An AI-assisted Method for Dementia Detection Using Images from the Clock Drawing Test

Samad Amini[a], Lifu Zhang[a], Boran Hao[a], Aman Gupta[a], Mengting Song[a], Cody Karjadi[c], HonghuangLin[b], Vijaya B. Kolachalama[b,d,e], Rhoda Au[f,c], Ioannis Ch. Paschalidis[a,d,g]

[a]*Department of Electrical & Computer Engineering, Division of Systems Engineering, and Department of Biomedical Engineering, Boston University*
[b]*Department of Medicine, Boston University School of Medicine*
[c]*Framingham Heart Study, Boston University*

[d]*Faculty of Computing & Data Sciences, Boston University*

[e]*Department of Computer Science, Boston University*

[f]*Departments of Anatomy & Neurobiology, Neurology, and Epidemiology, Boston University School of Medicine and School of Public Health*
[g]*Corresponding author: Ioannis Ch. Paschalidis, yannisp@bu.edu, 8 St. Mary's St Boston, MA 02215*

Abstract

Background: Widespread dementia detection could increase clinical trial candidates and enable appropriate interventions. Since the Clock Drawing Test (CDT) can be potentially used for diagnosing dementia-related disorders, it can be leveraged to develop a computer-aided screening tool.

Objective: To evaluate if a machine learning model that uses images from the CDT can predict mild cognitive impairment or dementia.

Methods: Images of an analog clock drawn by 3,263 cognitively intact and 160 impaired subjects were collected during in-person dementia evaluations by the Framingham Heart Study. We processed the CDT images, participant's age and education level using a deep learning algorithm to predict dementia status.

Results: When only the CDT images were used, the deep learning model predicted dementia status with anarea under the receiver operating characteristic curve (AUC) of $81.3\% \pm 4.3\%$. A composite logistic regression model using age, level of education, and the predictions from the CDT-only model, yielded an average AUC and average F1 score of $91.9\% \pm 1.1\%$ and $94.6\% \pm 0.4\%$, respectively.

Conclusion: Our modeling framework establishes a proof-of-principle that deep learning can be applied on images derived from the CDT to predict dementia status. When fully validated, this approach can offer a cost-effective and easily deployable mechanism for detecting cognitive impairment.

Keywords: Artificial Intelligence, Clock Test, Dementia, Deep Learning, Alzheimer's Disease.

Conflicts of Interest: Rhoda Au is a scientific advisor to Signant Health, consultant to Biogen, has receivedgrant support from Pfizer, and secured support to collect a subset of the data used in this study between 2017–2019; she states no conflict of interest with the present work. There is no declaration from other authors.

---

1. Introduction

In the United States, (*i*) the cost associated with *Alzheimer's Disease (AD) and Related Dementias (ADRD)* has been estimated to be $305 billion in 2020, expected to rise to as much as $1.1 trillion by 2050, and (*ii*) more than 5 million individuals are living with AD, with AD deaths increasing by 146% between 2000 and 2018 [1]. Worldwide, it is estimated that more than 50 million are living with dementia [2].

The standard approach to evaluating the severity of cognitive decline of a participant includes *Neuro-Psychological (NP)* exams, which have traditionally been conducted via in-person interviews to measure memory, thinking, language, and motor function. However, this approach can be expensive, time-consuming, and limited in availability to subjects with lower income and/or belonging to a racial or other underrepresented minority. With the ongoing COVID-19 pandemic, access to medical facilities for non-life-threatening conditions has been curtailed and medical care has shifted to virtual, online visits. A similar virtual approach is highly desirable for dementia screening. In addition to broader and more equitable access to care, virtual approaches can lead to accurate diagnosis and increase the pool of candidates for ADRD clinical trials, possibly accelerating the search for effective treatments.

We leveraged recent advances in machine learning and nonverbal screening tools such as the *Clock Drawing Test (CDT)* to determine dementia status. In the CDT test, subjects are asked to draw the face of an analog clock showing ten minutes past eleven. CDT is considered robust against cultural biases and language and provides insight into the mechanisms underlying cognitive dysfunction, including comprehension memory, numerical knowledge, and visuoperceptual skills [3,4]. Given the sensitivity of CDT in cognitive screening, numerous attempts have been made to exploit the full potential of CDT in identifying dementia. For instance, the sensitivity of the CDT was investigated in [5] to monitor and distinguish the evolution of cognitive decline in different cognitive domains. In [6], the authors found CDT useful in cognitive impairment screening using the fact that the CDT score correlates with the severity of global cognitive impairment, as assessed by the *Mini-Mental State Examination (MMSE)* score and the Hasegawa dementia scale. A low CDT score was also associated with progression to dementia, with the association being independent of the MMSE score [7,8].

Options such as digital Clock Drawing Test (dCDT) technology, where the drawing is traced by a digital pen, enable the examination of a detailed neurocognitive behavior as it unfolds in real-time; a capability that cannot

be obtained using a traditional pen and paper [9,10]. By using a dCDT, a machine learning approach based on non-interpretable boosted decision trees was able to outperform scoring systems used by clinicians [11], reaching an Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) of 93% using the entire battery of features provided by dCDT. The AUC drops to 83% for simpler, interpretable models. Moreover, a classification task to classify mild cognitive impairment subtypes and AD using 350 dCTD features has achieved accuracy ranging from 83% to 91% [12]. Others have leveraged medical imaging, which is expensive and requires an in-person visit to an imaging facility. Recently, deep learning algorithms have been successfully applied to AD detection, particularly using neuroimaging data [13]. Deep learning was applied to predict progression to AD based on hippocampal *Magnetic Resonance Imaging (MRI)* and other baseline clinical features [14], achieving an AUC of 86%. Also, deep learning frameworks that can process both imaging and non-imaging data revealed high AD classification accuracy across multiple dataset [15]. The authors in [16] have achieved high levels of accuracy (75% to 99%) in AD classification using a single MRI and a deep neural network. Furthermore, a deep learning classifier was adopted in [17] to identify the different stages of mild cognitive impairment based on MRI and *Positron Emission Tomography (PET)* [18,19], with accuracies ranging from 57% to 91%. It is important to note that some of these papers reported external validation results, thus underscoring the model's generalizability. In general, and as we elaborate on later, accuracy may not be the most appropriate metric for assessing performance when AD/ADRD datasets are imbalanced and when only a small fraction of subjects have dementia.

The above studies rely on a large collection of features available through NP tests, dCDT, blood biomarkers (e.g., apolipoprotein genes), or medical imaging, thus requiring expensive resources and in-person visits. These technologies, even the digital pen, make the cost prohibitive for low-resourced health care environments and perpetuate persistent health disparities in global testing. Consequently, in the context of using features from these tests for AI-assisted detection, they also embed inherent biases, further exacerbating the widening gap across global populations in health care knowledge and practice. The proposed approach uses digital images derived from CDT, the age, and education level of the participant, thus utilizing easily collectable information for dementia assessment. To that end, our method processes CDT images through a deep *Convolutional Neural Network (CNN*) [20,21] classifier and combines the output scores with age and education level in an ensemble, logistic regression-based classifier.

2. Materials and Methods

2.1. Clinical setting and data sources

The data have been collected by the Framingham Heart Study (FHS), the longest ongoing longitudinal study of chronic

disease [22]. Since 2011, the FHS has adopted digital pen technology to capture pen and paper NP tests, including the CDT. In the FHS dataset, two different clock images are collected: (i) one where the subjects are told to draw an analog clock showing ten minutes past eleven (command clock), and (ii) one where they are asked to copy the image of such a clock shown to them (copy clock). Additional information available includes self-reported gender, age, race, and education level, and the presence of Apolipoprotein E (ApoE) genes. All subjects were evaluated by trained examiners. For those subjects identified as showing symptoms of cognitive impairment or decline and flagged for diagnostic review, dementia diagnosis was reached by consensus of at least one neurologist and one neuropsychologist; the dementia surveillance, flagging, and diagnostic procedures have been previously outlined in [4,23] The dementia diagnosis referenced in this study occurred either before the CDT or within 180 days after the CDT. The entire dataset for all participants was anonymized prior to analysis. All participants have provided written informed consent and study protocols and consent forms were approved by the Boston University Medical Campus Institutional Review Board.

2.2. Data preparation

The original dataset contains information about 3,423 participants. The dataset attributes consist of participant demographic data, education level, ApoE gene information, command and copy clock drawings, as well as the dementia diagnosis. The clock drawings are stored in '*.csk'* format as they are recorded using the digital pen [24]. Therefore, a pre-processing pipeline was created to convert the '*.csk'* files into the clock images of size (128, 128, 3), which are three-channel images with 128 × 128 pixels. To normalize the data, the value of each pixel was divided by 255 to rescale pixel values into the [0, 1] range. We performed data augmentation on the original images by randomly applying ±10 degrees rotation, ±15 percent zoom, ±10 percent width and height shift, and ±10 percent shear. Data augmentation enables us to develop a deep learning model that is robust against image distortions. By disproportionally augmenting the non-dominant class of images, we also mitigate class imbalance – 95.3% of participants have no cognitive impairment – enabling training of the deep learning model in a balanced fashion without under-sampling the dominant class.

2.3. Statistical analysis

The composition of the dataset along with basic statistics is reported in Table 1. The 2nd column provides information on participants who were labeled as normal and the 3rd column corresponds to participants diagnosed as cognitively impaired (Cognitive Impairment, No Dementia – CIND), or with clinical dementia

(mild, moderate, severe). We included self-reported gender, education level, age statistics (mean ± standard deviation for each cohort), race, dementia diagnosis severity, and the type of ApoE (E2/E3/E4) genes for both copies of the gene. In the 4th column we report the *p*-value for each variable associated with the null hypothesis that the two cohorts have the same distribution of the variable. Hence, a low *p*-value implies that the distribution of the feature is different in each cohort, leading us to reject the null hypothesis. For age, we employed the Kolmogorov-Smirnov (K-S) test [25] whereas we used the Chi-square test for the categorical features [26]. It can be seen that age and education show significant difference among the two cohorts, leading us to use age, education and CDT images in the proposed ensemble model. An additional advantage of using only CDT images, age, and education is that these features are easily obtained remotely without the need to visit a clinic.

2.4. Model development

We formulated the dementia detection system as a classification task in which the model seeks to make a binary decision on the dementia status, i.e., if the diagnosis score is greater than zero, then we adjudicate the person to have cognitive impairment and be normal otherwise. Since there was limited data for subjects with either CIND or dementia, a single class was used to represent both CIND and dementia in our model.

Given that CNN models, which include many hidden layers and millions of parameters, require a large number of images to be trained, we adopted a transfer learning approach starting from a pre-trained CNN on the ImageNet dataset. Transfer learning is widely used in medical image analysis and natural language processing applications [27,28]. As the backbone network of our proposed method, we selected the lightweight MobileNet V2, which can be trained fast and is very suitable for embedded devices [29,30]. We fine-tuned the MobileNet V2 model using the CDT images in the training set. To that end, we detached the *fully-connected* layer and attached a global average pooling operation to convert the feature map into a smaller size by taking the average value from the spatial dimension of the feature map. Global average pooling avoids overfitting and provides a more robust model against spatial translations [31]. A *softmax* layer was attached to the deep learning model to predict the probability distribution of each class. A block diagram of the modified CNN can be found in Figure 1. In the training procedure, all the layers of the MobileNet V2 trained on the ImageNet [32] were frozen, except the softmax layer. Since the MobileNet V2 is based on a three-channel CNN, we used command CDT images of size (128, 128, 3) to train the modified deep learning model, adjusting the input size of MobileNet V2 as needed. The participants tend to make more mistakes in the command clock drawing task compared to the copying task (cf. Section 4), hence, the command clock images can reveal more types of image defects associated with cognitive impairments.

Once the training process of the deep learning model was completed, the scores corresponding to test command and

copy images were generated by feeding them into the model. Finally, image scores, the age, and the education level of every participant were used to train a logistic regression model to make predictions; a schematic of the approach is shown in Figure 2. The entire model was implemented using the Python deep-learning Keras library with a Tensorflow backend.

2.5. Validation and performance metrics

The hyperparameters of the model were selected on a validation set using random search. Learning rate, number of epochs, batch size, and regularization parameter of regression λ were set to 3.10^{-4}, 400, 32, and 1, respectively. A binary cross-entropy loss was selected for training the CNN model. After augmenting the CIND/dementia class 20 times more than normal images (400 and 20 images generated from data augmentation, respectively), the CNN model was trained in a balanced fashion where a subset of the data without augmentation was held out for the testing process. Data were randomly split into 5 folds using stratified *k*-fold cross validation. Specifically, a model was trained on the 4 folds and tested on the 5th – test – fold. The process was repeated five times, each time with a different fold retained for testing, and the average and standard deviation (std) of all performance metrics on the test set over these five runs was obtained.

Performance metrics included the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), the weighted F1 score, sensitivity, and specificity [33]. The ROC plots the True Positive Rate (TPR, given by the ratio of true positive cases over true positives and false negatives, a.k.a. recall or sensitivity) against the False Positive Rate (FPR, given by the ratio of false positives over false positives and true negatives, which is equal to 1 minus the specificity). The F1 score is the harmonic mean of precision and recall. Precision (or positive predictive value) is defined as the ratio of true positives over true and false positives. The F1 score is calculated by

$$\mathrm{F1} = 2\frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}},$$

and the closer it is to 1 the stronger the classification model. In this work, we reported the *weighted F1 score*, which is computed by weighting the F1-score of each class by the number of participants in that class.

3. Results

We trained and validated the proposed classifier, using 5-fold cross-validation to estimate out-of-sample performance. The results are summarized in Table 2. We report ROC AUC, weighted F1 score (W-F1), sensitivity, specificity, and accuracy (Acc), the latter mainly for comparison purposes with results reported in the literature and surveyed in Section 1. We note however, that accuracy in binary classification with a highly imbalanced dataset is not an appropriate metric, since predicting all subjects as being normal would lead to

a high accuracy. AUC does not depend on the decision threshold (the constant compared to the likelihood to make the classification decision) but the rest of the metrics do; we report the maximum W-F1 and maximum Acc over the decision threshold. We also select the threshold that maximizes the geometric mean of sensitivity and specificity and report the corresponding values in the last two columns of Table 2. The optimal threshold is not necessarily the same for these three maximization procedures.

In the first row of Table 2, we report the performance of the classifier that uses just the command CDT (Cmnd CDT) image of a subject. The second row reports the performance of our proposed *ensemble model* obtained by using logistic regression with features the deep learning scores of command and copy CDT images, the education level, and the subject's age. The third row, corresponds to the full model, also obtained by logistic regression, which uses command and copy CDT images, age, education, gender, ApoE, and race. The fourth row corresponds to a model, obtained by logistic regression, using all these features, except ApoE. The fifth column reports the performance of a logistic regression model using only age and the command CDT image. The sixth row reports the performance of a logistic regression model based solely on age. The last three rows report the performance of the models while considering only the subjects with age 60 and older. In all these models, education, gender, ApoE, and race features, were encoded using one-hot encoding, i.e., creating a binary variable for each category. Figure 3 plots the ROC curves for the two models with high average AUC (full and ensemble), and, for comparison purposes, the corresponding curve for the age-based model.

Since age in the dataset has a large range, [28, 100], and dementia is likely to be concentrated on those with older age, we performed an experiment excluding subjects less than 60 years old. Figure 4 reports the coefficients of the features used by the logistic regression ensemble model with subjects 60 years and older. The coefficients are comparable as the scores of the CDT images and age were normalized by subtracting the mean and dividing by the standard deviation. It can be seen that age and the command CDT image are contributing more to the decision than the copy CDT image. Although the education features have a negligible impact on the AUC of the model, the predictive importance of the education labels reveals that the higher the education level, the less likely a dementia diagnosis is. Moreover, based on Table 1, only 54% of subjects in the CIND/dementia class received a post-secondary education compared to 81% in the normal class. On the other hand, post-secondary education might have been less typical for the older generation.

4. Discussion

The proposed *ensemble* model can be employed to offer virtual cognitive impairment screening using only CDT images drawn on pen and paper, education, and the age of the subject. The out-of-sample AUC and weighted F1 scores we report indicate strong predictive power.

An important aspect of the method we used for training the deep learning models was *transfer learning*. We started from a deep learning image classifier previously trained on a large number of generic images, which, apparently, has the ability to identify useful features of presented images. Thus, with limited training using our CDT images, the deep learning model adapted quickly to score CDT images and produced scores representing the likelihood of the subject being cognitively impaired. Just using command CDT images yields a classifier of moderate strength (cf. first row of Table 2, AUC of 81.3%, on average). Combining command and copy CDT images with age and education using logistic regression yields a model with an AUC of 91.9%, on average, and a weighted F1 score of 94.6%, on average.

The ensemble model performs slightly better than the full model in terms of specificity (cf. Table 2). Furthermore, the full model is not amenable to online screening as ApoE genotyping requires laboratory testing (typically, a blood sample). From Figure 3, it can also be seen that the ROC curves of the full model and the ensemble model essentially coincide for low values of the FPR (below 15%), that is, within the range one may want to operate. Interestingly, adding to the ensemble model gender, education, and race (i.e., full model without ApoE), leads to the same performance, confirming the low discriminatory power of these additional features, which was also suggested by their *p*-values listed in Table 1.

Table 2 also indicates that a model based just on age performs relatively well; average AUC of 89.3% vs. 91.9% for the ensemble model, yielding a difference of 2.6%. This is consistent with related findings in [15] where a model based on age, gender, and MMSE had an average F1-score 1.4% lower on their internal validation dataset than the fusion model which also used a brain MRI. The ensemble model also outperforms the age model by 4.9% while removing subjects younger than 60 years of age. It is useful to compare the average TPR (sensitivity) of the ensemble model, the full model, and the age model for low values of the average FPR (high specificity). This comparison is shown in Table 3. For instance, at 10% FPR, the TPR of the ensemble model is 13.6% higher compared to the age model. Putting this difference in context, suppose we were interested in screening for a nationwide clinical trial all 5 million or so individuals in the U.S. estimated to be suffering from Alzheimer's [1]. Setting FPR to 10%, about 3.66 million would qualify with the ensemble model vs. 2.98 million with the age model, missing a non-trivial number of about 680,000 subjects with the latter. A similar perspective is gained by considering how many individuals one should screen to assemble a clinical trial with about 1600 subjects (similar in size to the EMERGE aducanumab trial by Biogen [34]). Using an FPR of 5%, and assuming the CIND and dementia incidence rate is 4.7% as in our dataset, it follows that one needs to screen 78,439 people with the age model compared to 64,110 with the ensemble model, namely, 14,329 less. Clearly, these differences imply significant differences in cost.

In the present study, the clock drawing images were collected from the FHS using a digital pen. The size of the images used for training was reduced to 128×128 pixels. Furthermore, the data augmentation described in

Section 2.2 empowers the deep learning features to become robust to various forms of image distortion that could be introduced by drawing the images using pen and paper and capturing them using a cell phone.

A limitation of the study is that we do not have access to actual cell phone-captured images, which would provide the ultimate test for the proposed screening approach. An additional limitation is that the FHS does not conduct a comprehensive dementia review on participants who are relatively younger or who may not exhibit severe signs of dementia; thus, it is possible that some subjects classified as cognitive normal are in the early CIND stages. In other words, FHS prioritizes full dementia assessment on higher-risk participants. As a result, our dataset likely contained a relatively higher number of participants with normal cognition.

In conclusion, our deep learning approach based on transfer learning allowed us to classify individuals with dementia based on CDT images. These frameworks can be explored further to assist dementia screening in limited resource settings. Future work entails testing the robustness of our modeling approach across various cohorts comprising individuals from multiple races and ethnicity as well as on CDT images captured via various devices.

Code Availability

We have made our code publicly available in https://github.com/noc-lab/CDT. Our long-term goal is to make such tools available online.

5. Acknowledgments

The research was partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under grant N00014-19-1-2571, by the NIH under grants R01 GM135930 and UL54 TR004130, by the DOE under grant DE-AR-0001282, by the Framingham Heart Study's National Heart, Lung, and Blood Institute contract (N01-HC-25195; HHSN268201500001I), by the NIH National Institute on Aging (AG008122, AG016495, AG033040, AG054156, AG049810, AG062109), by the Alzheimer's Association under grant AARG-NTF-20-643020, and by Pfizer. VBK acknowledges support from the Karen Toffler Charitable Trust, the American Heart Association (17SDG33670323, 20SFRN35460031) and the National Institutes of Health (R21-CA253498).

Rhoda Au is a scientific advisor to Signant Health and consultant to Biogen. There is no declaration from other authors.

6. References

[1] Alzheimer's Association (2020) 2020 Alzheimer's disease facts and figures. Alzheimers Dement 16, 391-460.
[2] DeTure MA, Dickson DW (2019) The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener* **14**, 1–18.
[3] Mahmood SS, Levy D, Vasan RS, Wang TJ (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet* **383**, 999–1008.
[4] Satizabal CL, Beiser AS, Chouraki V, Chêne G, Dufouil C, Seshadri S (2016) Incidence of dementia over three decades in the Framingham Heart Study. *N Engl J Med* **374**, 523–532.
[5] Allone C, Lo Buono V, Corallo F, Bonanno L, Palmeri R, Di Lorenzo G, Marra A, Bramanti P, Marino S (2018) Cognitive impairment in Parkinson's disease, Alzheimer's dementia, and vascular dementia: the role of the clock-drawing test. *Psychogeriatrics* **18**, 123–131.
[6] Ryu S-Y, Lee S-B, Kim Y-I, Lee K-S (2006) P2–113: The utility of the clock drawing test for cognitive impairment screening. *Alzheimers Dement* **2**, S266–S266.
[7] Umegaki H, Suzuki Y, Yamada Y, Komiya H, Watanabe K, Nagae M, Kuzuya M (2020) Association of the Qualitative Clock Drawing Test with Progression to Dementia in Non-Demented Older Adults. *J Clin Med* **9**, 2850.
[8] Cacho J, Benito-León J, García-García R, Fernández-Calvo B, Vicente-Villardón JL, Mitchell AJ (2010) Does the combination of the MMSE and clock drawing test (mini-clock) improve the detection of mild Alzheimer's disease and mild cognitive impairment? *J Alzheimers Dis* **22**, 889–896.
[9] Piers RJ, Devlin KN, Ning B, Liu Y, Wasserman B, Massaro JM, Lamar M, Price CC, Swenson R, Davis R (2017) Age and graphomotor decision making assessed with the digital clock drawing test: the Framingham Heart Study. *J Alzheimers Dis* **60**, 1611–1620.
[10] Cohen J, Penney DL, Davis R, Libon DJ, Swenson RA, Ajilore O, Kumar A, Lamar M (2014) Digital clock drawing: Differentiating 'thinking'versus 'doing'in younger and older adults with depression. *J Int Neuropsychol Soc JINS* **20**, 920.
[11] Souillard-Mandar W, Davis R, Rudin C, Au R, Libon DJ, Swenson R, Price CC, Lamar M, Penney DL (2016) Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Mach Learn* **102**, 393–441.
[12] Binaco R, Calzaretto N, Epifano J, McGuire S, Umer M, Emrani S, Wasserman V, Libon DJ, Polikar R (2020) Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease. *J Int Neuropsychol Soc* **26**, 690–700.
[13] Vieira S, Pinaya WH, Mechelli A (2017) Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev* **74**, 58–75.
[14] Li H, Habes M, Wolk DA, Fan Y, Alzheimer's Disease Neuroimaging Initiative (2019) A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimers Dement* **15**, 1059–1070.
[15] Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, Chang GH, Joshi AS, Dwyer B, Zhu S (2020) Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain* **143**, 1920–1933.
[16] Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, Filippi M, Alzheimer's Disease Neuroimaging Initiative (2019) Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage Clin* **21**, 101645.
[17] Li F, Tran L, Thung K-H, Ji S, Shen D, Li J (2015) A robust deep model for improved classification of AD/MCI patients. *IEEE J Biomed Health Inform* **19**, 1610–1616.
[18] Nordberg A, Rinne JO, Kadir A, Långström B (2010) The use of PET in Alzheimer disease. *Nat Rev Neurol* **6**, 78–87.
[19] Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* **32**, 2322-e19.
[20] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* **316**, 2402–2410.
[21] Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X (2018) Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Front Neurosci* **12**, 777.
[22] Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS (2019) 70-year legacy of the Framingham Heart Study. *Nat Rev Cardiol* **16**, 687–698.
[23] Au R, Piers RJ, Devine S (2017) How technology is reshaping cognitive assessment: Lessons from the Framingham Heart Study. *Neuropsychology* **31**, 846.

[24] Davis R, Penney D, Pittman D, Libon D, Swenson R, Kaplan E (2011) The Digital Clock Drawing Test (dCDT) I: Development of a new computerized quantitative system. *Int Neuropsychol Soc*.
[25] Massey Jr FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* **46**, 68–78.
[26] McHugh ML (2013) The chi-square test of independence. *Biochem Medica* **23**, 143–149.
[27] Shaha M, Pawar M (2018) Transfer learning for image classification. In IEEE, pp. 656–660.
[28] Rios A, Kavuluru R (2019) Neural transfer learning for assigning diagnosis codes to EMRs. *Artif Intell Med* **96**, 116–122.
[29] Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* **35**, 1285–1298.
[30] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. *Med Image Anal* **42**, 60–88.
[31] Ratul MAR, Mozaffari MH, Lee W, Parimbelli E (2020) Skin lesions classification using deep learning based on dilated convolution. *BioRxiv* 860700.
[32] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In Ieee, pp. 248–255.
[33] Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.
[34] Knopman DS, Jones DT, Greicius MD (2021) Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. *Alzheimers Dement* **17**, 696–701.

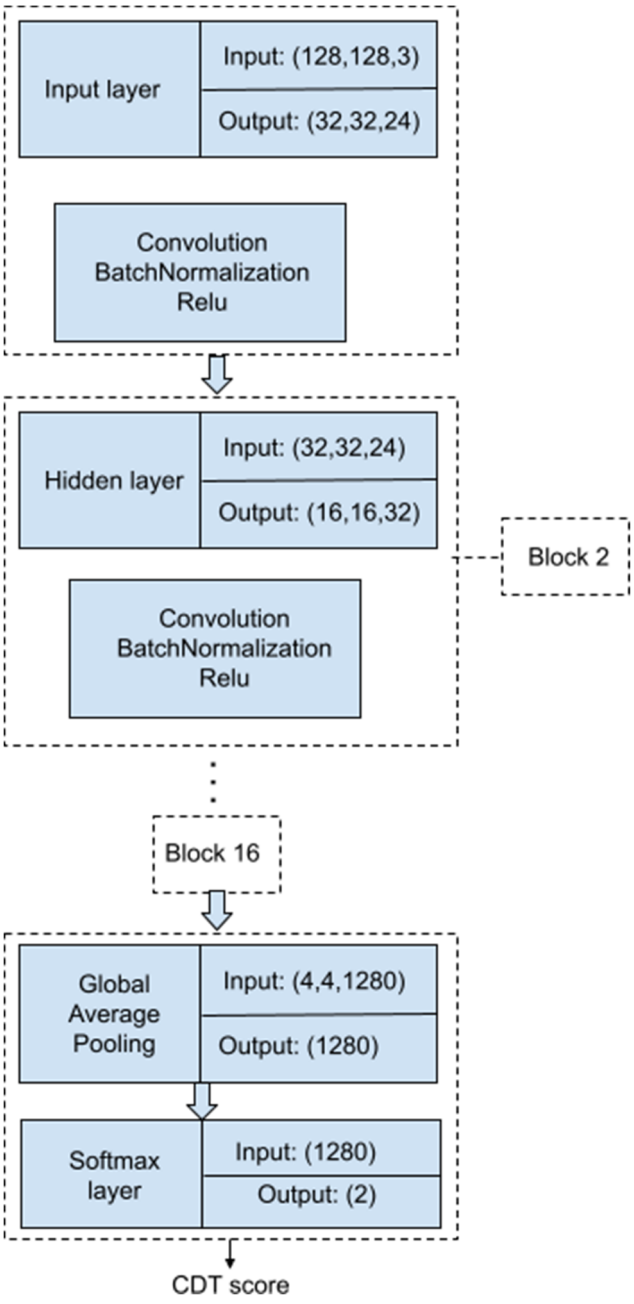

Figure 1: Schematic diagram of the CNN model with the MobileNet network.

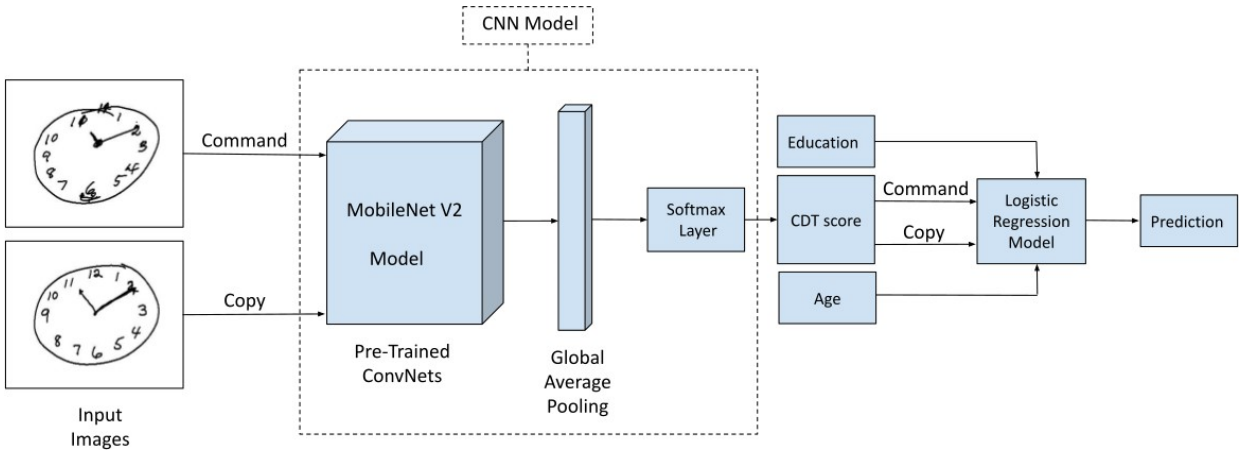

Figure 2: Online screening for cognitive impairment.

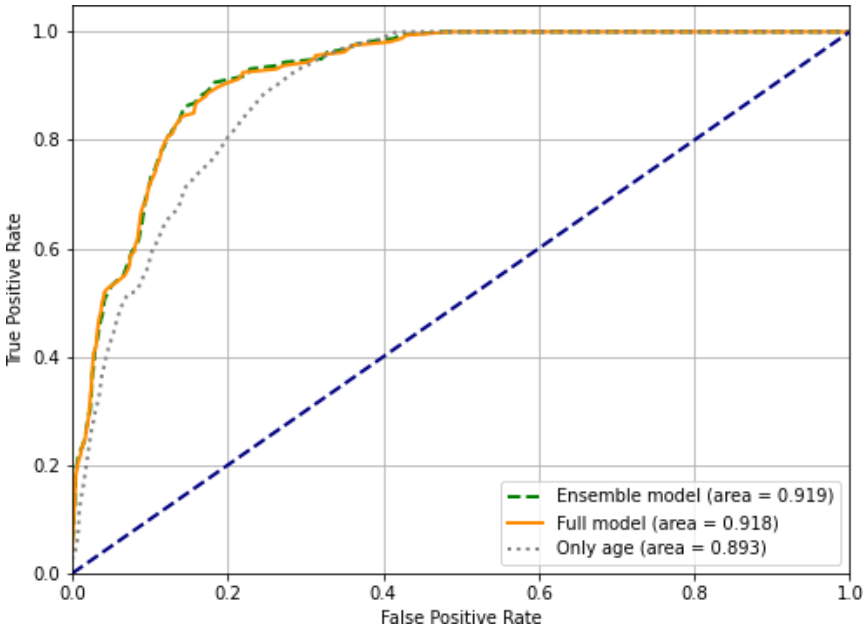

Figure 3: ROC curves for three models: the proposed ensemble model, the full model, and the model based only on age. Weplot the average TPR and FPR over the five folds.

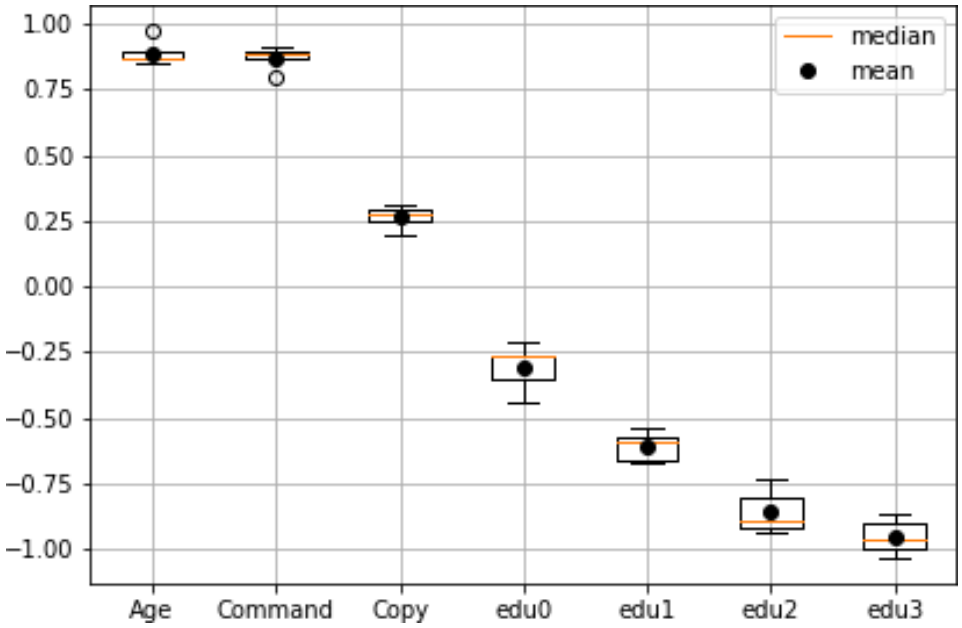

Figure 4: Logistic regression coefficients (mean, median, and 95% confidence intervals), indicating the relative predictive importance of the features in the Ensemble model with subjects 60 years and older. The features edu0, edu1, edu2, edu3 indicatethe education level of the subjects, corresponding to attending high school, graduating from high school, attending college, and graduating from college, respectively.

Table 1: Summary of the variables in the FHS dataset. Diagnosis scores 0, 0.5, 1–1.5, 2–2.5, and 3 are defined as normal cognition, CIND, mild dementia, moderate dementia, and severe dementia, respectively. Education labels 0, 1, 2, and 3 correspond to attending high school, graduating from high school, attending college, and graduating from college, respectively. Note that 52 education labels are missing among the normal group.

Variable	**Assessment**		***p*-value**
	Normal (n=3263)	CIND/Dementia (n=160)	
Diagnosis			-
0	3263 (100%)	0 (0%)	
0.5	0 (0%)	96 (60%)	
1-1.5	0 (0%)	38 (23.7%)	
2-2.5	0 (0%)	24 (15%)	
3	0 (0%)	2 (1.2%)	
Age	61.8 ± 13.2	82.1 ± 7.3	<0.0001
Gender			0.95
Female	1773 (54.3%)	86 (53.75%)	
Male	1490 (45.7%)	74 (46.25%)	
Education			<0.0001
0	45 (1.4%)	12 (7.5%)	
1	574 (17.9%)	61 (38.1%)	
2	761 (23.7%)	42 (26.2%)	
3	1831 (57.0%)	45 (28.1%)	
AopE			0.12
22	19 (0.6%)	1 (0.6%)	
23	399 (12.2%)	19 (11.9%)	
24	63 (1.9%)	6 (3.8%)	
33	2011 (61.6%)	87 (54.4%)	
34	593 (18.2%)	40 (25%)	
44	45 (1.4%)	4 (2.5%)	
Race			0.24
Asian	86 (2.6%)	1 (0.6%)	
Black	87 (2.7%)	1 (0.6%)	
Hispanic	80 (2.5%)	2 (1.2%)	
White	2957 (90.6%)	156 (97.5%)	
other	53 (1.6%)	0 (0%)	

Table 2: Results on the test set (mean ± std over the five runs). Acc stands for accuracy.

Methods	AUC %	W-F1 %	Acc %	Sensitivity %	Specificity %
Cmnd CDT	81.3 ± 5.8	94.3 ± 0.4	95.5 ± 0.3	73.8 ± 6.7	77.4 ± 5.8
Ensemble	91.9 ± 1.1	94.6 ± 0.4	95.7 ± 0.3	86.9 ± 1.2	84.5 ± 4.0
Full	91.8 ± 1.1	94.8 ± 0.3	95.7 ± 0.3	86.9 ± 2.3	83.8 ± 3.8
Full\ApoE	91.9 ± 1.2	94.7 ± 0.4	95.7 ± 0.3	85.6 ± 2.5	85.7 ± 1.9
Age, Cmnd CDT	91.9 ± 1.2	94.6 ± 0.7	95.8 ± 0.4	89.4 ± 1.5	82.9 ± 3.6
Age	89.3 ± 1.2	94.0 ± 0.5	95.4 ± 0.1	85.0 ± 5.4	77.9 ± 4.5
Cmnd CDT (age ≥ **60**)	77.7 ± 6.5	90.6 ± 0.7	92.6 ± 0.4	71.3 ± 8.9	73.6 ± 4.1
Ensemble (age ≥ **60**)	86.6 ± 2.0	91.3 ± 0.6	92.7 ± 0.2	85.0 ± 3.6	76.1 ± 3.2
Age (age ≥ **60**)	81.7 ± 2.0	90.1 ± 0.7	92.4 ± 0.2	67.5 ± 10.2	78.5 ± 8.7

Table 3: Average True Positive Rate (TPR, or sensitivity) for average False Positive Rates (FPR) at 5%, 10%, and 20%.

Methods	% TPR AT 5% FPR	% TPR AT 10% FPR	% TPR AT 20% FPR
Ensemble	53.1	73.1	91.2
Full	53.1	73.8	90.6
Age	43.4	59.5	81.9