A Preliminary Feasibility Study on Screening Cognitive Impairment based on Multi-Modal Biomarkers and Stacking Ensemble AI Approach

(Extended Abstract)

Whani Kim[1], Jin Sung Kim[2], 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,*], Geon Ha Kim[5,*], Jin Woo Kim[1,3]

[1]Department of Digital Therapeutics Research & Development, HAII Inc, Seoul, Republic of Korea

[2]Department of Human-Centered AI, Sangmyung University, Seoul, Republic of Korea

[3]HCI Lab, Department of Cognitive Science, Yonsei University, Seoul, Republic of Korea

[4]Ewha Medical Research Institute, Ewha Womans University, Seoul, Republic of Korea

[5]Department of Neurology, Ewha Womans University Mokdong Hospital, Ewha Womans University School of Medicine, Seoul, Republic of Korea

(*: Co-corresponding authors: geonha@ewha.ac.kr, jeehang@smu.ac.kr)

Background: Early screening of cognitive impairment is crucial for patients demanding timely treatment. The need for cost-effective and easily accessible tools to detect cognitive decline rapidly progressed with the breakout of the COVID-19 pandemic, which has brought attention to the potential benefits of remote means for cognitive assessment (Wind *et al.*, 2020). Modern digital-based screening tools subsequently make use of smart devices equipped with various onboard sensors (e.g., heart rate variability (HRV) sensors, accelerometers, gyroscope sensors). These digital sensors provide the capacity to collect a variety of health-related data at a higher frequency than traditional screening procedures (Kourtis *et al.*, 2019). This approach in turn accelerates advances in digital biomarker research showing promise in identifying cognitive impairment through both active and passive means of data collection.

Objective: In this work, we aimed to further increase the performance of mobile-app based cognitive assessment tools based on a digital biomarker approach by utilizing the onboard sensors of smartphones. To this end, we assessed speech and eye movement as digital biomarkers to see whether or not they would detect cognitive deficits, along with data collected from a classical cognitive task. Machine learning models were developed to classify the collected data and obtain accuracy on the prediction of the cognitive profiles along two different classes: Healthy Controls (HC) and cognitive Impairment (CI) individuals. Using the classifier, we primarily aimed at investigating to what extent the combination of digital biomarkers and cognitive tasks improves

classification accuracy. In addition, we attempted to probe to what extent digital biomarkers in classifiers enhance the performance in early screening of cognitive impairment as complementary to cognitive task data. We thus conducted an ablation study with three classifiers trained with (i) both two new digital biomarkers and cognitive task data (*Cog+Bio*), (ii) two digital biomarkers (*Bio*) and (i) cognitive task data (*Cog*). We anticipated that a hybrid form of the digital screening tool we proposed would be able to capture the various cognitive functions.

Methods: Nine *Alzguard-D* tasks were designed based on three biomarkers: Keystroke, speech/ language and eye movement (Table 1; Figure 1). The participants in this study were recruited from the national institute of dementia, welfare center, and nursing homes. A total of 48 CI and 241 HC were recruited according to detailed inclusion and exclusion criteria. The participants completed the second edition of the Korean Mini-Mental State Examination (*K-MMSE 2nd Ed.*; Kang, Y.) at the institution. The survey was conducted for about 15 minutes with a total of 26 questions. After completing *K-MMSE 2nd Ed.*, participants completed tasks in *Alzguard-D*, which took a total of 20 to 30 minutes approximately. Participants were categorized to CI when the scores of *K-MMSE 2nd Ed*. were one standard deviation below the age, education and gender-matched norm.

Experiments: Once collected the participants' data, we conducted exploratory experiments on the classification of CI and HC using a machine learning (ML) technique to probe the impact of biomarkers. To that end, we developed a stacking ensemble ML model (Figure 2). Since our data consisted of multi-modal data combined with (unstructured) biomarkers and (structured) cognitive task data, we first extracted the best feature sets, widely accepted in the clinical setting by experts, for each unstructured biomarker using deep neural networks at the first stack (Miner *et al.*, 2021). When all biomarkers' features were extracted so that the intermediate features were concatenated with the structured cognitive task data, the ML model at the second stack performed the classification on CI and HC. To train and test the model, we randomly split the participants' data into a training set (HC=217, CI=43) and a test set (HC=24, CI=5) based on *Repeated Stratified K-fold* scheme. We set the number of folds (K) to four, and the number of repetitions to five. We in the end built 20 models to investigate the impact of biomarkers on the classification of mild cognitive impairment. We note that pre-trained deep-learning models were used and only the final layer was trained for the feature extractor,

Results: To begin with, we measured the effect of biomarkers on the improvement of classification accuracy. Given the result of three classifiers trained with three sets of data (*Cog+Bio*, *Bio*, *Cog*), a paired *t*-test was conducted to test whether the performance increase due to the addition of digital biomarkers was statistically significant, which revealed ($p < 0.0001$) (Figure 3A). We note that a support vector machine (SVM) was employed for the three classifiers to make sure optimality in the classification. Secondly, we found the best classification models to complete the stacking ensemble ML model. *CatBoost* algorithm statistically outperformed all other candidate algorithms (Bagging, Logistic Regression, LGBM, Naïve Bayes, XGBoost, Random Forest, SVM, Gradient Boosting). We then evaluated the AUC of three different classifiers using *CatBoost*

trained with *Cog+Bio*, *Bio*, and *Cog*. Results showed that *Cog+Bio* AUC score achieved the highest score of 0.876 (max: 0.942), followed by *Bio* AUC with a score of 0.783 (max: 0.845), then *Cog* AUC with 0.677 (max: 0.726) (Figure 3B). Taken together, we were able to significantly improve the AUC of *Alzguard-D* with the inclusion of new digital biomarkers on top of existing cognitive task data. We note that SMOTE (Chawla *et al.*, 2002) was applied to avoid the data imbalance issue, and mean imputation was adopted for missing values.

Conclusions: We proposed a hybrid form of a digital screening tool for cognitive impairment by implementing cognitive tasks and digital biomarkers in the form of speech and eye movements. Instead of using the “End-to-End” paradigm, we employed an intermediate feature-level ensemble approach to effectively analyze the collected multi-modal data. AUC levels of models trained with three sets of data were evaluated to compare the performance of the digital biomarkers vs. cognitive task and cognitive+digital biomarkers. The result confirms that cognitive alongside digital biomarkers significantly increase the screening performance on cognitive impairment, which in turn achieved the best performance, a relatively high average AUC score of 0.876.

Table 1. Tasks in *Alzguard-D*

Biomarkers	Tasks
Cognitive task	Numeric stroop test (Executive function; Scarpina et al., 2017)
	Remembrance paired symbols (Associative recall; Troyer et al., 2008)
	Self-ordered pointing task (Visual working memory; Geva et al., 2016)
	Quantitative Comparison (Working memory; Kasai et al., 2020)
	Memorize sentence 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)
Speech	Memorize sentence 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 movement	Smooth pursuit(basic oculomotor)
	Saccade (Attention and inhibitory control; Crawford et al., 2005)
	Anti-Saccade(Inhibitory dysfunction, working memory)

Figure 1. ***AIzguard-D* overview. (A)** A screenshot of *Numeric stroop* test. **(B)** A screenshot of *Memorize sentence and speak* test to collect a speech biomarker. **(C)** A screenshot of *Smooth pursuit* test to collect an eye movement biomarker.

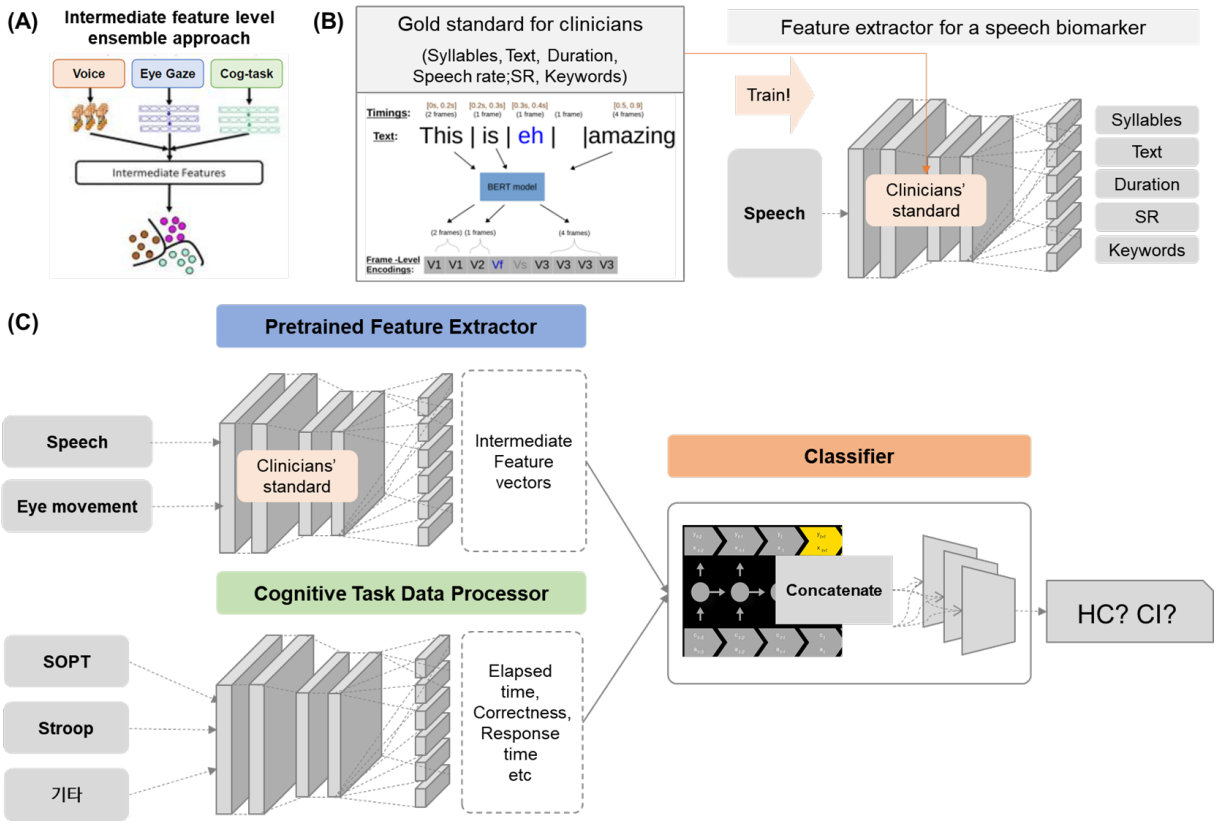

Figure 2. **Overview of the stacking ensemble model**. **(A)** A basic concept of the intermediate feature level ensemble approach. **(B)** An example of deep-learning based feature extractor. Instead of using 'end-to-end' paradigm, we used the feature extractor specifically trained with clinicians' favoured characteristics for screening HC and CI. The input was a speech, and the output was a set of features. **(C)** A schematic of the stacking ensemble model.

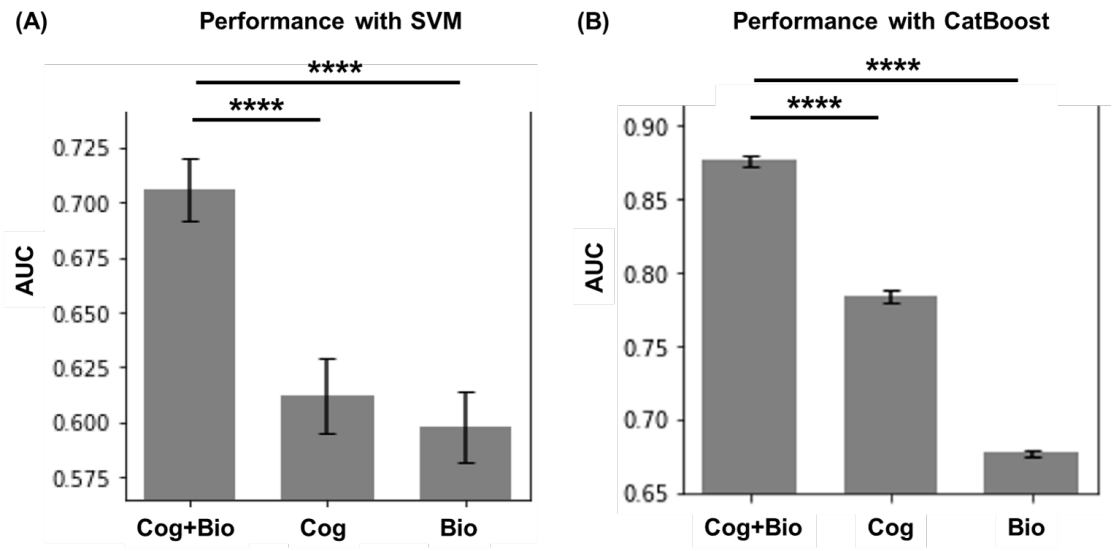

Figure 3. **Comparison of AUC scores in models of cognitive+digital biomarker, digital biomarker only, and cognitive task data only**. **(A)** Performance. SVM was the final stack of the stacking ensemble model. **(B)** Performance. CatBoost was the final stack.