Better Together: Combining Different Handwriting Input Sources Improves Dementia Screening

Nina Hosseini-Kivanani
University of Luxembourg, Luxembourg
nina.hosseinikivanani@uni.lu

Elena Salobrar-García, Lorena Elvira-Hurtado, Inés López-Cuenca, Rosa de Hoz, José M. Ramírez
Ramon Castroviejo Institute of Ophthalmologic Research, Universidad Complutense of Madrid, Spain
elenasalobrar@med.ucm.es, marelvir@ucm.es, inelopez@ucm.es, rdehoz@ucm.es, ramirezs@med.ucm.es

Pedro Gil, Mario Salas
Memory Unit, Geriatrics Service, Hospital Clínico San Carlos, Madrid, Spain
pgil@salud.madrid.org, mario.salas@salud.madrid.org

Christoph Schommer, Luis A. Leiva
University of Luxembourg, Luxembourg
christoph.schommer@uni.lu, luis.leiva@uni.lu

***Abstract*—Alzheimer's disease (AD) is a cognitive disorder, marked by memory loss and impaired reasoning, that requires early detection methods to better manage and potentially slow down the disease's progression. Recent advances in machine learning have offered new possibilities for AD detection using handwriting analysis, however previous work has considered only one type of input source, e.g. clock or pentagon drawings. Here we propose to develop an efficient method for detecting AD's early symptoms using Deep Feature Concatenation (DFC) models considering multiple handwriting sources: pentagon drawings, self-reported sentences, and signatures. Substantial performance improvements were observed when considering all input sources together with data augmentation techniques. For example, classification accuracy increased from 60% (best model, without data augmentation) to 80% (DFC and data augmentation). Our findings show that the use of diverse input sources can lead to an efficient and cost-effective method for early AD detection. Looking forward into the future, our study highlights the potential of DFC in supporting home-based healthcare diagnoses which is a crucial step in integrating artificial intelligence into healthcare practices.**

***Index Terms*—Deep Learning; Convolutional Neural Networks; deep feature concatenation; dementia; image processing.**

I. Introduction and Related work

Dementia is a condition primarily characterized by a gradual and irreversible decline in cognitive ability, which results in memory loss and language impairment that negatively affect the daily lives of the elderly [1]. The loss of neurons in various regions of the nervous system causes this neurodegenerative disorder, among which Alzheimer's disease (AD) is the most prevalent form.

Since currently there is no known cure for AD, the importance of an accurate and timely diagnosis of dementia cannot be overstated, as it is the cornerstone for implementing effective treatment and providing necessary support for patients and their families. Traditional diagnostic methods, on the other hand, can be subjective and time-consuming, since they require multiple tests, including expensive imaging techniques such as magnetic resonance imaging, invasive tests such as serological or cephalorachidian fluid analysis, and neuropsychological tests by a highly trained professional.

In light of these challenges, the development of automated, objective screening methods for dementia is a priority in the quest for efficient and precise diagnosis of dementia. Notably, the most beneficial methods would be non-invasive and user-friendly in order to minimize any additional burden on the individuals undergoing the diagnosis.

In digital medicine, there is growing interest in using deep learning (DL) models, in particular Convolutional Neural Networks (CNNs), to automatically score cognitive impairment tests instead of traditional manual methods. More concretely, transfer learning has been shown to enhance the performance of pre-trained CNNs in image classification tasks [2], [3], thus suggesting its potential for clinical applications.

Recent studies [4]–[6] have demonstrated the effectiveness of CNNs in analyzing patients' cognitive function assessment over various popular tests, including e.g. the Pentagon Drawing Test (PDT) [7], the Clock Drawing Test (CDT) [8]–[10], and the Rey–Osterrieth Complex Figure Test-copy (RCFT) [6]. Recent work proposed a CNN-based method for the automatic diagnosis of cognitive impairment based on CDT drawings [5] that were classified into healthy and non-healthy categories, demonstrating its potential for implementation in hospitals and clinics, particularly in resource-limited settings.

The ability of CNNs to learn complex image features and patterns associated with cognitive functioning makes these

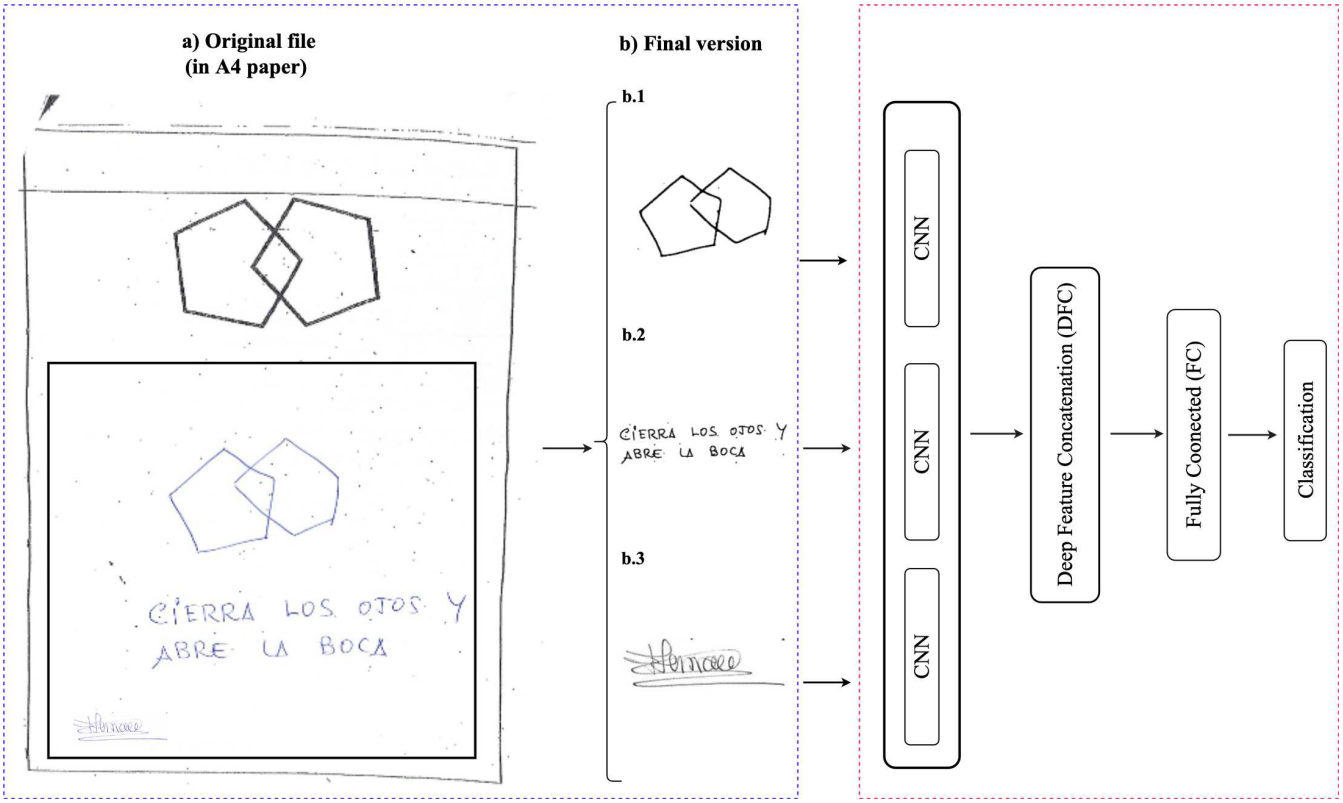


Fig. 1: Final output of image preprocessing which shows in the blue box (left side): a) prompting pentagon on top with participants' drawings at the bottom; b) final images after preprocessing. Structure of the proposed method on the red box (right side).

algorithms promising tools for developing objective and efficient diagnostic methods for dementia screening. Despite the drawing tests' proven ability to accurately predict cognitive impairment, they are not sufficient on their own to provide a comprehensive assessment of the user's overall cognitive state.

Despite the substantial progress made in this field on single input sources (e.g. CDT [5], [9] and PDT [4], [7]), to our knowledge, concatenation-based models (i.e. models considering multiple input sources) have not been considered so far in AD screening contexts. Concatenation-based models have primarily been used in clinical work that involves the analysis of medical images such as fMRI [11]. However, handling multiple input sources present unique challenges due to their variability in terms of styles, shapes, etc. Additionally, they may contain irrelevant or noisy information that could affect classification accuracy. It is therefore essential to develop methods for effectively extracting and combining relevant features from these images in order to ensure accurate classification.

In response to this need, we propose a deep feature concatenation (DFC) method which enables the efficient combination of features derived from various CNN models. Additionally, since clinical datasets are typically very small for DL standards, we propose simple data augmentation techniques to handle various handwriting data sources, and the use of the Structural Similarity Index Measure (SSIM) [12] to quantify the quality of the augmented data. Our experiments show significant improvements in classification performance in several scenarios, as presented and discussed later. Taken together, our contributions represent a potential pathway to assist practitioners in better detecting early symptoms of AD using handwriting data, potentially reducing the variability linked to human subjectivity when interpreting clinical data, and ultimately supporting home-based healthcare diagnosis.

II. Method

We studied state-of-the-art DL models (pre-trained CNNs and a custom CNN) for classifying AD disease according to various handwriting tasks. Building on previous work that showed the effectiveness of DFC in other medical domains [13], [14], we studied this approach in the context of handwriting. Figure 1 summarizes our method. In a nutshell, we use various CNN models to automatically extract features from handwritten images and then concatenate those features via DFC. The concatenated feature vectors are finally classified via a softmax function that predicts a probability distribution, followed by the argmax operation that selects the class with the highest probability.

A. CNN architectures

CNNs have been remarkably successful across a range of image-based classification tasks, particularly in the healthcare domain(e.g., [15]). Transfer learning, particularly *feature transfer* wherein features are extracted from pre-trained models, is a common approach for adopting these models for specific tasks [16]. For our study, we selected four pre-trained models that have proven their efficacy in the medical domain [17]. Further, to broaden the versatility of our approach, we also added a custom CNN model that we describe below.

- **VGG-16** [18]: Developed by the Visual Geometry Group (VGG) in Oxford, VGG-16 is a deep neural network comprising 16 CNN layers with a 3x3 kernel, followed by three fully connected (FC) layers. Noted for its simplicity and effectiveness in feature extraction.
- **ResNet-152** [19]: This is a deep architecture, with 152 CNN layers. It achieves an error rate of 3.5% and uses skip connections between CNN layers to achieve its excellent performance [19].
- **DenseNet-121** [20]: This is a 121-layer deep CNN model that uses a "dense" connectivity pattern between layers. This configuration allows each layer to have direct access to the output of all preceding layers. The architecture includes CNN layers with 7x7 kernels and DenseBlocks containing interconnected CNN layers with 1x1 and 3x3 kernels.
- **EfficientNet** [21]: This is a CNN architecture that employs a scaling technique to harmonize depth, width, and resolution, leading to improved performance and efficiency. With its ability to outperform with fewer parameters, EfficientNet is frequently chosen for applications with limited computational resources [21].
- **Custom CNN**: It comprises five CNN layers, each with 32 filters of size 5, followed by a pooling layer, an FC layer, and Rectified Linear Unit (RelU) layers for all layers except the output layer, which uses linear activation. To avoid overfitting, a dropout rate of 0.1 was implemented after each CNN layer. We explored various model configurations during our research, however, most of them did not show satisfactory results. For example, removing one or two layers did not result in better performance, it even led to overfitting issues, which we resolved by introducing the Dropout layers.

III. Experiments

A. Participants

We recruited 85 participants from the Memory Unit of the Hospital Clinico San Carlos (HCSC) in Madrid. The participants' ages ranged from 61 to 88 years, with a mean of 73.92 $\pm$ 6.78 years. Statistical analysis revealed no significant differences in age between the healthy group, the mild AD group, and the moderate AD group ($p > .05$). The study included 35 female and 50 male participants.

TABLE I: Demographics of Study Participants (mean$\pm$SD).

Attr.	**Healthy**	**Mild AD**	**Moderate AD**
Pentagon	30	3	3
Sentence	30	3	3
Signature	29	22	15
Total	**89**	**28**	**21**
Age	72.4 $\pm$ 6.07	81.33 $\pm$ 4.62	81.66 $\pm$ 2.32
MMSE	28.17 $\pm$ 2.15	23 $\pm$ 2.65	19 $\pm$ 2.01
Gender	Female = 35, Male = 50		

Following the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), the Alzheimer's Disease and Related Disorders Association (ADRDA) workgroup [22], and the Statistical Manual of Mental Disorders V (DSM V) guidelines [23], cognitively impaired patients were divided into two groups: mild AD and moderate AD. The demographic information, which included age and gender, along with the clinical information for each participant and Mini-Mental State Examination (MMSE) [24] scores, are presented in Table I.

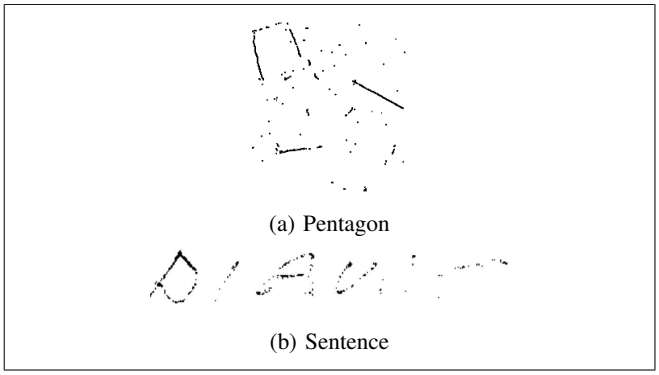
Participants with MMSE scores above 26 were considered cognitively healthy. The MMSE scores for participants diagnosed with cognitive impairment ranged from 25 to 17. Following standard practice [25], we excluded participants with a medical history of neurological or psychiatric disorders, serious medical conditions, or systemic disorders affecting vision. Additionally, we excluded those with ophthalmological conditions such as glaucoma or suspected glaucoma, media opacity, or retinal diseases to avoid biases caused by the vision problems of the participants.

B. Datasets

During the cognitive assessment tests, neuropsychologists used various tools to evaluate the cognitive performance of the participants. Among these, we rely on the PDT subtest of the MMSE results, which assessed visuoconstructional skills and cognitive impairment. The total number of collected images is 138, of which 30 images from the pentagon and sentence parts were removed from the healthy group because of the very low quality of the scanned images (see Figure 2).

Following the PDT protocol, participants were instructed to copy two overlapping pentagons with interlocking shapes to form a rhombus. In order to gain a more holistic understanding of the participants' cognitive abilities, the cognitive assessment incorporated further handwriting data (e.g., sentence and signature) from the same participant. Below each pentagon drawing, participants were requested to write a sentence of their choice and provide a signature on an A4-size blank paper (see Figure 1a).

The paper-and-pencil drawings of both healthy subjects and patients were scanned in PDF format and saved as PNG files. Subsequently, we converted the PNG images to grayscale



(a) Pentagon

(b) Sentence

Fig. 2: Example of badly scanned images.

format, resized them to the standard dimensions of 224×224 pixels, and removed noise using the "pad" (to get them in the same shape) and "canny" (for edge detection of images) operations as provided by the OpenCV library (see Figure 1b). We then removed low-quality images by manual inspection and the remaining images were labeled as either 'healthy', 'mild AD', or 'moderate AD'. The last two categories correspond to AD patients, which allows us to conduct two kinds of experiments: *binary* (patient vs healthy) and *multiclass* (healthy vs mild AD vs moderate AD) classification tasks.

One of the challenges in clinical settings is the lack of large datasets, which may potentially hinder the success of DL models. To address this limitation, we used the Albumentations open-source toolkit [26] to augment the handwritten images for model training and evaluation. We applied the following data augmentation techniques: elastic transformations, grid distortions, horizontal flipping, translation, and rotations to the image (see Figure 3). Note that not all data augmentations make sense in our data, given the nature of our grayscale handwritten images. For example, changing hue or inverting colors would do more harm than good.

The resulting dataset, after data augmentation, was partitioned into the train, validation, and test splits using stratified sampling, to ensure that each partition reflects the same data distribution as in the original dataset. The total size of our dataset increased to 240 images: 120 corresponding to healthy individuals and 120 to patients, among whom 66 represented mild AD and 54 moderate AD.

We used SSIM [12] to measure the quality between original and augmented images. This method uses sliding windows to compare structural distortions between two images. An SSIM value of 1 indicates that the two images are identical, whereas a value of 0 indicates that the two images are completely dissimilar. In our case, a high SSIM value between the original and augmented images implies that the augmentation process has successfully preserved the original images' structural information and visual quality. In contrast, a low SSIM value would suggest that the augmented image differs significantly from the original image and may not be of sufficient quality for the intended use.

As shown in Figure 4, our results reported SSIM values between 0.6 and 0.8, suggesting that the augmented images were reasonable variants (not near-duplicates) of the original data. In contrast, when using all available augmentation techniques provided by the Albumentations toolkit, the distribution of SSIM values is comprised of values between 0.1 and 0.7, indicating that the augmented images are much more different than the original images, which is not desirable in our research.

As hinted previously, it is important to note that not all data augmentation techniques will provide advantages. Each technique has its own strengths and weaknesses, and its utility often depends on the specific dataset and the task at hand. As illustrated in Figure 4 some augmentation methods produce images that correlate better with the original images. This usually enhances the learning process and potentially leads to more accurate and generalizable models [27]. Conversely, other methods might introduce noise or misleading patterns into the data, confounding the learning process and potentially leading to poorer performance. This shows that the chosen augmentation positively influences model learning and performance, while simultaneously reducing the likelihood of adverse or neutral impacts.

C. Model training

All models were trained on the training set for 100 epochs, using early stopping with 10 epochs as a form of regularization. Early stopping prevents overfitting, maintaining the optimal model weights before the model starts memorizing the training data. We used the popular Adam optimizer with variable learning rates (see next section) and categorical cross-entropy as a loss function in multiclass classification experiments. For binary classification experiments, we used the binary cross-entropy loss function.

D. Evaluation

The performance of our models was measured using two metrics: Accuracy and area under the receiver operating characteristic curve (AUC). On the one hand, Accuracy refers to the ratio of correctly classified patients to the overall number of participants. It serves as a straightforward measure of the overall performance of the classification models. On the other hand, AUC illustrates the relationship between sensitivity (the true positive rate) and specificity (the true negative rate) for any given classification model. High AUC values indicate that the model possesses a strong discriminatory capacity between classes (e.g., healthy subjects vs patients).

Continuing our discussion about the selected metrics, it is essential to note that each of them brings forth unique insights into the performance of our models. While Accuracy offers a general view of the model's capability to differentiate between classes, AUC provides a deeper understanding of the model's reliability across various classification thresholds, proving invaluable in scenarios where the cost of misclassification can be substantial. Hence, the integration of these two metrics serves as a comprehensive approach to evaluating and interpreting the performance of our models in a robust manner.

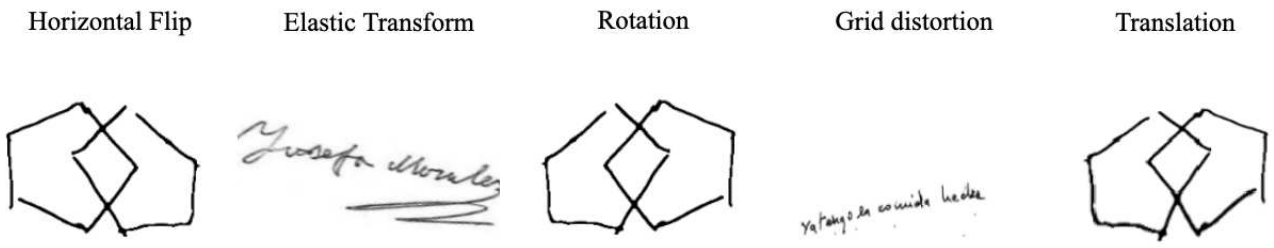


Fig. 3: Examples of applying selected data augmentation techniques on different input sources. From left to right: Horizontal Flip, Elastic transform, Rotation, Grid distortion, and Translation offset.

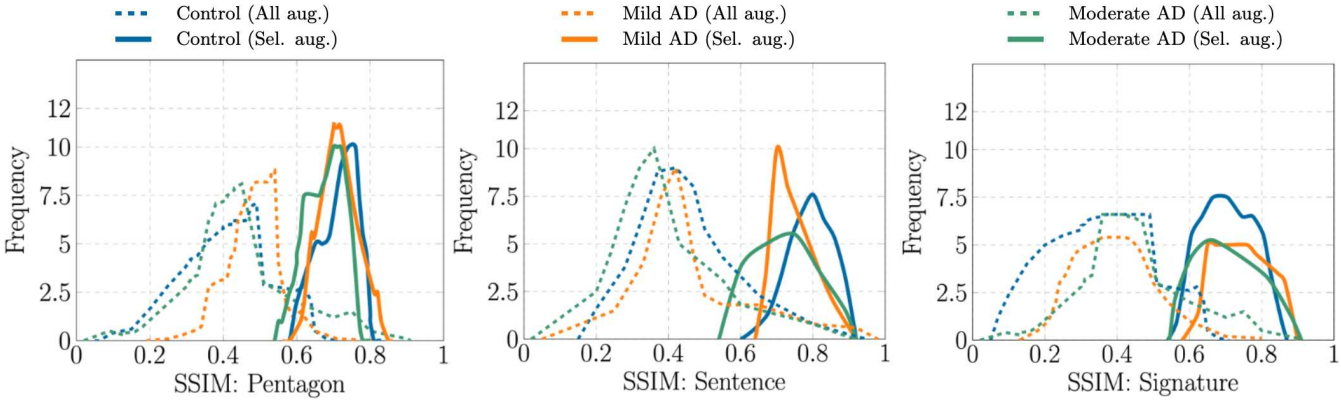


Fig. 4: SSIM distributions. Dashed plots correspond to the results considering all the augmentation techniques collectively (All aug.). The selected augmentation (Sel. aug.) techniques are Elastic transform (`e`), Rotation (`r`), Grid distortion (`g`), Horizontal flip (`f`), and Translation offset (`t`). We used `e+g+r+f+t` for Pentagon, `e+g+t` for Sentence, and `e+t` for Signature.

IV. Results and Discussion

We discuss the experimental results for both binary and multiclass classification settings. The results are based on CNN models that were trained with the following hyperparameters: number of epochs (100), batch size (16), dropout (0.1), weight decay (0.01), and learning rate (0.0001 to 0.1).

Figure 5 and Table II show the results for different model combinations. We use the following nomenclature: "model A + model B" to indicate that "model A" was used to process Pentagon images and that "model B" was used to process both sentences and signatures. As can be observed, it is clear that our DFC approach of concatenating all input sources is significantly superior to any model that considers fewer input sources; see Table II.

Overall, most of our DFC models are considered to improve substantially after data augmentation, except VGG-16 for both binary and multiclass classification, as observed in Figure 5. On the other hand, we can see that the combination of our custom CNN (for processing Pentagon images) and EfficientNet (for processing sentences and signatures) is the one that achieves the highest accuracy: 93% and 80% for binary and multiclass classification, respectively. In terms of AUC, the combination of EfficientNet models (EffNet+EffNet) is the best performer, although it is only 6 percentual points higher than our Custom+EffNet model.

Our results show that concatenated-based CNN models with augmented data outperform previous studies that did not use concatenated-based models with augmentation (e.g., [7]). These findings suggest that the combination of custom CNN and EfficientNet (Custom + EffNet) is a promising option for automatically evaluating handwriting tasks.

Overall, our findings have the potential to improve the accuracy of AD detection and treatment outcomes. Importantly, our DFC model does not need a large number of data, which means that we can reach a robust classifier for detecting AD patients. These findings underline the substantial role data augmentation plays in boosting model performance and further demonstrate the benefits of using diverse input sources in the performance of the models.

V. Conclusion and Future Work

We have successfully developed a robust and efficient model, capable of accurately classifying handwritten images into healthy individuals and AD patients ranging from mild to moderate severity. Our research primarily focused on evaluating the potential improvement in classification performance through the incorporation of handwriting data from diverse

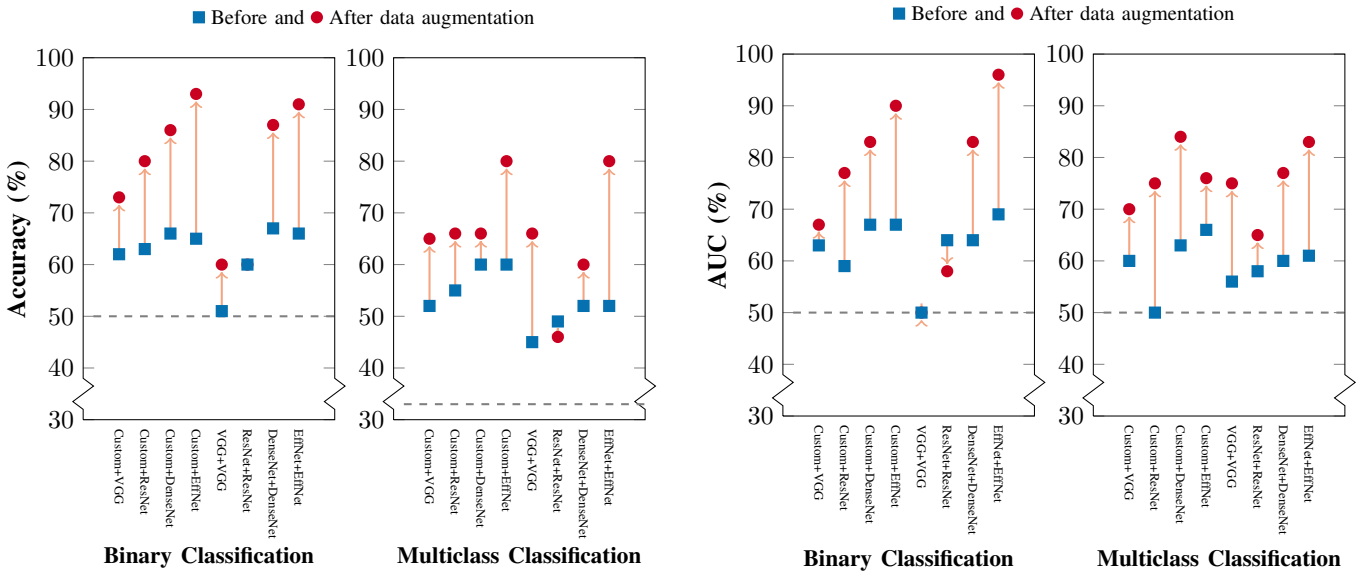


Fig. 5: Accuracy and AUC results before and after data augmentation considering all input sources (pentagon, sentence, and signature), both for binary (2 classes, leftmost plot) and multiclass (3 classes, rightmost plot) classification experiments. Dashed lines denote the performance of a random classifier, as a way to illustrate the empirical lower bound in classification performance.

TABLE II: Performance results of other input source combinations before and after data augmentation: Pentagon+Sentence (abbreviated as Pen+Sen), Pentagon+Signature (Pen+Sig), and Sentence+Signature (Sen+Sig). We use the following nomenclature: "model A + model B" to indicate that "model A" was used to process the first input source (e.g. Pentagon) and "model B" was used to process the second input source (e.g. Sentence).

			Custom + VGG	**Custom + ResNet**	**Custom + DenseNet**	**Custom + EffNet**	**VGG + VGG**	**ResNet + ResNet**	**DenseNet + DenseNet**	**EffNet + EffNet**
			Acc. | AUC	**Acc. | AUC**	**Acc. | AUC**	**Acc. | AUC**	**Acc. | AUC**	**Acc. | AUC**	**Acc. | AUC**	**Acc. | AUC**
Binary	**Before**	**Pen+Sen**	0.67 | 0.66	0.6 | 0.6	0.4 | 0.4	0.66 | 0.67	0.6 | 0.61	0.6 | 0.6	0.8 | 0.89	0.67 | 0.76
		Pen+Sig	0.6 | 0.83	0.6 | 0.6	0.8 | 0.83	0.8 | 0.79	0.6 | 0.59	0.6 | 0.6	0.73 | 0.73	0.93 | 0.92
		Sen+Sig	0.6 | 0.78	0.6 | 0.6	0.4 | 0.4	0.87 | 0.85	0.6 | 0.64	0.6 | 0.6	0.4 | 0.4	0.8 | 0.84
	After	**Pen+Sen**	0.8 | 0.79	0.6 | 0.64	0.4 | 0.4	0.87 | 0.93	0.4 | 0.4	0.93 | 0.93	0.8 | 0.8	0.93 | 0.93
		Pen+Sig	0.1 | 0.1	0.93 | 0.93	0.87 | 0.86	0.93 | 0.92	0.6 | 0.6	0.6 | 0.6	0.93 | 0.92	0.4 | 0.4
		Sen+Sig	0.4 | 0.43	0.53 | 0.51	0.93 | 0.93	0.93 | 0.92	0.6 | 0.6	0.47 | 0.42	0.6 | 0.6	0.93 | 0.93
Multiclass	**Before**	**Pen+Sen**	0.2 | 0.52	0.6 | 0.7	0.2 | 0.40	0.6 | 0.74	0.4 | 0.58	0.2 | 0.4	0.6 | 0.7	0.8 | 0.89
		Pen+Sig	0.27 | 0.48	0.6 | 0.7	0.2 | 0.59	0.73 | 0.78	0.53 | 0.63	0.2 | 0.4	0.6 | 0.72	0.87 | 0.9
		Sen+Sig	0.2 | 0.4	0.6 | 0.7	0.6 | 0.69	0.93 | 0.94	0.6 | 0.7	0.6 | 0.7	0.6 | 0.72	0.73 | 0.83
	After	**Pen+Sen**	0.6 | 0.78	0.73 | 0.81	0.87 | 0.91	0.87 | 0.92	0.8 | 0.89	0.8 | 0.88	0.93 | 0.95	0.93 | 0.96
		Pen+Sig	0.4 | 0.57	0.8 | 0.87	0.93 | 0.95	0.93 | 0.96	0.8 | 0.89	0.8 | 0.88	0.93 | 0.96	0.93 | 0.96
		Sen+Sig	0.4 | 0.58	0.8 | 0.87	0.8 | 0.9	0.93 | 0.95	0.8 | 0.89	0.8 | 0.88	0.93 | 0.95	0.93 | 0.95

sources and feature concatenation. Our findings put forward the efficacy of the combination of DFC and data augmentation techniques in developing more holistic and precise models for AD screening.

The potential implications of our study are manifold, with particularly important implications within clinical settings. The developed DFC model enhances healthcare providers' decision-making capabilities, especially for untrained professionals, fostering improved patient care and mitigating the likelihood of unnecessary procedures or subjective diagnoses. As a screening method, it can be used anywhere from primary care settings to daycare facilities. Looking forward, we suggest future research should focus on creating a smartphone app, grounded in the established framework, that can collect and analyze handwritten data on the go. This app could potentially integrate multiple models (e.g. binary and multiclass classifiers) in order to account for different practitioners' needs. Furthermore, our methodology exhibits promising potential for

the classification of other handwriting tasks, such as CDT or RCFT drawings. Hence, future work should focus on exploring and validating the model's proficiency in these tasks.

ACKNOWLEDGMENT

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] H. Hampel, D. Prvulovic, S. Teipel, F. Jessen, C. Luckhaus, L. Frölich, M. W. Riepe, R. Dodel, T. Leyhe, L. Bertram, W. Hoffmann, and F. Faltraco, "The future of Alzheimer's disease: The next 10 years," *Progress in Neurobiology*, vol. 95, no. 4, pp. 718–728, Dec. 2011.

[2] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[3] T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, and S. Kashem, "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, Jan. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/9/3233

[4] Y. Li, J. Guo, and P. Yang, "Developing an Image-Based Deep Learning Framework for Automatic Scoring of the Pentagon Drawing Test," *Journal of Alzheimer's disease: JAD*, vol. 85, no. 1, pp. 129–139, 2022.

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

[6] Y. C. Youn, J.-M. Pyun, N. Ryu, M. J. Baek, J.-W. Jang, Y. H. Park, S.-W. Ahn, H.-W. Shin, K.-Y. Park, and S. Y. Kim, "Use of the Clock Drawing Test and the Rey–Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment," *Alzheimer's Research & Therapy*, vol. 13, no. 1, p. 85, Apr. 2021.

[7] J. Maruta, K. Uchida, H. Kurozumi, S. Nogi, S. Akada, A. Nakanishi, M. Shinoda, M. Shiba, and K. Inoue, "Deep convolutional neural networks for automated scoring of pentagon copying test results," *Scientific Reports*, vol. 12, no. 1, p. 9881, Dec. 2022.

[8] S. Amini, L. Zhang, B. Hao, A. Gupta, M. Song, C. Karjadi, H. Lin, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test," *Journal of Alzheimer's Disease*, vol. 83, no. 2, pp. 581–589, 2021, publisher: IOS Press BV.

[9] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–11, Nov. 2020, publisher: Nature Publishing Group.

[10] R. Binaco, N. Calzaretto, J. Epifano, S. McGuire, M. Umer, S. Emrani, V. Wasserman, D. J. Libon, and R. Polikar, "Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease," *Journal of the International Neuropsychological Society*, vol. 26, no. 7, pp. 690–700, Aug. 2020.

[11] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor," *IEEE Access*, vol. 8, pp. 55 135–55 144, 2020.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[13] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Scientific Reports*, vol. 11, no. 1, p. 3254, Feb. 2021. [Online]. Available: https://www.nature.com/articles/s41598-020-74399-w

[14] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5, iSSN: 2379-447X.

[15] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, Jul. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522000913

[16] W. Zhu, J. Zhang, and J. Romagnoli, "General Feature Extraction for Process Monitoring Using Transfer Learning Approaches," *Industrial & Engineering Chemistry Research*, vol. 61, no. 15, pp. 5202–5214, Apr. 2022. [Online]. Available: https://doi.org/10.1021/acs.iecr.1c04565

[17] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," *Informatics in Medicine Unlocked*, vol. 19, p. 100360, Jan. 2020.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[21] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 6105–6114.

[22] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "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*, vol. 34, no. 7, pp. 939–939, Jul. 1984. [Online]. Available: https://n.neurology.org/content/34/7/939

[23] M. Guha, "Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th edition)," *Reference Reviews*, vol. 28, no. 3, pp. 36–37, Jan. 2014. [Online]. Available: https://doi.org/10.1108/RR-10-2013-0256

[24] T. Tombaugh and N. J. McIntyre, "The Mini-Mental State Examination: A Comprehensive Review - Tombaugh - 1992 - Journal of the American Geriatrics Society - Wiley Online Library," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[25] E. Salobrar-García, R. d. Hoz, A. I. Ramírez, I. López-Cuenca, P. Rojas, R. Vazirani, C. Amarante, R. Yubero, P. Gil, M. D. Pinazo-Durán, J. J. Salazar, and J. M. Ramírez, "Changes in visual function and retinal structure in the progression of Alzheimer's disease," *PLOS ONE*, vol. 14, no. 8, p. e0220535, Aug. 2019. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220535

[26] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.

[27] P. Bhardwaj and A. Kaur, "A novel and efficient deep learning approach for COVID-19 detection using X-ray imaging modality," *International Journal of Imaging Systems and Technology*, vol. 31, no. 4, pp. 1775–1791, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22627