Automating the Clock Drawing Test with Deep Learning and Saliency Maps

Violet Mayne[1(✉)], Harry Rogers[2], Saber Sami[2], and Beatriz de la Iglesia[2]

[1] School of Computer Science, University of Lincoln, Lincoln, UK
V.Mayne@uea.ac.uk
[2] School of Computing Sciences, University of East Anglia, Norwich, UK

Abstract. The Clock Drawing Test (CDT) is an important tool in the diagnosis of Cognitive Decline (CD). Using Deep Learning (DL), this test can be automated with a high degree of accuracy, more so where the medium of recording allows the use of temporal information on how the clock was drawn which may not be accessible to clinicians in traditional screening. The high-risk nature of this field makes understanding the reasoning for automated results imperative. A model's reasoning can often be described using saliency maps, however, there are a number of different methods for generating such maps. Therefore, we propose a methodology to train a DL classifier for use in the CDT which incorporates temporal information and use saliency maps to explain classification predictions. We find that our classifier achieves scores above 98% with F1 for clocks and over 96% F1 on average across a test set of 18 different classes. Our methodology also shows that Integrated Gradients using SmoothGrad produce the best saliency map results visually and statistically.

Keywords: Cognitive Decline · Deep Learning · eXplainable Artificial Intelligence

1 Introduction

Cognitive decline (CD) is a serious risk associated with aging that can lead to dementia [4]; therefore early diagnoses are imperative for interventions. There is a lifetime risk of cognitive impairment of 37% for women and 24% for men [10]. Therefore, development of CD testing is crucial.

One method of assessing CD is through the use of sketched drawings. The Clock Drawing Test (CDT) [14], is a popular test in which patients are asked to draw a clock to assess their state of CD. Within this, patients are typically asked to draw a circle and then instructed to set the hands at 11.10, the test is scored from 0–5 based on how well the clock is drawn [21]. The test is quick to administer and has been shown to differentiate at baseline between cognitively intact older adults who will develop dementia up to 2 years post baseline and also those that have mild cognitive impairments and will progress to dementia up to 6 years post baseline [7]. Evaluations of the test are typically performed by

qualified clinicians. The use of Deep Learning (DL) in diagnosis can be highly beneficial as it can reduce human error and increase speed of diagnosis [6]. This includes the application of DL to evaluate the results of the CDT which can be fully automated for patients [2].

Explainable AI (XAI) is a particularly important field within medical applications of DL. Using XAI allows for human-interpretable explanations of model predictions. XAI can increase trust in these systems, allowing for clearer explanations of why failure happens, and illustrate functionality to regulators [12].

Saliency maps, developed under XAI, can visualise classification decisions as a heatmap. Saliency heatmaps are generated using DL model parameters and an input image to then assign importance values to pixels [25]. This allows for further understanding with a human-interpretable visual representation. Adebayo et al. [1] present a method for evaluating how informative saliency maps are; this method is created and tested on several datasets of photographs. Here we extend the concept to evaluation on sketches recorded over a computer as is needed to automate the CDT and similar tests.

Our key contribution is to create a DL model to automate the CDT, incorporating all information introduced by the medium of recording - such as temporal information - then evaluating the series of saliency map methods in [1] to determine which are most informative within this use case.

The remainder of this paper is structured as follows. Section 2 introduces related work on automating the CDT, and saliency map methodologies including evaluation of saliency maps. Section 3 provides details of the methodology proposed to automate the CDT with dataset, DL models, evaluation metrics for classification and Saliency maps. The results of our findings are reported in Sect. 4. Finally, conclusions are drawn and future work is presented in Sect. 5.

2 Related Work

Our related work is split into works relating to the use of DL for sketch classification including the automation of the CDT, and the use and evaluation of saliency maps, particularly within a medical context.

2.1 Automation of CDT

Some attempts have been made to automate the completion of the CDT, for example Amini et al. [2] used a dataset of analog clock drawings of 3,263 cognitively intact and 160 cognitively impaired individuals to build a Convolutional Neural Network (CNN) which could predict whether an individual is undergoing CD. They further extended this model into an ensemble which takes in demographic information such as the age and education level of a patient to produce a model which classifies whether a patient shows signs of CD.

If prediction models are possible, a system to collect and classify sketch information more generally will aid with automating the assessment task. The model can later be deployed to collect data on a clinical setting. Pearson et al.

[13] trained a model to classify 24 different categories of sketch and produced a measure of certainty that a given sketch is of the instructed class - e.g. a clock. The expectation was that the certainty that the sketch belongs to a specific category may be correlated with some measurement of CD, although this was not tested.

A limitation of both approaches, in terms of how sketches are classified, is that they contain no temporal information and instead only assess the final completed sketch. However, information on how a patient goes about drawing a sketch over time can be recorded for a sketch drawing assessment tool. This has been found to be a useful features in recognising CD [5].

Several methods for sketch classification utilising temporal information are compared by Seddati et al. [17]. These methods rely on rasterising vector representations of the sketches at five stages of completion, i.e. 20% of time used, 40% used, etc. Five CNNs are then trained on the sketches, one for each stage of completion, these are fused using either a fully connected layer, hard-coded weights, or a Long-Short Term Memory network (LSTM). The LSTM is a recurrent network which modifies an internal state after seeing the outputs of each model where the other methods are fused based on all outputs simultaneously. Additionally, another model was created which stacks the five frames into a single input thus only requiring a single CNN. Using the TU-Berlin sketch benchmark [8], Seddati et al. found hard-coded values and a fully connected layer produced the best accuracies of 77.69%, and 77.53% whilst the LSTM and combined models produced lower results of 76.42%, and 74.07%, respectively. As a baseline comparison, the CNN trained on the fully completed dataset produced an accuracy of 75.61% when used in isolation.

Based on this, we will implement versions of the four models utilising temporal features using a generalised approach to clock classification which can later be deployed to collect clinical data as suggested in Pearson et al. [13].

2.2 Saliency Maps

Saliency maps are used in several medical contexts to better understand the reasons for classification from DL models. For example, in the work by Rajpurkar et al. [16] a CNN is created to identify pneumonia from chest X-rays and Class Activation Maps (CAMs) are used to identify areas that are important for a particular pathology classification. CAMs highlight large areas whereas Saliency maps use exact pixels, producing a higher granularity in the highlighted areas which could help with explaining CD as manifested in sketch drawing.

Saliency methods are known to sometimes produce misleading or unclear results [9]. Arun et al. [3] applied eight saliency map techniques based on basic gradients, integrated gradients, guided backpropagation, and GradCAM to two pneumonia datasets. This found that, when applied to these datasets, all techniques failed at least one of their criteria for utility and robustness. This will diminish the ability of clinicians to verify that the model is identifying relevant features within the sketches and reasons for false positives and false negatives.

Due to the high-risk nature of the medical diagnosis, the accuracy and usefulness of these methods needs to be ensured.

Adebayo et al. [1] compares the performance of saliency maps across three datasets. One of these consists of photographed objects across many categories, one of greyscale photographs of items of clothing, and one of photographed and pre-processed hand-drawn digits. Their evaluation of performance works by randomising the parameters of trained models and calculating the correlation between the saliency maps produced before and after randomisation, this should assess how much of a given map is based on the form of the image instead of on the model's calculations. When applied to models trained on these datasets the qualities of the saliency map methods vary by dataset whilst maintaining some similarities, such as the ineffectiveness of Guided GradCAM.

As different saliency map methods vary in how informative they are across differing domains, an evaluation of saliency map methods specifically within the domain of the CDT will be undertaken in this paper, using the same randomisation methodology and including all available methods for completeness.

3 Methods

Our methodology is broken down into several steps to ensure clarity and effectiveness within the classification and XAI evaluation for the CDT.

3.1 Dataset

The Google 'Quick, Draw!' dataset [11] was used to develop the concepts in this paper. This is a dataset of 50 million sketch style drawings across 345 drawing categories. The sketches in this dataset were captured as timestamped vectors and publicly released alongside a simplified 28×28 rasterised version.

To ensure temporal information can be factored into analysis, one of the contributions of our work, the timestamped vectors were used and a modified version of Quick Draw's simplification methodology [11] was used to allow sketches to be reconstructed at any stage of drawing. This was completed as follows:

1. Remove curve points with timestamps beyond the current maximum time.
2. Rescale vectors to fit within a 28×28 grid, preserving aspect ratio.
3. Rasterise, antialiasing using Xiaolin Wu's line algorithm [26].

The reconstruction at different time points is demonstrated in Fig. 1. It may be for example, that a pause while trying to work out where the clock hands should be drawn is associated with CD.

To simplify the computational complexity of training and deployment, a subset of 18 classes were used for training. These were selected for their simplicity of drawing, so as to minimise the confusion between CD and poor drawing skills, and are shown in Fig. 2.

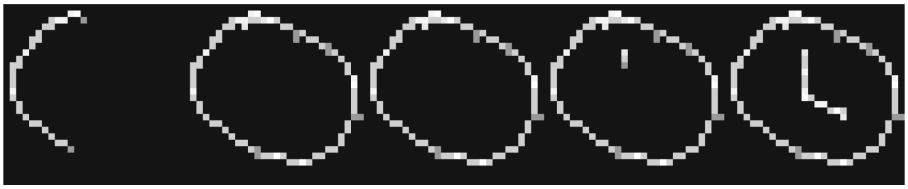

Fig. 1. Example of a clock sketch rendered at five points in time. Note that the drawer paused after drawing the circle so images 2 and 3 are identical

Fig. 2. The 18 chosen target classes. In order from left to right, top to bottom, these are: Bucket, Butterfly, Candle, Clock, Cloud, Envelope, Eye, Fence, House, Ice Cream, Ladder, Mushroom, Paper Clip, Pizza, Rainbow, Snowflake, Star, and Wine Bottle

3.2 Model Selection

The four model architectures created by Seddati et al. [17] as described in Sect. 2 were implemented. Sketches were rendered at five equidistant stages of completion and used to train five CNNs alongside a fully connected layer and a LSTM layer to fuse them. They were also fused using the hardcoded values $1, ..., 5$. Finally a model which stacks the five frames into a single input, thus, only requiring a single CNN was implemented.

All CNNs were created with the same structure of a batch normalisation layer followed by four Convolutional layers and four fully connected layers. To reduce overfitting, three dropout layers are used. Layers were connected using the Rectified Linear Unit activation function (ReLU).

Additional features calculated from the vectors were passed to the dense layer consisting of the number of lines in the sketch and the mean, variance, range, and quartiles of the lines' lengths, distances (from start and end of line), and duration spent drawing. The full models are visualised in Fig. 3.

1000 instances from each class were reserved as test data with all remaining instances being used as training data. Due to the massive size of training data, sets of $12,600$ new training instances (700 for each instance) were each trained upon for 5 epochs after which a new set of instances would be loaded. This was completed 40 times, using a total of $28,000$ instances of each class.

3.3 Model Evaluation

Performance on a single class is calculated using the F1-score, we will look at clocks in particular as they are used in the CDT. Overall performance is calcu-

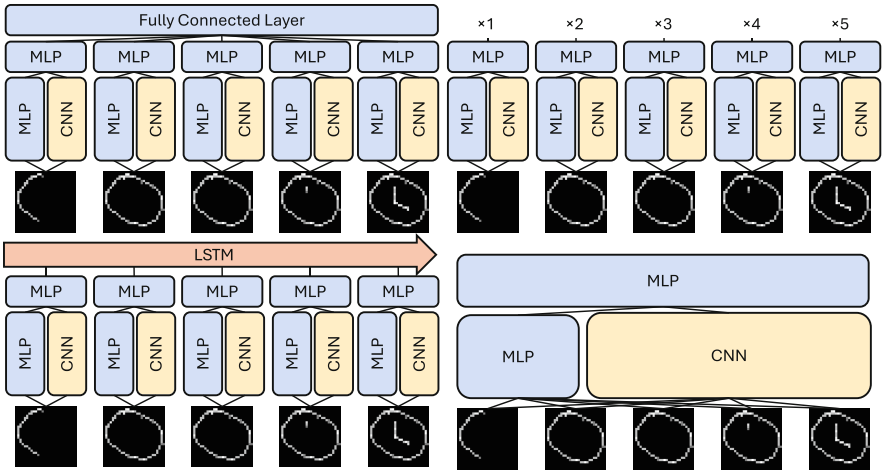

Fig. 3. Diagrams of the four models, from left to right, top to bottom, fusion using a fully connected layer, hard-coded values, LSTM, and single input

lated using overall accuracy and average F1-score across all classes. These are implemented as in Rainio et al. [15]. The F1-scores for clocks and metrics across all classification are presented and discussed in Sect. 4.1.

3.4 Saliency Map Methods and Evaluation

Saliency maps when visualised show a heatmap of importance values where red indicates positive importance (this area pushes towards the class that was classified) and blue indicates negative importance. The methods used in this paper are summarised as follows.

- **Basic Gradient Method (Gradient)**: Calculates the saliency of a pixel x as $\text{Gradient}(x) = \frac{\partial S}{\partial x}$. This indicates the rate at which a change in x changes the prediction $S(x)$.
- **SmoothGrad (SG)**: Smooths the noisy results of basic gradients by averaging gradients calculated with added Gaussian noise $\mathcal{N}(0, \sigma^2)$. This is computed as $\text{SG}(x) = \frac{1}{n}\sum_1^n \text{Gradient}(x + \mathcal{N}(0, \sigma^2))$ [22].
- **Gradient-Input (Ipt-Grad)**: Multiplies the basic gradient by the input to reduce noise, calculated as $\text{Ipt-Grad}(x) = x\frac{\partial S}{\partial x}$ [20].
- **Integrated Gradients (IG)**: Uses a baseline input x' and sums gradients over different input scalings, computed as $\text{IG}(x) = (x - x')\int_{\alpha=0}^{1} \frac{\partial S(x' + \alpha(x - x'))}{\partial x}$ [24].
- **Guided Backpropagation (GBP)**: Modifies backpropagation to set negative gradients to zero and multiplies gradients with the activation gradient. This ensures only positive gradients contribute to the saliency map [23].

– **GradCAM**: Calculates pixel saliency as the gradients of the classified concept passing into a convolutional layer, usually the final layer [18].
– **Guided GradCAM (GBP-GC)**: Combines GBP and GradCAM, producing a saliency map that preserves information from both methods [19].

To evaluate the quality of the saliency mapping method, weights of layers are re-initialised to their untrained (random) values. Layers are randomised sequentially - starting from the first layer - and saliency maps calculated at each stage of randomisation. When comparing trained layers against randomised there should be a significant difference; if there is not, the saliency map method is not representative of the trained layers. To compare we use Spearman rank correlation using the magnitude of each pixel where the smallest value is 1 and the largest n, where n is the number of values. The difference between the flattened maps d is then calculated and the correlation ρ is calculated as in Eq. 1. In Sect. 4 we report the correlation at each stage of randomisation as the mean of the correlations across 100 randomly selected test instances.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

4 Results and Discussion

4.1 Model Evaluation

The evaluation metrics for the four models are shown in Table 1 alongside the metrics for the CNN trained on completed sketches, extracted from the larger models. The highest average F1 score was produced by fusion using hard coded values (in bold).

The models are all highly accurate, producing results over 95% in all metrics. Temporal features are clearly useful in classification, improving F1 from 95.89% to at least 96.69%. This trend is matched in the F1 scores for clocks specifically.

Going forwards the CNNs trained in the fusion using hard coded values model will be used due to their higher F1 value.

Table 1. Evaluation Metrics for the Trained Models

Model	Accuracy (%)	Average F1 (%)	Clock F1 (%)
Hard Coded Values	**96.794**	**96.797**	**98.443**
Fully Connected Layer	96.778	96.779	98.395
LSTM	96.767	96.767	98.348
Early Fusion	96.689	96.693	98.100
Without Temporal Information	95.889	95.891	97.571

4.2 Saliency Method Evaluation

Each saliency map method was applied to a random set of 100 instances from the test data using the five CNNs at differing stages of completion. A comparison of saliency maps generated using 100% completed sketches is shown in Fig. 4. The impact randomisation had on the saliency maps across all five CNNs is plotted in Fig. 5.

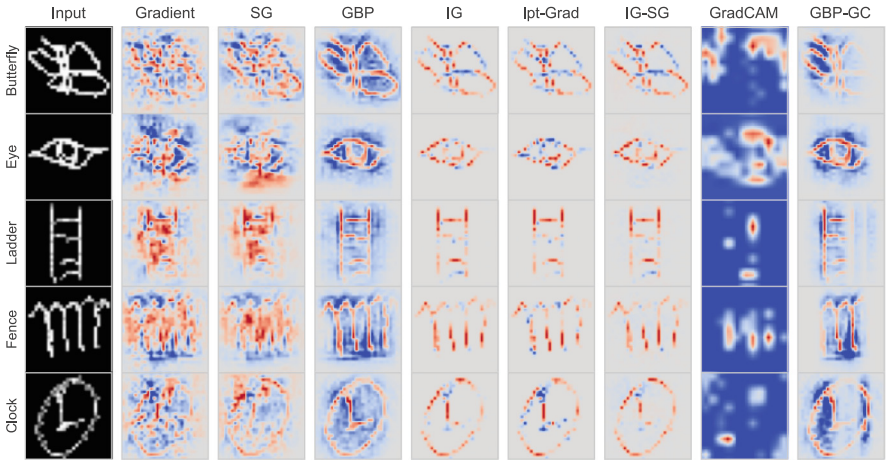

Fig. 4. Examples of all saliency map types used (described in Section 3.4). Red shows positive importance and blue shows negative importance

All saliency map generation methods had a lowered correlation upon randomisation as shown on Fig. 5. This demonstrates that the learned values within the models were having an impact on the results of the saliency map generation methods.

The highest correlation was produced by GBP with a mean correlation of 0.562 and a standard deviation of 0.0558. This indicates that the produced saliency map mostly consists of information irrelevant to the model's reason for classifying an image, instead producing values akin to edge detection. The combination of GBP and GradCam, GBP-GC, produced similar, though slightly better, results.

All gradient based approaches, Gradient, SG, IG, Ipt-Grad, IG-SG, produced reasonably high quality results in terms of correlations. Visually some of these are quite explanatory, for example in Fig. 4 they highlight where the minute hand shows a gap to the circle which is how a clock might appear. The best method, with the lowest correlation, was IG-SG with a mean of 0.00956 and a standard deviation of 0.0599. In its basic form, IG produces a higher correlation than Gradient, however, smoothing (SG) has a positive effect. This may be because

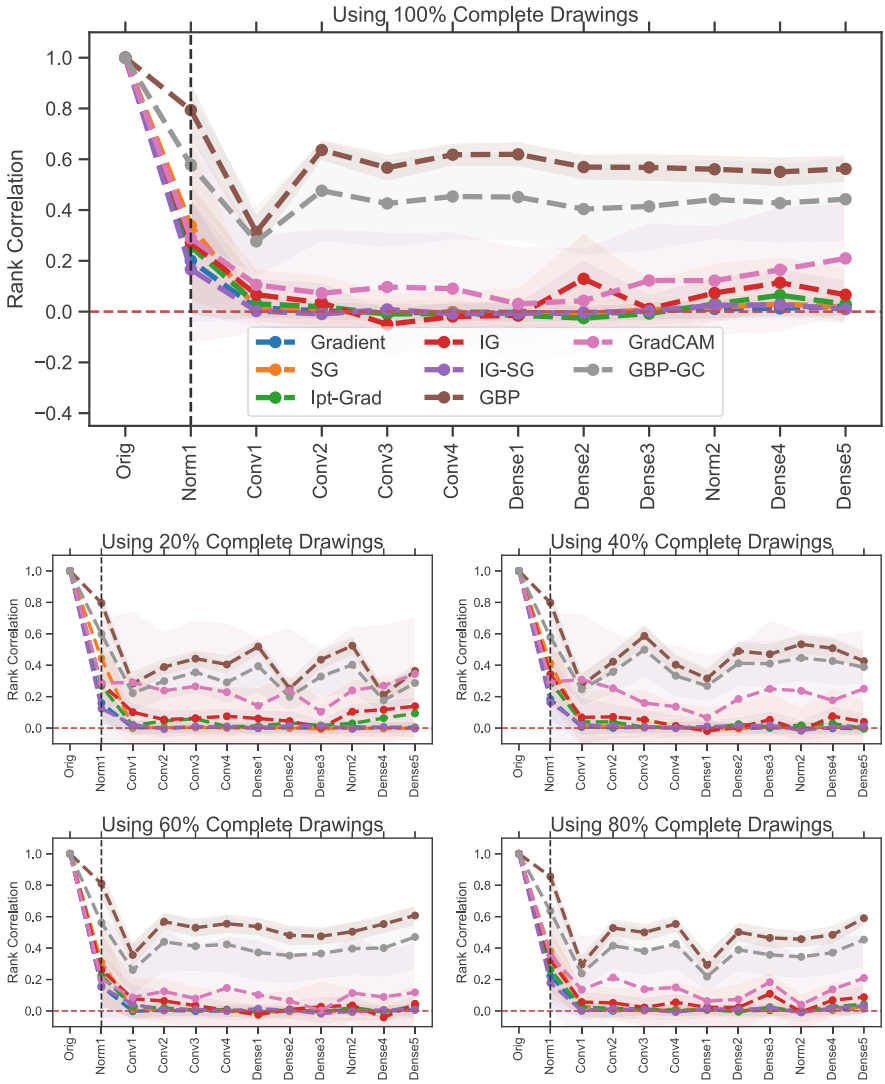

Fig. 5. Charts showing the rank correlations for each saliency map method (calculated as the average rank correlation using 100 test instances) at each step of cascading randomisation. At Orig all weights are preserved, at Dense5 all have been randomised. Charts are shown for the five CNNs at differing stages of sketch completion (Color figure online)

multiplying by the input sketch, as is done with IG, ensures that only the pixels of the saliency map laying atop the drawn lines show saliency information.

These results are relatively consistent across all five models trained at differing stages of completion with some minor differences. Models using less completed sketches tended to find GradCAM less useful. As shown in Fig. 5 (shown

in pink), the rank correlation increases from 0.220 after randomisation using 100% of sketches to 0.344 when using 20%, making it worse than GBP which decreased from 0.562 to 0.364. This may simply be a result of these images having less detail, therefore, any CNN trained may not be able to learn enough from the data. This could also be from GradCAM generating a heatmap that is not representative of the CNN predictions more often. It should be noted that the gradient based methods do produce slightly lower correlations using the less completed sketches for example, Gradient decreases from 0.021 to 0.004. Since, by nature of taking gradients from a convolutional layer, GradCAM does not detect these edges it may not be enhanced by having less sketch data. Regardless, IG-SG still produces the smallest correlation (-0.0031) with SG producing the second smallest (-0.000051).

5 Conclusion

The CDT plays an important part in identifying CD; it can be automated with a high degree of accuracy [2]. However, due to the high-risk nature of this domain deployed models need to be human-interpretable.

We created a model to automate the CDT utilising all information available within the medium of computer recorded sketches, including temporal information, then evaluated a series of eight saliency map methods for how informative they are.

Among various methods for fusing CNNs trained on sketches at five stages of completion, the best combination used a set of hard-coded weights. This seems unusual as theoretically using a fully connected layer should learn whatever values are most optimal; however this result is consistent with existing literature. The results produced an accuracy and average F1 score over 96% with the clock class specifically having an F1 score over 98%. The temporal information increased performance with all models incorporating temporal information producing a higher accuracy, average F1, and clock F1 than the model without.

Our saliency map results find GBP to be heavily influenced by the shape of the image, not providing useful information. This result matches those found in the literature. Gradient based approaches using SG, most notably IG-SG, produce the most informative results.

Future work can focus on the ability for saliency maps to highlight known features within drawings that may be helpful to clinicians, including for example temporal aspects of sketch drawing.

Acknowledgments. This work was supported by the Engineering and Physical Sciences Research Council and AgriFoRwArdS CDT [EP/S023917/1]

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adebayo, J., (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
2. Amini, S., et al.: An artificial Intelligence-Assisted method for dementia detection using images from the clock drawing test. J. Alzheimers Dis. **83**(2), 581–589 (2021)
3. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol. Artif. Intell. **3**(6), e200267 (2021). https://doi.org/10.1148/ryai.2021200267
4. Busse, A., Hensel, A., Gühne, U., Angermeyer, M.C., Riedel-Heller, S.G.: Mild cognitive impairment. Neurology **67**(12), 2176–2185 (2006). https://doi.org/10.1212/01.wnl.0000249117.23318.e1
5. Davoudi, A., et al.: Phenotyping cognitive impairment using graphomotor and latency features in digital clock drawing test. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5657–5660 (2020). https://doi.org/10.1109/EMBC44109.2020.9176469
6. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr. Cardiol. Rep. **16**(1), 441 (2013). https://doi.org/10.1007/s11886-013-0441-8
7. Ehreke, L., et al.: Is the clock drawing test appropriate for screening for mild cognitive impairment? Results of the German study on ageing, cognition and dementia in primary care patients (Agecode). Das Gesundheitswesen **72** (2010). https://doi.org/10.1055/s-0030-1266578
8. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. Graph. **31**(4), 1–10 (2012). https://doi.org/10.1145/2185520.2185540
9. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digital Health **3**(11), e745–e750 (2021). https://doi.org/10.1016/S2589-7500(21)00208-9
10. Hale, J.M., Schneider, D.C., Mehta, N.K., Myrskylä, M.: Cognitive impairment in the U.S.: Lifetime risk, age at onset, and years impaired. SSM Popul Health **11**, 100577 (2020)
11. Jongejan, J., Rowley, H., Kawashima, T., Kim, J., Fox-Gieg, N.: The quick, draw! (2016). https://quickdraw.withgoogle.com/
12. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 1–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
13. Pearson, C., De La Iglesia, B., Sami, S.: Detecting cognitive decline using a novel doodle-based neural network. In: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 99–103 (2022). https://doi.org/10.1109/MetroXRAINE54828.2022.9967549
14. Pinto, E., Peters, R.: Literature review of the clock drawing test as a tool for cognitive screening. Dement. Geriatr. Cogn. Disord. **27**(3), 201–213 (2009). https://doi.org/10.1159/000203344
15. Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Sci. Rep. **14**(1), 6086 (2024). https://doi.org/10.1038/s41598-024-56706-x

16. Rajpurkar, P., et al.: CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning (2017)
17. Seddati, O., Dupont, S., Mahmoudi, S.: Deepsketch 3. Multimedia Tools Appl. **76**(21), 22333–22359 (2017). https://doi.org/10.1007/s11042-017-4799-2
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
19. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? (2017)
20. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (2017). https://proceedings.mlr.press/v70/shrikumar17a.html
21. Shulman, K.I.: Clock-drawing: is it the ideal cognitive screening test? Int. J. Geriatr. Psychiatry **15**(6), 548–561 (2000)
22. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise (2017)
23. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net (2015)
24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017)
25. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. **32**(11), 4793–4813 (2021). https://doi.org/10.1109/TNNLS.2020.3027314
26. Wu, X.: An efficient antialiasing technique. SIGGRAPH Comput. Graph. **25**(4), 143–152 (1991). https://doi.org/10.1145/127719.122734