Developing an Accessible Dementia Assessment Tool: Leveraging a Residual Network, the Trail Making Test, and Demographic Data

Authors: Jingmei Yang[1], Samad Amini[1], Boran Hao[1], Seho Park[2], Cody Karjadi[3], Lance San Souci[2], Vijaya B. Kolachalama[2,4,5], Stephanie Cosentino[6], Stacy L. Andersen[2], Rhoda Au[2,3,7], and Ioannis Ch. Paschalidis[1,4,*]

[1] Department of Electrical & Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University

[2] Department of Medicine, Chobanian & Avedisian School of Medicine, Boston University

[3] Framingham Heart Study, Boston University

[4] Faculty of Computing & Data Sciences, Boston University

[5] Department of Computer Science, Boston University

[6] Department of Neurology, Columbia University Irving Medical Center

[7] Departments of Anatomy & Neurobiology, Neurology, and Epidemiology, Chobanian & Avedisian School of Medicine and School of Public Health, Boston University

Corresponding Author:

Ioannis Ch. Paschalidis

Rafik B. Hariri Institute for Computing and Computational Science & Engineering,

Boston University, 665 Commonwealth Ave., Boston, MA 02215, USA

e-mail: yannisp@bu.edu, Tel: 617-353-0434, http://sites.bu.edu/paschalidis

Abstract

Background: The global burden of Alzheimer's disease and related dementias is rapidly increasing, particularly in low- and middle-income countries where access to specialized healthcare is limited. Neuropsychological tests are essential diagnostic tools, but their administration requires trained professionals, creating screening barriers. Automated computational assessment presents a cost-effective solution for global dementia screening.

Objective: To develop and validate an artificial intelligence-based screening tool using the Trail Making Test (TMT), demographic information, completion times, and drawing analysis for enhanced dementia detection.

Methods: We developed: (1) non-image models using demographics and TMT completion times, (2) image-only models, and (3) fusion models. Models were trained and validated on data from the Framingham Heart Study (FHS) (N = 1,252), the Long Life Family Study (LLFS) (N = 1,613), and the combined cohort (N = 2,865).

Results: Our models, integrating TMT drawings, demographics, and completion times, excelled in distinguishing dementia from normal cognition. In the LLFS cohort, we achieved an Area Under the Receiver Operating Characteristic Curve (AUC) of 98.62%, with sensitivity/specificity of 87.69%/98.26%. In the FHS cohort, we obtained an AUC of 96.51%, with sensitivity/specificity of 85.00%/96.75%.

Conclusions: Our method demonstrated superior performance compared to traditional approaches using age and TMT completion time. Adding images captures subtler nuances from the TMT drawing that traditional methods miss. Integrating the TMT drawing into cognitive assessments

enables effective dementia screening. Future studies could aim to expand data collection to include more diverse cohorts, particularly from less-resourced regions.

Keywords: Alzheimer's disease, Artificial Intelligence, Dementia, Trail Making Test

Introduction

Dementia, of which *Alzheimer's Disease (AD)* is the most common form, profoundly impacts memory, thinking, and daily functioning. The global prevalence of dementia is increasing rapidly and is projected to reach 139 million cases by 2050.[1] Currently, about 60% of dementia cases are in low- and middle-income countries, a figure projected to rise to 71% by 2050.[2] Dementia not only impacts individuals but also imposes substantial financial and emotional strains on families and societies, with costs exceeding 1.3 trillion dollars annually and anticipated to rise to 2.8 trillion dollars by 2030.[3] Dementia, with its substantial global economic impact, often remains undiagnosed, with only 20-50% of cases identified in high-income countries and fewer in low-income regions.[4] The growing global population of older adults intensifies the need for large-scale screening to effectively manage, monitor, and even predict this age-related disease.[5,6]

To address the need for large-scale dementia screening, pen-and-paper drawing tasks within neuropsychological test batteries have become essential. These include the *Clock Drawing Test (CDT)*,[7] *Trail Making Test (TMT)*, *Pentagon Drawing Test (PDT)*,[8] and *Rey-Osterrieth Complex Figure Test (RCFT)*, which are extensively used in detecting neurocognitive disorders such as *Alzheimer's Disease and Related Dementias (ADRD)*. Although these drawing tests can be easily implemented in a pen-and-paper format, they still require administration and interpretation by trained professionals. Recent research has utilized digital pen technology to collect drawing tests electronically. Leveraging digital data, studies have used deep learning techniques to develop automated scoring systems, demonstrating promising performance in the PDT,[9–12] RCFT,[13,14] and CDT.[15,16]

However, research on developing an automated diagnostic tool using the TMT has been relatively

limited. The TMT consists of two parts: (*i*) in Part A (TMT-A), numbered circles are displayed on a piece of paper, and participants are instructed to use a pen to draw lines to connect the numbers in sequential order as quickly as possible; (*ii*) in TMT Part B (TMT-B), a series of numbers and letters are displayed and participants are instructed to connect the numbers and letters in alternate sequences.[17] Over the past decades, research has shown that the TMT is a useful tool for detecting cognitive impairment, with completion time being a strong indicator. This has led to the development and validation of adapted TMT versions across many countries.[18–20] A major area of research on TMT involves establishing normative data for older adult populations, investigating differences based on education level, gender, and age across culturally and geographically diverse countries.[21–23]

Additionally, digital adaptations of the TMT, known as dTMT, have introduced more nuanced measurements. Studies comparing these digital iterations with the traditional pen-and-paper format affirm their efficacy in cognitive assessment.24 The dTMT's key advantage is its capability to gather additional data, offering a richer, more comprehensive analysis than its original counterpart.25–27 While there are studies demonstrating the utility of TMT in discriminating between cognitive impairment and normal cognition utilizing geographically diverse datasets, a significant portion of prior studies have relied on data from a single center or limited geographical area with small sample sizes. Moreover, the majority of these studies utilized only the completion time of the TMT as a predictor variable, with few incorporating the actual drawings into their models. As a result, the predictive potential of TMT drawings remains largely unexplored. To fully explore the diagnostic capabilities of the TMT for dementia, large-scale, multi-center studies are needed that analyze both the TMT completion time and the drawings, in conjunction with demographic information.

To accelerate the screening and diagnosis of cognitive impairment globally, our study aimed to develop an accessible detection tool that can be readily implemented worldwide, including in resource-limited regions. Accordingly, our model utilized easily collectible information such as age, gender, education level, the completion time of TMT-A, and the TMT-A drawing. Specifically, TMT-A was selected over TMT-B given its use of universally recognized Arabic numerals, as opposed to alphabet letters used in TMT-B which may not be as widely recognized, especially in non-English speaking countries. Using data from the *Framingham Heart Study (FHS)* and *Long Life Family Study (LLFS)*, our multifaceted study first explored the ability of demographics and completion time, individually and in combination, to distinguish individuals at risk for developing dementia. We then fine-tuned two vision networks to evaluate the predictive potential of TMT-A drawings. Ultimately, we developed a fusion model that integrated three key components: (1) the probability score output by the fine-tuned vision model when using a TMT-A drawing as input, (2) demographic characteristics, and (3) the TMT-A completion time. By combining these signals in our fusion model, we significantly enhanced the model's overall performance both within and across studies.

Methods

Study cohorts

We used data collected from 1,252 participants in the FHS cohort and 1,613 participants in the LLFS cohort. FHS is the longest ongoing longitudinal transgenerational cohort study of chronic disease.[28] LLFS is a multicenter (Boston, New York, Pittsburgh, and Denmark) longitudinal study of human longevity and healthy aging.[29] Data used in our study were collected across several geographic regions which are different in environment, culture, and demographics. All participants provided written informed consent. Study protocols and consent forms were approved by the Boston University Medical Campus Institutional Review Board and the Institutional Review Boards of the LLFS field sites as well as the LLFS coordinating center at Washington University St. Louis.

The cognitive status was provided by each study cohort. In the FHS cohort, a participant's cognitive status was determined by the dementia diagnostic review panel.[30] A dementia diagnosis for those showing signs of cognitive decline was reached by consensus between at least one neurologist and one neuropsychologist, based on neurology exams, medical records, and brain imaging.[31] In the FHS dataset, there were 1,233 participants with normal cognition (NC) and 19 with dementia. Similar to the FHS cohort, an impaired cognitive status in the LLFS cohort was determined by a dementia diagnostic review panel based on cognitive testing and informant interviews. Specifically, participants from Denmark who were flagged for a dementia review were omitted due to the absence of case adjudication in Denmark, with only those classified as normal being included. Given the gradual progression of cognitive decline and the limited available samples of Mild Cognitive Impairment (MCI) cases, we focused our analysis on distinguishing

between normal cognition and dementia cases. Consequently, 1,548 participants were identified with normal cognition, and 65 participants were diagnosed with dementia. The participant selection process for both the FHS and LLFS datasets is illustrated in Figure 1.

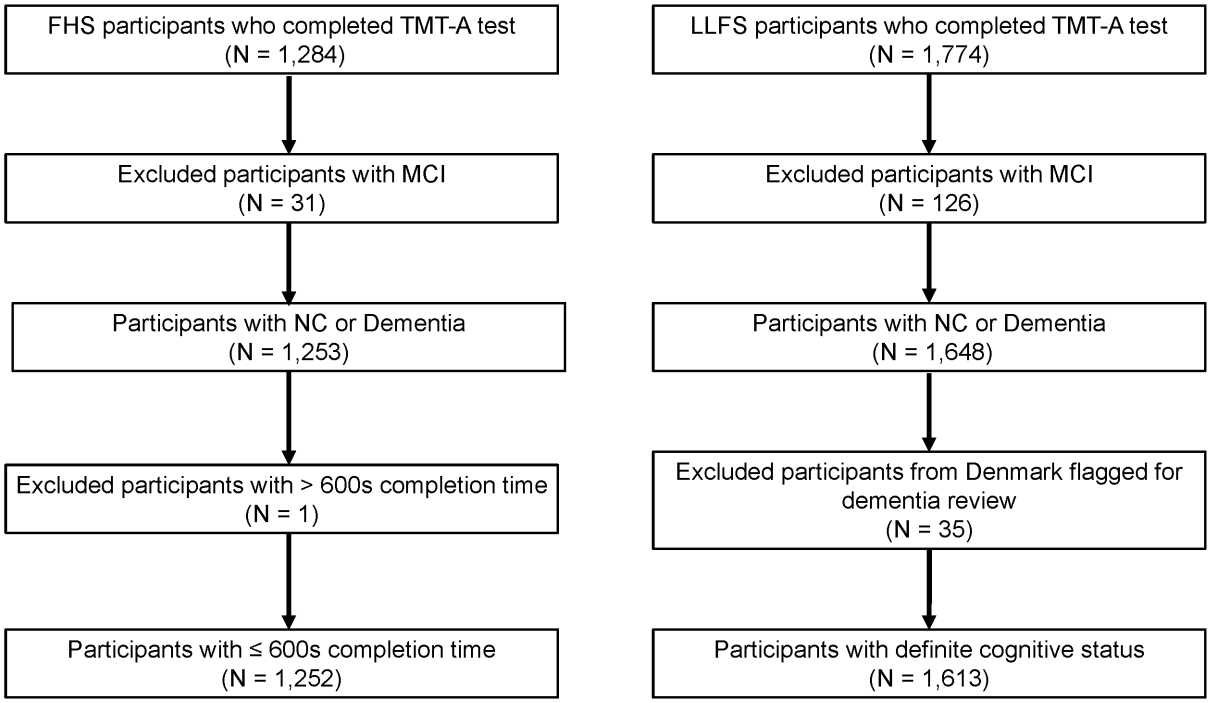


Figure 1: Flow diagram for the participant selection process in the Framingham Heart Study (FHS) and the Long Life Family Study (LLFS). Participants with mild cognitive impairment (MCI), those with greater than 600s completion time on the Trail Making Test Part A, and those without a definite cognitive status were excluded.

Data preparation

During the in-person TMT, information such as gender, age, education, and type of Apolipoprotein E (ApoE) alleles was documented. A digital pen recorded x and y coordinates approximately every 13 milliseconds. The data corresponding to each pen stroke, which is defined as continuous drawing without lifting the pen, was saved in a separate section of a *.txt* file.

We deliberately chose not to use the raw data collected by the digital pen, as digital pens are not

readily available, and older adults are often unfamiliar with such technology. Our goal was to develop an automated tool that does not heavily depend on resources or expertise predominantly available in developed countries and regions. While digital pens capture rich information including pressure sensitivity, precise temporal dynamics, and stroke-level metrics, these advantages come at the cost of accessibility and widespread applicability. There were additional reasons for not relying on the digital pen detailed trajectory data. Specifically, LLFS test administrators recorded participant names during the TMT administration. By converting to a static image format, we could easily anonymize the data through targeted image cropping, while preserving the essential elements of the TMT drawing itself. Further, an image-based approach allows our method to be applied to TMT drawings collected through various means, including those administered with traditional paper-and-pencil methods and subsequently digitized. This universality enhances the potential for retrospective analyses of existing datasets and enables wider implementation across diverse clinical environments without requiring specialized equipment.

To diminish our reliance on digital pens, we developed a preprocessing pipeline. This pipeline was designed to extract x and y coordinates, compute the duration of each stroke, and derive the overall completion time. We then plotted these extracted coordinates from a participant on a blank canvas to create an image and stored it as a *.png* file. This approach preserves the essential spatial and visuomotor patterns evident in TMT performance while allowing for standardized processing across diverse collection methods.

Our datasets are highly imbalanced, with a significantly smaller proportion of dementia cases compared to normal cognition cases. To address this imbalance and mitigate potential overfitting, we implemented a comprehensive set of strategies. Following our previous work, we generated additional training data through image augmentation, applying transformations more frequently to

the minority class.[15] These transformations include rotations of ±10 degrees, zooming in/out by ±15%, width and height shifts of ±10%, shearing of ±10%, and image resizing to 224 × 224 pixels. We applied these transformations disproportionately, based on the ratio of positive to negative cases. This approach enabled us to expand our training dataset with more varied positive cases. For non-image and fusion models, we adopted oversampling in the training sets, where samples from the minority class were randomly duplicated to match the majority class distribution. Additionally, we implemented stratified 5-fold cross-validation to ensure that each fold maintained the same proportion of dementia and normal cognition cases as the overall dataset, preventing potential sampling biases during model evaluation. To further optimize model performance on imbalanced data, we refined the classification threshold based on F1 scores for each fold, typically resulting in thresholds greater than the standard 0.5. This approach helped achieve a more appropriate balance between sensitivity and specificity in our predictions. We also applied regularization techniques to prevent overfitting. Specifically, we used weight decay (ℓ_2-norm penalty with weight $\lambda = 0.001$) during model training, which penalizes large weights and encourages the model to learn more generalizable patterns. Detailed data preprocessing steps and imputation procedures are provided in Appendix A.

Statistical analysis and performance metrics

We performed the Kolmogorov-Smirnov test for continuous variables and the χ^2 test for categorical variables to determine whether the distributions across the *Normal Cognition (NC)* and Dementia groups are significantly different.[32,33] A significance level of 0.05 was used. A p-value less than 0.05 indicates that the distribution of a given feature is significantly different across cognitive statuses.

The data were randomly divided using stratified five-fold cross-validation. A model was trained on four folds and tested on the remaining fold. This training process was repeated five times. Performance metrics for all models were reported as the mean across the five runs, along with the standard deviation. The performance metrics included Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, weighted F1 score, and accuracy.

Models

As demonstrated in Figure 2, we developed three types of binary classification models aimed at identifying participants with dementia from those with normal cognition: (*i*) image-only models, (*ii*) non-image models, and (*iii*) fusion models.

The TMT-A drawings were used as inputs for our image-only models and two backbones were employed: *Residual Network (ResNet)* and *Vision Transformer (ViT)*.[34,35] Specifically, we selected the *ResNet-50* variant and the *ViT* base variant. ResNet-50 is a 50-layer convolutional neural network, which includes 48 convolutional layers, one *MaxPool* layer, and one average pooling layer. Conversely, the ViT base variant has 12 transformer layers, a hidden size of 768 dimensions, and 12 attention heads; it is designed to process 196 patches of 16 × 16 pixels each. In our study, the ResNet-50 and ViT base variants are referred to as ResNet and ViT, respectively. Given the relatively small sample size, we applied transfer learning and initialized the selected backbones with weights pre-trained on ImageNet.[36] We fine-tuned these two backbones by appending a fully connected layer to each and training only that layer while freezing all other layers. Notably, this approach significantly reduces the complexity of training; only 1,538 trainable parameters in ViT and 4,098 in ResNet need to be trained. In the model training process, a batch size of 32 and a total of 50 epochs were used to train each vision model. Images were resized to 224 × 224 pixels

and were normalized by dividing the value of each pixel by 255, thus rescaling the pixel values to a range between 0 and 1. The Adam optimizer was utilized for updating model parameters, with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-3}. A dynamic adjustment mechanism for the learning rate was implemented, reducing the learning rate by a factor of 0.2 if no improvement in validation loss was observed for 5 consecutive epochs. The best model state was saved based on the lowest validation loss observed over the training course.

We used factors such as age, gender, education, and completion time to develop our non-image models. Although ApoE is known for its significant predictive value in dementia, we intentionally excluded it from our model development since it isn't routinely assessed in dementia evaluations. We developed a wide range of traditional machine learning classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).[37] These classifiers were trained solely on demographic features (referred to as baseline), specifically age, gender, and education, to determine the best model for subsequent analysis. As the LR classifier showed the highest overall performance, we selected the LR algorithm for our fusion model (cf. Appendix B). To understand how non-imaging features affected our model's performance, we used the following sets of features as inputs to logistic regression models for comparison: (1) age alone, (2) education alone, (3) gender alone, (4) completion time alone, (5) baseline (age, education, and gender), (6) age and completion time, and (7) baseline and completion time.

The fusion model integrates multiple features through a logistic regression model. Specifically, we combine: (1) the probability score derived from the fine-tuned vision model's analysis of TMT-A drawings, (2) demographic information, and (3) completion time. All these features serve as inputs to the logistic regression model, which then predicts the final probability of dementia. Figure 3

209 illustrates how the best-performing fusion model combines age, completion time, and the ResNet-
210 50 derived image probability to generate predictions, alongside participants' ground truth
211 cognitive status.

212

213

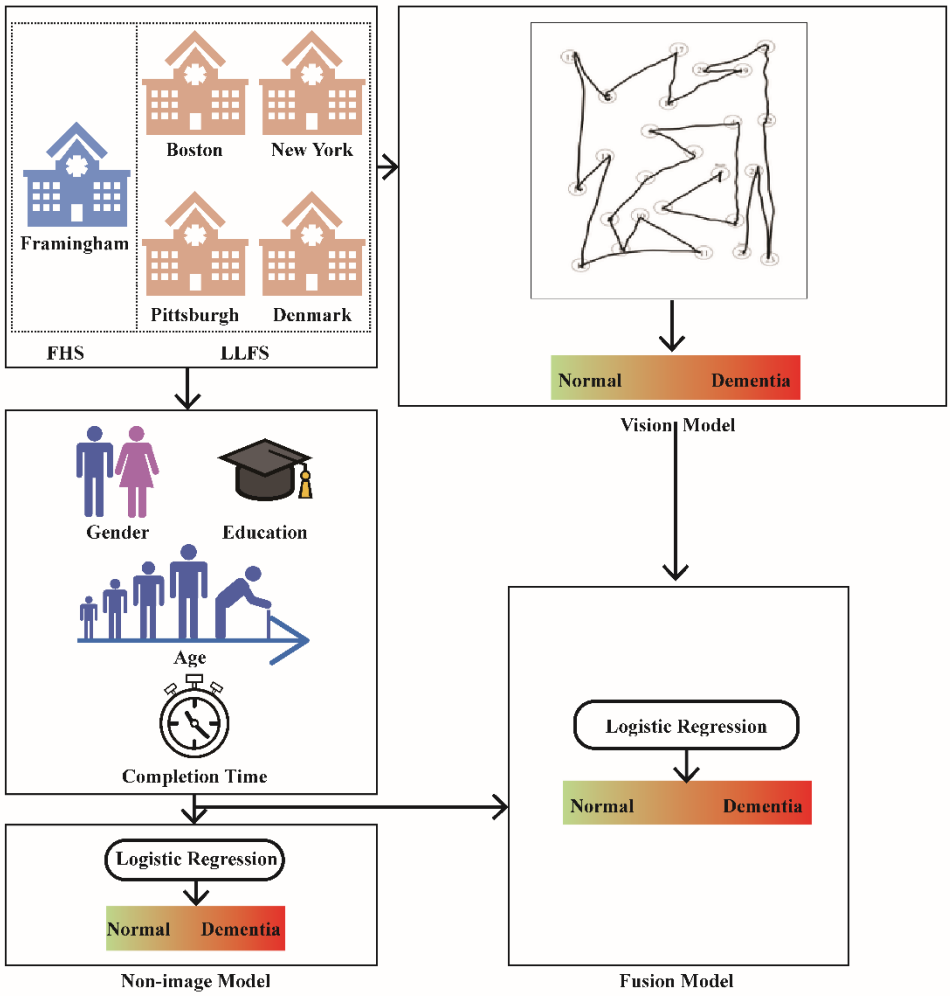


Figure 2: Dementia detection framework. This study leveraged data from two cohorts: the Framingham Heart Study (FHS) cohort (N = 1,252, single-center in Framingham, MA) and the Long Life Family Study (LLFS) cohort (N = 1,613, multi-center in Boston, New York, Pittsburgh, and Denmark). Our study used two forms of data collected during an in-person neuropsychological test: demographic characteristics (age, gender, and education) and digital pen data from the Trail Making Test Part A (TMT-A), including the TMT-A completion time and x and y coordinates. These coordinates were plotted and saved as *.png* images. We developed three types of dementia detection models: (*i*) image-only models, where the *Vision Transformer (ViT)* and *Residual Network (ResNet)* were fine-tuned using TMT-A drawings; (*ii*) non-image models, utilizing demographic data and the TMT-A completion time; and (*iii*) fusion models, which combine non-image features with the image scores derived from the fine-tuned vision models to differentiate individuals with dementia from their normal counterparts.

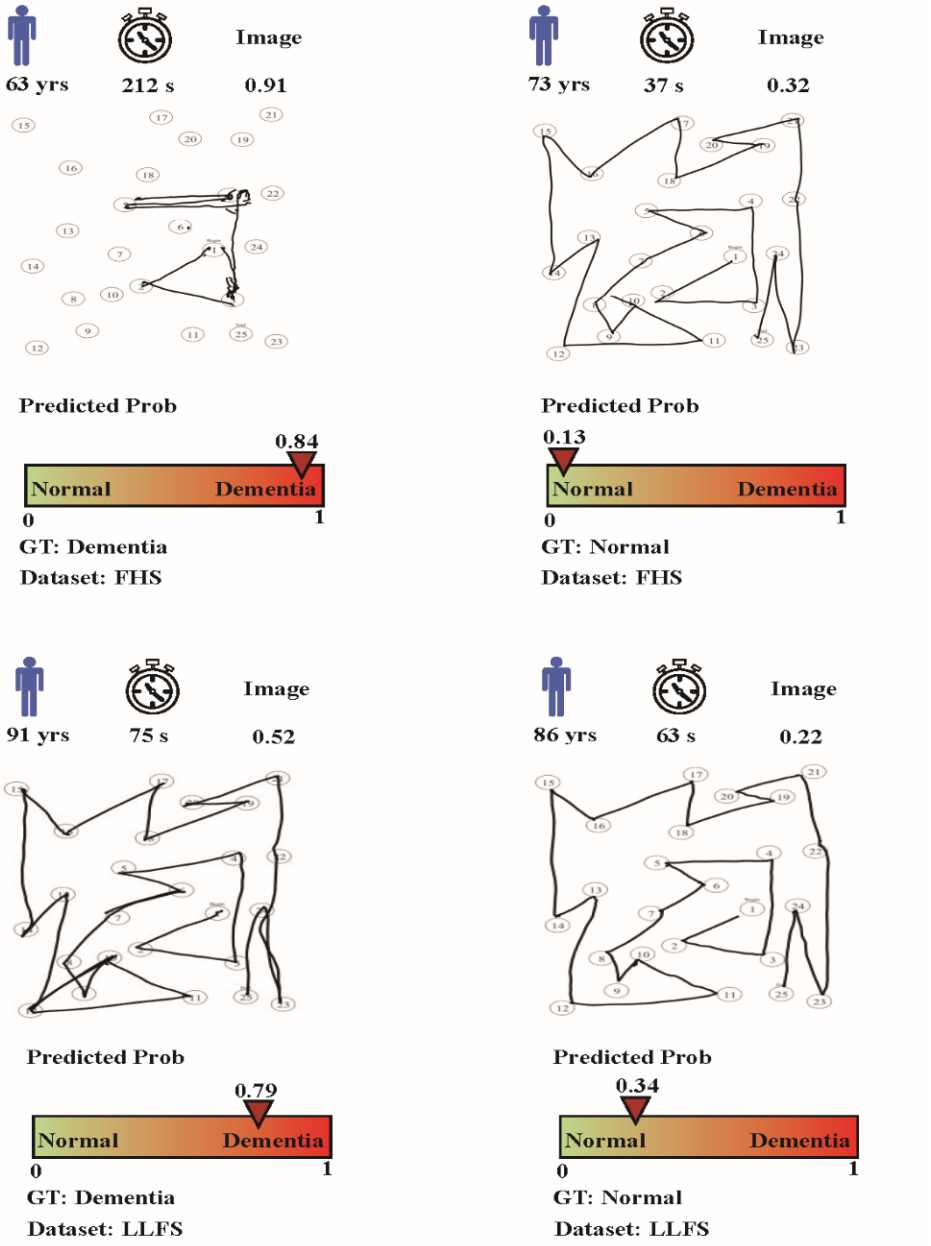


Figure 3: Visualization of the proposed fusion model. The figure includes four subfigures, each representing a participant from either the FHS or LLFS cohort. Two of the participants are diagnosed with dementia, while the other two are cognitively normal. For each participant, their age, the TMT-A completion time, and the probability of dementia derived from the ResNet-50 model based on the TMT-A image are displayed. The bar below each subfigure shows the predicted probability of dementia when combining age, the TMT-A completion time, and the image-derived probability as predictors in the proposed fusion model. The Ground Truth (GT) label indicates the actual cognitive status of each participant.

Results

Characteristics of the participants

Descriptive statistics based on cognitive status (NC and dementia) are presented in Table 1 and Table 2. The feature distributions for the FHS and LLFS datasets are provided in Appendix C. In both datasets, several differences emerged when comparing the characteristics of participants in the dementia group with those in the normal cognition group. Participants flagged as demented were generally older. They took longer to complete the TMT-A. The dementia group also had a higher percentage of females than males. Additionally, a greater proportion of participants in the dementia group had an education of high school level or lower compared to those in the normal group.

When comparing the overall characteristics across datasets, distinct differences become evident. Specifically, participants from LLFS are older than those from FHS on average. Additionally, LLFS has a higher proportion of individuals with an education level of high school or lower. Typically, individuals from LLFS take a longer time to complete the test.

Table 1: Descriptive characteristics and statistical analysis in the FHS dataset. This table compares both numerical and categorical features between the Normal Cognitive (NC) and dementia groups within the FHS cohort. For Age and Completion time, this table shows mean values for the NC, dementia, and overall groups. Kolmogorov-Smirnov tests were conducted to assess the significance of differences in these features between NC and dementia groups. For Gender, Education, and Apolipoprotein E (ApoE), column percentages are presented for each group. Chi-square (χ^2) tests were conducted to determine the significance of differences in these features between NC and dementia groups. A significance level of $\alpha = 0.05$ was used, with a *p*-value less than 0.05 indicating a significant difference in feature distribution by cognitive status.

Feature	**NC**	**Dementia**	**Overall**	***p*-value**
	N = 1233 (98.48%)	N = 19 (1.52%)	N = 1252 (100.00%)	
Age	69.74 ± 12.31	84.79 ± 8.71	70.25 ± 12.42	< 0.01
Gender				0.37
Female	713 (57.83%)	13 (68.42%)	741 (57.76%)	
Male	520 (42.17%)	6 (31.58%)	542 (42.24%)	
Education				0.24
High school or lower	228 (18.49%)	5 (26.32%)	242 (18.86%)	
College or above	1005 (81.51%)	14 (73.68%)	1041 (81.14%)	
Completion time	37.56 ± 21.89	105.44 ± 72.9	39.14 ± 25.20	< 0.01
ApoE				0.09
ε2/ε2	7 (0.57%)	0 (0.00%)	7 (0.55%)	
ε2/ε3	148 (12.00%)	5 (26.32%)	157 (12.24%)	
ε2/ε4	23 (1.87%)	0 (0.00%)	23 (1.79%)	

ε3/ε3	812 (65.86%)	7 (36.84%)	840 (65.47%)
ε3/ε4	226 (18.33%)	7 (36.84%)	237 (18.47%)
ε4/ε4	17 (1.38%)	0 (0.00%)	19 (1.48%)

254

255

Table 2: Descriptive characteristics and statistical analysis in the LLFS dataset. This table compares both numerical and categorical features between the Normal Cognitive (NC) and dementia groups within the LLFS cohort. For Age and Completion time, this table shows mean values for the NC, dementia, and overall groups. Kolmogorov-Smirnov tests were conducted to assess the significance of differences in these features between NC and dementia groups. For Gender, Education, and Apolipoprotein E (ApoE), column percentages are presented for each group. Chi-square (χ^2) tests were conducted to determine the significance of differences in these features between NC and dementia groups. A significance level of $\alpha = 0.05$ was used, with a p-value less than 0.05 indicating a significant difference in feature distribution by cognitive status.

Feature	**NC**	**Dementia**	**Overall**	***p*-value**
	N = 1548 (95.97%)	N = 65 (4.03%)	N = 1613 (100.00%)	
Age	69.67 ± 8.68	94.97 ± 7.38	71.74 ± 10.74	< 0.01
Gender				0.64
Female	841 (54.33%)	35 (53.85%)	939 (54.00%)	
Male	707 (45.67%)	30 (46.15%)	800 (46.00%)	
Education				0.02
High school or lower	479 (30.94%)	29 (44.62%)	556 (31.97%)	
College or above	1069 (69.06%)	36 (55.38%)	1183 (68.03%)	
Completion time	38.68 ± 17.57	135.05 ± 91.24	44.36 ± 33.05	< 0.01
ApoE				0.21
ε2/ε2	9 (0.58%)	0 (0.00%)	9 (0.52%)	
ε2/ε3	214 (13.82%)	5 (7.69%)	241 (13.86%)	
ε2/ε4	30 (1.94%)	1 (1.54%)	37 (2.13%)	

ε3/ε3	999 (64.53%)	50 (76.92%)	1129 (64.92%)
ε3/ε4	274 (17.7%)	9 (13.85%)	299 (17.19%)
ε4/ε4	22 (1.42%)	0 (0.00%)	24 (1.38%)

265

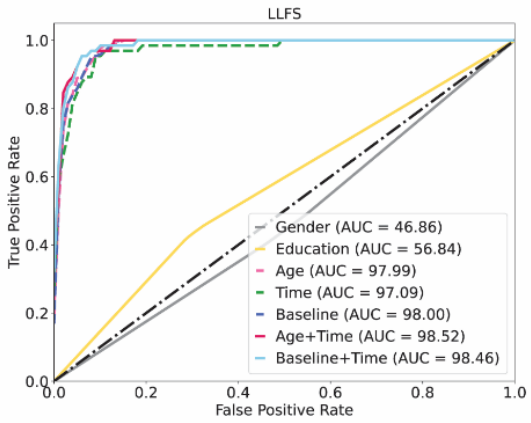
Non-image models

Table 3 presents the mean performance metrics, across five runs, of the non-image models specifically designed for distinguishing participants with dementia from those with normal cognition. Both gender and education, when evaluated individually, showed poor predictive power. Their AUC values were just around 50%, indicating that their performance was nearly equivalent to random guessing. In contrast, age and completion time were strong individual predictors. The higher AUC scores for age and completion time suggest that these features considerably contributed to the model's ability to discern individuals with dementia. When we integrated the completion time with demographics, we observed increases in both AUC and sensitivity across all datasets. Since completion time is the time taken to complete the TMT-A, this finding suggests that effective screening for potential dementia requires the inclusion of some form of cognitive testing. The risk of developing dementia cannot be inferred from demographics alone.

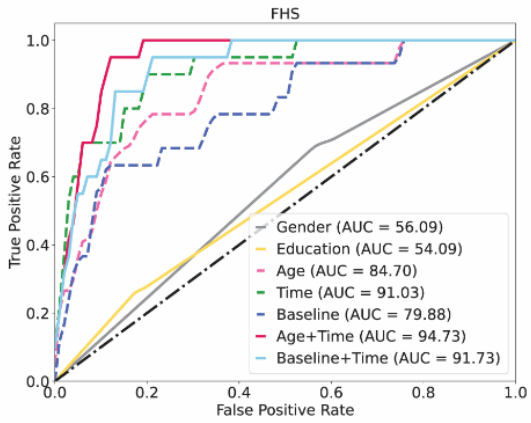
The best-performing non-image models are the models that used both age and completion time as features. Specifically, compared to the metrics of the baseline model, this combination enhanced the AUC by 14.96% and sensitivity by 18.33% in the FHS dataset. For the LLFS dataset, this combination achieved an AUC of 98.50%, a sensitivity of 81.54%, and a specificity of 98.64%. Like the results for the individual datasets, the top-performing non-image model for the combined dataset utilized both age and completion time as predictors. As shown in Figure 4, mixing the two datasets led to decreases in AUC values compared to using the LLFS dataset alone, but resulted in increases in AUC values for the FHS dataset.

Table 3: Performance metrics (in %) of non-image models using demographics and completion time. The models were trained and evaluated on the LLFS, FHS, and combined datasets, respectively, to differentiate participants with dementia from those with normal cognition. This table summarizes the mean values ± standard deviations in percentages for a 5-fold cross-validation process. Evaluation metrics included AUC, sensitivity, specificity, F1 score, and accuracy.

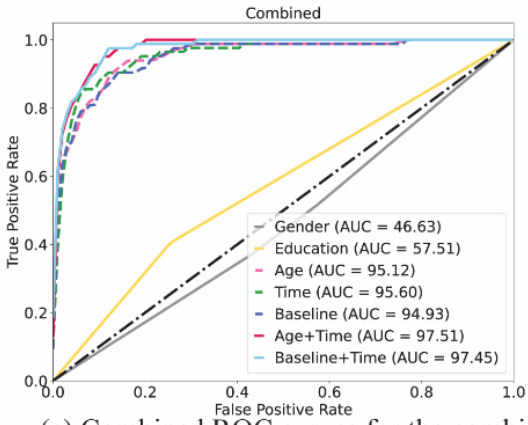
Dataset	Feature	AUC	Sensitivity	Specificity	F1 score	Accuracy
	Gendser	46.86 ± 2.07	43.07 ± 4.21	50.65 ± 4.86	63.68 ± 4.10	50.34 ± 4.56
	Education	56.84 ± 6.37	44.61 ± 11.41	69.06 ± 1.75	77.74 ± 1.43	68.07 ± 2.04
	Age	97.98 ± 0.78	78.46 ± 10.03	97.81 ± 0.66	97.21 ± 0.59	97.02 ± 0.67
LLFS	Time	97.09 ± 2.01	73.85 ± 11.67	98.19 ± 1.49	97.35 ± 0.93	97.21 ± 1.18
	Baseline	97.97 ± 0.69	78.46 ± 13.76	98.06 ± 0.68	97.40 ± 0.23	97.27 ± 0.26
	Age+Time	**98.50 ± 0.59**	**81.54 ± 15.95**	98.64 ± 0.48	98.00 ± 0.42	97.95 ± 0.35
	Baseline+Time	98.46 ± 0.68	81.54 ± 15.95	98.77 ± 0.84	98.11 ± 0.42	98.08 ± 0.46
	Gender	56.09 ± 21.09	70.00 ± 41.08	42.17 ± 1.50	58.27 ± 1.73	42.57 ± 1.98
	Education	54.09 ± 13.07	26.67 ± 25.28	89.22 ± 9.87	92.15 ± 4.99	88.26 ± 9.29
	Age	84.72 ± 10.68	46.67 ± 31.51	92.55 ± 14.64	94.20 ± 8.70	91.87 ± 13.96
FHS	Time	91.00 ± 6.11	46.67 ± 19.19	98.70 ± 1.12	98.08 ± 0.76	97.92 ± 1.03
	Baseline	79.87 ± 13.16	46.67 ± 31.51	95.46 ± 3.60	96.14S ± 2.17	94.73 ± 3.64
	Age+Time	**94.83 ± 2.30**	**65.00 ± 33.54**	96.59 ± 3.99	97.07 ± 2.01	96.08 ± 3.47
	Baseline+Time	91.75 ± 7.49	55.00 ± 27.39	96.18 ± 5.19	96.72 ± 2.99	95.52 ± 5.06
	Gender	46.62 ± 4.73	46.40 ± 13.87	46.85 ± 5.97	61.23 ± 4.95	46.84 ± 5.45
	Education	57.51 ± 2.88	40.44 ± 6.11	74.58 ± 0.49	82.33 ± 0.26	73.58 ± 0.36
	Age	95.19 ± 2.82	57.13 ± 8.71	98.99 ± 0.37	97.71 ± 0.30	97.76 ± 0.31
Combined	Time	95.63 ± 2.65	70.29 ± 9.09	97.91 ± 1.26	97.36 ± 0.74	97.1 ± 1.04



(a) Combined ROC curves for the LLFS dataset.



(b) Combined ROC curves for the FHS dataset.



(c) Combined ROC curves for the combined dataset.

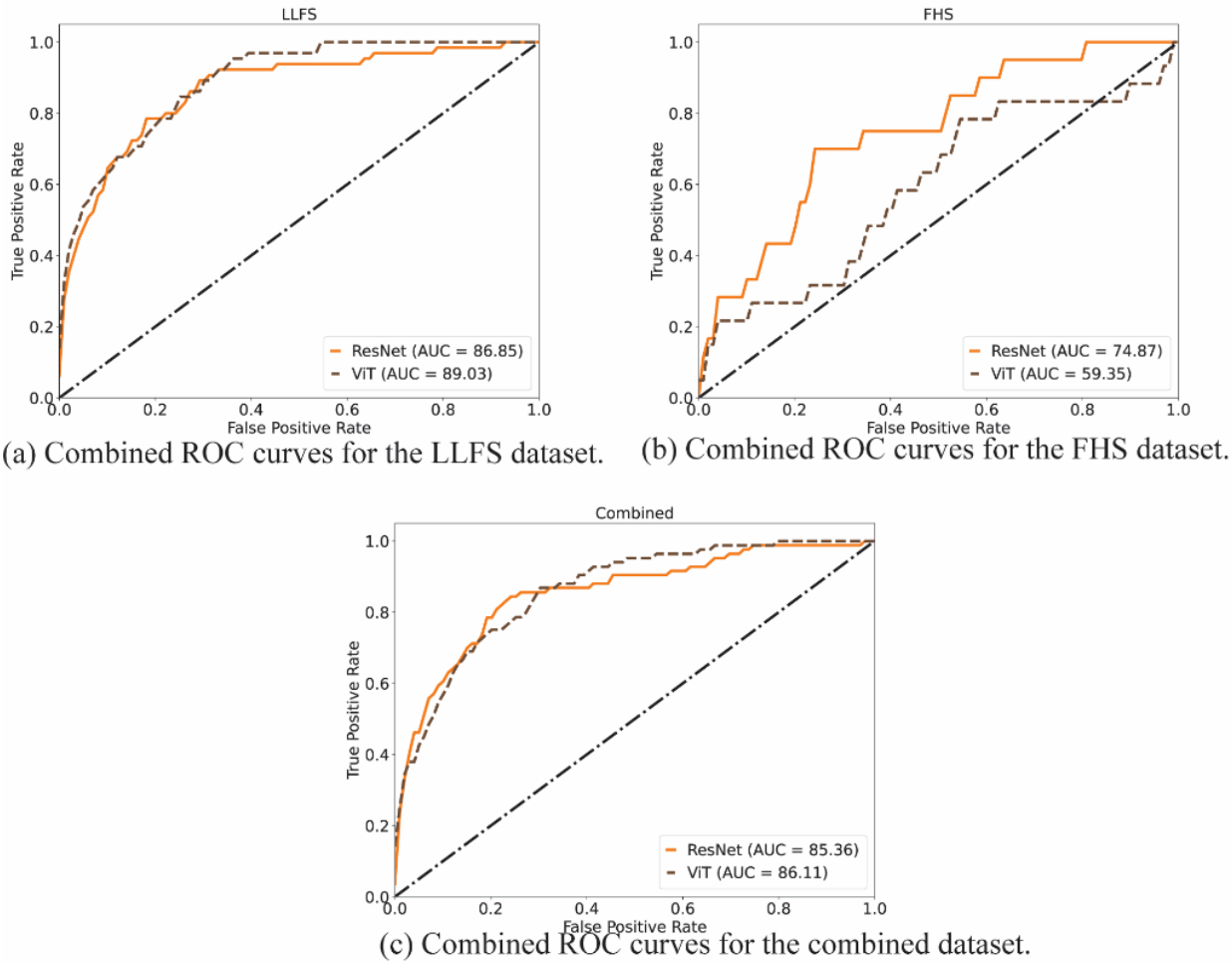
Baseline	94.96 ± 2.72	60.59 ± 12.75	98.60 ± 0.39	97.52 ± 0.37	97.49 ± 0.36
Age+Time	**97.56 ± 1.41**	**71.32 ± 15.51**	98.78 ± 0.62	98.02 ± 0.39	97.97 ± 0.45
Baseline+Time	97.44 ± 1.40	70.15 ± 15.25	98.92 ± 0.36	98.09 ± 0.26	98.08 ± 0.21

Time: completion time; Baseline: age, gender, and education.

Figure 4: Mean ROC curves of non-image models using demographics and completion time. The models were trained and evaluated on the LLFS, FHS, and combined datasets, respectively, to differentiate participants with dementia from those with normal cognition.

Image models

Our comparison of two image models, ViT and ResNet, on the LLFS and FHS datasets revealed performance disparity (cf. Figure 5). In the LLFS dataset, ViT outperformed ResNet, achieving a higher AUC of 89.03% compared to 86.85% from ResNet. In contrast, in the FHS dataset, ResNet demonstrated significantly better performance with an AUC of 74.87% versus 59.35% from ViT. When evaluating the combined dataset of LLFS and FHS, the models achieved more comparable performance. ViT achieved an AUC of 86.11%, while ResNet achieved 85.36%. The variability between the datasets appears to balance out when examining the aggregated dataset. Additionally, the image-only models achieved AUCs between 59.35% to 89.03%, indicating room for improvement in predictive performance by incorporating additional features like demographics and completion time.



(a) Combined ROC curves for the LLFS dataset. (b) Combined ROC curves for the FHS dataset.

(c) Combined ROC curves for the combined dataset.

Figure 5: Mean ROC curves of ViT and ResNet models using the TMT-A drawings. The models were trained and evaluated on the LLFS, FHS, and combined datasets, respectively, to differentiate participants with dementia from those with normal cognition.

Fusion models

The fusion models, detailed in Table 4, incorporated non-imaging features along with probability scores derived from fine-tuned vision models to enhance dementia identification. Across all datasets, the AUC exceeded 96% for each fusion model, demonstrating strong discriminative performance in differentiating between demented and cognitively healthy individuals. The best AUC (98.74%) was achieved with the LLFS dataset using age, completion time, and the probability score predicted by ViT. The highest sensitivity (92.32%) was obtained from the LLFS dataset using age, completion time, and the probability score derived from ResNet. Overall, ResNet outperformed ViT in terms of sensitivity, given that there was no significant difference in AUCs and other metrics. Models utilizing age, completion time, and the image score from ResNet achieved better AUCs and sensitivities compared to models that additionally incorporate gender and education level, alongside age, completion time, and image scores from ResNet, in both the FHS and the combined datasets. Therefore, the optimal feature combination is age, completion time, and the probability score from ResNet. For the LLFS dataset, this combination yielded an AUC of 98.62%, sensitivity of 87.62%, and specificity of 98.26%. Similarly, for the FHS dataset, it achieved an AUC of 96.51%, sensitivity of 85.00%, and specificity of 96.75%.

Figure 6 and Table 5 show the performance improvements by adding completion time alone and both completion time and ResNet-derived image score to the models using only age as a single predictor. Adding completion time alone demonstrated moderate gains while incorporating both temporal data and visual signals led to substantial improvements across all datasets. Notably, the addition of temporal information and visual signals to the demographic data resulted in marked increases in sensitivity across all datasets. Specifically, for the LLFS dataset, we observed a 3.08%

improvement by adding temporal information to age, and a 9.23% improvement by incorporating both temporal and visual signals in terms of sensitivity. For the FHS dataset, the improvements were 18.33% and 38.33%, respectively; for the combined dataset, the increases were 14.19% and 16.55%, respectively. Moreover, AUC also exhibited a pronounced boost with the full fusion model for the FHS dataset. These gains indicated that fusing the visual signal from a participant's drawing and the time taken to complete the TMT-A could considerably enhance predictive performance.

The ablation study in Appendix D compared different model configurations to assess the impact of various features on model performance. The study revealed that while the Age+Image model improved sensitivity compared to the traditional Age+Time model, it slightly compromised other metrics. The proposed Age+Time+Image model consistently outperformed others across most metrics, demonstrating the value of including completion time. Further analysis of error-related features collected by trained examiners during the test showed no significant improvement when added to the Age+Time+Image model. This suggests that the ResNet-derived image probability likely captures this information. In conclusion, the Age+Time+Image model provides the best balance of predictive power and ease of administration, aligning with the goal of developing an accessible dementia screening tool.

Since the risk of developing Alzheimer's disease and related dementias increases with aging, we reapplied our pipeline to a subset of the combined dataset that included only participants aged 65 years and older. Table 6 reports the results for this subsample.

As depicted in Figure 7, age was the dominant factor with a significantly higher coefficient than all other features. After excluding younger participants, age remained the top contributing feature, but a decrease in the coefficient was observed. Although the probability score derived from ResNet

ranks as the least important factor in distinguishing individuals with cognitive dysfunction across most datasets, its coefficients are notably above zero. This suggests that the visual subtleties captured by ResNet offer a complementary signal.

Table 4: Performance metrics of fusion models using combinations of demographics, completion time, and probability score derived from ResNet and ViT. The models were trained and evaluated on the LLFS, FHS, and combined datasets, respectively, to differentiate participants with dementia from those with normal cognition. This table summarized the mean values ± standard deviations in percentages for a 5-fold cross-validation process. Evaluation metrics included AUC, sensitivity, specificity, F1 score, and accuracy.

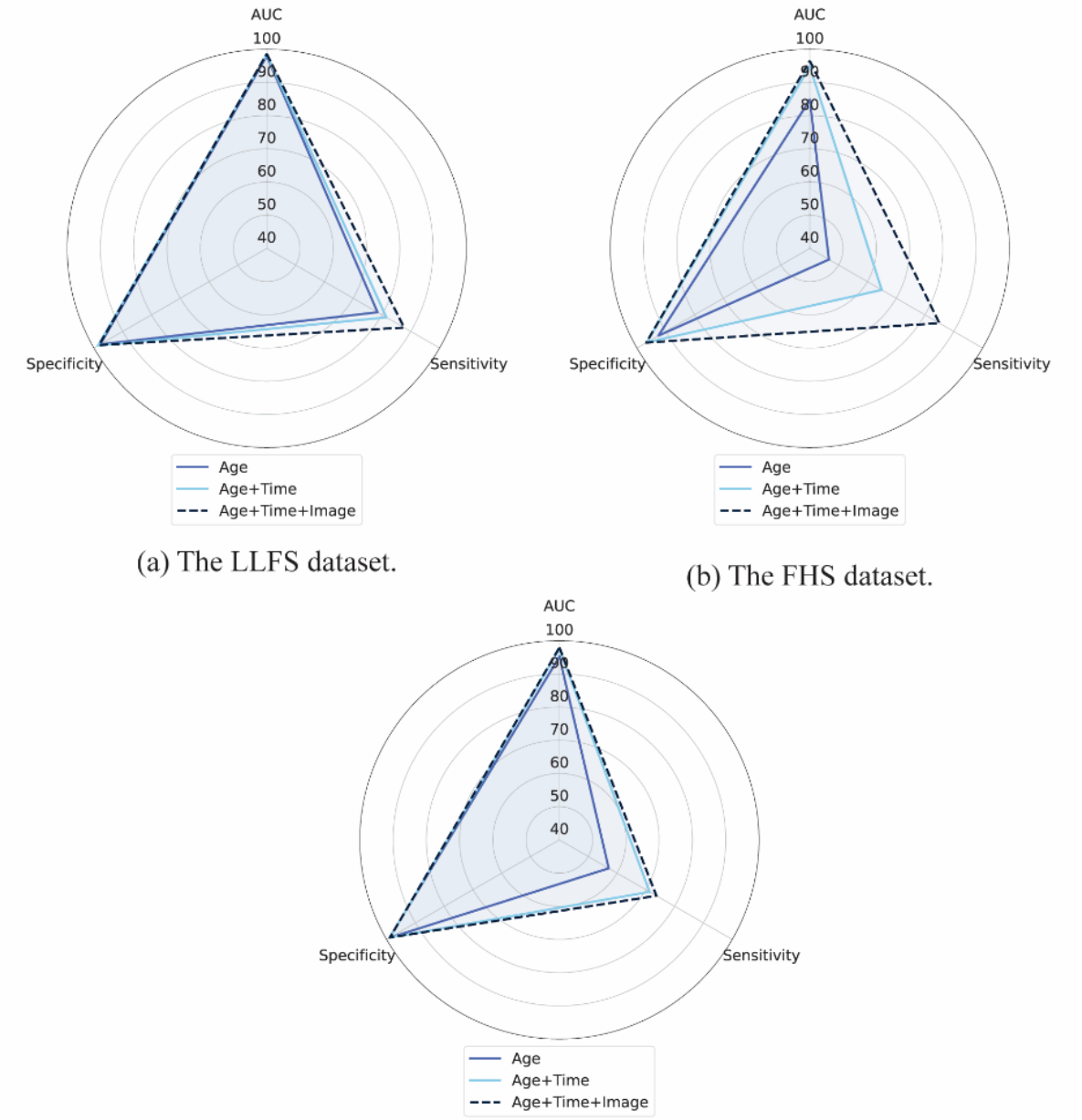
	Dataset	AUC	Sensitivity	Specificity	F1 score	Accuracy
LLFS	Age+Time+Image (ResNet)	**98.62 ± 0.91**	87.69 ± 8.77	98.26 ± 0.67	97.97 ± 0.58	97.83 ± 0.66
	Baseline+Time+Image (ResNet)	98.59 ± 0.97	**92.31 ± 7.69**	98.25 ± 0.93	98.17 ± 0.57	98.01 ± 0.72
FHS	Age+Time+Image (ResNet)	**96.51 ± 2.37**	**85.00 ± 22.36**	96.75 ± 3.40	97.47 ± 1.93	96.56 ± 3.12
	Baseline+Time+Image (ResNet)	94.26 ± 6.09	80.00 ± 20.92	96.10 ± 4.11	97.03 ± 2.44	95.84 ± 3.97
Combined	Age+Time+Image (ResNet)	**97.95 ± 1.02**	**73.68 ± 7.50**	98.81 ± 0.52	98.14 ± 0.31	98.08 ± 0.39
	Baseline+Time+Image (ResNet)	97.80 ± 1.26	70.15 ± 9.69	98.88 ± 0.68	98.08 ± 0.37	98.04 ± 0.48
LLFS	Age+Time+Image (ViT)	**98.74 ± 0.75**	80.00 ± 16.85	98.90 ± 0.54	98.14 ± 0.48	98.14 ± 0.38
	Baseline+Time+Image (ViT)	98.66 ± 0.89	**81.54 ± 13.97**	98.97 ± 0.62	98.28 ± 0.66	98.26 ± 0.65
FHS	Age+Time+Image (ViT) FHS	**97.71 ± 1.21**	61.67 ± 19.19	98.62 ± 1.67	98.32 ± 0.95	98.08 ± 1.48
	Baseline+Time+Image (ViT)	97.29 ± 1.95	**71.67 ± 24.01**	97.81 ± 2.50	97.98 ± 1.59	97.44 ± 2.44
Combined	Age+Time+Image (ViT)	**98.15 ± 0.68**	66.55 ± 19.64	99.10 ± 0.72	98.11 ± 0.25	98.15 ± 0.26
	Baseline+Time+Image (ViT)	98.10 ± 0.84	**70.15 ± 16.86**	98.63 ± 0.86	97.87 ± 0.33	97.80 ± 0.47

Time: completion time; Baseline: age, gender, and education; Image (ResNet): probability score derived from ResNet; Image (ViT): probability score derived from ViT.

Table 5: Performance metrics for models using age only, age with completion time, and age with both completion time and the probability score generated by ResNet. This table compares the performance of models that use a single demographic feature without any cognitive assessment, models that incorporate the demographic feature and completion time (commonly used in traditional cognitive assessments), and models that use demographic, temporal information, and TMT-A drawing from digital cognitive assessments.

Dataset	Feature	AUC	Sensitivity	Specificity	F1 score	Accuracy
LLFS	Age	97.98 ± 0.78	78.46 ± 10.03	97.81 ± 0.66	97.21 ± 0.59	97.02 ± 0.67
	Age+Time	98.50 ± 0.59	81.54 ± 15.95	98.64 ± 0.48	98.00 ± 0.42	97.95 ± 0.35
	Age+Time+Image (ResNet)	**98.62 ± 0.91**	**87.69 ± 8.77**	98.26 ± 0.67	97.97 ± 0.58	97.83 ± 0.66
FHS	Age	84.72 ± 10.68	46.67 ± 31.51	92.55 ± 14.64	94.20 ± 8.70	91.87 ± 13.96
	Age+Time	94.83 ± 2.30	65.00 ± 33.54	96.59 ± 3.99	97.07 ± 2.01	96.08 ± 3.47
	Age+Time+Image (ResNet)	**96.51 ± 2.37**	**85.00 ± 22.36**	96.75 ± 3.40	97.47 ± 1.93	96.56 ± 3.12
Combined	Age	95.19 ± 2.82	57.13 ± 8.71	98.99 ± 0.37	97.71 ± 0.30	97.76 ± 0.31
	Age+Time	97.56 ± 1.41	71.32 ± 15.51	98.78 ± 0.62	98.02 ± 0.39	97.97 ± 0.45
	Age+Time+Image (ResNet)	**97.95 ± 1.02**	**73.68 ± 7.50**	98.81 ± 0.52	98.14 ± 0.31	98.08 ± 0.39

Time: completion time; Image (ResNet): probability score derived from ResNet.



(a) The LLFS dataset.

(b) The FHS dataset.

(c) The combined dataset.

372 **Figure 6**: Comparison of key performance metrics: from age to integrating temporal and visual information. This
373 figure presents radar plots for the LLFS, FHS, and combined datasets, respectively. Each plot highlights the
374 improvements in three key metrics — AUC, sensitivity, and specificity — when augmenting the baseline model first
375 with the completion time alone, and then with both completion time and ResNet-derived image score. Other metrics,
376 not showing significant improvements, were excluded from this visualization.

Table 6: Performance metrics of models trained and evaluated on the subset of older adults aged 65 and over from both cohorts. This table summarizes mean values ± standard deviations in percentages for a five-fold cross-validation process.

Feature	AUC	Sensitivity	Specificity	F1 score	Accuracy
Gender	47.42 ± 4.34	45.95 ± 11.78	48.89 ± 6.44	62.14 ± 5.36	48.76 ± 5.87
Education	56.28 ± 4.44	40.96 ± 9.80	71.61 ± 2.05	79.34 ± 1.12	70.37 ± 1.75
Age	94.28 ± 2.95	56.62 ± 9.17	98.38 ± 0.92	96.68 ± 1.04	96.7 ± 1.16
Time	94.38 ± 1.61	65.00 ± 11.12	97.67 ± 0.72	96.51 ± 0.62	96.36 ± 0.70
Baseline	94.15 ± 2.69	60.15 ± 12.87	98.28 ± 0.77	96.76 ± 0.97	96.74 ± 1.03
Age+Time	96.57 ± 0.97	71.18 ± 7.40	98.38 ± 1.05	97.35 ± 0.50	97.28 ± 0.72
Baseline+Time	96.41 ± 1.31	72.28 ± 3.28	98.43 ± 0.61	97.44 ± 0.43	97.38 ± 0.53
Age+Time+Image (ResNet)	97.04 ± 1.49	69.85 ± 7.28	98.63 ± 0.63	97.49 ± 0.31	97.47 ± 0.41
Baseline+Time+Image (ResNet)	97.09 ± 1.90	68.60 ± 8.01	98.63 ± 0.73	97.44 ± 0.32	97.43 ± 0.44

Time: completion time; Baseline: age, gender, and education; Image (ResNet): probability score derived from ResNet.

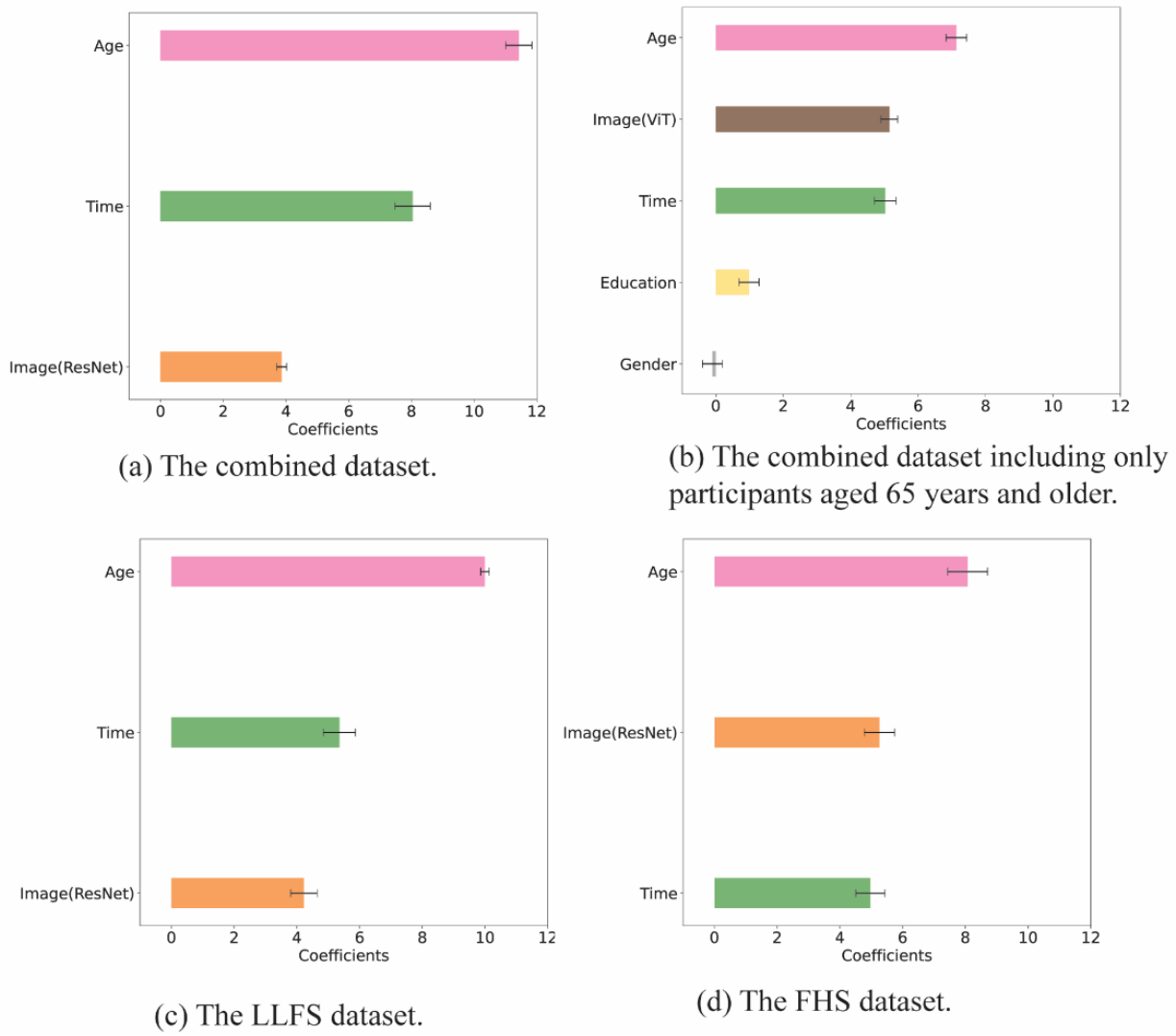


Figure 7: Feature importance for the best-performing fusion models. This figure shows the mean coefficients of the fusion models using the combination of age, completion time, and probability score derived from ResNet. These coefficients were obtained from the fusion models trained on the combined dataset, the subset of older adults aged 65 and over, the LLFS dataset, and the FHS dataset, respectively

.

Discussion

This work demonstrated the ability of multimodal AI models to distinguish individuals with dementia from cognitively healthy controls. We found that integrating visual, temporal, and demographic data significantly enhances performance, surpassing single-modality models. That is, models using age alone are less predictive. Incorporating completion time with age adds value, showing substantially improved predictive performance. Furthermore, the inclusion of the visual predictor, TMT-A image, enhances discriminative capability beyond age and completion time. As shown in Table 5, incorporating the predictor of completion time, used in traditional cognitive assessments, along with age, results in sensitivity improvements of 3.08% and 17.33% for the LLFS and FHS datasets, respectively, compared to age alone. Moreover, the inclusion of the visual predictor, TMT-A image, significantly increases sensitivity by 6.15% and 20% when integrated with the traditional cognitive predictors, age, and completion time, for the LLFS and FHS datasets, respectively. In addition, we showed that the discriminative capability of these fusion models remains consistent across various cohorts. Our findings highlight the potential of utilizing geographically diverse data sources to automate large-scale dementia screening. Importantly, our evaluation of the clinical significance of these improvements reveals meaningful real-world impact (see Appendix E). Using established metrics such as clinical significance ratio and Number Needed to Diagnose (NND), we demonstrate that the fusion model captures a substantial proportion of previously missed cases, with particularly strong performance for the FHS dataset where an NND of only 3.33 indicates a good diagnostic efficiency in clinical practice.

One strength of our framework is its accessibility. While neuroimaging and biofluid analyses are sensitive and accurate for diagnosing Alzheimer's disease, their limited accessibility and reliance

on specialized personnel make them less practical, particularly in resource-limited regions.[38,39] In contrast, our proposed framework addresses these challenges through deliberate design choices. We opted not to use ApoE. Individuals in resource-limited regions face challenges in obtaining genetic information, such as limited access to genetic testing facilities, the costs associated with testing, and the availability of trained professionals to administer a blood test. Similarly, we made a strategic decision not to use raw digital pen recordings directly. Instead, we extracted the overall completion time and converted the raw data into images, then fine-tuned deep learning models to learn spatial information from TMT-A drawings. This approach represents a deliberate prioritization of accessibility over the capture of fine-grained data. Digital pen technology, while capable of recording nuanced information about drawing dynamics, timing, and pressure, remains inaccessible in many settings due to cost, technological barriers, and unfamiliarity among older adults – our primary target population. Our best-performing model utilizes age, TMT-A completion time, and TMT-A drawings – all readily obtainable variables that eliminate the need for specialized equipment. This approach enables diverse implementation pathways: individuals can complete the test using standard pen and paper, time it with a basic stopwatch, and then digitize the drawing through various means. For example, drawings can be collected through traditional paper-and-pencil methods in clinical settings, completed at home and photographed using a smartphone camera, or captured via telehealth platforms during remote assessments. This versatility extends the potential reach of automated cognitive assessment by supporting both prospective implementation across diverse clinical environments and retrospective analysis of existing datasets. This flexibility addresses a critical gap in cognitive assessment infrastructure globally, particularly in settings where specialized neuropsychological expertise is limited. The performance of our image-only models suggests that essential diagnostic information is preserved

in the spatial patterns of TMT-A drawings. This finding is consistent with a growing body of evidence supporting image-based approaches for cognitive assessments. In the clock drawing tests, several works have demonstrated that deep learning models can effectively screen for dementia and identify cognitive decline using the visual features of drawings.[15,40,41] This pattern extends to other neuropsychological instruments, including the Rey-Osterrieth Complex Figure Test, where image-based deep learning approaches have shown comparable efficacy in predicting cognitive impairment.[13] These findings across multiple drawing-based assessments suggest that spatial patterns and visual features captured in static images contain sufficient information for effective cognitive evaluation.

Another advantage of our framework is its scalability. We selected TMT-A over the more challenging TMT-B to facilitate large-scale screening. TMT-A, which simply involves connecting Arabic numerals, is easier and more universally recognized than TMT-B, which requires alternating between numerals and alphabetic letters. This difference is particularly relevant for older adults, non-English speaking populations in regions with lower education levels, where the complexity of TMT-B may lead to a higher failure rate due to misunderstood instructions or the difficulty of the task, as demonstrated in previous studies.[42,43] The simplicity and broader global recognition of TMT-A make it a more practical choice for implementing our framework widely.

In addition, our framework can adapt to other cohorts, as we use multi-center, multi-cohort data from geographically diverse populations across various states and nations. Our framework was applied to and validated on both individual and combined cohorts, demonstrating its generalizability across different cohorts. This versatility indicates the potential for global adaptation of our method for large-scale cognitive impairment screening, particularly in settings where dementia expertise is limited or unavailable.

Our framework demonstrates broad applicability to general populations, as it was developed using community-based cohorts comprising participants with diverse cognitive profiles. Unlike dementia-specific cohorts, which primarily include individuals with diagnosed dementia or elevated risk factors, our community-based approach captures a more representative spectrum of cognitive function. Consequently, our model provides a more ecologically valid approach for identifying dementia cases within heterogeneous populations. The performance characteristics observed in our models, particularly the trade-off between sensitivity and specificity, should be interpreted within this community-based research context. While specialized clinical settings such as Alzheimer's Disease Research Centers (ADRCs) or memory clinics offer comprehensive assessments with higher diagnostic accuracy and potentially more balanced case distributions, they introduce selection bias. Our approach reflects the natural prevalence of cognitive impairment in general populations, where most individuals are cognitively normal. This better represents the intended application environment for accessible screening tools and minimizes potential distributional shift during real-world implementation. A better dementia care pathway would likely combine approaches like ours with comprehensive assessment, using community-based digital screening for initial identification and referral of high-risk individuals to specialized centers, allowing for more efficient allocation of resources while maximizing population coverage.

The performance metrics of models trained on the FHS dataset demonstrate greater variability compared to those trained on the LLFS dataset, as indicated by higher standard deviations across evaluation metrics. This performance discrepancy can be attributed to fundamental differences in cohort characteristics and study designs (detailed in Appendix F). Age distribution represents a primary factor influencing model performance, with FHS participants being younger on average (70.25±12.42 years) compared to those in LLFS (71.74±10.74 years). This younger demographic

profile corresponds with a substantially lower prevalence of dementia in FHS (1.52% vs. 4.03% in LLFS), resulting in fewer dementia cases available for model training (19 cases vs. 65 cases), and thereby creating a significant class imbalance challenge for deep learning algorithms. Variations in performance may also stem from differences in cohort selection criteria. While the FHS primarily examines cardiovascular health in a community-based population, the LLFS cohort consists of individuals selectively recruited from families with exceptional longevity, emphasizing factors associated with healthy aging.[44] These divergent selection criteria result in populations with distinct cognitive aging profiles. Data collection settings also differ meaningfully between the two studies. LLFS conducted in-home assessments, potentially including participants who may be unable to travel to clinical settings due to health or mobility limitations. This approach may capture a broader spectrum of cognitive impairment compared to FHS, where all assessments required clinic visits. The cumulative effect of these differences manifests in the increased variability observed in FHS model performance, reflecting the inherent challenges of detecting subtle cognitive changes in a relatively younger population with fewer overt cases of dementia.

Although this study offers valuable insights, it also has several limitations. Our study is limited by the small sample size and highly imbalanced dataset. The area under the precision-recall curve (AUPRC) revealed low values, likely due to the scarcity of dementia cases. While our models achieved decent sensitivity, this came at the cost of precision, which lowered the AUPRC. These limitations highlight the need for larger datasets with more dementia cases. However, addressing this challenge is inherently difficult as our datasets are derived from community-based studies where the prevalence of dementia is naturally low.

We compromised on some clinically valuable information to enhance the feasibility of implementation in different settings by choosing not to use the time-series data collected by digital

pens, despite its potential to provide additional clinically meaningful insights. Features such as stroke-by-stroke timing, pauses between strokes, and writing pressure variations contain information about cognitive processing speed, motor control, and executive function that is not fully captured in static images.[45–47] As digital pen technology becomes more widely available and affordable, future research could compare the performance of models trained on TMT images versus those utilizing digital pen data. Such studies would help quantify the practical impact of this tradeoff between data richness and accessibility. Additionally, future work could explore hybrid approaches that maintain accessibility while incorporating selective temporal features that can be captured without specialized equipment, such as developing smartphone applications that capture basic temporal dynamics during test administration.

A key limitation of our study was the exclusion of TMT-B data. While TMT-A primarily measures processing speed through a simple number-sequence task, TMT-B offers a more comprehensive evaluation of executive functions through its complex alternation between numbers and letters, making it particularly sensitive for dementia detection. TMT-A's simpler design means its performance might be more influenced by motor impairments related to factors like frailty, rather than cognitive decline alone. Due to limited TMT-B data availability in our dataset for comprehensive experimentation, we focused our analysis on TMT-A. Future studies should incorporate both tests to develop more comprehensive screening models that can better differentiate between motor and cognitive impairments.

Another limitation is that both FHS and LLFS do not conduct a comprehensive review to determine the cognitive status of all participants. Instead, the FHS focuses on those at higher risk, prioritizing dementia reviews for them. Participants who are at lower risk, show no signs of cognitive concerns, or are relatively young, are presumed to have normal cognition without undergoing a detailed

consensus review. It is important to note that the FHS confirms cognitive status, including normal cognition, at the time of death. Similarly, in the LLFS cohort, participants were selected for review if they had a clinical dementia rating score greater than 0 or if they were labeled as having cognitive impairment consistent with dementia by a diagnostic algorithm that considered sex and specific cognitive scores.[48] Consequently, both datasets include both confirmed cases of normal cognition, determined through a dementia review, and presumed cases of normal cognition.

While our study focuses on distinguishing normal cognition from dementia, we recognize the clinical importance of detecting earlier stages of cognitive decline. Our preliminary investigations explored multiple classification scenarios across the cognitive spectrum, including MCI detection (see Appendix G for comprehensive analysis). These preliminary analyses revealed that our fusion model performed significantly better when distinguishing between the more distinct cognitive states of normal cognition and dementia, compared to classifying the more subtle differences between normal cognition and MCI. This guided our decision to focus the current study on normal vs. dementia classification, establishing a strong methodological foundation. Future research should build upon this foundation to develop specialized approaches for the more challenging task of early MCI detection, potentially incorporating longitudinal data and additional neuropsychological measures to capture subtle cognitive changes.

Future research directions could explore several key areas to enhance model performance. More extensive fine-tuning of both ResNet and ViT models, allowing additional layers to be updated during training, may improve the models' ability to learn from limited datasets. For ViT specifically, investigating the impact of different patch sizes, particularly smaller patches, might capture more fine-grained information. Given the sparse nature of TMT drawings, which consist primarily of line segments, future work should also explore architectures specifically designed for

sketch analysis, such as graph neural networks. These specialized architectures might better capture the sequential and structural aspects of TMT drawings.

Although our data collection spans geographically diverse locations, there remains a need to include more demographically and ethnically diverse populations to enhance the generalizability of our framework. We have already begun addressing this limitation by expanding beyond the initial FHS cohort through integration with the LLFS dataset, which introduced greater demographic variability. We also explored the National Alzheimer's Coordinating Center's Uniform Data Set (NACC UDS) for its potential to enhance ethnic diversity through multi-center cohorts, but encountered methodological challenges as UDS contains TMT completion times without corresponding drawings. Looking ahead, we are implementing several practical validation strategies. First, we have developed a web-based platform (developed in our previous CDT work) that targets global participation across varied demographic and ethnic backgrounds, allowing us to gather TMT drawings from diverse populations worldwide.[15] As this dataset reaches sufficient sample size, we will implement external validation and periodic model retraining to improve generalizability. Additionally, we plan to establish collaborative partnerships with international research centers to validate our models, particularly focusing on underrepresented populations and resource-limited environments. Finally, integrating factors such as race, ethnicity, and other sociodemographic characteristics into our models will further enhance their predictive power across diverse populations.

Acknowledgements

None

Ethical Considerations

All participants provided written informed consent. Study protocols and consent forms were approved by the Boston University Medical Campus Institutional Review Board and the Institutional Review Boards of the Long Life Family Study field sites as well as the Long Life Family Study coordinating center at Washington University St. Louis.

Consent to Participate

Written informed consent was obtained from all participants.

Consent for Publication

All authors reviewed and approved the final manuscript for publication.

Funding

This research was partially supported by the NSF under grants CCF-2200052, DMS-1664644, ECCS-2317079, and IIS-1914792; by the ONR under grant N00014-19-1-2571; by the DOE under grant DE-AC02-05CH11231; by the NIH under grant UL54 TR004130; by Boston University; by the Karen Toffler Charitable Trust; by National Institute on Aging's Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research program under grant P30-AG073104; by the American Heart Association under grant 20SFRN35460031; by Gates Ventures and National Institutes of Health under grants RF1-AG062109, U19-AG068753; by the Framingham Heart

584 Study's National Heart, Lung, and Blood Institute contract under grants N01-HC-25195,
585 HHSN268201500001I, and by the NIH National Institute on Aging under grants AG008122,
586 AG016495, AG062109, and AG068753. The LLFS was supported by National Institute on Aging
587 under grants U01AG023746, U01AG023712, U01AG023749, U01AG023755, U01AG023744,
588 and U19 AG063893; SLA was supported by National Institute on Aging under grant K01AG057798.

589 Declaration of Conflicting Interests

590 The authors declare no conflicts of interest.

591 Data Availability Statement

592 The data for the study is available after a reasonable request to the authors and approval from the
593 FHS and LLFS.

594 Author Contributions

595 Jingmei Yang (Formal analysis; Software; Investigation; Visualization; Writing - Original draft;
596 Writing - Review & Editing); Samad Amini (Formal analysis; Investigation; Writing - Review &
597 Editing); Boran Hao (Formal analysis; Investigation; Writing - Review & Editing); Seho Park
598 (Data curation; Writing - Review & Editing); Cody Karjadi (Data curation; Writing - Review &
599 Editing); Lance San Souci (Data curation; Writing - Review & Editing); Vijaya B. Kolachalama
600 (Conceptualization; Formal analysis; Writing - Review & Editing); Stephanie Cosentino
601 (Conceptualization; Formal analysis; Writing - Review & Editing); Stacy Andersen
602 (Conceptualization; Data curation; Formal analysis; Investigation; Writing - Review & Editing);
603 Rhoda Au (Conceptualization; Data curation; Formal analysis; Investigation; Methodology;

Writing - Review & Editing; Project administration; Supervision; Funding acquisition); Ioannis Ch. Paschalidis (Conceptualization; Formal analysis; Investigation; Methodology; Writing - Review & Editing; Project administration; Supervision; Funding acquisition).

References

1. International AD. Dementia Statistics, https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ (2023).

2. International AD. World Alzheimer Report 2022, https://www.alzint.org/resource/world-alzheimer-report-2022/ (2022).

3. Shin J-H. Dementia epidemiology fact sheet 2022. *Ann Rehabil Med* 2022; 46: 53.

4. Organization WH. Ageing and health, https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (2022).

5. Amini S, Hao B, Zhang L, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. *Alzheimers Dement* 2023; 19: 946–955.

6. Amini S, Hao B, Yang J, et al. Prediction of Alzheimer's disease progression within 6 years using speech: a novel approach leveraging language models. *Alzheimers Dement J Alzheimers Assoc* 2024; 20: 5262–5270.

7. Sunderland T, Hill JL, Mellow AM, et al. Clock drawing in Alzheimer's disease: a novel measure of dementia severity. *J Am Geriatr Soc* 1989; 37: 725–729.

8. Rabin LA, Barr WB, Burton LA. Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Arch Clin Neuropsychol* 2005; 20: 33–65.

9. Li Y, Guo J, Yang P. Developing an image-based deep learning framework for automatic

scoring of the pentagon drawing test. *J Alzheimers Dis* 2022; 85: 129–139.

10. Tasaki S, Kim N, Truty T, et al. Explainable deep learning approach for extracting cognitive features from hand-drawn images of intersecting pentagons. *NPJ Digit Med* 2023; 6: 157.

11. Maruta J, Uchida K, Kurozumi H, et al. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Sci Rep* 2022; 12: 9881.

12. Ruengchaijatuporn N, Chatnuntawech I, Teerapittayanon S, et al. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. *Alzheimers Res Ther* 2022; 14: 1–11.

13. Park JY, Seo EH, Yoon H-J, et al. Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline. *Alzheimers Res Ther* 2023; 15: 1–11.

14. Cheah W-T, Hwang J-J, Hong S-Y, et al. A digital screening system for alzheimer disease based on a neuropsychological test and a convolutional neural network: System development and validation. *MIR Med Inf* 2022; 10: e31106.

15. Amini S, Zhang L, Hao B, et al. An artificial intelligence-assisted method for dementia detection using images from the clock drawing test. *J Alzheimers Dis* 2021; 83: 581–589.

16. Handzlik D, Richmond LL, Skiena S, et al. Explainable automated evaluation of the clock drawing task for memory impairment screening. *Alzheimers Dement* 2023; 15: e12441.

17. Tombaugh TN. Trail Making Test A and B: normative data stratified by age and education. *Arch Clin Neuropsychol* 2004; 19: 203–214.

18. Wei M, Shi J, Li T, et al. Diagnostic accuracy of the Chinese version of the trail-making test for screening cognitive impairment. *J Am Geriatr Soc* 2018; 66: 92–99.

19. Inomoto A, Deguchi J, Fukuda R, et al. Gender-specific factors associated with the Japanese version of the trail making test among Japanese workers. *J Phys Ther Sci* 2023; 35: 547–552.

20. Hashimoto R, Meguro K, Lee E, et al. Effect of age and education on the Trail Making Test and determination of normative data for Japanese elderly people: the Tajiri Project. *Psychiatry Clin Neurosci* 2006; 60: 422–428.

21. Specka M, Weimar C, Stang A, et al. Trail Making Test Normative Data for the German Older Population. *Arch Clin Neuropsychol* 2022; 37: 186–198.

22. Suzuki H, Sakuma N, Kobayashi M, et al. Normative Data of the Trail Making Test Among Urban Community-Dwelling Older Adults in Japan. *Front Aging Neurosci* 2022; 14: 832158.

23. Cangoz B, Karakoc E, Selekler K. Trail Making Test: normative data for Turkish elderly population by age, sex and education. *J Neurol Sci* 2009; 283: 73–78.

24. Fellows RP, Dahmen J, Cook D, et al. Multicomponent analysis of a digital Trail Making Test. *Clin Neuropsychol* 2017; 31: 154–167.

25. Du M, Andersen SL, Cosentino S, et al. Digitally generated trail making test data: analysis using hidden Markov modeling. *Alzheimers Dement* 2022; 14: e12292.

26. Dahmen J, Cook D, Fellows R, et al. An analysis of a digital variant of the Trail Making Test using machine learning techniques. *Technol Health Care* 2017; 25: 251–264.

27. Zhang W, Zheng X, Tang Z, et al. Combination of Paper and Electronic Trail Making Tests

for Automatic Analysis of Cognitive Impairment: Development and Validation Study. *Med Internet Res* 2023; 25: e42637.

28. Mahmood SS, Levy D, Vasan RS, et al. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet* 2014; 383: 999–1008.

29. Wojczynski MK, Jiuan Lin S, Sebastiani P, et al. NIA Long Life family study: Objectives, design, and heritability of cross-sectional and longitudinal phenotypes. *J Gerontol Biol Sci Med Sci* 2022; 77: 717–727.

30. Au R, Piers RJ, Devine S. How technology is reshaping cognitive assessment: Lessons from the Framingham Heart Study. *Neuropsychology* 2017; 31: 846.

31. Satizabal CL, Beiser AS, Chouraki V, et al. Incidence of dementia over three decades in the Framingham Heart Study. *N Engl J Med* 2016; 374: 523–532.

32. Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951; 46: 68–78.

33. Greenwood PE, Nikulin MS. *A guide to chi-squared testing*. John Wiley & Sons, 1996.

34. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016, pp. 770–778.

35. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Prepr ArXiv201011929*.

36. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. IEEE, 2009, pp. 248–255.

37. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer, 2009.

38. Olsson B, Lautner R, Andreasson U, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol* 2016; 15: 673–684.

39. Tartaglia MC, Rosen HJ, Miller BL. Neuroimaging in dementia. *Neurotherapeutics* 2011; 8: 82–92.

40. Chen S, Stromer D, Alabdalrahim HA, et al. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Sci Rep* 2020; 10: 20854.

41. Sato K, Niimi Y, Mano T, et al. Automated evaluation of conventional clock-drawing test using deep neural network: Potential as a mass screening tool to detect individuals with cognitive decline. *Front Neurol* 2022; 13: 896403.

42. Smith Watts AK, Ahern DC, Jones JD, et al. Trail-making test part B: evaluation of the efficiency score for assessing floor-level change in veterans. *Arch Clin Neuropsychol* 2019; 34: 243–253.

43. Papandonatos GD, Ott BR, Davis JD, et al. Clinical utility of the Trail-Making test as a predictor of driving performance in older adults. *J Am Geriatr Soc* 2015; 63: 2358–2364.

44. Newman AB, Glynn NW, Taylor CA, et al. Health and function of participants in the Long Life Family Study: a comparison with other cohorts. *Aging* 2011; 3: 63.

45. Prange A, Barz M, Heimann-Steinert A, et al. Explainable automatic evaluation of the trail

making test for dementia screening. In: *Proc SIGCHI Conf Hum Factor Comput Syst*. 2021, pp. 1–9.

46. Prange A, Sonntag D. Modeling users' cognitive performance using digital pen features. *Front Artif Intell* 2022; 5: 787179.

47. Kobayashi M, Yamada Y, Shinkawa K, et al. Automated early detection of Alzheimer's disease by capturing impairments in multiple cognitive domains with multiple drawing tasks. *J Alzheimers Dis* 2022; 88: 1075–1089.

48. Cosentino S, Schupf N, Christensen K, et al. Reduced prevalence of cognitive impairment in families with exceptional longevity. *JAMA Neurol* 2013; 70: 867–874.

Appendix A: Data preprocessing

In the FHS dataset, one participant, who took over 600 seconds, was identified as an outlier and excluded from our analysis due to the completion time from that participant significantly deviating from the distribution, while no outlier was identified in the LLFS dataset. Regarding education data, nine participants from FHS lacked information. We filled in these missing values using the most common education level, i.e., college or above. The LLFS dataset had seven such cases, and the missing values were addressed in a similar manner. In both datasets, all participants reported their gender. As for the ApoE genotype, the FHS dataset had 45 missing entries, while the LLFS dataset had 143. These missing entries were replaced with the most common genotype, i.e., $\varepsilon 3/\varepsilon 3$. In the FHS cohort, all participants reported their ages. In the LLFS cohort, two participants did not provide their ages. For those with missing ages, we imputed them using the mean age. Furthermore, we excluded participants identified with mild cognitive impairment from both the LLFS and FHS datasets. The missing values in our datasets are relatively low. Specifically, for education, the missing values are only 0.72% in the FHS dataset and 0.43% in the LLFS dataset. In terms of age, there are no missing values in the FHS dataset and only 0.12% in the LLFS dataset. Additionally, there are no missing values for gender in both datasets. For completion time, only a single case was excluded from the LLFS dataset. Given the relatively low percentages of missing values for these variables, it is suggested that the missing data and imputation would likely not have a significant impact on model performance.

Appendix B: Model selection

In this study, a wide range of traditional machine learning models were developed, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). These models were built utilizing baseline demographic features, namely age, education, and gender. As shown in Table A1, the LR model was the best-performing model. Therefore, it was adopted as the base model for both non-image models and fusion models.

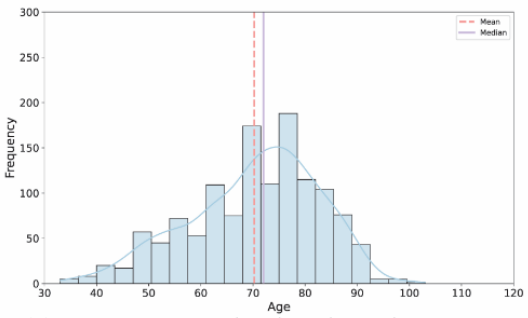
Table A1: Model selection.

Model	AUC (%)	Sensitivity (%)	Specificity (%)	F1 score (%)	Accuracy (%)
DT	78.9 ± 4.35	56.92 ± 8.77	98.04 ± 1.29	96.47 ± 0.61	96.42 ± 0.94
LR	97.71 ± 0.94	75.39 ± 10.03	98.29 ± 1.52	97.53 ± 1.23	97.39 ± 1.47
RF	89.81 ± 2.89	63.08 ± 12.64	97.41 ± 1.63	96.28 ± 0.73	96.05 ± 1.14
SVM	97.45 ± 0.91	75.39 ± 10.03	97.85 ± 1.5	97.17 ± 0.98	96.96 ± 1.27
XGBoost	95.74 ± 1.03	61.54 ± 7.69	98s.11 ± 1.05	96.73 ± 0.66	96.66 ± 0.86

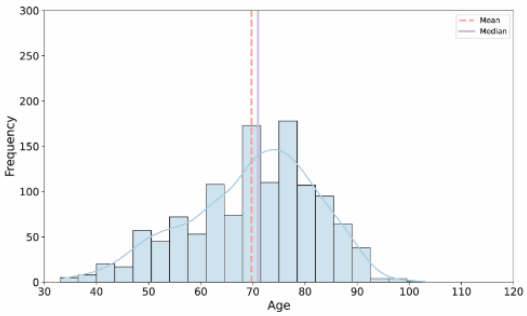
LR: Logistic Regression; SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; XGBoost: Extreme Gradient Boosting.

Appendix C: Feature distribution

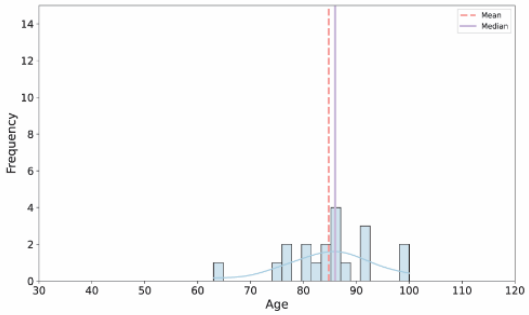
This section presents the feature distributions for the FHS and LLFS datasets. Figure A1 shows the age distribution for the FHS dataset, while Figure A2 displays the age distribution for the LLFS dataset. The completion time distributions for the FHS and LLFS datasets are presented in Figure A3 and Figure A4.



(a) Overall age distribution histogram.

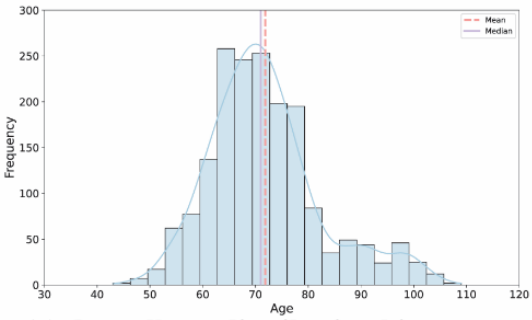


(b) Age distribution for the normal cognition group.

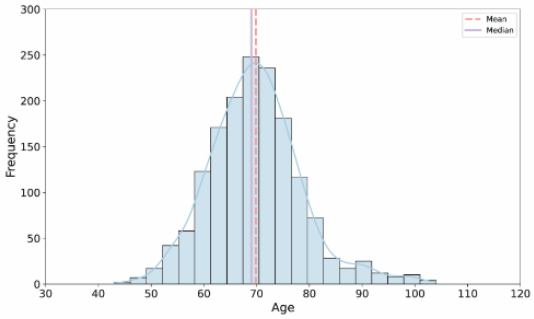


(c) Age distribution for the dementia group.

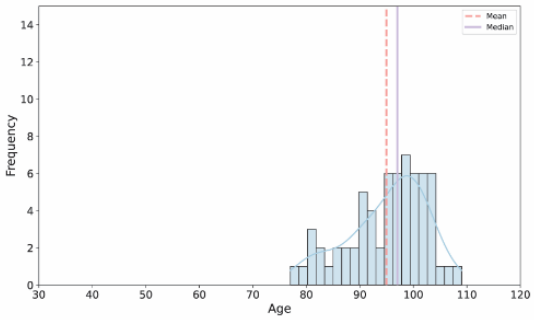
Figure A1: Age distribution for the FHS dataset. (a) Histogram displaying the overall age distribution. (b-c) Histograms showing the age distributions for the normal cognition and dementia groups, respectively.



(a) Overall age distribution histogram.

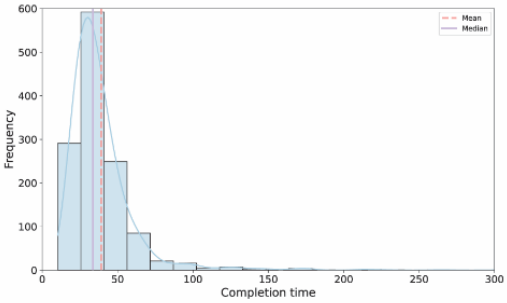


(b) Age distribution for the normal cognition group.

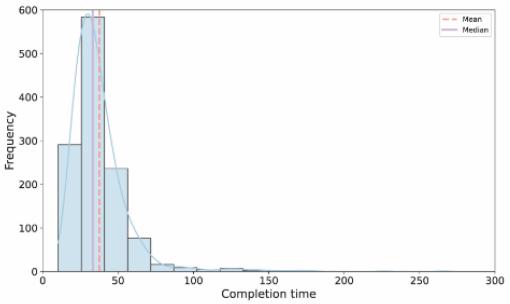


(c) Age distribution for the dementia group.

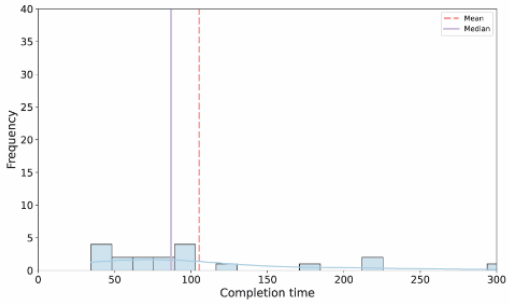
Figure A2: Age distribution for the LLFS dataset. (a) Histogram displaying the overall age distribution. (b-c) Histograms showing the age distributions for the normal cognition and dementia groups, respectively.



(a) Overall completion time distribution histogram.

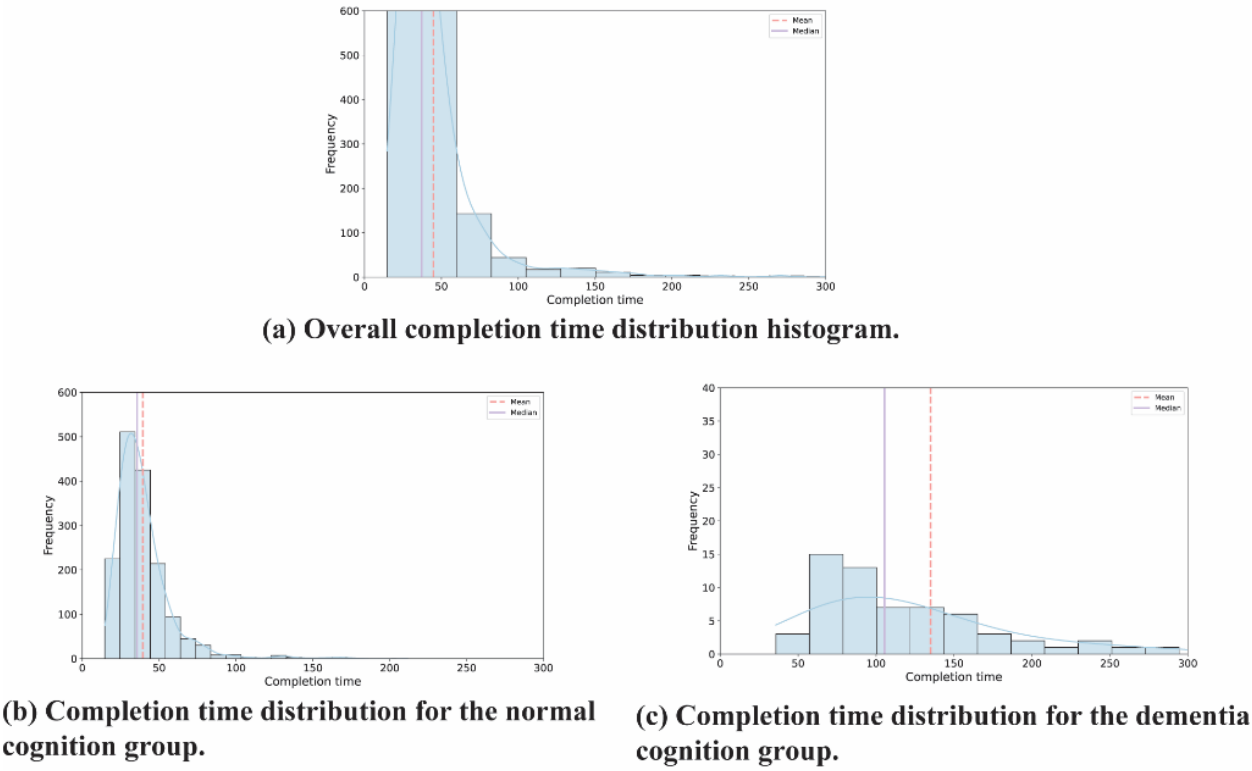


(b) Completion time distribution for the normal cognition group.



(c) Completion time distribution for the dementia cognition group.

Figure A3: Completion time distribution for the FHS dataset. (a) Histogram displaying the overall distribution of completion times. (b-c) Histograms showing the completion time distributions for the normal cognition and dementia groups, respectively. The completion time represents the duration taken by participants to complete the tasks.



(a) Overall completion time distribution histogram.

(b) Completion time distribution for the normal cognition group.

(c) Completion time distribution for the dementia cognition group.

Figure A4: Completion time distribution for the LLFS dataset. (a) Histogram displaying the overall distribution of completion times. (b-c) Histograms showing the completion time distributions for the normal cognition and dementia groups, respectively. The completion time represents the duration taken by participants to complete the task.

Appendix D: Comparative analysis of feature combination

We conducted an experiment comparing three models: Age+Time (traditional), Age+Image (excluding Time), and Age+Time+Image (our proposed model). This experiment aimed to explore the impact of including completion time. As shown in Table A2, the Age+Image model demonstrated substantially better sensitivity compared to the Age+Time model across both LLFS and FHS datasets. However, this improvement came at the cost of slightly lower AUC, specificity, F1 score, and accuracy. Across all datasets, the Age+Time+Image model consistently outperformed the Age+Image model in terms of AUC, specificity, F1 score, and accuracy. Additionally, the Age+Time+Image model achieved higher sensitivity across the FHS and combined datasets. While the Age+Image model offers decent discriminative power and reduces the participant effort for timing, the inclusion of completion time in the Age+Time+Image model offers superior overall performance across most metrics. We propose that the image can be captured as a snapshot and the time can be tracked using a stopwatch. In settings where approximate timing is feasible, we recommend using the Age+Time+Image model. Although this approximation of completion time using a simple stopwatch by the participant may not be as accurate as that tracked by a digital pen or trained examiner, it still contributes valuable information. In extremely resource-limited settings or where timing is impractical, the Age+Image model provides an alternative for initial dementia screenings.

We also examined error-related features documented during the TMT administration. We focused on two key error features: LiftPen, which records the number of times participants lifted their pen from the paper, and Tremor, indicating the presence of obvious tremor. While examiners also noted self-corrected and examiner-corrected perceptual errors, these were excluded from our

analysis due to lack of statistical significance. We compared our proposed model (Age+Time+Image) against models incorporating these error-related features to assess their contribution to predictive performance. As shown in **Table 3,** adding error-related features to our proposed model (Age+Time+Image) did not improve the AUC and sensitivity. We hypothesize this may be because the error information is already captured by ResNet. Furthermore, tracking these features requires documentation by trained professionals, which deviates from our goal of developing an accessible dementia screening tool and limits the applicability of our approach. Our findings indicate that the most effective model, combining age, completion time, and image probability derived from ResNet, captures the most relevant information for discriminating between individuals with and without dementia. This approach balances predictive power and ease of administration.

Table A2: Performance metrics for models using age only, age with completion time, age with image probability score derived from ResNet, and age with both completion time and the probability score generated by ResNet.

Dataset	Feature	AUC	Sensitivity	Specificity	F1 score	Accuracy
	Age	97.98 ± 0.78	78.46 ± 10.03	97.81 ± 0.66	97.21 ± 0.59	97.02 ± 0.67
	Age+Time	98.50 ± 0.59	81.54 ± 15.95	98.64 ± 0.48	98.00 ± 0.42	97.95 ± 0.35
	Age+Image (ResNet)	98.44 ± 0.82	92.31 ± 7.69	97.87 ± 1.09	97.86 ± 0.71	97.64 ± 0.90
LLFS	Age+Time+Image (ResNet)	98.62 ± 0.91	87.69 ± 8.77	98.26 ± 0.67	97.97 ± 0.58	97.83 ± 0.66
	Age	84.72 ± 10.68	46.67 ± 31.51	92.55 ± 14.64	94.20 ± 8.70	91.87 ± 13.96
	Age+Time	94.83 ± 2.30	65.00 ± 33.54	96.59 ± 3.99	97.07 ± 2.01	96.08 ± 3.47
	Age+Image (ResNet)	94.82 ± 3.13	73.33 ± 18.07	95.94 ± 5.02	96.87 ± 2.86	95.60 ± 4.73
FHS	Age+Time+Image (ResNet)	96.51 ± 2.37	85.00 ± 22.36	96.75 ± 3.40	97.47 ± 1.93	96.56 ± 3.12
	Age	95.19 ± 2.82	57.13 ± 8.71	98.99 ± 0.37	97.71 ± 0.30	97.76 ± 0.31
	Age+Time	97.56 ± 1.41	71.32 ± 15.51	98.78 ± 0.62	98.02 ± 0.39	97.97 ± 0.45
	Age+Image (ResNet)	97.26 ± 1.03	67.65 ± 13.15	98.74 ± 0.85	97.89 ± 0.50	97.83 ± 0.64
Combined	Age+Time+Image (ResNet)	97.95 ± 1.02	73.68 ± 7.50	98.81 ± 0.52	98.14 ± 0.31	98.08 ± 0.39

Time: completion time; Image(ResNet): probability score derived from ResNet.

Table A3: Performance metrics for models using age, the completion time, and the probability score generated by ResNet and models using additional error-related features. This table compares the performance of our proposed models with the models that use extra error-related features.

Dataset	Feature	AUC	Sensitivity	Specificity	F1 score	Accuracy
LLFS	Age+Time+Image (ResNet)	98.62 ± 0.91	87.69 ± 8.77	98.26 ± 0.67	97.97 ± 0.58	97.83 ± 0.66
	Age+Time+Image (ResNet)+Error	98.37 ± 1.49	84.62 ± 7.70	98.36 ± 1.48	97.98 ± 1.08	97.81 ± 1.30
FHS	Age+Time+Image (ResNet)	96.51 ± 2.37	85.00 ± 22.36	96.75 ± 3.40	97.47 ± 1.93	96.56 ± 3.12
	Age+Time+Image (ResNet)+Error	95.79 ± 2.52	68.33 ± 20.75	97.24 ± 2.74	97.52 ± 1.53	96.80 ± 2.38
Combined	Age+Time+Image (ResNet)	97.95 ± 1.02	73.68 ± 7.50	98.81 ± 0.52	98.14 ± 0.31	98.08 ± 0.39
	Age+Time+Image (ResNet)+Error	97.76 ± 1.18	71.54 ± 15.60	98.86 ± 0.79	98.12 ± 0.74	98.07 ± 0.79

Time: completion time; Image (ResNet): probability score derived from ResNet. Error: the number of times participants lifted their pen from the paper) and the presence of obvious tremor.

Appendix E: Clinical relevance of sensitivity improvements

We conducted statistical analysis of our models using the Wilcoxon signed-rank test. While the improvements did not reach conventional statistical significance, we argue that in clinical diagnostic contexts, practical significance can be more relevant than statistical analysis—particularly with limited sample sizes. We therefore present four complementary metrics specifically designed to quantify clinical impact:

$$\text{Absolute Improvement} = Fusion - Baseline + Time,$$

$$\text{Relative Improvement} = \frac{Fusion - Baseline + Time}{\text{Baseline} + \text{Time}} \times 100\%,$$

$$\text{Clinical Significance Ratio (CSR)} = \frac{Fusion - Baseline + Time}{100\% - \text{Baseline} + \text{Time}},$$

$$\text{Number Needed to Diagnose (NND)} = \frac{1}{Fusion - Baseline + Time}.$$

The CSR measures the percentage of previously missed cases now correctly identified, while the NND quantifies the screening effort required to benefit one additional patient—metrics widely accepted in clinical research for assessing practical utility.

As shown in Table A4, our analysis reveals substantial clinical relevance across all datasets, with particularly notable results for the FHS cohort. The FHS dataset shows remarkable improvements with an absolute sensitivity increase of 30 percentage points (from 55.00% to 85.00%). This translates to a 54.55% relative improvement, with the fusion model correctly identifying 66.67% of cases that would have been missed by the baseline model (CSR). Most importantly, the NND of only 3.33 indicates that for approximately every 3-4 patients screened, one additional case will

be correctly identified that would otherwise be missed—an efficiency level considered highly significant in clinical practice.

Table A4: Sensitivity performance comparison and clinical impact metrics for Baseline+Time vs. Fusion models.

Dataset	Baseline+Time	Fusion	Absolute Improvement	Relative Improvement	Clinical Significance Ratio	Number Needed to Diagnose
LLFS	81.54%	87.69%	6.15%	7.54%	33.29%	16.26
FHS	55.00%	85.00%	30.00%	54.55%	66.67%	3.33
Combined	70.15%	73.68%	3.53%	5.03%	11.83%	28.33

The Fusion model incorporates age, completion time, and image probability (ResNet).

Appendix F: Comparison of cohort characteristics

The comparison between the FHS and LLFS cohorts reveals notable differences that help explain the performance variations observed in our models (Table A5). Despite similar average ages (70.25 years in FHS vs. 71.74 years in LLFS), the LLFS cohort contains a substantially higher proportion of dementia cases (4.03% compared to just 1.52% in FHS), providing more balanced training examples for our deep learning models. Educational backgrounds also differ considerably, with FHS participants demonstrating higher educational attainment (81.14% with college education or above compared to 68.03% in LLFS). Additionally, the LLFS cohort shows longer average Trail Making Test (TMT) completion times (44.36 seconds vs. 39.14 seconds in FHS), suggesting potential underlying differences in cognitive performance distributions.

The FHS primarily focused on cardiovascular health with clinic-based assessments, while LLFS specifically selected families demonstrating exceptional longevity and conducted in-home assessments. This design difference may have resulted in the inclusion of participants with mobility issues in LLFS who might also exhibit higher rates of cognitive impairment.

These cohort differences, particularly the lower prevalence of dementia cases in FHS (only 19 cases compared to 65 in LLFS), create a more challenging classification task due to fewer positive examples available for training. This disparity helps explain the higher standard deviations and lower performance metrics observed in our FHS models.

Table A5: Overall cohort comparison between FHS and LLFS datasets. This table highlights key demographic differences between FHS and LLFS cohorts that may influence model performance.

Feature	FHS (N = 1, 252)	LLFS (N = 1, 613)
Age (mean ± SD)	70.25 ± 12.42	71.74 ± 10.74
Gender		
Female (%)	57.76%	54.00%
Male (%)	42.24%	46.00%
Education		
High school or lower (%)	18.86%	31.97%
College or above (%)	81.14%	68.03%
Completion time (mean ± SD)	39.14 ± 25.20	44.36 ± 33.05
Dementia cases (%)	1.52%	4.03%

Appendix G: Evaluation of fusion models across the cognitive impairment spectrum

During our initial experimental design phase, we systematically explored model performance across four distinct classification tasks representing different clinical objectives in cognitive assessment (Table A6).

Normal vs. MCI: This task focused on detecting early cognitive decline by differentiating individuals with normal cognition from those with MCI. This represents the earliest stage of detection, aimed at identifying subtle cognitive changes before dementia develops.

Normal/MCI vs. Dementia: This task aimed to identify established dementia regardless of whether individuals were cognitively normal or had intermediate impairment. This classification approach aligns with the clinical need to identify patients requiring dementia-specific interventions and care.

Normal vs. MCI/Dementia: This task focused on detecting any cognitive impairment, grouping together individuals with either MCI or dementia. This classification mirrors screening approaches that aim to identify patients requiring further clinical evaluation.

Normal vs. Dementia: This task specifically differentiated individuals with normal cognition from those with dementia, excluding the intermediate MCI category.

To systematically evaluate these classification approaches, we analyzed performance using our fusion model (Age + Time + Image (ResNet)) on the LLFS dataset, as shown in Table A6. Analysis of these performance metrics reveals several important patterns. The *Normal vs. Dementia* task demonstrated consistently superior performance across all evaluation metrics, with the highest AUC (98.62%), F1 score (97.97%), and accuracy (97.83%). In contrast, tasks involving MCI

158 classification showed relatively lower performance, particularly in terms of sensitivity. The
159 *Normal vs. MCI* task presented the greatest challenge, with substantially lower sensitivity
160 (60.98%) compared to other tasks. This indicates that approximately 39% of MCI cases were
161 misclassified as normal cognition using our fusion model. Similarly, the *Normal vs. MCI/Dementia*
162 task showed lower sensitivity (69.16%), primarily due to the difficulty in correctly identifying MCI
163 cases. The *Normal/MCI vs. Dementia* task performed better than the MCI-specific classifications
164 but still showed lower sensitivity (80.00%) compared to the *Normal vs. Dementia* task (87.69%).
165 This suggests that while our fusion model can effectively identify dementia, the inclusion of MCI
166 cases introduces significant classification challenges.

167 These performance differences likely stem from several factors. MCI represents a heterogeneous
168 intermediate stage with subtle cognitive changes that may not be as consistently reflected in TMT
169 performance as the more pronounced deficits seen in dementia. Additionally, the boundaries
170 between normal aging and MCI are less distinct than those between normal cognition and dementia,
171 leading to greater classification uncertainty. MCI detection typically requires more extensive
172 neuropsychological assessment or longitudinal monitoring in clinical practice, suggesting that
173 single timepoint TMT data may have inherent limitations for identifying this intermediate stage.

Table A6: Performance of Fusion model (Age + Time + Image) across different classification tasks on LLFS dataset.

Task	AUC	Sensitivity	Specificity	F1	Accuracy
Normal vs. MCI	86.38	60.98	94.69	92.59	92.22
Normal/MCI vs. Dementia	97.39	80.00	97.95	97.48	97.29
Normal vs. MCI/Dementia	90.20	69.16	95.20	92.53	92.39
Normal vs. Dementia	98.62	87.69	98.26	97.97	97.83

Appendix H: Oversampling effects

Table A7 presents the results of our ablation study examining the effects of oversampling on our best-performing non-image model (Age+Time). The results demonstrate that oversampling significantly improves model sensitivity for detecting the minority class (dementia). For the Combined dataset, oversampling increases recall from 13.02% to 71.32%, while maintaining high specificity (98.78% vs 99.89%). Similarly, for the FHS dataset, recall improves from 0% to 65%, albeit with increased variability, and for the LLFS dataset, recall enhances from 24.62% to 81.54%. These findings underscore the critical role of oversampling in addressing the inherent class imbalance in our dementia detection task. Without oversampling, models exhibit strong bias toward the majority class (normal cognition). The observed trade-off between improved recall and marginally reduced specificity aligns with our clinical objective of prioritizing dementia case identification. These results validate our methodological choice of employing oversampling techniques to enhance model performance.

Table A7: Performance comparison of the best-performing non-image model with and without oversampling (mean ± std across 5-fold cross validation).

Dataset	Oversampling	AUC	Sensitivity	Specificity	Accuracy
Combined	No	97.35 ± 1.00	13.02 ± 6.35	99.89 ± 0.16	97.38 ± 0.23
	Yes	97.56 ± 1.41	71.32 ± 15.51	98.78 ± 0.62	97.97 ± 0.45
FHS	No	95.36 ± 2.50	0.00 ± 0.00	100.00 ± 0.00	98.48 ± 0.18
	Yes	94.83 ± 2.30	65.00 ± 33.54	96.59 ± 3.99	96.08 ± 3.47
LLFS	No	98.21 ± 0.98	24.62 ± 6.44	99.81 ± 0.28	96.84 ± 0.35
	Yes	98.50 ± 0.59	81.54 ± 15.95	98.64 ± 0.48	97.95 ± 0.35