International Journal of Neural Systems, Vol. 0, No. 0 (2021) 1–18

Using XAI in the Clock Drawing Test to reveal the cognitive impairment pattern

CARMEN JIMÉNEZ-MESA[1,2], JUAN E ARCO[1,2,3], MERITXELL VALENTÍ-SOLER[4], BELÉN FRADES-PAYO[4], MARÍA A. ZEA-SEVILLA[4], ANDRÉS ORTIZ[2,3], MARINA ÁVILA-VILLANUEVA[4], DIEGO CASTILLO-BARNES [1,2], JAVIER RAMÍREZ[1,2], TEODORO DEL SER-QUIJANO [4],CRISTÓBAL CARNERO-PARDO[5], JUAN M GÓRRIZ[1,2]

[1]*Data Science and Computational Intelligence (DASCI) Institute, Spain*

[2]*Department of Signal Theory, Networking and Communications, University of Granada, 18010 Spain*

[3]*Department of Communications Engineering, University of Malaga, 29010 Spain*

[4]*Alzheimer Disease Research Unit, CIEN Foundation, Carlos III Institute of Health, Queen Sofía Foundation Alzheimer Center, Madrid, Spain.*

[5]*FIDYAN Neurocenter, Spain*

E-mail: gorriz@ugr.es

The prevalence of dementia is currently increasing worldwide. This syndrome produces a deterioration in cognitive function that cannot be reverted. However, an early diagnosis can be crucial for slowing its progress. The Clock Drawing Test (CDT) is a widely used paper-and-pencil test for cognitive assessment in which an individual has to manually draw a clock on a paper. There are a lot of scoring systems for this test and most of them depend on the subjective assessment of the expert. This study proposes a computer-aided diagnosis (CAD) system based on artificial intelligence (AI) methods to analyze the CDT and obtain an automatic diagnosis of cognitive impairment (CI). This system employs a preprocessing pipeline in which the clock is detected, centered and binarized to decrease the computational burden. Then, the resulting image is fed into a Convolutional Neural Network (CNN) to identify the informative patterns within the CDT drawings that are relevant for the assessment of the patient's cognitive status. Performance is evaluated in a real context where patients with CI and controls have been classified by clinical experts in a balanced sample size of 3282 drawings. The proposed method provides an accuracy of 75.65% in this classification task, with an AUC of 0.83. These results overcome previous studies, showing that the method proposed has a high reliability to be used in clinical contexts. The large size of the sample and the performance obtained despite being applied to the classic version of the CDT demonstrate the suitability of CAD systems in the CDT assessment process. Explainable AI (XAI) methods are applied to identify the most relevant regions during classification. Finding these patterns is extremely helpful to understand the brain damage caused by cognitive impairment. A validation method using resubstitution with upper bound correction in a machine learning approach is also discussed.

Keywords: Clock Drawing Test, Cognitive Impairment, Clinical diagnosis, Computer-aided diagnosis, Deep learning, Explanaible AI, Image processing, Machine Learning, Alzheimer's disease

1. Introduction

Dementia, which is most frequently caused by Alzheimer's disease (AD), is one of the most common neurological syndromes in the world.[1] Its diagnostic process begins with the use of a test for evaluating the cognitive state of the patient. The Clock Drawing

Test (CDT) is a common paper-and-pencil screening tool for the identification of cognitive changes related to visuospatial functions, frontal lobe execution or memory, among others.[2] During the test, patients are said to draw a clock including the numbers from 1 to 12 and a specific position of the clock hands: ten past eleven. After that, a physician evaluates the resulting drawing and establishes a score, which reflects the patient's cognitive status and detects an eventual cognitive impairment (CI).

This test is widely employed given its simplicity and high sensitivity,[3] which can be improved by including additional cognitive tests such as the Mini-Cog,[4,5] to assess memory and other cognitive domains.[6] However, the CDT scoring task performed by the physician is manual, time-consuming and based on a subjective decision. In recent years, the emergence of machine learning (ML) and deep learning (DL) techniques[7] has provided solutions for the automation of the diagnostic process of a high number of diseases. In the field of neuroimaging, the use of computer aided diagnosis (CAD) systems has been successfully applied in the study of diseases such as Alzheimer's disease,[8–10] Parkinson's disease,[11–13] Autism[14,15] or Dyslexia.[16,17]

This kind of intelligent systems have also been used for the automatic evaluation of the CDT. In fact, the number of works involving machine learning[18–21] or deep learning[22–24] has increased substantially in recent years. Most of these works were focused on a digital version of the CDT, in which a digital ballpoint pen was used instead of a pencil. This allows the acquisition of additional information such as pressure on surface or air-time during drawing.[19] According to previous studies the digital version of the CDT provides higher diagnostic accuracy than the standard one[25] and better sensitivity and specificity to detect mild cognitive impairment (MCI) or demented patients.[26] Despite this boost in performance, the digital version of the test requires expensive equipment compared to the standard one, in which only a pencil and a paper is needed. This can be problematic in scenarios where this technology is not available. Thus, it would be quite interesting to find a methodology that performs similarly to digital CDTs but applies only on the classical version of the CDT.

In this work, we propose an alternative for automatically identifying patients with CI using the classical version of the CDT. Specifically, our proposal relies on the use of a preprocessing to isolate the regions of interest from all images that are subsequently entered into a convolutional neural network (CNN). The model is trained in order to find the relationship between the drawings and the diagnosis of all patients. We aim to demonstrate that our method is able to identify spatial patterns in the drawings that are relevant in the early diagnosis of mild cognitive impairment.

The rest of the paper is organized as follows. Section 3 provides a detailed description of the database, whereas Section 4 indicates the methods developed in this work. First, the preprocessing steps are explained. Then, the classification algorithms proposed, based on CNN and Support Vector Machines (SVM), are described. Afterwards, in Section 5 we evaluate the applicability of our proposal to find differences in the draws depending on the patient cognitive status. Finally, results are discussed in Section 6, whereas conclusions and future works are available in Section 7.

2. Related works

The use of CAD systems for the classification of medical imaging is widespread. Ref.[27] proposed a method based on sparse coding in order to automate the diagnosis of pneumonia. To do so, images were first partitioned into different files. Then, a dictionary was built after applying Principal Component Analysis (PCA) to these files. After that, the images were reconstructed from an iterative deactivation process of the different elements of the dictionary, and entered into a linear SVM classifier, leading to a high performance in a 4-class context. Other studies have focused on the early diagnosis of AD. Ref.[28] presented a multiclass classification approach for predicting the conversion from mild cognitive impairment to AD. Specifically, they proposed a method for addressing the outlier detection problem based on pairwise t-test for feature selection. Then, the selected features projected by a new subspace after applying Partial-Least-Squares (PLS). Finally, classification was performed after using one-versus-one error correction output codes. Results in the multiclass scenario (67%) outperformed similar works, evidencing the high applicability of the proposal as an aid for clinicians. Ref.[29] proposed a solution for early diagnosis in neuroimaging based on deep learn-

ing architectures. Specifically, the brain was parcellated according to an anatomical atlas and entered into an individual deep belief network. The final prediction was determined by combining the individual outputs of all neural networks following different voting schemes. Frameworks based on deep learning have been successfully employed in the detection of Parkinson's disease,[30,31] epilepsy[32,33] or multiple sclerosis.[34,35]

All these works have in common that they use direct measures from the brain (such as different images modalities or data from electroencephalography) to establish the diagnosis of a specific disorder. Although complexity of the classification task mainly depends on the differences between the two groups to be classified, patterns extracted from direct measures are usually more informative, which means that classification based on them should be easier than when it relies on indirect measures. such as behavioral or test measures. However, the use of indirect measures can also be extremely interesting because of its reduced cost compared to direct ones, as the topic of this work shows. The paper-and-pencil test is much less expensive than acquiring an MRI or a PET scan, which means that it would be extremely relevant to develop a method for detecting cognitive impairment or dementia. In fact, recent works have demonstrated the importance of behavioral data as an indicator of a specific disorder. Ref.[36] showed that the way we interact with computer keyboards can be used to detect motor signs associated with the early stages of Parkinson's disease. Ref.[37] presented an approach based on a web-game for universal screening of dyslexia. Data from auditory and visual perception were employed for training a machine learning model, leading to an F1-score of 0.75 for Spanish speakers. Ref.[38] developed a method that evaluates different social, psychological and biological risk factors in order to identify the presence of a specific conduct disorder in children. To do so, a feed-forward neural network was used in addition to an algorithm for estimating the conditional density underlying the different classes in order to preserve the data distribution. Results predicted the presence of conduct disorders with 91.18% accuracy, identifying and ranking certain factors according to their relationship with these disorders.

3. Materials

The database employed in this work was collected from volunteers in the Multidisciplinary Unit of CIEN Foundation (Madrid, Spain) and the Department of Neurology of FIDYAN Neurocenter (Granada and Malaga, Spain). Cognitive status of every participant was diagnosed by consensus of a team of experienced neurologist and neuropsychologist, taking into account their age, functional status, clinical data and performance in an extensive neuropsychological battery. The criteria from the National Institute on Aging-Alzheimer´s Association (NIA-AA),[39] and from the fourth edition, text revised, of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)[40] were used to diagnose mild cognitive impairment and dementia, respectively. This dataset consists of 7009 CDT drawings; 5368 of them were drawn by individuals with normal cognition (healthy controls, HC), and 1641 by individuals with CI including mild cognitive impairment or dementia. The average age of the participants is 73.30 years, whereas 51.73% of them have superior education. All the information regarding demographics is summarized in Table 1.

Regarding the process of the clock drawings collection, participants were given an A4 size paper and a pencil and asked to draw a clock with the clock hands pointing to ten past eleven. Once the clock drawing was finished, physicians assigned a score to the resulting drawing from 0 to 7, according to standard rules.[41] The participant got the maximum score when a perfect clock was drawn, which usually means that the person does not suffer any relevant cognitive impairment. By contrast, a score of 0 indicates that the subject is unable to draw the clock, and it is highly likely that he/she suffers a severe CI. Figure 1 illustrates the diversity of the drawings in the database. The associated scores from these drawings range from the lowest score (0, clock in the left) to the highest (7, clock in the right).

4. Methods

4.1. *Image preprocessing*

The paper-and-pencil draws of all patients were scanned in order to obtain a digital version of the images. They can contain not only the clock drawings but another non-relevant information such as previous drawing attempts, comments from the clini-

Table 1. Demographics of the subjects contained in the database. The acronym CI stands for cognitive impairment subject, S denotes superior education, NS stands for non-superior education, M represents male and F stands for female.

	CI	Controls	Total
Number of participants	1641	5368	7009
Age	74.36 8.21	72.98 6.06	73.30 6.65
Education (S/NS)	671/970	2955/2413	3626/3383
Sex (M/F)	709/932	2104/3264	2813/4196

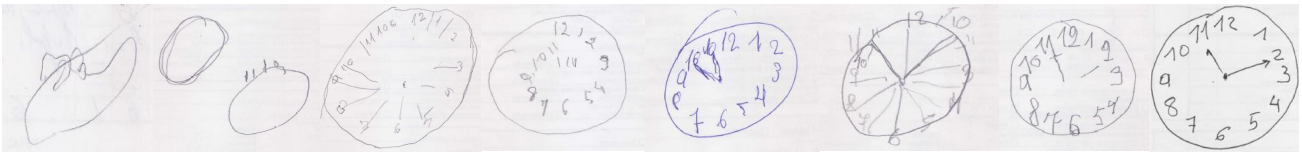

Figure 1. Examples of drawings made by the subjects of the dataset. From left to right, their associated scores range from the lowest (0) to the highest (7) possible score.

cians and numerical identifiers related to the subject. For this reason, we applied a preprocessing process in order to isolate the region of interest (the clock drawing), eliminating the non-relevant information for further analysis.

Figure 2 summarizes the different stages of the preprocessing pipeline. First, the original scanned images were converted to grayscale in order to reduce the number of channels to be processed (from three to one). After that, the resulting images were binarized to isolate the pixels contained in the clock from those that form the background. Then, an edge filling process[42] was applied to detect the objects contained in the image and to identify if they belong to the region of interest (see Figure 2-c). Our algorithm properly recognizes elements even when they are drawn outside the clock face, which is not unusual for numbers 12, 3, 6 and 9. Finally, the images were cropped and downsampled to a final size of 224x224 to reduce the computational burden while preserving their quality. The resulting images were binary, which means that the intensity of the pixels was 1 for the informative ones and 0 for the rest, as depicted in Figure 2-e.

4.2. *Deep learning approach*

After preprocessing the images, the resulting versions were entered into a deep learning model based on a CNN. This architecture has become the standard one in image processing. The application of CNN to neuroimaging has revolutionized the field, addressing problems in a more efficient way, such as in brain's tumor detection[43] or in the identification of patterns associated with Autism.[44] CNNs are usually formed by several layers, from the first related to the extraction of informative patterns to the last ones whose purpose is perform classification. This architecture can be used individually or as a part of a more complex network, such as U-net,[45] DenseNet-121[46] or Mobilenetv2.[47]

The architecture of our CNN is depicted in Figure 3, and includes four 2D convolutional blocks: convolutional layer, batch normalization, rectified linear (ReLU) activation function and a maxpooling layer; and three fully connected layers. Dropout[48] was applied in combination with the linear layers to prevent overfitting, whereas a final softmax layer was added to the model to predict the probability of each sample belonging to the two classes under analysis (CI and HC). Regarding the hyperparameters associated with the CNN, we employed a dropout of 0.5 and an Adam optimization algorithm with a learning rate of 0.001. Besides, we used a Binary Cross-Entropy loss function, whereas the system was trained during 70 epochs employing a batch size of 1.

4.2.1. *Visual explainability*

Despite the great performance that CNNs offer, a clear disadvantage is that they work as black boxes, which means that it is not easy to explain what the network is basing its decisions on. This is especially problematic in the biomedical field, since any

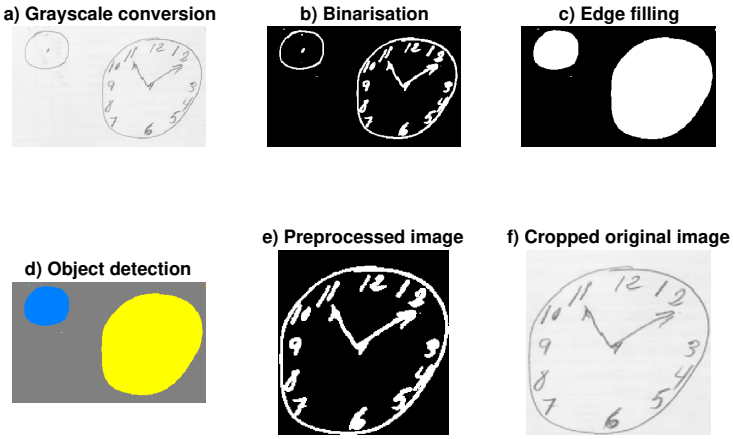

Figure 2. Steps involved in the preprocessing of the images: a) greyscaling of the original image, b) binarisation with manually selected threshold (the same for all images), c) filling of existing elements in the image, d) detection of objects located in the image, e) cropping the image to only the clock and standardise its dimensions ($224x224$) and f) image reconverted to greyscale for comparative reasons with the original image. The image obtained in step e) is the one that is fed into the classification algorithm.

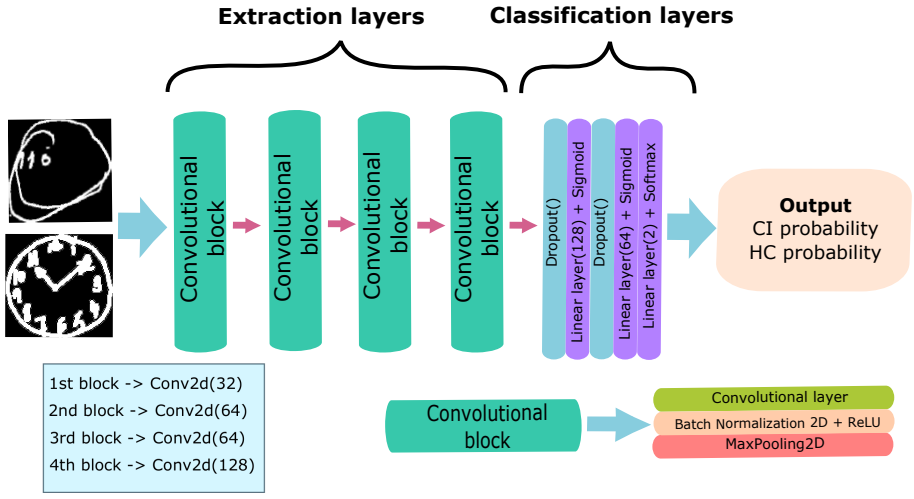

Figure 3. Design of the convolutional neural network used in this work. The architecture consists of four convolutional blocks, including a convolutional layer, bach normalisation and maxpooling, as well as fully-connected layers for the classification stage with dropout.

CAD system to be implemented in a clinical environment must be understandable by clinicians and apply trustworthy criteria.[49] For this reason, it is extremely important to provide models that are interpretable in order to widen our knowledge about the reasons why the different classes can be distinguished. To do so, we used backpropagation-based saliency maps[50] and the Guided Gradient Class Activation Map (Grad-CAM) algorithm[51] in order to identify which areas of the patients drawings are more relevant in classification. Both approaches are explainable methods that assist in the interpretation of the CNN's predictions. Neither of them require configuration changes or re-training.

Saliency map is one of the oldest and more common interpretation method. A saliency map represents the parts of the image that contribute most to the network's decision. Given an image I and a class score function $S_c(I)$, which depends on the vector weights and bias of the model, a saliency map is computed by obtaining the derivative w calculated via backpropagation at a given point p:[50]

$$w = \frac{\partial S_c}{\partial I}_{p} \tag{1}$$

Then, the saliency map is finally obtained by rearranging the elements of w, i.e, according to the pixels distribution in the final score.

Grad-CAM is another of the most commonly used methods of visual interpretation. It was originally designed as an improvement of the CAM algorithm[52] and it can be applied to networks that include fully-connected layers. Once the class c under analysis is selected, Grad-CAM computes the gradient of the score for c, y^c according to the activations maps of the final convolutional layer. Then, gradients flowing back are global-average-pooled to obtain the neuron importance weights α_k^c:[51]

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

where A_{ij}^k represents the activation map k in the convolution layer over the indexes i and j related to width and height, respectively. The first part of the equation represents the global averaged pooling. Once the importance weights are computed, they are multiplied by its associated activation map and all are summed. Finally, the final heatmap is obtained after applying the ReLU nonlinearity.

$$\text{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{3}$$

We applied both methods to all the images of the database in two different ways. In the first one, a study was performed for each image individually. In the second, a collective study was performed over the images that were correctly classified. First, images were separated according to their label. Then, an averaged saliency map was generated. This class activation map analysis reveals the patterns associated with each class that guided the classification.

4.3. *Machine Learning approach*

The preprocessed images were also entered into an alternative based on machine learning that was used as baseline, i.e. a performance to compare with. Following the usual pipeline in classification contexts,[28,53] a method based on Partial Least Squares[54] was employed to reduce the dimensionality of the input data while extracting informative patterns.[55] This statistical method establishes a relationship among observed variables by means of latent variables. Therefore, given an input data $\mathbf{X}_{l \times m}$ and its set of labels $\mathbf{Y}_{l \times 1}$, where l represents the number of samples and m the number of features, PLS computes linear combinations of the score matrices, $\mathbf{X}_s$ via matrices of loadings, $\mathbf{X}_l$, assuming an error matrix $\mathbf{E}$:[56]

$$\mathbf{X} = \mathbf{X}_s \mathbf{X}_l^T + \mathbf{E} \tag{4}$$

The size of the matrix of loadings is mxd, where d is the reduced number of components ($m > d$). Thus, $\mathbf{X}_l$ allows the reduction of the m original features to a new d-dimensional space which contains the original information.

The resulting d features, $d = 5$ in this work, were then used as input of a Support Vector Machines (SVM) classifier with a linear kernel.[57] This classification algorithm estimates the maximum-margin hyperplane to separate the existing classes in the dataset. In a linear binary problem, this hyperplane could be described as the sets of points $\mathbf{x}$ that meet:

$$\mathbf{w}^T \mathbf{x} \quad b = 0 \tag{5}$$

where $\mathbf{w}$ represents the normal vector to the hyperplane and b is the error term. There are two parallel hyperplanes associated with the main one to maintain the largest possible distance between the two classes:

$$\begin{aligned} \mathbf{w}^T \mathbf{x} \quad b &= 1 \\ \mathbf{w}^T \mathbf{x} \quad b &= \quad 1 \end{aligned} \tag{6}$$

Thus, elements above the first hyperplane are considered to be of one class, and those below the second hyperplane are considered to be of the other class. In neuroimaging, the use of SVM as a classification algorithm is widely adopted when a small sample set is involved.[28,58]

4.4. Validation procedure

To assess the reliability of our results, we applied two different validation methods. In the CNN-based approach we employed a 5-fold stratified cross-validation scheme[59] in order to guarantee the independence between the samples used to train the classification model and the ones used for estimating its generalization ability. As the database was split randomly over $K = 5$ iterations, in each iteration 80% of the database was used as the training set. The remaining 20% was used as the test set, each time using a different fold as a test set. This flowchart is shown in Figure 4 (left). The mean and standard deviation (std) of all performance metrics were calculated from the values obtained in the five iterations.

In contrast to the previous scenario, we used an upper bound-corrected resubstitution as a validation method for the ML approach. The upper bound can be seen as the difference between empirical and actual errors, $\mu \quad jE_{act}(f(x)) \quad E_{emp}(f(x))j$. Thus, the actual accuracy obtained using the whole dataset as training and test set could be limited by the upper bound proposed in,[60] as follows:

$$\mu_{VC} \quad \sqrt{\frac{h \quad \ln \quad \frac{2n}{h} \quad + 1 \quad \ln \quad \frac{\eta}{4}}{n}} \tag{7}$$

where η is the significance level and n is the size of the training set. The VC dimension is represented by h and is equal to $d + 1$ for linear functions, as in this case, being d the features dimension. This upper bound could be seen as a theoretical classification limit, in this case for linear classifiers, which allows the use of all accessible data to establish the metrics of interest. Besides, accuracy, sensitivity and specificity can be limited by this value considering that its associated errors are partial errors of the classification one.

The flowchart related to this scenario is depicted in Figure 4 (right). We used the five main components extracted by PLS as the input features of the classifier. Regarding the upper bound-corrected resubstitution method, we set a significance level of 0.05.

4.4.1. Dataset validation

To analyse the separability between the classes in the dataset, the divergence between them was estimated by means of covariance matrices. This allows to validate the results previously obtained by means of classification algorithms.

First, the images were reduced to a size of 12x12 for computational reasons. Then, the HC samples set was partitioned into training and test subsets using K-fold with the condition that each partition had the same number of samples as the CI set. The covariance matrix for each subset was estimated and the divergence between them was estimated as follows:

$$Div^{a;b} = \sum_i \sum_j \mathbf{S}^a_{ij} \quad \mathbf{S}^b_{ij} \quad ^2 /n \tag{8}$$

where $\mathbf{S}^a$ and $\mathbf{S}^b$ are any two covariance matrices of n samples which are compared pixel by pixel. The scenarios analysed according to these criteria are CI set *vs* HC test subset, CI set *vs* HC training subset and HC training *vs* HC test subsets.

4.5. Performance evaluation

The performance metrics employed for evaluating the results include accuracy, specificity (true negative rate) and sensitivity (true positive rate), as follows:

$$\begin{aligned} Bal\ Acc &= \frac{1}{2} \quad \frac{T_P}{P} + \frac{T_N}{N} \\ Spec &= \frac{T_N}{T_N + F_P} \\ Sens &= \frac{T_P}{T_P + F_N} \end{aligned} \tag{9}$$

where T_P refers to the number of patients correctly classified as CI (true positives), T_N corresponds to the number of controls properly identified (true negatives), F_P quantifies the number of controls labelled as CI (false positives), whereas F_N quantifies the number of CI patients incorrectly classified as controls.

We included two additional metrics: the positive predictive value (precision) and the negative predictive value, as follows:

$$PPV = \frac{T_P}{T_P + F_P} \qquad NPV = \frac{T_N}{T_N + F_N} \tag{10}$$

The relevance of these two metrics is that while positive and negative predictive values depend on the prevalence of the condition in a specific population, whereas sensitivity and specificity depend on the test conducted. Additionally, the area under the receiver

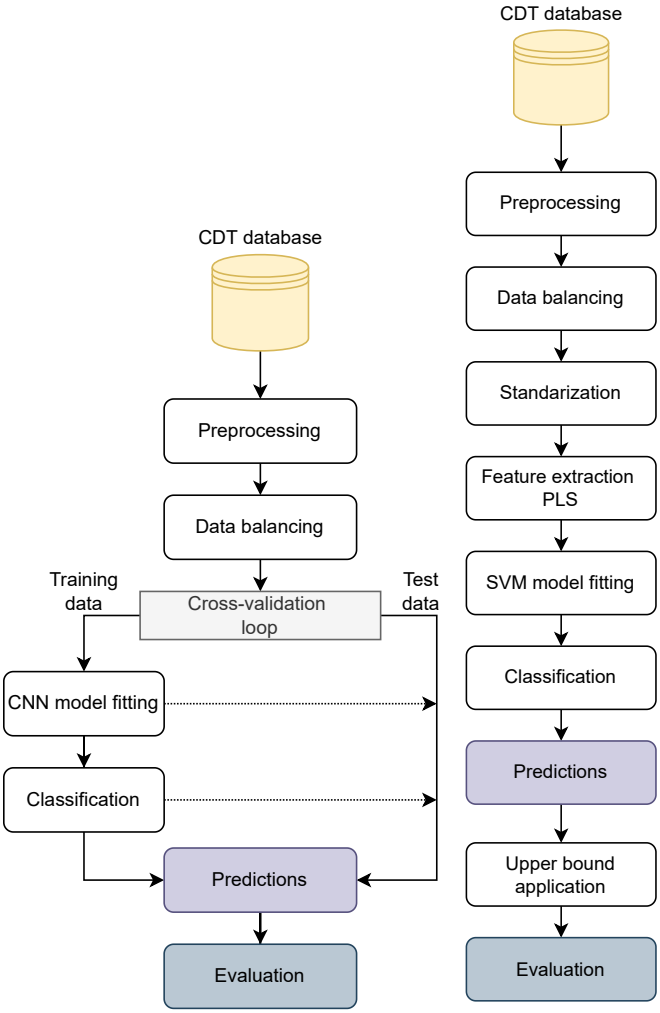

Figure 4. Flowchart of the analysed models. On the left, the flowchart associated with the deep learning-based model, which is based on cross-validation (K-fold). On the right, the flowchart associated with the machine learning-based model, based on resubstitution with upper bound correction as validation procedure.

operating characteristic (ROC) curve was employed as an additional measure for evaluating the ability of the model to identify the different classes.[61,62]

4.6. *Experimental setup*

Since the number of HC is much higher than the CI patients and the sample of cases is not recruited as a population-based cohort, all the experiments were performed with a balanced version of the database where the condition has an a priori probability of 0.5. Thus, the number of drawings finally included in this work was 3282. Table 2 shows the demographic data of the samples contained in this subset.

To estimate the classification performance of the proposed method in the CDT database, two strategies were applied. A 5-fold cross validation strategy in conjunction with the CNN model, and a resubstitution validation strategy based on the VC dimension and a ML approach. Both strategies are illustrated in Figure 4. To do so, we developed custom code written in Python3.6, in addition to employ a number of libraries such as Numpy 1.19.5 and Scikit-Learn 1.0. The experiments were carried out on a cluster with the following hardware specifications: two Intel® Xeon® E5-2630 node 2.40GHz processors, with 10 cores per processor. The total RAM memory capacity of the system is 128 GB.

Table 2. Demographics of the balanced version of the database. The acronym CI stands for cognitive impairment subject, S denotes superior education, NS stands for non-superior education, M represents male and F stands for female.

	CI	Controls	Total
Number of participants	1641	1641	3282
Age	74.36 8.21	72.99 5.51	73.68 7.02
Education (S/NS)	671/970	914/727	1585/1697
Sex (M/F)	709/932	660/981	1369/1913

5. Results

As a preliminary analysis to the classification, the knowledge of the score obtained in the drawing clock test by the 1520 samples from FIDYAN Neurocenter was used to relate different demographic characteristics to this score. Figure 5 contains different box plots that reflect the connection between educational level, gender or age and the quality of the drawing made by the subject. Using the 7009 samples, the complexity of identifying the classes from the estimated correlation matrices for each class is analysed. The estimated correlation matrices of case-control sets are depicted in Figure 6 (top). The experiment was conducted using subsets of samples from healthy subjects, in order to be able to compare matrices generated with the same number of samples in both classes. For this purpose, the HC sample set was split by K-fold cross validation ($K = 3$), so that the so-called test set contained a similar number of samples as the CI set, while the training set was composed of the remaining drawings. The results were randomised through 100 iterations and subsequently averaged as shown in Figure 6 (bottom right), where the thickness of the lines represents the standard deviation. It can be observed that the divergence between the correlation matrices of the two classes (blue lines) is much larger than the estimated divergence between samples of the same class (green line). This is graphically supported by the divergence matrix of both correlation matrices, depicted in Figure 6 (bottom left) where slight differences can be observed. This indicates that classification algorithms should be able to establish separability criteria between classes, since lower dimensionality drawings generate distinguishable correlation matrices.

Final classification results obtained by each approach are shown in Table 3. In the CNN scenario, the values for PPV and NPV are 76.86 1.36 and 74.66 1.84, respectively. The ROC curve obtained is depicted in Figure 7 with an AUC value of 0.8337. In the resubstitution with upper bound correction approach, the accuracy obtained was 72.01% when all the dataset is trained and evaluated, whereas the mean accuracy obtained using the same training subsets as in the CV scenario was 73.37%. In the first case, the upper bound applied was 0.1263 per unit since the sample size was 3282. In the latter case, the upper bound was 0.1394 since the number of samples is lower. In addition, the full unbalanced database was analysed in the CV scenario, where the accuracy obtained was 70.04% with an AUC value of 0.8322. Table 4 provides a summary of the performance metrics obtained in recent works focused in classification systems for diagnosis of CI based on the CDT and our results applying the DL approach.

In order to verify whether the neural network's learning about the distribution of the drawings is correct and makes sense, saliency maps and Grad-CAMs were obtained given an image. Figure 8 shows an example of drawing along with its associated saliency map, illustrated as a heat map (top left) and the different Grad-CAMs depending on whether the label under analysis is that of normal or cognitively impaired subjects (bottom left and right, respectively).

In addition, the saliency maps of the correctly classified images have been averaged, separating them into normal and cognitively impaired samples. This was repeated for different sample sizes, 50, 100, 800 and 1250 samples. These results can be seen in Figure 9.

Finally, based on the probabilities of the last layer of the network, the difficulty of classifying the samples according to class was analysed. A representative graph of this is shown in Figure 10.

6. Discussion

In this work, a classification framework for the automatic classification of CDT is proposed. This ap-

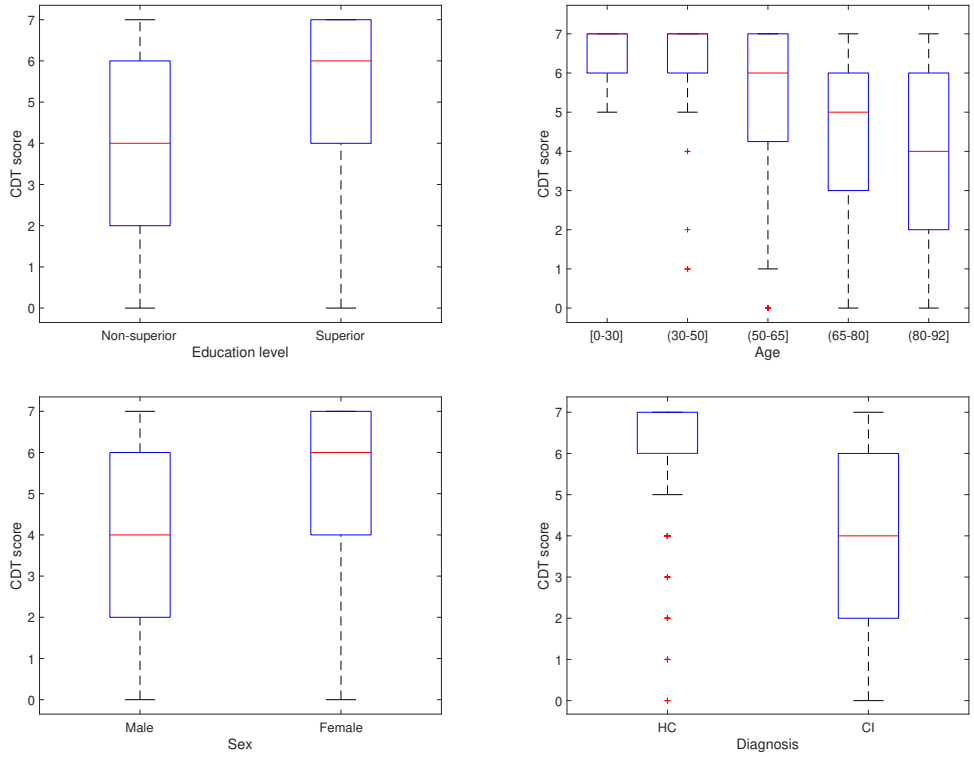

Figure 5. Relationship between (from left to right, from top to bottom) Cdt score and a) education level, b) age, c) sex, d) diagnosis, given the data subset from FYDIAN neurocenter (1520 samples).

Table 3. Classification results obtained using CNN and SVM with their different validation methods. **Experiment where the same 2625 subsets of samples than in the training set of the 5-Fold experiment are evaluated. Its upper bound is 0.1394.

Experiment	CNN		SVM+PLS	
Validation	5-Fold CV		Resubstitution with upper bounding	
Dataset	Test set		All	CV training set
Acc (%)	75.65	1.10	72.01	73.37 0.22
Spec (%)	77.82	2.13	70.67	72.11 0.76
Sens (%)	73.49	2.98	73.35	74.65 0.58
PPV (%)	74.66	1.85	72.97	74.36 0.46
NPV (%)	76.86	1.36	71.11	72.46 0.58
AUC (%)	0.8337	0.0143	0.8013	0.8074 0.0029

proach is based on a preprocessing pipeline where images are cropped, centered and binarized. Once these images are standardized, they are entered into both a ML and a DL approaches that identify the most relevant features of each individual class. The performance in the CDT is evaluated to differentiate between patients diagnosed with cognitive impairment and healthy controls. Besides, we employed activation maps in order to visually verify that the training of the deep learning model is performed correctly. Moreover, we proved reliable results by implementing a model based on theoretical limits with similar results. The underlying hypothesis for this classification is that there are differences between the drawings made by healthy subjects and those with cognitive impairment. As shown in Figure 5d, those with cognitive impairment tend to score lower on the test. Furthermore, this score tends to decrease with age (Figure 5b), which is closely related to the development of cognitive dysfunction.

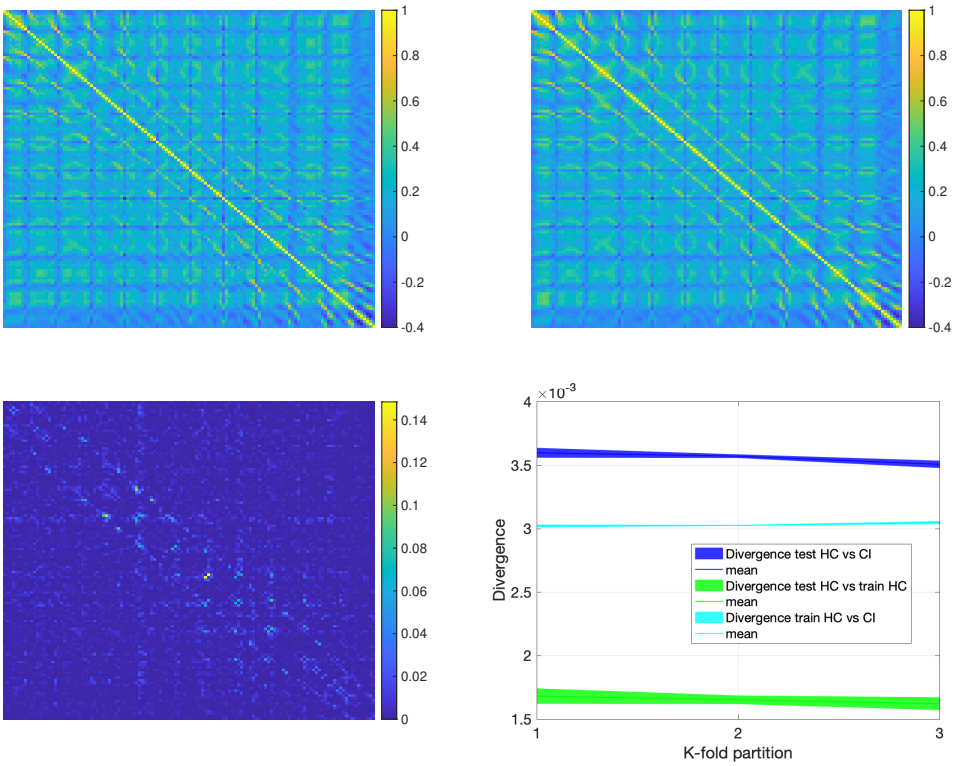

Figure 6. Correlation matrix of the HC (top left) and CI (top right) sample set using images scaled down to 12x12. Case-Control divergence matrix (bottom left). Divergence between the estimated correlation matrices of subsets of samples of healthy subjects and the set related to subjects with cognitive impairment (bottom right). The test HC set contains the same number of samples as the CI set, while the training HC contains the remaining samples. A 3-fold was performed to divide the set of healthy subjects and the results were randomised by 100 iterations. The thickness of the lines represents the standard deviation obtained by averaging the 100 values.

Table 4. Summary of previous works focused on CDT classification automatic in addition to our performance metrics. *Labels in this work were "pass" or "fail" the test, all subjects had at least one positive diagnosis. Symbol - stands for unknown information.

	Reference	Is the face clock preprinted?	Methodology	Patients (CI/HC)	Accuracy	AUC
Digital Clock Drawing Test	63	No	ML methods (best SVM)	2169 (1763/406)	-	0.91
	23	No	Neural networks	198 (163/35)	83.69	-
	24	No	Pretrained MobileNet V2	3423 (160/3263)	95.50	0.8130
	21	No	Random forest	231 (56/175)	90.48	0.8976
Clock Drawing Test	22	Yes	Pretrained DenseNet-121	1315*	96.65	-
	64	No	CNN	747 (293/454)	77.37	-
	Our method	**No**	**CNN**	**3282 (1641/1641)**	**75.65**	**0.8337**
	Our method	**No**	**CNN**	**7009 (1641/5368)**	**70.04**	**0.8322**

The results obtained indicate that the methodology proposed outperforms the expected performance according to Chan et al.,[26] who establish a mean sensitivity and specificity of 0.63% and 0.77% when the paper-and-pencil CDT is analysed. The results are supported by the divergence analysed between classes in Figure 6. Moreover, results from our previous work have been surpassed by more than 7

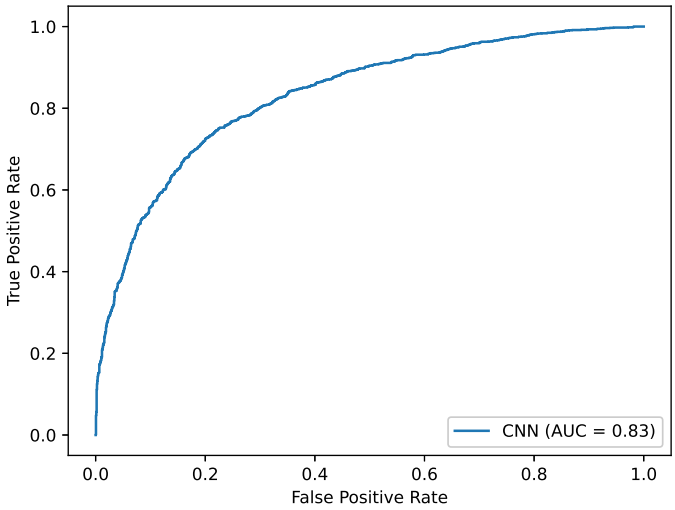

Figure 7. ROC curve obtained by our deep learning approach (CNN model in conjunction with 5-fold cross-validation). The AUC value is also displayed.

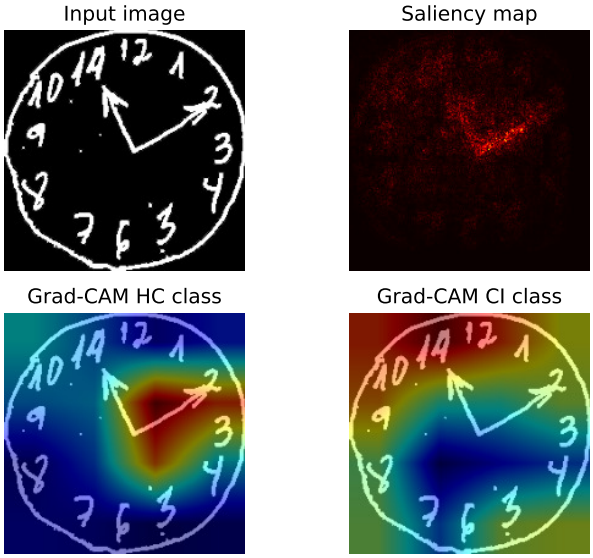

Figure 8. Comparison of different activation maps for a specific input image. The particular draw (top left). The saliency map (top right) is represented like a hot map. Its associated grad-CAMs maps are located in the bottom row, one for each class, HC (left) and CI (right).

points in accuracy.[65] The reason for this improvement is the increase in the number of samples in the database from less than 1000 to more than 3000 samples in the balanced case. Previous studies[23,66,67] have developed systems for the automatic diagnosis of cognitive impairment from the clock-drawing test with better results than those obtained in this work, as it can be seen in Table 4. Nevertheless, these works rely on the use of a digital version of the CDT. This leads to a higher variety of features to be employed, resulting in a considerable increase in performance. We would like to highlight that these devices are not

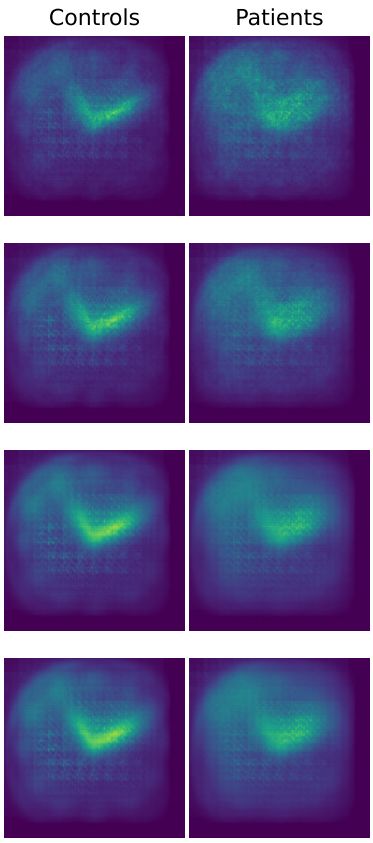

Figure 9. Average saliency maps obtained for each class, controls (left) and patients with cognitive impairment (right), for different samples sizes (from top to bottom n = [50,100,800,1265]). The images used for averaging are those correctly classified in both the training and test set in the fourth fold of the cross-validation approach.

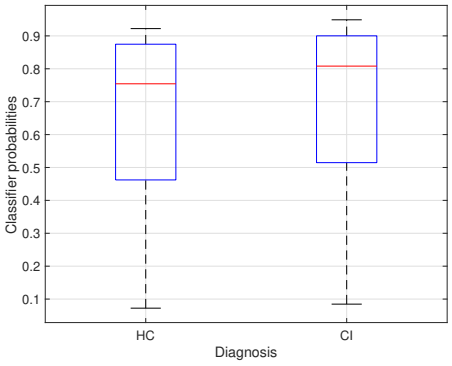

Figure 10. Distribution of the class probabilities obtained as classifier output of well-classified test samples in its corresponding class in the first fold of the cross-validation approach.

common in clinical centers, and even unaffordable for hospitals of some regions. Therefore, it is also necessary to continue the development of automatic classification systems for the original CDT. With respect to the results obtained in other studies using the paper-and-pencil CDT, it should be pointed out

that our method led to a lower accuracy than the one obtained in Ref.[64] However, one of the limitations of that study was the small unbalanced sample size, 747, whereas our sample size is 3282. Therefore, our results have a more reliable generalizability. In addition, as far we know we have implemented the study with the largest number of real samples so far, 7009, although the accuracy rate is lower due to the high unbalance.

Another relevant aspect of our work is that we did not provide a preprinted face clock to patients to make them fill the rest of information.[22,68] This has as an important consequence that all the resulting drawings have a common part: the circumference used as the face clock. When trying to learn the relevant aspects of a clock, using a preprinted circumference decreases the variability between the different drawings. This has two main consequences: first, the differences (if so) between the drawings of both classes (CI and controls) must be inside the face clock. Second, the model can detect easier the presence of the clock since all of them have the same structure. However, the way patients draw the face clock can also contain vital information about their cognitive state. The main drawback is that modeling and identifying the informative patterns associated with each individual drawing is not a straightforward task, since the variability between clocks is much higher. The large performance obtained in this work manifests the ability of our method for accurately extracting the drawing pattern of a person with cognitive impairment, regardless of individual differences in the way the face clock is drawn.

The comparative study conducted on the different validation methods applied suggests that the performance associated with cross-validation and resubstitution with upper bound correction is similar, as has already been proven before.[69] The latter offers the advantage of being able to use the complete database for the evaluation of the results. The upper bound correction implies not to consider the empirical error obtained but the actual one, setting an upper limit on the theoretical accuracy the classifier is able to perform. In this work, we chose Vapnik's bound[60] because it is the best known, but there is a high number of bounds proposals that can be applied.[56,70]

Comparing the results obtained in our previous work[65] with those obtained in this study, it can be seen that the accuracy has increased using the cross validation approach (from 68.62% to 75.65%) while with resubstitution-based approach the results have slightly decreased (from 74.25% to 73.37%). This is due to the increase of the sample size. On the one hand, a methodology based on deep learning requires the use of a large sample size in order to learn and generalise well.[71] Therefore, increasing the database has improved the generalization ability of the model. On the other hand, the machine learning model applied is a linear classifier. Therefore, it is not uncommon that the extension of the database does not lead to an improvement in the results since the classification ability of linear kernels can be limited. But this is must not be considered as a drawback, but a demonstration of the advantages of applying a simpler structure in conjunction with a resubstitution-based method when the sample size is very small, which is common in the field of neuroimaging. This approach has allowed us to obtain similar results with different sample sizes, thus demonstrating the validity of the results from the outset.

From a visual perspective, the results reflect what can be expected from the realisation of the drawing. Figure 8 clearly shows that the neural network focuses on the position of the clock hands during classification. Moreover, while in controls the relevant features are located in central positions, in cognitively impaired subjects this information is around the edges. This is due to the high variability between subjects in these areas, especially those who do not draw a clock correctly. This is supported by Figure 9, where the clock hands are easily identified in the average activation map of healthy subjects but in the map associated with CI patients the zone of interest is much imprecise. Moreover, the hand-clock zone becomes more intense as the sample size increases. Finally, the results shown in Figure 10 reflect the reality that it is easier to classify drawings of patients with cognitive impairment than those of healthy subjects, as they can be more variable.

The fact that a subject is healthy does not imply that their drawing is perfect. This can be observed in the preliminary study shown in Figure 5. For example, educational level leads to a better score (Fig 5a), or even the gender (Figure 5c). The latter may be due to the fact that historically women have tended to be more involved in more educational activities, such as drawing or reading, while men are more associated

with physical and social activities from an early age.

Therefore, the model presented in this work offers reliable results that would allow the CAD system to be implemented as a method to help specialists in clinical tasks. The method performs both the preprocessing and the classification stages and has been tested using a large sample size. Besides, the large performance obtained in the analysis of the paper-and-pencil based clock test demonstrates that is a reliable and cost-effective method for being used as an aid for clinicians in any hospital and research centre.

7. Conclusion

In this work, we propose a method for the automatic diagnosis of cognitive impairment based on the clock-drawing test. This is addressed by employing a preprocessing in which the clock is detected and centered, in addition to binarized in order to reduce the computational cost of the subsequent mathematical operations. Then, a CNN is used to find the relevant patterns of information that characterize CI patients and controls. To this end, graph-theoretical methods are applied, which also allows us to analyse the capacity for generalisation of the CNN. The performance of the model was compared with other approaches, both for classification (support vector machines) and validation (resubstitution with upper bound correction). The performance achieved is in line with what is expected to be obtained using an analogical version of the CDT. The large number of real samples used guarantees the reliability of the results, overcoming most of previous studies where the number of samples was extremely reduced. It is important to note that our method was applied to the most difficult classification scenario: the one based on the analogical CDT without any element of the printed drawing. Thus, our results manifest the suitability of our method in hospitals and medical clinics worldwide, especially in those regions with low resources. The use of the paper-and-pencil based clock test is much cheaper and easier to perform than those based on ballpoint pens, which validates its use in a wide range of scenarios.

Future work will focus on the implementation of more sophisticated classification systems based on ensemble frameworks[29,33,44] in order to obtain a similar performance than when using digitised versions of the test. Besides, the database will be expanded fruit of the collaboration with the clinical entities mentioned in this paper.

Acknowledgments

This work was supported by the MCIN/ AEI/10.13039/501100011033/ and FEDER "Una manera de hacer Europa" under the RTI2018-098913-B100 project, by the Consejeria de Economia, Innovacion, Ciencia y Empleo (Junta de Andalucia) and FEDER under CV20-45250, A-TIC-080-UGR18, B-TIC-586-UGR20 and P20-00525 projects, and by the Ministerio de Universidades under the FPU18/04902 grant given to C. Jimenez-Mesa and the Margarita-Salas grant to J.E. Arco.

Bibliography

1. D. S. Knopman, H. Amieva, R. C. Petersen, G. Chételat, D. M. Holtzman, B. T. Hyman, R. A. Nixon, and D. T. Jones, "Alzheimer disease," *Nature Reviews Disease Primers*, vol. 7, no. 1, may 2021.
2. M. Freedman, L. Leach, E. Kaplan, G. Winocur, K. Shulman, and D. C. Delis, *Clock drawing: A neuropsychological analysis*. Oxford University Press, USA, 1994.
3. K. I. Shulman, "Clock-drawing: is it the ideal cognitive screening test?" *International Journal of Geriatric Psychiatry*, vol. 15, no. 6, pp. 548–561, 2000.
4. C. Carnero-Pardo, I. Rego-García, J. Barrios-López, S. Blanco-Madera, R. Calle-Calle, S. López-Alcalde, and R. Vílchez-Carrillo, "Assessment of the diagnostic accuracy and discriminative validity of the clock drawing and mini-cog tests in detecting cognitive impairment," *Neurología (English Edition)*, vol. 37, no. 1, pp. 13–20, 2022.
5. S. Borson, J. M. Scanlan, P. Chen, and M. Ganguli, "The mini-cog as a screen for dementia: validation in a population-based sample," *Journal of the American Geriatrics Society*, vol. 51, no. 10, pp. 1451–1454, 2003.
6. D. Palsetia, G. P. Rao, S. C. Tiwari, P. Lodha, and A. De Sousa, "The clock drawing test versus mini-mental status examination as a screening tool for dementia: a clinical comparison," *Indian journal of psychological medicine*, vol. 40, no. 1, pp. 1–10, 2018.
7. J. M. Górriz, J. Ramírez, A. Ortíz, F. J. Martinez-Murcia, F. Segovia, J. Suckling, M. Leming, Y.-D. Zhang, J. R. Álvarez-Sánchez, G. Bologna *et al.*, "Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications," *Neurocomputing*, vol. 410, pp. 237–270, 2020.
8. F. C. Morabito, M. Campolo, D. Labate, G. Morabito, L. Bonanno, A. Bramanti, S. De Salvo, A. Marra, and P. Bramanti, "A longitudinal eeg

study of alzheimer's disease progression based on a complex network approach," *International journal of neural systems*, vol. 25, no. 02, p. 1550005, 2015.

9. A. Ortiz, J. Munilla, J. M. Gorriz, and J. Ramirez, "Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease," *International journal of neural systems*, vol. 26, no. 07, p. 1650025, 2016.
10. I. Beheshti, H. Demirel, H. Matsuda, A. D. N. Initiative *et al.*, "Classification of alzheimer's disease and prediction of mild cognitive impairment-to-alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm," *Computers in biology and medicine*, vol. 83, pp. 109–119, 2017.
11. F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, and A. Ortiz, "Convolutional neural networks for neuroimaging in parkinson's disease: is preprocessing needed?" *International journal of neural systems*, vol. 28, no. 10, p. 1850035, 2018.
12. D. Castillo-Barnes, F. J. Martinez-Murcia, A. Ortiz, D. Salas-Gonzalez, J. RamÍrez, and J. M. Górriz, "Morphological characterization of functional brain imaging by isosurface analysis in parkinson's disease," *International journal of neural systems*, vol. 30, no. 09, p. 2050044, 2020.
13. F. Segovia, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, and D. Castillo-Barnes, "Assisted diagnosis of parkinsonism based on the striatal morphology," *International Journal of Neural Systems*, vol. 29, no. 09, p. 1950011, 2019.
14. J. M. Górriz, J. Ramírez, F. Segovia, F. J. Martínez, M.-C. Lai, M. V. Lombardo, S. Baron-Cohen, M. A. Consortium, and J. Suckling, "A machine learning approach to reveal the neurophenotypes of autisms," *International journal of neural systems*, vol. 29, no. 07, p. 1850058, 2019.
15. O. Dekhil, M. Ali, Y. El-Nakieb, A. Shalaby, A. Soliman, A. Switala, A. Mahmoud, M. Ghazal, H. Hajjdiab, M. F. Casanova *et al.*, "A personalized autism diagnosis cad system using a fusion of structural mri and resting-state functional mri data," *Frontiers in psychiatry*, vol. 10, p. 392, 2019.
16. A. Ortiz, F. J. Martinez-Murcia, J. L. Luque, A. Giménez, R. Morales-Ortega, and J. Ortega, "Dyslexia diagnosis by eeg temporal and spectral descriptors: an anomaly detection approach," *International Journal of Neural Systems*, vol. 30, no. 07, p. 2050029, 2020.
17. F. J. Martinez-Murcia, A. Ortiz, J. M. Gorriz, J. Ramirez, P. J. Lopez-Abarejo, M. Lopez-Zamora, and J. L. Luque, "Eeg connectivity analysis using denoising autoencoders for the detection of dyslexia," *International Journal of Neural Systems*, vol. 30, no. 07, p. 2050037, 2020.
18. R. Davis, D. Penney, D. Pittman, D. Libon, R. Swenson, and E. Kaplan, "The digital clock drawing test (dcdt) i: Development of a new computerized quantitative system," *The International Neuropsychological Society*, 2011.
19. W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, C. C. Price, M. Lamar, and D. L. Penney, "Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test," *Machine Learning*, vol. 102, pp. 393–441, 2015.
20. Z. Harbi, Y. Hicks, and R. Setchi, "Clock drawing test interpretation system," *Procedia computer science*, vol. 112, pp. 1641–1650, 2017.
21. A. Davoudi, C. Dion, S. Amini, P. J. Tighe, C. C. Price, D. J. Libon, and P. Rashidi, "Classifying non-dementia and alzheimer's disease/vascular dementia patients using kinematic, time-based, and visuospatial parameters: The digital clock drawing test," *Journal of Alzheimer's Disease*, vol. 82, no. 1, pp. 47–57, Jun 2021.
22. 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*, vol. 10, no. 1, nov 2020.
23. 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, mar 2020.
24. 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, Sep 2021.
25. S. Müller, O. Preische, P. Heymann, U. Elbing, and C. Laske, "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," *Frontiers in Aging Neuroscience*, vol. 9, apr 2017.
26. J. Y. C. Chan, B. K. K. Bat, A. Wong, T. K. Chan, Z. Huo, B. H. K. Yip, T. C. Y. Kowk, and K. K. F. Tsoi, "Evaluation of digital drawing tests and paper-and-pencil drawing tests for the screening of mild cognitive impairment and dementia: A systematic review and meta-analysis of diagnostic studies," *Neuropsychology Review*, oct 2021.
27. J. E. Arco, A. Ortiz, J. Ramírez, Y.-D. Zhang, and J. M. Górriz, "Tiled sparse coding in eigenspaces for image classification," *International Journal of Neural Systems*, vol. 32, no. 03, p. 2250007, 2022.
28. C. Jimenez-Mesa, I. A. Illan, A. Martin-Martin, D. Castillo-Barnes, F. J. Martinez-Murcia, J. Ramirez, and J. M. Gorriz, "Optimized one vs one approach in multiclass classi-

fication for early alzheimer's disease and mild cognitive impairment diagnosis," *IEEE Access*, vol. 8, pp. 96 981–96 993, 2020.
29. A. Ortiz, J. Munilla, J. M. Górriz, and J. Ramírez, "Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease," *International Journal of Neural Systems*, vol. 26, no. 07, p. 1650025, aug 2016.
30. F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, and A. Ortiz, "Convolutional neural networks for neuroimaging in parkinson's disease: Is preprocessing needed?" *International Journal of Neural Systems*, vol. 28, no. 10, p. 1850035, 2018.
31. P. Vuttipittayamongkol and E. Elyan, "Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease," *International Journal of Neural Systems*, vol. 30, no. 08, p. 2050043, 2020.
32. A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *International Journal of Neural Systems*, vol. 32, no. 02, p. 2150058, 2022.
33. A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International Journal of Neural Systems*, vol. 29, no. 04, p. 1850011, 2019.
34. F. Cruciani, L. Brusini, M. Zucchelli, G. R. Pinheiro, F. Setti, I. B. Galazzo, R. Deriche, L. Rittner, M. Calabrese, and G. Menegaz, "Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis," *Journal of Neural Engineering*, vol. 18, no. 4, p. 0460a6, jul 2021.
35. D. Viatkin, B. Garcia-Zapirain, and A. Méndez Zorrilla, "Deep learning techniques applied to predict and measure finger movement in patients with multiple sclerosis," *Applied Sciences*, vol. 11, no. 7, 2021.
36. L. Giancardo, A. Sánchez-Ferro, T. Arroyo Gallego, I. Butterworth, C. Mendoza, P. Montero-Escribano, M. Matarazzo, J. Obeso, M. Gray, and R. Estepar, "Computer keyboard interaction as an indicator of early parkinson's disease," *Scientific Reports*, vol. 5, p. 34468, 10 2016.
37. M. Rauschenberger, R. Baeza-Yates, and L. Rello, "A universal screening tool for dyslexia by a web-game and machine learning," *Frontiers in Computer Science*, vol. 3, 2022.
38. L. Chan, C. Simmons, S. Tillem, M. Conley, I. A. Brazil, and A. Baskin-Sommers, "Classifying conduct disorder using a biopsychosocial model and machine learning method," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2022.
39. M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen *et al.*, "The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270–279, 2011.
40. S. B. GUZE, "Diagnostic and statistical manual of mental disorders, 4th ed. (DSM-IV)," *American Journal of Psychiatry*, vol. 152, no. 8, pp. 1228–1228, aug 1995.
41. P. R. Solomon, A. Hirschoff, B. Kelly, M. Relin, M. Brush, R. D. DeVeaux, and W. W. Pendlebury, "A 7 minute neurocognitive screening battery highly sensitive to alzheimer's disease," *Archives of neurology*, vol. 55, no. 3, pp. 349–355, 1998.
42. P. Soille *et al.*, *Morphological image analysis: principles and applications*. Springer, 1999, vol. 2, no. 3.
43. J. Seetha and S. S. Raja, "Brain tumor classification using convolutional neural networks," *Biomedical & Pharmacology Journal*, vol. 11, no. 3, p. 1457, 2018.
44. M. Leming, J. M. Górriz, and J. Suckling, "Ensemble deep learning on large, mixed-site fmri datasets in autism and other tasks," *International journal of neural systems*, vol. 30, no. 07, p. 2050012, 2020.
45. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
46. 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.
47. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
48. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
49. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.
50. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
51. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.
52. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*,

2016, pp. 2921–2929.
53. J. E. Arco, J. Ramírez, J. M. Górriz, and M. Ruz, "Data fusion based on searchlight analysis for the prediction of alzheimer's disease," *Expert Systems with Applications*, vol. 185, p. 115549, 2021.
54. S. Wold, A. Ruhe, H. Wold, and I. W. J. Dunn, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, sep 1984.
55. J. M. Górriz, J. Ramírez, J. Suckling, I. A. Illan, A. Ortiz, F. J. Martínez-Murcia, F. Segovia, D. Salas-Gonzalez, and S. Wang, "Case-based statistical learning: a non-parametric implementation with a conditional-error rate svm," *IEEE Access*, vol. 5, pp. 11 468–11 478, 2017.
56. J. M. Górriz, J. Ramirez, and J. Suckling, "On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging," *Pattern Recognition*, vol. 93, pp. 1–13, sep 2019.
57. B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
58. A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, A. D. N. Initiative *et al.*, "Lvq-svm based cad tool applied to structural mri for the diagnosis of the alzheimer's disease," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1725–1733, 2013.
59. R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.
60. V. Vapnik, E. Levin, and Y. L. Cun, "Measuring the vc-dimension of a learning machine," *Neural computation*, vol. 6, no. 5, pp. 851–876, 1994.
61. J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, 2010.
62. K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013.
63. W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, C. C. Price, M. Lamar, and D. L. Penney, "Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test," *Machine Learning*, vol. 102, no. 3, pp. 393–441, oct 2015.
64. 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, apr 2021.
65. C. Jimenez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. Zea-Sevilla, A. Ortiz, M. Ávila-Villanueva, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano *et al.*, "Automatic classification system for diagnosis of cognitive impairment based on the clock-drawing test," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2022, pp. 34–42.
66. W. Souillard-Mandar, D. Penney, B. Schaible, A. Pascual-Leone, R. Au, and R. Davis, "Dctclock: Clinically-interpretable and automated artificial intelligence analysis of drawing behavior for capturing cognition," *Frontiers in Digital Health*, vol. 3, 2021.
67. X. Feng, Q. Zou, Y. Zhang, Y. Tang, J. Ding, and X. Wang, "Clock drawing test evaluation via object detection for automatic cognitive impairment diagnosis," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020, pp. 1229–1234.
68. Z. Harbi, Y. Hicks, and R. Setchi, "Clock drawing test interpretation system," *Procedia computer science*, vol. 112, pp. 1641–1650, 2017.
69. C. Jiménez Mesa, J. Ramírez, J. Suckling, J. Vöglein, J. Levin, and J. M. Górriz, "Deep learning in current neuroimaging: a multivariate approach with power and type i error control but arguable generalization ability," *arXiv preprint arXiv:2103.16685*, 2021.
70. F. Laviolette, E. Morvant, L. Ralaivola, and J.-F. Roy, "Risk upper bounds for general ensemble methods with an application to multiclass classification," *Neurocomputing*, vol. 219, pp. 15–25, 2017.
71. C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.