Introducing An Automated Coding Procedure Using Deep Learning Neural Networks to Score the Clock Drawing Test in the National Health and Aging Trends Study

June 2023

Suggested citation: Hu, Mengyao, Maureen Skehan, Yi Lu Murphey, Song Wang, Tian Qin & Vicki A. Freedman. 2023. Introducing an Automated Coding Procedure Using Deep Learning Neural Networks to Score the Clock Drawing Test in the National Health and Aging Trends Study. NHATS Technical Paper #35. Baltimore: Johns Hopkins Bloomberg School of Public Health. Available at www.nhats.org. We thank Richard Gonzalez for his contributions to the section on deep learning neural networks model selection. This technical paper was prepared with funding from the National Institute on Aging (U01AG032947 & R21AG073971).

Overview

The National Health and Aging Trends Study (NHATS) conducts a clock-drawing test (CDT) annually to evaluate older adults' executive function. The CDT score is included as a criterion in the NHATS dementia classification (Kasper et al. 2013). For the first 11 rounds of NHATS, CDT images were manually coded. Starting in Round 12, NHATS switched to using deep learning neural networks (DLNN) to code the CDT images. This technical paper describes the methodology used in developing and validating the automated coding as well as procedures for implementing in Round 12.

Clock-Drawing Test Administration

The CDT is administered according to the following procedures. The respondent is given a sheet of paper and an erasable pen and asked to draw a clock. The interviewer says, "Start by drawing a large circle. Put all of the numbers in the circle and set the hands to show 11:10 (10 past 11)." The respondent has 2 minutes to complete the activity. Interviewers may repeat the instructions as needed. Image files of clock drawings are available to download from the NHATS website as part of the Public Use files: www.nhats.org/researcher/data-access.[1]

Manual Coding in Rounds 1-11

Clock drawing tests in prior rounds of NHATS were manually coded. Clocks were scored on a scale from 0, not recognizable as a clock, to 5, an accurate depiction of a clock. Scoring guidelines for the clock drawing test were reproduced for use in NHATS by special permission of the Publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, from the Calibrated Neuropsychological Normative System, by David J. Schretlen, PhD, S. Marc Test, PhD, and Godfrey D. Pearlson, MD, Copyright 2010 by Psychological Assessment Resources, Inc.

Criteria for scores for manual coding were as follows:

- 5 (accurate depiction)—numbers in correct quadrants; hands pointing to the numbers 11 and 2; minute hand longer than the hour hand.
- 4 (reasonably accurate depiction)—numbers in roughly correct quadrants; hands reasonably close to the numbers 11 and 2; hands could be of equal length or the minute hand could be shorter than the hour hand; numbers may be outside the perimeter of the clock face.
- 3 (mildly distorted depiction)—some numbers may be missing or disoriented; there may be a few extra numbers. Hands may be incorrectly drawn or pointing to wrong number combinations; a hand may be missing.
- 2 (moderately distorted depiction)—several numbers are missing, repeated, or drawn in reverse order; there were more than two hands or no hands.
- 1 (severely distorted depiction)—viewer might be able to tell that the drawing was a clock but could not tell the time shown.
- 0 (not recognizable as a clock)—viewer would not be able to tell drawing was supposed to be a clock.

[1] The clock drawing activity was administered in the NHATS Cognition (CG) section in Rounds 1–10. In Round 10, materials were mailed to the SP and the clock drawing activity was attempted over the phone and returned by mail for scoring. In Round 11, it was permanently moved from the CG section to a new Clock Drawing (CD) section. See the NHATS User Guide (Freedman et al. 2022) for details.

Clocks were scanned into an online database for coding. If a participant drew more than one clock, coders were instructed to score the best of the clocks drawn. If participants clearly marked out something they had drawn, like an extra hand, coders were instructed to score as if two hands had been drawn not three. Clocks that were difficult to read such as clocks missing a part of the outside or a clock that was too small to see were scored according to what could be seen by the coder. In addition to scoring the accuracy of the clock drawing (variable name=cg#dclkdraw) coders also rated the clarity of the clock image (variable name=cg#dclkimgcl for Rounds 1 - 11).

Clock Coder Training and Selection of Coders in Rounds 1-11

Training included a presentation on test administration, a review of the scoring guide, and a review of 20-25 clocks and discussion of coding. A neuropsychologist was consulted in the development of the training. Following training, each coder was given 220 clock drawings from the NHATS Validation Study conducted in the Spring of 2010, which were also coded by two neuropsychology fellows. The performance of each lay coder was evaluated against the two clinically trained neuropsychology coders based on a weighted kappa statistic. Coders with higher levels of agreement with the neuropsychology coders were selected to code the clocks drawn by NHATS respondents. Weighted kappa of over 0.60 – 0.77 were used as the criterion for coder selection across the years. Detailed descriptions of clock coder training and selection of coders can be found in the NHATS User Guide (Freedman et al. 2022).

Using Deep Learning Neural Networks (DLNN) to code CDT

To improve accuracy and efficiency of coding the CDT in NHATS, we explored the use of deep learning neural networks (DLNN) to automate the scoring (Hu et al., 2023).

Data

Clock images from NHATS Rounds 1–9 were used for training and testing deep learning models. In total, more than 47,000 CDT images were available for the 9 rounds. To ensure the training and test data had better coding quality (i.e., less coding error), we used images with high clarity (cg#dclkimgcl = 1) and coded by the top 8 coders with highest average Kappa scores evaluated based on inter-coder reliability with the two neuropsychology fellows. In total, 25,872 images were selected, with a random subset of 90% as training data and the remaining 10% as test data.

Clock extraction

The first step was to extract the clock from the image. Note that the original CDT images included masked IDs, which in this application could be considered extraneous markings. To accurately extract the clock from the image, we developed a robust clock image segmentation system that combines classic image processing and computer vision techniques including line removal, clustering, connected component analysis, and digit detection. The overall accuracy of the clock image segmentation system is 95%. The 5% of CDT images that were not correctly extracted were extracted manually.

Model selection

We examined three types of deep learning architecture used in computer vision tasks: ResNet101, EfficientNet and Vision Transformers (ViT). Both ResNet101 and EfficientNet are Convolutional Neural Networks (CNN), which were found to outperform other CNN models in a number of computer vision tasks (He et al. 2016; Tan & Le, 2019). The third approach, ViT, was adapted from a deep learning model originally designed for natural language processing (Dosovitskiy et al. 2020).

For each of the deep learning models, we compared two different outcomes: 1) classification of the CDT score into a nominal variable; and 2) treatment of the score as an ordinal variable. For each model and outcome, we generated three evaluation metrics: accuracy, root mean square error (RMSE), and γ coefficient (capturing how closely two pairs of data points match). For both nominal classification and ordinal outcomes, we found ViT had the highest accuracy, lowest RMSE, and highest γ coefficient, compared to the other two method. Further, we found that ViT nominal classification and ViT ordinal led to similar results (see Table 1).

Table 1. Comparisons between DLNN models by nominal classification vs. ordinal approaches

	ViT	
	Nominal Classification	Ordinal
Accuracy	78.8%	76.3%
RMSE	0.56	0.56
γ coefficient	0.91	0.93

Validation

To compare ViT approaches to manual coding, we did the following. We computed the average weighted Kappa against the two neuropsychology fellows (gold standard) for manual coding and ViT approaches for the 220 training images. We found that both ViT approaches aligned more closely than manual codes to scores assigned by the neuropsychology fellows. The average weighted kappa was 0.81 for the ViT ordinal approach and 0.83 for the ViT nominal classification approach, and 0.76 for manual coding.

To further validate DLNN-generated CDT scores, we used both ViT approaches to code clock images from NHATS Round 11. Weighted kappa between the two ViT approaches was 0.88 and between the ViT methods and manual coding was about 0.70.

We also examined whether and to what extent using ViT approaches to code CDT images would change the NHATS dementia classification (Kasper et al. 2013). Using dementia classification constructed using manually coded-CDT as the benchmark, the accuracy of ViT-based dementia classifications were both 99.4% and the weighted kappa between the ViT-based dementia classification and manually coded-CDT-based dementia classification equal to 0.99 for both ViT approaches. This suggests minimal changes in dementia classification can be expected if researchers replace manually-coded CDT with ViT-coded CDT.

CDT-Coding in NHATS Round 12

In total, 5,591 CDT images were collected in NHATS Round 12. Clocks were extracted from the image files using the clock extraction system. Most of the CDT images (96.0%) were extracted correctly, remaining clocks (n=225) were extracted manually. In cases where respondents drew more than one CDT image the best one was selected and manually extracted.

The two ViT approaches described above were used to code R12 CDT images. Among the 5,591 images, 85.4% were given the same score by both ViT approaches and the remaining 14.6% (n=817) images were assigned different codes by the two methods.

We undertook the following steps to develop procedures for assigning scores when the two ViT approaches disagreed. An experienced NHATS clock coder scored a subset of 817 images. Specifically, the coder manually coded the following three groups of CDT images:

1) All 63 CDT images where the two ViT approaches lead to a different conclusion about whether the respondent scored <1.5 SD below the mean (e.g., 0,1 vs. 2+, criteria used as part of the dementia classification);
2) An additional 38 CDT images where the two ViT approaches coded CDT with 2 or more points apart;
3) A random subset of 100 of the remaining 716 images, which were coded only one point apart by the two ViT approaches.

For Groups 1 and 2, manually coded scores were used as final codes. For Group 3, we compared which ViT method (nominal classification or ordinal) was closer to the manually coded scores. We found that for 90% of images (n=90), one of the ViT approaches matched the manually coded score: n=59 matched ViT ordinal and n=31 matched the nominal classification approach. We also found that the weighted kappa between the manual coding and ViT approach was better for the ordinal (0.61) than the nominal classification (0.37) approach. Based on these results, we assigned the ViT ordinal scores to the 716 images in group 3.

CDT Variables

Starting in Round 12, two new CDT-related derived variables are included in the public SP file. The variable cg#dclkcoding describes the certainty of automated scoring, with 1 = Machine-coded high certainty (two ViT approaches agree); 2 = Machine-coded moderate certainty (two ViT approaches coded with one-point apart and did not change the CDT component of the dementia classification); 3 = Manually coded. The variable cg#dclkextract describes whether images were automatically extracted by machine or manually extracted. As in prior rounds, the variable cg#dclkdraw provides the CDT score.

Table 2. Clock-drawing test variables in Round 12

Variable Name VARIABLE LABEL	CODING SPECIFICATIONS	VALUES and VALUE LABELS
cg#dclkcoding R# D CLOCK CODING CERTAINTY MACHINE OR MANUAL	-1 if r12dresid = 3 or 5 or 6 or 7 or 8 1 if CDT score coded by VIT classification is the same as score coded by VIT ordinal approach: vit_classification = vit_ordinal 2 if CDT score coded by VIT classification is one point away from the score coded by VIT ordinal approach, and it does not change the CDT component of the NHATS dementia classification: absolute value of (vit_classification - vit_ordinal) = 1 and ((vit_classification !=1 & vit_classification != 2) or (vit_ordinal !=1 & vit_ordinal != 2) or (vit_classification == 1 & vit_ordinal == 0) or (vit_classification == 2 & vit_ordinal == 3) or (vit_classification == 0 & vit_ordinal == 1) or (vit_classification == 3 & vit_ordinal == 2)) 3 if (absolute value of (vit_classification - vit_ordinal) > 1) or (vit_classification == 1 & vit_ordinal == 2) or (vit_classification == 2 & vit_ordinal == 1) Else -4 if cg12atdrwclck = 2 Else -7 if cg12atdrwclck = 97Else cg12dclkdraw = -9	-9 = Missing (no clock) -7 = SP refused to draw clock -4 = SP did not attempt to draw clock -1 = Inapplicable 1 = Machine-coded high certainty 2 = Machine-coded moderate certainty 3 = Manually coded
cg#dclkextract R# D CLOCK IMAGE EXTRACTED MACHINE OR MANUAL	-1 if r12dresid = 3 or 5 or 6 or 7 or 8 1 if CDT image extracted by machine (manually_extracted = No) 2 if CDT image manually extracted (manually_extracted = Yes) Else -4 if cg12atdrwclck = 2 Else -7 if cg12atdrwclck = 97 Else cg12dclkdraw = -9	-9 = Missing (no clock) -7 = SP refused to draw clock -4 = SP did not attempt to draw clock -1 = Inapplicable 1 = Machine extracted 2 = Manually extracted
cg#dclkdraw R# D SCORE OF CLOCK DRAWING TEST	-1 if r12dresid = 3 or 5 or 6 or 7 or 8 if cg12dclkcoding = 1 or cg12dclkcoding = 2 then cg12dclkdraw = vit_ordinal; Else if cg12dclkcoding = 3 then cg12dclkdraw = manually coded score. Else -4 if cg12atdrwclck = 2 Else -7 if cg12atdrwclck = 97 Else cg12dclkdraw = -9	-9 = Missing (no clock) -7 = SP refused to draw clock -4 = SP did not attempt to draw clock -1 = Inapplicable 0 = Not recognizable as a clock 1 = Severely distorted depiction of a clock 2 = Moderately distorted depiction of a clock 3 = Mildly distorted depiction of a clock 4 = Reasonably accurate depiction of a clock 5 = Accurate depiction of a clock (circular or square)

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hu, M., Murphey, Y.L., Wang, S., Qin T., Zhao Z., Gonzalez, R., Freedman V. A. & Zahodne L. (2023) Exploring the Use of Deep Learning Neural Networks to Improve Dementia Detection: Automating Coding of the Clock-Drawing Test. *Survey Research Center Seminar Series*, Institute for Social Research, University of Michigan, Ann Arbor.

Kasper, J. D., Freedman, V. A., & Spillman, B. C. (2013). Classification of persons by dementia status in the National Health and Aging Trends Study. *Technical paper*, *5*, 1-4.

Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Freedman, V. A., Schrack, J.A., Skehan, M. E., & Kasper, J. D. (2022). National Health and Aging Trends Study User Guide: Rounds 1-11 Final Release. Baltimore: Johns Hopkins University School of Public Health. Available at www.NHATS.org.