Blueprint of Tomorrow: Contrasting Off-line and On-line Drawing Tasks for Alzheimer's Disease Screening

Nina Hosseini-Kivanani[1], Elena Salobrar-García[2,3], Lorena Elvira-Hurtado[3,4], Mario Salas[5], Christoph Schommer[1], and Luis A. Leiva[1]

[1] University of Luxembourg, Luxembourg
[2] Ramon Castroviejo Institute of Ophthalmologic Research, Spain
[3] Universidad Complutense of Madrid, Spain
[4] Memory Unit, Geriatrics Service, Hospital Clínico San Carlos, Spain
nina.hosseinikivanani@uni.lu, elenasalobrar@med.ucm.es, marelvir@ucm.es, mario.salas@salud.madrid.org, christoph.schommer@uni.lu, luis.leiva@uni.lu

Abstract. Alzheimer's disease (AD) is the leading cause of dementia. Although there is currently no cure for AD, early detection of cognitive decline can help clinicians mitigate its impact. Recently, Machine Learning (ML) approaches have been developed to automatically analyze handwriting and hand-drawing tasks to support the early diagnosis of AD. In this paper, we study pentagon and clock drawing tests using both off-line (scanned image pixels) and on-line (discrete point sequences) data as input to several ML models (i.e., DensNet, ResNet, EfficientNet, RNN, LSTM, and GRU). Our study is the first to determine the most effective modality (on-line vs. off-line) and drawing tasks to distinguish healthy controls from AD patients (binary classification) as well as two stages of AD severity (multi-class classification). Our results suggest that, contrary to other domains, the off-line modality outperforms the on-line one, sometimes by a large margin: 90% vs. 60% accuracy in binary classification and 53% vs. 82% accuracy in multi-class classification. This suggests that, for drawing tasks and small-scale datasets, image-based representations may be more effective in predicting AD than those relying on more complex data representations.

Keywords: Alzheimer disease · Off-line handwriting · On-line handwriting · Deep learning · Data augmentation · Classification

1 Introduction

Neurodegenerative disorders are among the primary causes of disability worldwide, marked by an irreversible loss of neurons that culminates in progressive neurological decline, manifesting as motor and cognitive impairments. Alzheimer's disease (AD) and Parkinson's disease (PD) are particularly notable for their widespread prevalence, which affects approximately 50 million and 10 million individuals, respectively. AD, in particular, is closely linked with cognitive deficits,

affecting memory, attention, language comprehension, and spatial awareness [32]. The global demographic trend toward an older population underscores the critical need for early AD detection and intervention as the most prevalent form of dementia. However, the present diagnostic landscape reveals a concerning trend, with an estimated 75% of dementia cases worldwide going undiagnosed and rates of early-stage detection being considerably lower [10]. Enhancing screening methodologies in accessible settings, particularly in primary healthcare, emerges as a strategic response to improve diagnostic rates [8]. Research indicates that primary healthcare practitioners face substantial challenges in the early detection of dementia and in the timely referral of patients to specialized care [10]. Thus, the development and implementation of accessible and efficient screening tools for use in primary healthcare or by individuals at home are crucial steps toward closing the diagnostic gap, potentially leading to improved detection rates of AD.

Handwriting and hand-drawing tasks[5] entail the coordination of fine motor movements and cognitive processes, making them popular as a psychometric tool to evaluate and diagnose AD [35], leveraging the correlation between declining drawing abilities and the onset of AD. Deterioration in handwriting skills, characterized by inconsistencies in size, spacing, and letter formation, indicates a progression of the disease [7]. Recent studies (e.g., [17, 23]) have shown the potential of handwriting-related tasks to reveal specific cognitive deficits indicative of AD. Researchers have explored various automated methods, including drawing tasks [17], neuroimaging [31], and gait assessments [11], to capture cognitive impairments across multiple domains. However, the current need for healthcare professionals' reliance on manual analysis highlights a significant bottleneck. This puts forward the importance of developing automated tools to make the AD screening process easier, quicker, and more affordable, particularly in non-specialist settings.

Our contributions are straightforward yet significant. Firstly, we gathered handwriting data from both AD patients and healthy individuals. The data includes two types of drawings—pentagons and clocks—captured *simultaneously* in two formats: off-line, as scanned images, and on-line, as sequences of discrete points. This dual-method collection allows us to: (i) Compare how different hand-drawing tasks perform in classification tests, (ii) Assess various machine learning models to see which best classifies the data, and (iii) Examine the differences in using static images versus dynamic, point-by-point data in model performance.

Secondly, our experiments focus on distinguishing AD (mild AD and moderate AD) patients from healthy controls using state-of-the-art neural network models. We used pre-trained convolutional neural networks (CNNs) for analyzing the off-line data and recurrent neural networks (RNNs) for the on-line data. Additionally, we used data augmentation techniques to enhance the models' ability to generalize. Interestingly, our findings reveal that the off-line modality consis-

[5] We consider 'handwriting' and 'hand-drawing' synonymous because both tasks involve the same neurophysiological and peripheral processes involved in motor control.

tently outperformed the on-line one, achieving higher accuracy in both binary (90% versus 60%) and multi-class (53% versus 82%) classification tasks. This suggests that simpler, static image-based approaches may be more effective for tasks like drawing analysis in AD research than those relying on more complex, temporal data, at least when working with small-scale datasets.

2 Related Work

Technological advancements in ML and computer vision have significantly enhanced the efficiency and objectivity of remote patient monitoring systems by providing real-time data for improved care beyond conventional healthcare settings [6]. Handwriting analysis has been shown to be effective in detecting cognitive decline and changes in motor skills in AD, thus serving as an effective diagnostic tool [19]. However, despite these advances, there is no research aimed at comparing model performance using *both* off-line and on-line data from *identical* patient/healthy cohorts [5, 30].

Recent studies have used CNNs (e.g., [18, 19, 3]) and RNNs (e.g., [18, 2, 6]) for early detection of AD, showing that Deep Learning (DL) models can significantly enhance the accuracy of diagnosing AD in its early stages. We can find notable works that studied each modality separately (cf. off-line [17, 15, 3] and on-line [6, 23, 21]), suggesting that on-line data are preferred over off-line data, given that on-line handwriting provides a feature-rich representation, including, e.g., temporal and spatial sequences of discrete points that are not available in the off-line representation. Collectively, these studies highlight the potential of ML technologies not only to revolutionize AD diagnosis but also to enable more personalized and timely therapeutic interventions, ultimately improving patient outcomes. Most of these studies have relied on some form of data augmentation, given the limited number of samples in clinical datasets. In this regard, Dao et al. [6] used Generative Adversarial Networks (GANs) as an alternative to data augmentation, and trained AD classifiers with RNNs that achieved 89% accuracy. However, GANs require a significant amount of data to begin with, which is often not available in most cases.

Finally, we should mention relevant studies that have compared different drawing symbols for AD screening, such as clocks drawngs [1] using CNN models, which achieved an AUC score of 81%. By combining clock drawing with age and education using logistic regression, their model improved to 91%. Pentagon drawings [26] reached an accuracy of 93% using GoogLeNet for binary classification, distinguishing between correct and incorrect pentagon drawings from patients only. Another study focused on letters [6], and obtained high accuracy by using DL models for detecting and classifying early-stage of ADs patients based on on-line handwriting loop patterns. In sum, it remains unclear which is the most adequate input modality for AD screening and also what the most adequate drawing symbols are to achieve competitive performance.

By systematically addressing this gap in the research literature, our study paves the way for a more holistic understanding of AD classification models,

opening promising directions to more accurate AD screening approaches in the future. For example, [23] reported that combining multiple drawing tasks improves detection accuracy by capturing different cognitive impairments, achieving a classification accuracy of 75.2%.

3 Materials and Methods

3.1 Participants

Thirty-three individuals were recruited from the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid between January 2023 and January 2024. The group consisted of 22 patients and 11 healthy controls (HCs), all aged between 70 and 89. Participants were asked to both clocks and pentagons, which are well-established symbols in cognitive assessment tasks [9]. All participants had normal vision and hearing. They underwent a neuropsychological assessment of their drawing tasks in a clinical setting to minimize distractions and reduce background noise. Each participant was individually assessed, beginning with an informed consent form. Cognitive status was evaluated using the Mini-Mental State Examination (MMSE) [39]. Patients with AD were classified into mild AD and moderate AD, based on guidelines from the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), the Alzheimer's Disease and Related Disorders Association (ADRDA) workgroup [28], and the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [13]. Statistical analysis (ANOVA test, after verification of normality and homoscedasticity) confirmed that there were no significant age differences between the healthy, mild AD, and moderate AD groups ($F(2, 55) = 2.04, p > .139$).

Table 1. Demographics and the Number of Drawing Tasks of this Study.

Drawing Task	**Num. of drawing**	**HC**	**Mild AD**	**Moderate AD**
Pentagon	33	11	8	14
Clock	33	11	8	14
Total	**66**	**22**	**16**	**28**
Gender (F & M)		8F, 3M	5F, 3M	12F, 4M
Age (Mean ± SD)		82.64±2.46	76.5±5.75	78.94±4.78
MMSE (Mean ± SD)		29.9±0.83	25.33±1.21	22.37±3.58

3.2 Drawing tasks and preprocessing

Participants were instructed to draw the pentagons and clocks using a Repaper tablet (size: 10.9-inch) [6] with a blank sheet attached and a regular pen that had

[6] https://www.iskn.co/eu

an accelerometer connected to the Repaper app via Bluetooth for data capturing (see Figure 1). This setup was designed to provide a familiar pen-and-paper experience to participants while being able to capture on-line and off-line data simultaneously. The participants were asked to draw each symbol from memory. We collected 66 drawings in total.

The on-line data (discrete point sequences) were stored as SVG files (as per the Repaper app) and then converted to JSON format, comprising sequences of `{x,y,t}` points. The off-line data (image pixels) were stored as PDF files (scanned with the HP Color LaserJet Pro scanner) and then converted to PNG images and resized to square size (224×224 px) as this is standard for CNN models. We applied the canny edge detector to enhance the quality of the scanned images.

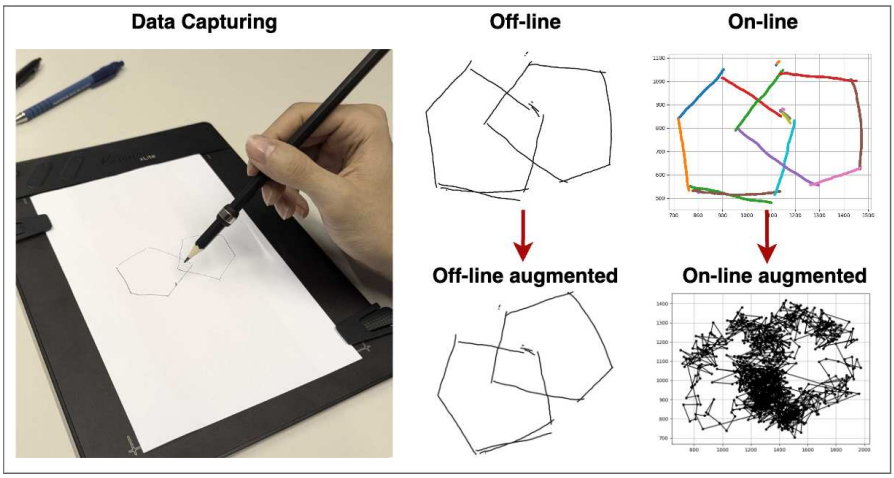

Fig. 1. Example of pentagon drawing on a tablet (left) and drawing samples in off-line and on-line version (right), before and after applying data augmentation.

3.3 Data augmentation

We created synthetic samples to make the models more robust and generalizable [12]. For off-line data, we applied the usual geometric transformations, where suitable:[7] For on-line data, jittering, scaling, and warping have been proposed [22]; however, more recently, Maslych et al. [27] found that an "All Variability Chain" (AVC) of transformations (gaussian, frame-skip, spatial, perspective, rotate, scale) provided a significant boost in classification performance with

[7] For example, pentagons can be flipped in horizontal or vertical axes, whereas clocks cannot be flipped because it would destroy their semantics.

RNNs, achieving state-of-the-art accuracy in gesture recognition. Therefore, we adopt the AVC approach to augment our on-line data; see Figure 1. After data augmentation, we concluded to a dataset consisting of 300 images (off-line data) and 300 point sequences (on-line data), where all groups were balanced up to 100 observations; e.g., 92 variations of the 8 pentagons from the Mild AD group were created; see Figure 2.

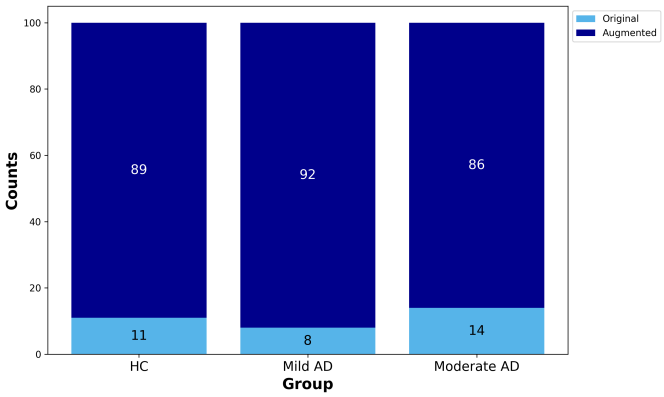

Fig. 2. Class distribution of Pentagon drawings before and after data augmentation.

To evaluate the quality of the augmented data, we used the Structural Similarity Index Measure (SSIM) [41] for off-line data and Dynamic Time Warping (DTW) [34] for on-line data. We use SSIM to compare augmented data against original images to ensure that key structural details are preserved even as variations are introduced. This balance is critical since, while the augmented data are inherently different, maintaining structural similarity ensures that the variations remain realistic and relevant for training robust models. By using SSIM, we can confirm that the augmentation process does not distort the data to the extent that it loses its representative characteristics. In this sense, structural similarity is beneficial, as it ensures that the augmented data faithfully represents the original data. SSIM values range from 0.7 to 0.8 (M=0.75, SD=0.03), whereas DTW values range between 123 and 5678 (M=2000, SD=850), indicating that augmented images are not near-duplicates of the original data but rather new images that eventually should help improving model performance.

3.4 Models

Convolutional Neural Nets We use three state-of-the-art pre-trained CNNs for analyzing the off-line data: ResNet50 [14], DenseNet121 [20], and EfficienNet [37]. ResNet and DenseNet use residual connections, which are instrumental to train very deep models. While ResNet performs an element-wise addition to

pass the output to the next layer, DenseNet connects all layers directly to each other through concatenation. However, EfficienNet uses a uniform compound scaling technique that achieves the same performance as state-of-the-art CNNs but with much better efficiency. These CNNs were trained on the large ImageNet dataset, and we fine-tuned them to our AD dataset by transfer learning [42].

Recurrent Neural Nets Since transfer learning for on-line data is not currently possible, as there are no public pre-trained models available, we train three RNNs from scratch: Vanilla RNN, LSTM [16], and GRU [4]. LSTM is an improvement over vanilla RNNs by adding long-term memory, making them ideal for complex sequences. GRU is a simplification of LSTM while retaining the same performance, making them ideal for cases where computational efficiency is crucial. These RNN models include a hidden layer of 100 units with hyperbolic tangent activation and 0.1 dropout, followed by a softmax output layer. We experimented with other combinations of layers and different hidden units, but we did not observe improvements with regard to this configuration.

Training and evaluation All CNNs and RNNs are trained with the popular Adam optimizer, with a learning rate of 0.001 and decay rates $\beta_1 = \beta_2 = 0.99$. The loss function is categorical cross-entropy, consistent with our binary and multi-class classification tasks. All models use a batch size of 32 (images or sequences) and use up to 50 epochs for training with early stopping (patience of 10 epochs) to prevent overfitting. We train each model on 80% of the data and test on the remaining 20% of both on-line and off-line data. We then used stratified 5-fold cross-validation on the training set only, ensuring that each fold was representative of the whole by maintaining approximately the same percentage of samples of each class as in the training subset. We computed the classification accuracy (Acc) and Area Under the ROC curve (AUC) to assess model performance.

4 Results and discussion

Figure 3 summarizes the results of our experiments, depicting the differences between CNN classifiers (top row) for off-line data (ResNet, DenseNet, and EfficienNet) and RNN classifiers (bottom row) for on-line data (vanilla RNN, LSTM, and GRU). These results are instrumental for assessing the efficacy of binary (distinguishing AD patients from HC) and multi-class classification (distinguishing HC, mild AD, and moderate AD) tasks in the context of AD screening using hand-writing tasks (i.e., pentagon and clock).

In binary classification of pentagon drawings, the before-augmentation scenario showed modest performance across models. After-augmentation, however, there was an increase in classifier accuracy, particularly with EfficienNet, which improved substantially, from 60% to 90%. A similar trend was observed in

the classification of clock drawings, where the after-augmentation results underscored the effective impact of our data augmentation strategies on model performance.

For multi-class classification, the differential impact of data augmentation was again significant. EfficientNet was the best performer, especially for pentagon drawings, confirming the model's ability to differentiate between various stages of AD severity. The trend of improved after-augmentation accuracy was consistent across other models, although the degree of improvement varied.

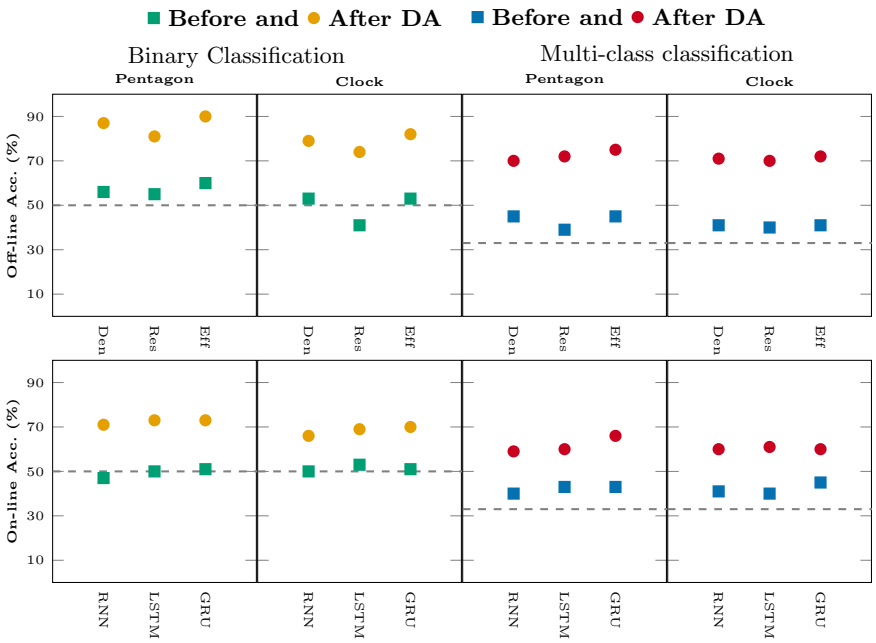

Fig. 3. Experiment results. CNN classifiers for off-line data (Res: ResNet, Den: DenseNet, Eff: EfficienNet) are depicted in the top row, whereas RNN classifiers for on-line data (vanilla RNN, LSTM, GRU) are depicted in the bottom row. The dashed lines represent the performance of a random classifier, illustrating the empirical lower bound.

Our results indicate that off-line data, when enhanced through strategic data augmentation, provides a more stable and consistent basis for AD classification compared to on-line data. This stability can be attributed to the static nature of off-line data, which, unlike on-line data, is less affected by the variabilities introduced by the temporal and dynamic components of on-line drawing. In non-clinical domains, researchers have shown that on-line data is preferred over off-line data (e.g., [38, 24]), given the rich patterns and movement dynamics involved [25]. In our experiments, however, we observed that pre-trained CNNs

outperformed RNNs trained from scratch due to the size and variability in on-line data that our RNNs were not able to capture as effectively as the CNNs. This variability also seems to affect tasks differently; for example, more cognitively demanding drawing tasks allowed for an easier distinction of AD patient handwriting from HCs; see e.g., Figure 3.

Our study builds upon previous research that focused on one drawing type for AD screening [19, 3]. However, our findings suggest that both pentagon and clock drawings are suitable for AD screening, with EfficientNet achieving the highest performance in binary classification (90% accuracy, 92% AUC) and in multi-class classification settings (75% accuracy, 79% AUC) for using pentagons, followed by clocks (Binary classification: 82% accuracy and 79% AUC, multi-class classification: 70% accuracy and 75% AUC), highlighting the effectiveness of these simple tasks.

Our results also indicate that the performance gap before and after data augmentation differs across tasks, with larger differences in off-line data. Previous results by Maslych et al. [27] reported improved performance of an RNN model in several handwriting tasks using AVC for data augmentation, although all the tasks were focused on gesture recognition and a specific dataset. In any case, it seems clear that data augmentation is necessary for both off-line and on-line drawing tasks.

Without data augmentation, most models behaved like a random classifier; see the dashed lines in Figure 3. After data augmentation, however, we observed a significant improvement in model performance across all tasks for both binary and multi-class classification scenarios. These improvements were more apparent for CNN models, which somehow disagrees with previous findings in AD screening that reported similar performance for RNNs [2, 29].

Interestingly, Souillard-Mandar et al. [36] found that the digital Clock Drawing Test had superior diagnostic performance compared to traditional paper-and-pencil methods for differentiating healthy individuals from cognitive impairment subjects (only binary classification), using traditional ML models without any data augmentation. Previous studies have demonstrated the effectiveness of various augmentation strategies in similar contexts for off-line data [17, 33]. In our work, we designed and optimized our augmentation strategy to ensure its suitability for our specific datasets. This involved iterative testing and refinement of different augmentation techniques, such as geometric transformations.

5 Limitations and future work

The main limitation of our study is the small sample size of the original dataset, which had to be augmented with suitable variations in order to fine-tune the CNN models and, more importantly, to train the RNN models from scratch. However, we should remark that dealing with small sample sizes is a well-known and pervasive issue among clinical studies [21, 40]. Recruiting participants with AD is very challenging due to strict criteria and ethical concerns. Despite our best efforts, it took us one year to recruit 33 suitable participants. On the other hand,

future work should consider different data augmentation approaches for on-line trajectories, since we have observed that the AVC method [27] is suboptimal for AD screening. Additionally, an interesting direction for future research could be assessing whether simpler, custom-built CNN models can achieve comparable or superior results with reduced complexity.Ultimately, despite these shortcomings, our findings show promise and could lead to practical clinical applications.

6 Conclusion

We have analyzed the impact of off-line vs. on-line handwriting data for AD screening using pentagon and clock drawings with suitable data augmentation techniques. We trained several CNNs and RNNs for binary and multi-class classification settings. Our results show that data augmentation is always beneficial and that pentagons have better discriminative power than clocks. Our results also show that CNNs outperform RNNs in all settings, contradicting what was previously known in non-clinical work.

Our observed improvements in performance suggest that while our current strategy enhances model robustness, further optimizations could indeed yield even better results, a possibility that we aim to explore in future work. We acknowledge that there are numerous opportunities for additional refinement, and future work will continue to explore and optimize these techniques to maximize their efficacy in enhancing model robustness and accuracy. Our code and models are available upon reasonable request.

Acknowlegments

Work supported by the UCM research group (Grupo de investigacion básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

References

1. Amini, S., Zhang, L., Hao, B., Gupta, A., Song, M., Karjadi, C., Lin, H., Kolachalama, V.B., Au, R., Paschalidis, I.C.: An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. J. Alzheimers Dis. **83**(2) (2021)
2. Bensalah, A., Parziale, A., De Gregorio, G., Marcelli, A., Fornés, A., Lladós, J.: I can't believe it's not better: In-air movement for alzheimer handwriting synthetic generation. In: Proc. IGS (2023)
3. Chen, S., Stromer, D., Alabdalrahim, H.A., Schwab, S., Weih, M., Maier, A.: Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. Scientific Reports **10**(1) (2020)
4. Cho, K., Merrienboer, B.V., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proc. EMNLP (2014)

5. Cilia, N.D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., Parziale, A.: Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. J. Eng. Appl. Artif. Intell. **111** (2022)
6. Dao, Q., El-Yacoubi, M.A., Rigaud, A.S.: Detection of Alzheimer Disease on Online Handwriting Using 1D Convolutional Neural Network. IEEE Access **11** (2023)
7. Delazer, M., Zamarian, L., Djamshidian, A.: Handwriting in Alzheimer's Disease. J. Alzheimer's Disease **82**(2) (2021)
8. Eichler, T., Thyrian, J.R., Hertel, J., Michalowsky, B., Wucherer, D., Dreier, A., Kilimann, I., Teipel, S., Hoffmann, W.: Rates of formal diagnosis of dementia in primary care: The effect of screening. Alzheimers Dement. Diagn. Assess. Dis. Monit. **1**(1) (2015)
9. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. **12**(3) (1975)
10. Gauthier, S., Rosa-Neto, P., Morais, J.A., Webster, C.: World Alzheimer Report 2021: Journey through the diagnosis of dementia. Alzheimers Dis. Int. **2022** (2021)
11. Ghoraani, B., Boettcher, L.N., Hssayeni, M.D., Rosenfeld, A., Tolea, M.I., Galvin, J.E.: Detection of mild cognitive impairment and alzheimer's disease using dual-task gait assessments and machine learning. Biomed. Signal Process. Control **64** (2021)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
13. Guha, M.: Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). Reference Reviews **28**(3) (2014)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. CVPR (2016)
15. Higaki, Y.: Clock-Drawing Test and Cube-Copying Test to Quickly Screen Dementia: In Combination with the Mini-Mental State Examination Scores. Intern. Med. (2023)
16. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8) (1997)
17. Hosseini-Kivanani, N., Salobrar-García, E., Elvira-Hurtado, L., López-Cuenca, I., de Hoz, R., Ramírez, J.M., Gil, P., Salas-Carrillo, M., Schommer, C., Leiva, L.A.: Ink of insight: Data augmentation for dementia screening through deep learning. In: Proc. ICMHI (2024)
18. Hosseini-Kivanani, N., Salobrar-Gracía, E., Elvira-Hurtado, L., López-Cuenca, M., Schommer, C., Leiva, L.A.: Predicting alzheimer's disease and mild cognitive impairment with off-line and on-line house drawing tests. In: Proc. e-Science. IEEE (2024)
19. Hosseini-Kivanani, N., Schommer, C., Leiva, L.A.: The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images. In: Proc. Healthcom (2023)
20. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proc. CVPR (2017)
21. Impedovo, D., Pirlo, G.: Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. IEEE Rev. Biomed. Eng. **12** (2018)
22. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLOS ONE **16**(7) (2021)
23. Kobayashi, M., Yamada, Y., Shinkawa, K., Nemoto, M., Nemoto, K., Arai, T.: Automated Early Detection of Alzheimer's Disease by Capturing Impairments in Mul-

tiple Cognitive Domains with Multiple Drawing Tasks. J. Alzheimers Dis. **88**(3) (2022)
24. Leiva, L.A., Alabau, V., Romero, V., Toselli, A.H., Vidal, E.: Context-aware gestures for mixed-initiative text editing UIs. Interact. Comput. **27**(6) (2015)
25. Leiva, L.A., Diaz, M., Ferrer, M.A., Plamondon, R.: Human or Machine? It Is Not What You Write, But How You Write It. In: Proc. ICPR (2021)
26. Maruta, J., Uchida, K., Kurozumi, H., Nogi, S., Akada, S., Nakanishi, A., Shinoda, M., Shiba, M., Inoue, K.: Deep convolutional neural networks for automated scoring of pentagon copying test results. Scientific Reports **12**(1) (2022)
27. Maslych, M., Taranta, E.M., Aldilati, M., Laviola, J.J.: Effective 2D Stroke-based Gesture Augmentation for RNNs. In: Proc. CHI (2023)
28. McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M.: Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology **34**(7) (1984)
29. Mwamsojo, N., Lehmann, F., El-Yacoubi, M.A., Merghem, K., Frignac, Y., Benkelfat, B.E., Rigaud, A.S.: Reservoir Computing for Early Stage Alzheimer's Disease Detection. IEEE Access **10** (2022)
30. Müller, S., Preische, O., Heymann, P., Elbing, U., Laske, C., Popa-Wagner, A., Yu, J.: Increased Diagnostic Accuracy of Digital vs. Conventional Clock Drawing Test for Discrimination of Patients in the Early Course of Alzheimer's Disease from Cognitively Healthy Individuals. Front. Aging Neurosci. **11** (2017)
31. Odusami, M., Maskeliūnas, R., Damaševičius, R.: An intelligent system for early recognition of alzheimer's disease using neuroimaging. Sensors **22** (2022)
32. Perry, R.J., Hodges, J.R.: Attention and executive deficits in Alzheimer's disease: A critical review. Brain **122**(3) (1999)
33. Raksasat, R., Teerapittayanon, S., Itthipuripat, S., Praditpornsilpa, K., Petchlorlian, A., Chotibut, T., Chunharas, C., Chatnuntawech, I.: Attentive Pairwise Interaction Network for AI-Assisted Clock Drawing Test Assessment of Early Visuospatial Deficits (2023)
34. Senin, P.: Dynamic time warping algorithm review. Inf. Comput. Sci. Dep. Univ. Hawaii **855**(1-23) (2008)
35. Smith, A.D.: On the use of drawing tasks in neuropsychological assessment. Neuropsychology **23**(2) (2009)
36. Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D., Swenson, R., Price, C., Lamar, M., Penney, D.L.: Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. Mach. Learn. **102** (2015)
37. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proc. ICML (2019)
38. Tappert, C.C., Mosley, P.H.: Recent advances in pen computing. Tech. Rep. 166, Pace University (2001)
39. Tombaugh, T., McIntyre, N.J.: The Mini-Mental State Examination: A Comprehensive Review. J. Am. Geriatr. Soc. **40**(9) (1992)
40. Vessio, G.: Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review. Appl. Sci. **9**(21) (2019)
41. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. **13**(4) (2004)
42. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1) (2016)