHAT	Microphone	Speech/Voice	Fraser et al., 2016
Wearable device	Microphone	Speech/Voice	Cay et al., 2024

Virtual kiosk	Camera	Eye-movements	Kim et al., 2023
	Hand controller	Hand-movements	

Table S2. Tasks in *Digital Assessment of Cognitive Impairment*

Cognitive Task	Stroop test (**Executive** function; Scarpina et al., 2017)
	Symbol association (**Associative recall**; Troyer et al., 2008)
	Self-ordered pointing task (**Visual working memory**; Geva et al., 2016)
	Arithmetic (**Working memory**; Kasai et al., 2020)
Voice	Sentence memorize and speak (**Logical memory**; Wechsler Memory Scale; Sullivan et al., 2018)
	Picture Description (**Language**; Rentoumi et al., 2014; Ahmed et al., 2013; Nicholas et al., 1985; Tomoeda et al., 1996)
Eye tracking	Smooth pursuit (**basic oculomotor**)
	Saccade (**Attention and inhibitory control**; Crawford et al., 2005)
	Anti-Saccade (**Inhibitory dysfunction**, **working memory**)

Table S3. Comparison of performance metrics (AUC, Accuracy, Sensitivity, Specificity) across ten candidate ML classifiers.

	AUC	Accuracy	Sensitivity	Specificity
CatBoost	**0.65**	**0.67**	0.67	0.67
LightGBM	0.61	0.61	0.70	0.60
XGBoost	0.59	0.64	0.65	0.64
SVM_RBF	0.59	0.63	0.64	0.62
RandomForest	0.59	0.65	0.61	0.65
GradientBoosting	0.58	0.57	**0.71**	0.55
NaiveBayes	0.56	**0.67**	0.55	0.69
Bagging	0.55	**0.67**	0.48	**0.70**
LogisticRegression	0.52	0.65	0.53	0.67
SVM_Linear	0.48	**0.67**	0.48	0.69

We compared the performance of the ten candidate machine-learning algorithms that were described in the Method section using repeated stratified K-Fold cross-validation. As we explained in Methods section 2.4, the input consisted of all 725 features, and the output was the prediction on MCI. Table S3 shows the average AUC of each algorithm. *CatBoost* achieved the highest mean AUC across 30 simulation runs, which outperformed all other models (see Supplement Table 3 for more details). Based on this result, we selected *CatBoost* as the classifier for the final layer (which is denoted as "classifier" in Figure 2B).

Table S4: Results of Ablation Study – Changes in Classification Performance When Recursively Eliminating Specific DBMs

We fixed the total number of features at 28, and recursively eliminated the set of features in each DBM. The baseline for performance comparison was the model trained with the NEV combination (Neuropsychological tasks, Eye movements, Voice). For the elimination process, we removed seven features when eliminating the Voice DBM, seven features when eliminating the Neuropsychological tasks, and 14 features when eliminating the Eye movements DBM.

	Number of Features				Performance							
Task	Neuropsychological tasks	Eye-movement	Voice	Total	AUC	ΔAUC	Accuracy	ΔAccuracy	Sensitivity	ΔSensitivity	Specificity	ΔSpecificity
NEV (Base)	7	14	7	28	0.83	0%	0.81	0%	0.87	0%	0.72	0%
EV	-	14	7	21	0.81	-2%	0.82	+1%	0.71	-16%	0.85	+13%
NE	7	14	-	21	0.77	-6%	0.79	-2%	0.68	-19%	0.81	+9%
NV	7	-	7	14	0.73	-10%	0.78	-3%	0.69	-18%	0.79	+7%
V	-	-	7	7	0.73	-10%	0.74	-7%	0.78	-9%	0.73	+1%
E	-	14	-	14	0.71	-12%	0.7	-11%	0.75	-12%	0.7	-2%
N	7	-	-	7	**0.63**	**-20%**	**0.67**	**-14%**	**0.65**	**-22%**	**0.68**	**-4%**

Table S5: Results of Ablation Study – Changes in Classification Performance When Recursively Eliminating Specific DBMs

We in this case performed Recursive Feature Elimination with Cross-Validation (RFECV) when building machine learning (ML) models using various combinations of DBMs. We trained each ML model with a specific combination of DBMs where we conducted an optimal feature selection process. For example, to train the EV model, we first excluded all features of Neuropsychological tasks (N) while including all features of Eye movements (E) and Voice (V). We then performed optimal feature selection for this model with the E and V combination. This approach resulted in different numbers of selected features for each combination of DBMs.

	Number of Features				Performance							
Task	N	E	V	Total	AUC	ΔAUC	Accuracy	ΔAccuracy	Sensitivity	ΔSensitivity	Specificity	ΔSpecificity
NEV (Base)	7	14	7	28	0.83	0%	0.81	0%	0.87	0%	0.72	0%
EV	-	15	8	23	0.81	-2%	0.83	+2%	0.8	-7%	0.75	+3%
NE	6	18	-	24	0.81	-2%	0.77	-4%	0.86	-1%	0.68	-4%
E	-	30	-	30	0.76	-7%	0.83	+2%	0.75	-12%	0.76	+4%
NV	12	-	35	47	0.75	-8%	0.8	-1%	0.74	-13%	0.73	+1%
V	-	-	40	40	0.74	-9%	0.82	+1%	**0.73**	**-14%**	0.75	+3%
N	13	-	-	13	**0.71**	**-12%**	**0.75**	**-6%**	0.78	-9%	**0.66**	**-6%**

Table S6. *Three-Way ANOVA Results on AUC scores Variation by Task Combination*

Source	sum_sq	df	F	PR(>F)
N	0.3019	1	72.31	2.E-15
E	0.9648	1	231.09	1.E-36
V	0.4829	1	115.67	4.E-22
N * E	0.0988	1	23.66	2.E-0.6
N * V	0.1953	1	46.78	7.E-11
E * V	0.1835	1	43.95	2.E-10
N * E * V	0.107	1	25.62	8.E-0.7
Residual		232		

Table S7. *Optimized Feature Sets Selected for Each Task Combination*

	NEV	EV	NE	NV	V	E	N
1	Reaction_time_A4	PD1.Pause:total	Reaction_time_A4	Reaction_time_A4	PD1.Pause:rate	SP1.Duration1	Reaction_time_A4
2	Reaction_time_A5	PD1.Pause:rate	Reaction_time_A5	Reaction_time_A5	PD1.CIU:rate	SP1.InitAccel:min	Reaction_time_A5
3	Reaction_time_A8	PD1.CIU:rate	Reaction_time_cal2	Reaction_time_A8	PD1.Keyword:rate	SP2.InitAccel1	Reaction_time_A8
4	Reaction_time_cal2	PD1.Keyword:rate	Reaction_time_cal4	Correct_count.2	PD2.Pause:rate	SP2.InitAccel:min	Reaction_time_cal2
5	Reaction_time_cal4	PD2.Pause:rate	Correct_count.3	Reaction_time_cal2	PD3.Pause:rate	SP3.Gain1	Reaction_time_cal4
6	Total_time.7	PD3.Pause:std	Total_time.7	Reaction_time_cal4	WMS1.Phonation:min	SP3.Velocity1	Correct_count.3
7	PD1.Pause:total	PD3.Pause:rate	SP1.Duration1	Total_time.7	WMS3.CIU:rate	Saccade1.Velocity:mean	Total_time.7
8	PD1.Pause:rate	WMS1.Pause:rate	SP1.InitAccel:min	PD1.Pause:total	-	Saccade1.Velocity:median	-
9	PD1.Keyword:count	WMS1.CIU:rate	SP3.Velocity1	PD1.Pause:rate	-	Saccade1.Velocity:std	-
10	PD1.Keyword:rate	WMS2.CIU:rate	SP3.InitAccel2	PD1.Keyword:rate	-	Saccade4.Velocity:min	-
11	PD2.Pause:rate	SP1.InitAccel:min	Saccade1.Velocity:mean	PD2.Pause:rate	-	Saccade5.Velocity2	-
12	PD3.Pause:std	SP2.InitAccel1	Saccade1.Velocity:median	PD3.Pause:rate	-	Saccade5.Velocity:std	-
13	PD3.Pause:rate	SP3.Velocity1	Saccade1.Velocity:std	PDX.Phonation:std	-	Saccade5.Velocity:max	-
14	WMS1.Phonation:min	Saccade1.Velocity1	Saccade1.Velocity:min	WMS1.Phonation:min	-	AntiSaccade5.Velocity:max	-

15	SP1.Duration1	Saccade1.Velocity:mean	Saccade4.Velocity:min	-	-	-	-
16	SP1.InitAccel:min	Saccade1.Velocity:median	Saccade5.Velocity2	-	-	-	-
17	SP3.Gain1	Saccade1.Velocity:std	Saccade5.Velocity:std	-	-	-	-
18	SP3.Velocity1	Saccade4.Velocity:min	Saccade5.Velocity:max	-	-	-	-
19	Saccade1.Velocity1	Saccade5.Velocity2	AntiSaccade1.Duration1	-	-	-	-
20	Saccade1.Velocity:mean	Saccade5.Velocity:std	AntiSaccade2.Velocity:std	-	-	-	-
21	Saccade1.Velocity:median	AntiSaccade1.Duration1	AntiSaccade5.Velocity:max	-	-	-	-
22	Saccade1.Velocity:std	-	-	-	-	-	-
23	Saccade1.Velocity:min	-	-	-	-	-	-
24	Saccade4.Velocity:min	-	-	-	-	-	-
25	Saccade5.Velocity2	-	-	-	-	-	-
26	Saccade5.Velocity:std	-	-	-	-	-	-
27	Saccade5.Velocity:max	-	-	-	-	-	-
28	AntiSaccade1.Duration1	-	-	-	-	-	-