Ink of Insight: Data Augmentation for Dementia Screening through Handwriting Analysis

NINA HOSSEINI-KIVANANI, Department of Computer Science, University of Luxembourg, Luxembourg
ELENA SALOBRAR-GARCÍA, LORENA ELVIRA-HURTADO, INÉS LÓPEZ-CUENCA, ROSA DE HOZ, and JOSÉ M. RAMÍREZ, Ramon Castroviejo Institute of Ophthalmologic Research, Universidad Complutense de Madrid, Spain
PEDRO GIL and MARIO SALAS-CARRILLO, Memory Unit, Geriatrics Service, Hospital Clínico San Carlos, Madrid, Spain
CHRISTOPH SCHOMMER and LUIS A. LEIVA, Department of Computer Science, University of Luxembourg, Luxembourg

We investigate the use of handwriting data as a means of predicting early symptoms of Alzheimer's disease (AD). Thirty-six subjects were classified based on the standardized pentagon drawing test (PDT) using deep learning (DL) models. We also compare and contrast classic machine learning (ML) models with DL by employing different data augmentation (DA) techniques. Our findings indicate that DA greatly improves the performance of all models, but the DL-based ones are the ones that achieve the best and highest results. The best model (EfficientNet) achieved a classification accuracy of 87% and an area under the receiver operating characteristic curve (AUC) of 91% for binary classification (healthy or AD patients), whereas for multiclass classification (healthy, mild AD, or moderate AD) accuracy was 76% and AUC was 77%. These results underscore the potential of DA as a simple, cost-effective approach to aid practitioners in screening AD in larger populations, suggesting DL models are capable of analyzing handwriting data with a high degree of accuracy, which may lead to better and earlier detection of AD.tempate

Additional Key Words and Phrases: Alzheimer's Disease; Screening; Pentagon Drawing Test; Data Augmentation; Image Classification; Machine Learning; Deep Learning

1 INTRODUCTION AND RELATED WORK

Alzheimer's disease refers to a dementia syndrome characterized by primary impairments of cortical cognitive functions, including memory, language, and praxis, that gradually progress over time [15]. These impairments have a high functional impact and are often accompanied by various neuropsychiatric symptoms [7]. As the disease progresses, the number of damaged neurons and the extent of affected brain regions increases, leading to a greater need for assistance from family members, friends, and professional caregivers for daily tasks [1]. The early stages of AD are characterized by memory loss, recognition problems (such as an object or face recognition [8]), visual impairments, and deficits in spatial perception [21], despite relatively normal visual acuity values and intact visual fields [23]. Recent research has shown that assessing visuospatial function, in addition to brain scanning, can aid in the early detection of impairments. Effective screening tests can identify visuospatial dysfunction, which may manifest years before the onset of clinical symptoms [18]. However, existing screening measures for cognitive changes face challenges, particularly with regard to their limited intra-individual reliability, which hinders accurate tracking of cognitive changes over time.

Authors' addresses: Nina Hosseini-Kivanani, nina.hosseinikivanani@uni.lu, Department of Computer Science, University of Luxembourg, Luxembourg; Elena Salobrar-García, elenasalobrar@med.ucm.es; Lorena Elvira-Hurtado, marelvir@ucm.es; Inés López-Cuenca, inelopez@ucm.es; Rosa de Hoz, rdehoz@ucm.es; José M. Ramírez, ramirezs@med.ucm.es, Ramon Castroviejo Institute of Ophthalmologic Research, Universidad Complutense de Madrid, Spain; Pedro Gil, pgil@salud.madrid.org; Mario Salas-Carrillo, mario.salas@salud.madrid.org, Memory Unit, Geriatrics Service, Hospital Clínico San Carlos, Madrid, Spain; Christoph Schommer, christoph.schommer@uni.lu; Luis A. Leiva, luis.leiva@uni.lu, Department of Computer Science, University of Luxembourg, Luxembourg.

Drawing tests, frequently used in dementia screening, can reflect the presence of the condition through changes in a person's drawing ability [22]. However, the subjectivity in scoring systems used in these tests and their limited scope in capturing a range of drawing attributes often result in missing subtle yet clinically significant indicators of cognitive decline. This means that no single scoring system is reported as the most effective and reliable evaluation method (e.g., [12]). This highlights the need for more comprehensive and objective screening methods. There is a growing interest in exploring more advanced analytical approaches, such as the integration of machine learning (ML) techniques, to augment the diagnostic effectiveness of cognitive screening tools.

Recent advancements in artificial intelligence (AI), particularly in deep learning (DL), have significantly impacted healthcare, especially when it comes to diagnosing neurodegenerative diseases like AD (e.g., [6, 12, 24]). DL models have played an instrumental role in the analysis of neuroimaging, detecting complex patterns in brain scans that are imperceptible to the human eye. Our study focuses on refining DL models for dementia screening and emphasizing the importance of data augmentation (DA) techniques in contexts with limited high-quality and diverse data. This approach is vital for improving model robustness, especially in applications like automated analysis of scanned paper-based handwriting and drawings, which are crucial in AD screening. Recent research has highlighted DL's transformative role in healthcare, particularly in the early detection and management of cognitive impairments [3, 6, 11, 14].

Relevant studies (e.g., [6, 11]) have highlighted the precision of DL models, particularly convolutional neural networks (CNNs). However, the effectiveness of these models is often limited by the small size of available datasets. Maruta et al. [19] demonstrated that the fine-tuned GoogleNet CNN outperforms other CNN models like VGG-16, ResNet-50, and Inception-v3 in automatically evaluating the pentagon drawings for constructional apraxia. Additionally, Tasaki et al. [28] conducted a study on the usage of a DL model called PentaMind. which analyzes hand-drawn images of intersecting pentagons to extract cognition-related features. The model was trained on 13,777 images and successfully extracted features such as line waviness, which shows improvement over conventional visual assessment methods. Jiménez-Mesa et al. [16] proposed a CNN-based method for diagnosing cognitive impairment through the Clock Drawing Test (CDT), effectively classifying drawings as healthy or patient, indicating its potential for hospital and clinic use, particularly in resource-limited areas. The use of DL in cognitive impairment tests is not without limitations, primarily due to the limited dataset sizes and variability. DA emerges as a pivotal solution to enhance model robustness and accuracy. It involves generating additional training data from existing datasets, increasing size, diversity, and quality. However, challenges exist in preserving clinical relevance and avoiding artificial biases.

Summary of Contributions

Our research builds upon significant advancements in ML for cognitive impairment screening, aiming to tackle the existing challenges. This brings us to the core objectives of our research. Firstly, we aim to develop robust DL Models for AD screening to refine and enhance the existing models. Secondly, our study focuses on the importance of DA in clinical settings, emphasizing the preservation of data integrity and reliability. Thirdly, we explore the comparative advantages of DL over classic ML in the context of AD screening, providing a comprehensive insight into the future of digital screening in AD.

2 MATERIALS AND METHODS

Our goal is to improve the performance of ML models in classifying handwriting data by implementing suitable DA techniques. Although DA has shown advantages in other scientific domains, its application to handwriting data in clinical contexts has received little attention. This is primarily because the augmented data is often either too similar

to the original data or too distorted for the models to learn effectively from it (e.g., [25]). This study compares classic ML models (SVMs, RFs, k-NNs) and DL models (CNNs) in the context of classifying binary (healthy vs. patient) and multiclass (healthy, mild AD, and moderate AD) classification tasks, both with and without applying DA.

2.1 Data Collection and Tasks

The study recruited 36 subjects (13 female and 23 male) from the Memory Unit of the Hospital Clinico San Carlos (HCSC) for a study on cognitive and neurophysiological characteristics of individuals at high risk of dementia. Subjects were categorized according to the guidelines of the National Institute of Neurological and Communicative Disorders and Stroke-AD and Related Disorders Association (NINCDS-ADRDA) [20] and the Statistical Manual of Mental Disorders V (DSM V) [9]. Based on these guidelines, the subjects were classified into two groups of patients (mild AD, n=3, and moderate AD, n=3) and one group of healthy subjects (control, n=30). All the subjects provided informed consent prior to participation. The participants' ages ranged from 61 to 88 years old, with a mean age of 73.92 ± 6,78 years old. No significant differences in age were observed among the healthy group, mild AD, or moderate AD based on p-value > .05. The study included 30 individuals aged between 61 to 84 years who were cognitively healthy with no evidence of brain injury and had MMSE score above 26. Non-healthy participants had Mini-Mental State Examination (MMSE) scores between 25 and 17.

2.2 Image Preprocessing and Data Augmentation

Participants were given a blank A4-sized paper and asked to copy a figure of two overlapping pentagons with an interlocking shape (as shown in Figure 1a). The paper-and-pencil drawings were converted from PDF files to image format (PNG format) (Figure 1a and b) to be processed with our classic ML and DL models. The PNG images were then converted from color images (three channels) to grayscale (one channel) (Figure 1c). The resulting images were resized to standard dimensions (224×224 px). Any nonrelevant information, such as the original printed images from the clinicians, that appeared on the top side of the original file (Figure 1a), was removed during the preprocessing pipeline. Finally, images were padded to remove noise from the image and make them in the same shape, and the canny edge detector from OpenCV library [4] was used to improve the resulting image (Figure 1d). Low-quality and noisy images (in total, 14 images from the healthy group) were manually filtered out.

ML (and, particularly, DL) models perform better when trained on large datasets; however, obtaining such large-scale datasets is really challenging in clinical fields. To address this, DA techniques can be used to artificially increase the size of the dataset. By generating additional images from the input images, these techniques can help reduce the risk of overfitting and increase the model's generalizability, leading to better overall performance. These techniques include applying geometric transformations (such as flipping, cropping, rotating, and translating), changing the color space of the images, mixing images, or even using generative adversarial networks [25]. In this study, we only applied geometric transformations to images for DA, carefully avoiding transformations that would potentially destroy the semantics of the original image and are not suitable for our grayscale handwritten images. Therefore, techniques commonly used in broader computer vision applications, such as hue adjustments or color inversion, were deliberately excluded from our process. Our approach was to maintain the integrity of the original handwritten samples, ensuring that the essential characteristics of these images were preserved.

To determine the quality of the augmented data, we used the structural similarity index (SSIM) [29]. SSIM measures the similarity between two images by considering the human visual perception of differences in terms of luminance, contrast, and structure. SSIM is a widely used measurement tool because of its low computational complexity and

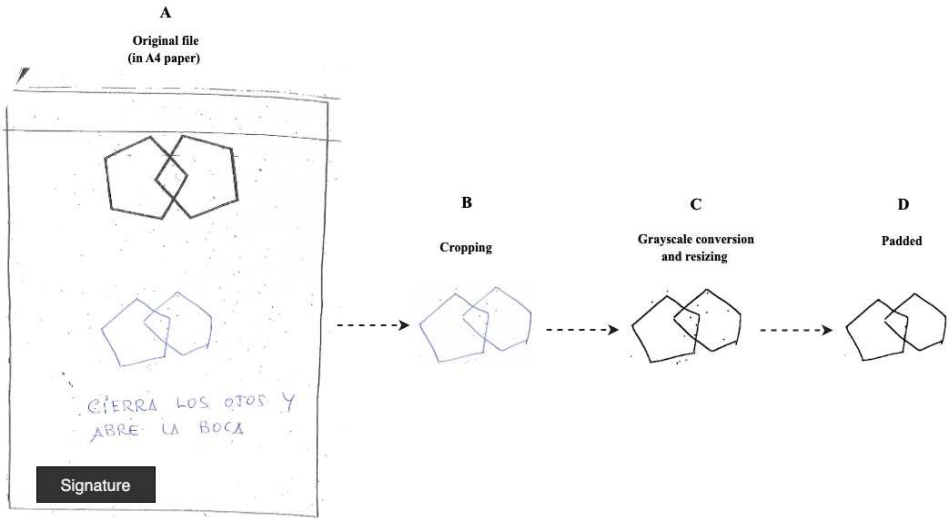

Fig. 1. Example of the preprocessing steps for Pentagon Drawing Test (PDT) images: prompting pentagon (A) on top with participants' drawings at the bottom; image processing (B, C); and final image (D) for model input.

ability to compare synthetic and original images. The SSIM method uses a sliding window to analyze the structural distortion between two similar images. The SSIM score ranges from 0 to 1, with a score of 1 indicating that the images are the same and a score of 0 indicating that the images are totally different. For applying DA techniques such as elastic transformation, grid distortion, and rotation to the images in our training set, we used the Albumentations open-source toolkit [5]. These DA techniques were applied to the images in our training set, which resulted in an increased sample size. Crucially, we allocated all original images from patient subjects exclusively to the test set to ensure a robust testing protocol. The training set consisted of 60 images for healthy and 60 for patient classes. The test set comprised 6 images for healthy and 6 images for patient classes.

After DA, as shown in Figure 2, the SSIM values ranged from 0.6 to 0.7, indicating that the augmented images are not near-duplicates of the original data but are rather new images. However, when all DA techniques from the Albumentations toolkit were applied, the distribution of SSIM values was from 0.1 to 0.7, indicating that the augmented images are much more different than the original images, which is not desirable in our research.

2.3 Classic ML and DL models for AD screening

We selected classic ML and DL models based on their proven strengths and applicability to medical image analysis. Classic ML models were support vector machines (SVM), random forest (RF), and k-nearest neighbors (k-NN). They require manual feature extraction, whereas DL models automatically identify and optimize relevant features from data. Among DL models, CNNs are the most popular and widely used in image-related tasks [31], due to their ability to automatically detect features by using a composition of the different types of layers: (i) Convolution layers (CONV) are the primary building blocks of a CNN model for extracting features such as colors, edges, and corners from the input by applying the convolution operation through a sliding kernel, (ii) Pooling layers (POOL) are used to reduce the

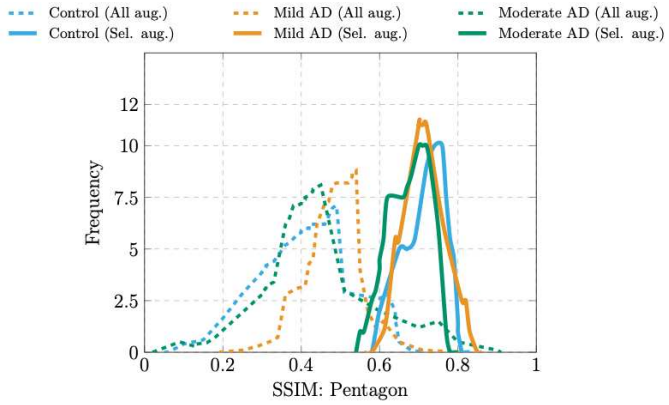

Fig. 2. SSIM distributions for Pentagon drawings. Dashed lines represent the SSIM results of all augmentation techniques (All aug.), while solid lines correspond to the selected augmentation techniques (Sel. aug).

dimensionality of the feature maps computed by the CONV layers, and (iii) Fully-connected layers (FC) are placed at the end of the model's architecture to flatten the output of the previous layer and to add non-linearities to the model.

We evaluated various state-of-the-art CNN architectures for AD screening. All models have the same input layer (224x224 px grayscale images) and the same output layer (with Softmax activation function):

VGG-16 [26] features 16 CONV layers with 3×3 kernels, followed by 3 FC layers before the output layer.
ResNet-152 [10] is a deep residual network architecture with 152 CONV layers. It uses skip connections between CONV layers, a kernel size of 3×3, and batch normalization. The model has two FC layers before the output layer.
DenseNet-121 [13] is a deep CNN composed of 121 layers, including CONV layers with 7×7 kernels, and DenseBlocks, which are groups of CONV layers with 1×1 and 3×3 kernels interconnected through transition layers, and finally an FC layer followed by the output layer.
EfficientNet [27] has multiple CONV layers with a mix of different kernels, followed by corresponding POOL layers and a single FC layer before the output layer.
Custom CNN that we designed with five CONV layers with 3×3 kernels, followed by two POOL layers and one FC layer before the output layer.

Except for our proposed Custom CNN model, the other CNNs are pre-trained on the ImageNet dataset, which contains 1M images distributed over 1000 classes. Therefore, we used transfer learning to finetune those architectures on our dataset. Accordingly, the dimensionality of the output layer is reduced from 1000 classes to 2 or 3, depending on the classification experiments. We used 2 classes in binary classification experiments and 3 classes in multiclass classification experiments.

2.4 Model training

To train the classic ML models (SVM, *k*-NN, RF), we used 5-fold cross-validation, which involves randomly dividing the dataset into 5 groups or folds. For the SVM classifier, a "C" value of 0.1, a "gamma" of 0.0001, and a "linear" kernel were determined to be best. For the *k*-NN classifier, the "manhattan" metric with "n_neighbors" set to 3 and "weights"

configured as “distance” was used. The RF classifier, on the other hand, used a “max_depth” of 15, “max_features” of 9, a “min_impurity_decrease” of 1e-05, and “n_estimators” set at 70.

To train the DL models (CNNs), we used grid search to find the optimal parameters for each model. The learning rate varied between 0.0001 and 0.1, weight decay was fixed at 0.01, and the Adam optimization was employed. The models were trained over 50 epochs, using a batch size of 16, and the Cross-Entropy loss function was applied to optimize classification performance.

3 RESULTS AND DISCUSSION

The efficiency of ML and DL models was evaluated using accuracy and area under the receiver operating characteristic curve (AUC). Accuracy represents the ratio of correct classifications to the total number of samples. AUC reports the performance of a classifier as a trade-off between the True Positive Rate and False Positive Rate, ranging from 0.5 (indicating random performance) to 1 (indicating perfect performance).

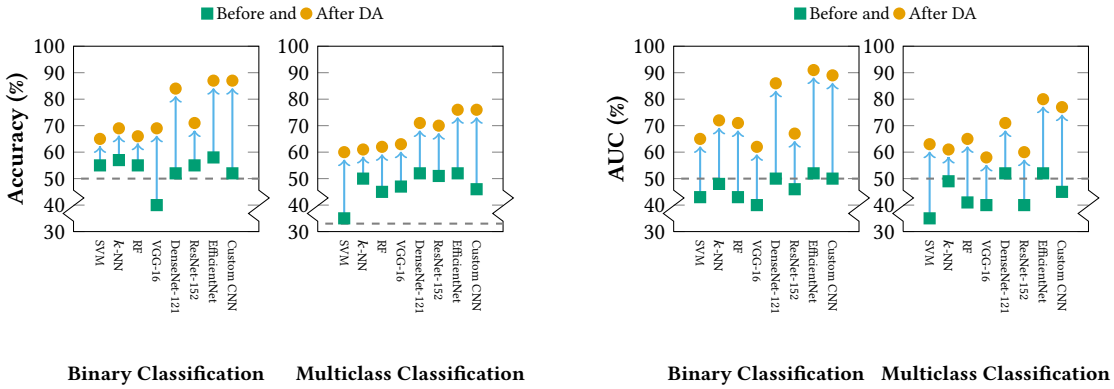

Fig. 3. Accuracy and AUC values of classic ML and DL models, both before and after DA and for both binary class and multiclass classification experiments. Dashed lines represent the performance of a random classifier, illustrating the empirical lower bound in classification performance.

We have explored various classic ML and DL models for binary (healthy and patients) and multiclass (healthy, mild AD, and moderate AD) classification. The results are presented in Figure 3. Both classic ML and DL models showed an increase in accuracy after DA. This improvement was significant when compared to the baseline model (without DA). The results obtained from the classifiers that employed EfficientNet and our custom CNN outperformed all the other models, with 0.87 accuracy and 0.91 AUC scores for binary classification and 0.76 accuracy and 0.8 AUC for multi-class classification. In sum, DA led to a 10 to 30% increase in binary classification experiments and to a 10 to 20% increase in multi-class classification experiments.

Our work showcases the ability of classic ML and DL models to accurately classify AD patients, with a particular emphasis on integrating DA techniques. These DA methods were carefully selected based on their suitability in analyzing cognitive assessment tests used in AD diagnosis, addressing the limitations of current approaches in the existing literature (e.g., [30]). According to the SSIM results (Figure 2), the most appropriate DA techniques for PDT images include elastic transformation, grid distortion, horizontal flipping, translation offset, and rotations. Hosseini-Kivanani et al. [12] highlighted the importance of accurately choosing the DA techniques, showing that flipping and

rotation can destroy the semantics of a CDT image. In contrast, in this work, flipping and rotation are appropriate DA techniques for PDT images, given their symmetry.

DL models have been used in various research for different types of cognitive assessments such as the paper-and-pencil CDT or cube drawing (e.g., [2, 6, 16, 24]). However, none of these studies have specifically focused on the use of DA. Furthermore, while there have been a few efforts to apply DL to automatically screen PDT images, these have not included the use of DA, as seen in [17, 19]. Our Custom CNN model, enhanced with DA, demonstrated exceptional proficiency in evaluating PDT images and outperformed previous studies' results with fewer data used in their studies.

After benchmarking our custom CNN against other state-of-the-art models, we found that it performs better in many cases, particularly when the data has a simple underlying pattern. The simpler structure of our Custom model allows it to learn and generalize these patterns more effectively, leading to higher performance. This suggests that our Custom CNN model with well-designed augmented images excels at certain tasks, such as simple drawings by patients, and is valuable for detecting AD patients from healthy individuals. Furthermore, it outperforms recent work that used pre-trained CNN models in similar task [19].

Our findings underscore the transformative potential of DA in enhancing the DL model's performance. By artificially increasing the dataset's size and diversity, both ML and DL models can be trained to be more robust and accurate, ultimately leading to improved patient outcomes in clinical settings. This research lays the groundwork for future advancements in AD treatment and care, aiming to ultimately improve the quality of life for those affected by AD.

4 CONCLUSION

This work provides valuable insights into the effectiveness of using DA on small clinical datasets for AD screening through handwriting analysis. Both classic ML and DL models were able to achieve better performance than they could without DA. Our method, which is practical for clinical use, offers a cost-effective solution to assist healthcare professionals in patient screening and minimizes subjectivity in interpreting clinical data, particularly in resource-limited settings. It can have a significant impact by helping doctors make more informed decisions and eventually provide better treatment options for patients.

ACKNOWLEDGMENTS

Work supported by the UCM research group (Grupo de Investigación 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] Alzheimer's Association. 2022. 2022 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 18, 4 (April 2022), 700–789.

[2] Samad Amini, Lifu Zhang, Boran Hao, Aman Gupta, Mengting Song, Cody Karjadi, Honghuang Lin, Vijaya B. Kolachalama, Rhoda Au, and Ioannis Ch Paschalidis. 2021. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *Journal of Alzheimer's Disease* 83, 2 (2021), 581–589. https://doi.org/10.3233/JAD-210299 Publisher: IOS Press BV.

[3] Sabyasachi Bandyopadhyay, Jack Wittmayer, David J. Libon, Patrick Tighe, Catherine Price, and Parisa Rashidi. 2023. Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands. *Scientific Reports* 13, 1 (May 2023), 7384. https://doi.org/10.1038/s41598-023-34518-9

[4] Gary Bradski. 2000. The openCV library. *Miller Freeman Inc* 25, 11 (2000), 120–123.

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information* 11, 2 (Feb. 2020), 125.

[6] Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas Maier. 2020. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports 2020 10:1* 10, 1 (Nov. 2020), 1–11. Publisher: Nature Publishing.

[7] Jeffrey L. Cummings. 2004. Alzheimer's disease. *The New England Journal of Medicine* 351, 1 (July 2004), 56–67. https://doi.org/10.1056/NEJMra040223

[8] John D. W. Greene and John R. Hodges. 1996. Identification of famous faces and famous names in early Alzheimer's disease: Relationship to anterograde episodic and general semantic memory. *Brain* 119, 1 (Feb. 1996), 111–128. https://doi.org/10.1093/brain/119.1.111

[9] Martin Guha. 2014. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition). *Reference Reviews* 28, 3 (Jan. 2014), 36–37.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Nina Hosseini-Kivanani, Elena Salobrar-García, Lorena Elvira-Hurtado, Inés López-Cuenca, Rosa de Hoz, José M. Ramírez, Pedro Gil, Mario Salas, Christoph Schommer, and Luis A. Leiva. 2023. Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening. In *IEEE 19th International Conference on e-Science (e-Science)*. IEEE, Cyprus, 1–7.

[12] Nina Hosseini-Kivanani, Christoph Schommer, and Luis. A Leiva. 2023. The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images. In *International Conference on E-health Networking, Application & Services (Healthcom)*. IEEE, China, 23–28.

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[14] Donato Impedovo and Giuseppe Pirlo. 2018. Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective. *IEEE reviews in biomedical engineering* 12 (2018), 209–220. Publisher: IEEE Rev Biomed Eng.

[15] Jessica J Jalbert, Lori A Daiello, and Kate L Lapane. 2008. Dementia of the Alzheimer Type | Epidemiologic Reviews | Oxford Academic. *Epidemiologic reviews* 30, 1 (2008), 15–34. https://academic.oup.com/epirev/article/30/1/15/623289

[16] C. Jiménez-Mesa, Juan E. Arco, M. Valentí-Soler, B. Frades-Payo, M. A. Zea-Sevilla, A. Ortiz, M. Ávila Villanueva, Diego Castillo-Barnes, J. Ramírez, T. del Ser-Quijano, C. Carnero-Pardo, and J. M. Górriz. 2022. Automatic Classification System for Diagnosis of Cognitive Impairment Based on the Clock-Drawing Test. *Lecture Notes in Computer Science* 13258 LNCS (2022), 34–42.

[17] Yike Li, Jiajie Guo, and Peikai Yang. 2022. Developing an Image-Based Deep Learning Framework for Automatic Scoring of the Pentagon Drawing Test. *Journal of Alzheimer's disease: JAD* 85, 1 (2022), 129–139.

[18] José Eduardo Martinelli, Juliana Francisca Cecato, Marcos Oliveira Martinelli, Brian Alvarez Ribeiro de Melo, and Ivan Aprahamian. 2018. Performance of the Pentagon Drawing test for the screening of older adults with Alzheimer's dementia. *Dementia & Neuropsychologia* 12, 1 (Jan. 2018), 54–60. Publisher: Academia Brasileira de Neurologia, Departamento de Neurologia Cognitiva e Envelhecimento.

[19] Jumpei Maruta, Kentaro Uchida, Hideo Kurozumi, Satoshi Nogi, Satoshi Akada, Aki Nakanishi, Miki Shinoda, Masatsugu Shiba, and Koki Inoue. 2022. Deep convolutional neural networks for automated scoring of pentagon copying test results. *Scientific Reports* 12, 1 (Dec. 2022), 9881.

[20] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. 1984. 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 (July 1984), 939–939. https://doi.org/10.1212/WNL.34.7.939

[21] Mario F Mendez, Monique M Cherrier, and Robert S Meadows. 1996. Depth Perception in Alzheimer's Disease. *Perceptual and motor skills* 83, 3 (1996), 987–995.

[22] Gabriel Poirier, Alice Ohayon, Adrien Juranville, France Mourey, and Jeremie Gaveau. 2021. Deterioration, Compensation and Motor Control Processes in Healthy Aging, Mild Cognitive Impairment and Alzheimer's Disease. *Geriatrics* 6, 1 (2021), 33. Publisher: Geriatrics (Basel).

[23] Elena Salobrar-García, Rosa de Hoz, Ana I. Ramírez, Inés López-Cuenca, Pilar Rojas, Ravi Vazirani, Carla Amarante, Raquel Yubero, Pedro Gil, María D. Pinazo-Durán, Juan J. Salazar, and José M. Ramírez. 2019. Changes in visual function and retinal structure in the progression of Alzheimer's disease. *PLOS ONE* 14, 8 (Aug. 2019), e0220535. https://doi.org/10.1371/journal.pone.0220535

[24] Kenichiro Sato, Yoshiki Niimi, Tatsuo Mano, Atsushi Iwata, and Takeshi Iwatsubo. 2022. Automated Evaluation of Conventional Clock-Drawing Test Using Deep Neural Network: Potential as a Mass Screening Tool to Detect Individuals With Cognitive Decline. *Frontiers in Neurology* 13 (2022), 831–831. https://www.frontiersin.org/articles/10.3389/fneur.2022.896403

[25] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1 (July 2019).

[26] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition.

[27] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 6105–6114.

[28] Shinya Tasaki, Namhee Kim, Tim Truty, Ada Zhang, Aron S. Buchman, Melissa Lamar, and David A. Bennett. 2023. Interpretable deep learning approach for extracting cognitive features from hand-drawn images of intersecting pentagons in older adults.

[29] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (April 2004), 600–612.

[30] Victor Wasserman, Sheina Emrani, Emily F. Matusz, Jamie Peven, Seana Cleary, Catherine C. Price, Terrie Beth Ginsberg, Rodney Swenson, Kenneth M. Heilman, Melissa Lamar, and David J. Libon. 2020. Visuospatial performance in patients with statistically-defined mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology* 42, 3 (April 2020), 319–328. https://doi.org/10.1080/13803395.2020.1714550

[31] Guangle Yao, Tao Lei, and Jiandan Zhong. 2019. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters* 118 (Feb. 2019), 14–22. https://doi.org/10.1016/j.patrec.2018.05.018