



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A decision support system for Rey–Osterrieth complex figure evaluation

Davide Di Febbo^a, Simona Ferrante^{a,*}, Marco Baratta^a, Matteo Luperto^b, Carlo Abbate^c, Pietro Davide Trimarchi^c, Fabrizio Giunco^c, Matteo Matteucci^a^a Politecnico di Milano, DEIB, via Ponzio 34/5 20133, Milan, Italy^b Università degli Studi di Milano, via Celoria 18 20133, Milan, Italy^c IRCCS Fondazione Don Carlo Gnocchi, Via Capecelatro 66 20148, Milan, Italy

ARTICLE INFO

Keywords:

Cognitive decline
Rey Osterrieth complex figure
Expert systems
eHealth
Computer vision
Deep learning

ABSTRACT

Objective: The Rey Osterrieth complex figure (ROCF) is one of the most used neuropsychological tests for the assessment of mild cognitive impairment (MCI) and dementia. In the *copy* test, the patient has to draw a replica of a 18-pattern image and the outcome is a score based on the accuracy of the overall drawing. The standard scoring system however have limitations related to its subjective nature and its inability to evaluate other cognitive domains than constructional abilities. Previous works addressed those problems by proposing tablet-based automated evaluation systems. Even promising, such methods are still far away from clinical validation and translation. In this work, we developed a decision support system (DSS) for the evaluation of the ROCF copy test in the common practice using retrospective information from previously performed drawings. The goal of our system was to support the professionals providing a qualitative judgement for each of the 18 patterns, estimating the most probable diagnosis for the patient, and identifying the main signs associated to the obtained diagnosis.

Methods: A total of 250 human evaluated ROCF copies were scanned from 57 healthy subjects, 131 individuals with MCI, and 62 individuals with dementia. The images were pre-processed and analysed using both computer vision and deep learning techniques to assign a qualitative label to the 18 patterns. Then, the 18 labels were used as features in 3 binary (healthy VS MCI, healthy VS dementia, MCI VS dementia) and a 3-class classifications with model explanation (SHAP).

Results: Very good to excellent performance were obtained in all the diagnosis classification tasks. Indeed, an accuracy of about 85%, 91%, and 83% was obtained in discriminating healthy subjects from MCI, healthy subjects from dementia and MCI from dementia respectively. An accuracy of 73% was achieved in the 3-class classification. The model explanation showed which patterns are responsible for each prediction and how the importance of some patterns changes according to the severity of the cognitive decline.

Significance: The proposed DSS enriches the standard evaluation and interpretation of the ROCF copy test. Being trained with retrospective knowledge, the performance of the DSS can be further enhanced by extending the dataset with existing ROCF copies.

1. Introduction

Neuropsychological assessment is the main non-invasive instrument for the diagnosis of mild cognitive impairment (MCI) and dementia in older adults (Zucchella et al., 2018). It is performed by administering a set of paper-and-pencil tests during in-person visits which prove various cognitive domains (visuospatial function, memory, attention, executive function, and language). The patient's performance in each of the tests

is determined by a numerical score which correlates with the entity of the cognitive impairment (Donders, 2019). Among the clinical neuropsychological tests, the Rey Osterrieth Complex Figure (ROCF) (Rey & Osterrieth, 1993) is widely used. The ROCF is a drawing test based on reproducing a complex geometrical figure using paper-and-pencil and under the supervision of a clinician. The template figure is shown in Fig. 1 and will be noted with F in the following. The ROCF test is articulated in two variants, which could be administered in sequence

* Corresponding author.

E-mail addresses: davide.difebbo@polimi.it (D. Di Febbo), simona.ferrante@polimi.it (S. Ferrante), marco.baratta@mail.polimi.it (M. Baratta), matteo.luperto@unimi.it (M. Luperto), cabbate@dongnocchi.it (C. Abbate), ptrimarchi@dongnocchi.it (P.D. Trimarchi), fgiunco@dongnocchi.it (F. Giunco), matteo.matteucci@polimi.it (M. Matteucci).

<https://doi.org/10.1016/j.eswa.2022.119226>

Received 5 July 2021; Received in revised form 27 October 2022; Accepted 3 November 2022

Available online 9 November 2022

0957-4174/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

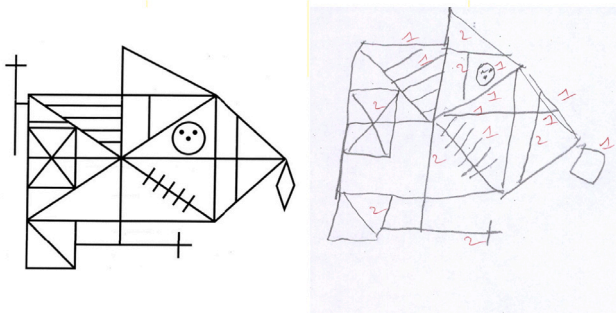


Fig. 1. Example of a ROCF test. A template ROCF on the left and the patient's replica, with the clinician's annotation, on the right.

or mutually, the *copy* and the *recall*. In the copy test, the subject has to sketch the ROCF image while looking at a template figure F as reference. In the recall test, which is sometimes performed after the copy one, the subject has to sketch what he recalls of the ROCF figure F after 30 min and without looking at the reference figure. The complex figure F subject of the tests is composed of 18 geometrical patterns. The score of the test is computed by identifying how the image drawn by the subject (noted with \hat{F}) is similar, in all 18 patterns, to the original one F . The total score is computed summing up all the pattern scores and it ranges from 0 (worst performance) to 36 (best performance) (Bertolani, Renzi, & Faglioni, 1993).

However, the sole ROCF test score in the evaluation of the cognitive decline in older adults presents some limitations as the normal values range between 27 and 36 (Elderkin-Thompson, Boone, Kumar, & Mintz, 2004; Merten & Blaskewitz, 2008). Elderly individuals can have difficulties reproducing the ROCF so that the normal values of coping the figure are generally highly variable. This problem makes hard the definition of a baseline performance and the detection of changes in the long term evaluation (Rasmussen et al., 2001). In addition, some studies investigating the reliability of the ROCF test reported poor inter-rater agreement with a score variation approaching the 20% (Lieberman, Stewart, Seines, & Gordon, 1994; Tupler, Welsh, Asare-aboagye, & Dawson, 1995), mostly due to the operators' subjective experience in the evaluation process.

To copy the ROCF a person should engage several cognitive abilities: attention and concentration, visuo-spatial perception to identify elements and process the visual information; organisational abilities, and executive skills (Broderick, Van Gemmert, Shill, & Stelmach, 2009; Shin, Park, Park, Seol, & Kwon, 2006; Strauss, Sherman, & Spreen, 2006). Such abilities are completely hidden inside the actual overall scoring of the test (Scarpina, Ambiel, Albani, Pradotto, & Mauro, 2016; Westin et al., 2010). The multi-dimensionality of the ROCF test was demonstrated by the multitude of qualitative scoring systems proposed since the creation of the test (Shin et al., 2006). Given the lack of precision and accuracy of scoring and interpretation of the ROCF test, there is the need of developing computer-based systems to help neuropsychologists during the ROCF evaluation.

1.1. Objective

In this study, we aimed at developing a decision support system (DSS) to improve the evaluation and interpretation of the paper-and-pencil ROCF copy test. To design the DSS, we considered the following practical requirements: (1) the DSS should be based on the paper and pencil tests. Indeed, the executions of writing and drawing tasks on a tablet is not ecological and may have an influence on the final performance (Gerth et al., 2016); (2) the DSS should enrich the already available standard methodology for the test evaluation in order to facilitate the translation into clinical practice; (3) the DSS should

exploit explainable AI methods in order to let the clinicians interpret the results obtained by the AI algorithms. To meet the first requirement, the proposed DSS used retrospective data i.e., scanned images of ROCF copy tests, to build the knowledge base of the DSS. In addition, the main goal of our system was not to replicate the overall scoring of the experts, but to enrich the standard evaluation of the test (requirement 2). Indeed, the proposed DSS involved a quick human-in-the-loop initial setup process to select initial control points in the image. Then, it included both computer vision (CV) and deep learning (DL) to produce a qualitative evaluation of each image pattern based on the type of error (omitted, distorted, misplaced and correct) used in the standard evaluation of the test (Rey & Osterrieth, 1993). Starting from the 18 pattern evaluations as input, a machine learning algorithm classified the patient's diagnosis in healthy, MCI and dementia. Finally, to meet the third requirement, we adopted explainable methods to support clinician in identifying which drawing pattern is more important to discriminate between groups of diagnosis. The proposed DSS was calibrated and trained using retrospective copy-test records from 250 mid-aged to older individuals (healthy, MCI, and with dementia). The system performance was assessed by computing the accuracy in the pattern evaluation and various classification metrics in the formulation of the diagnosis.

This paper is organised as follows: Section 2 presents the state of the art analysis, Section 3 the data collection and Section 4 describes the DSS including the whole process of data analysis including: (i) the pre-processing, (ii) the methods applied for patterns evaluation and (iii) the diagnosis formulation. Results are shown in Section 5 and discussed in Section 6. At last, Section 7 discusses the limitations and the possible improvements of the decision support system.

2. Related work

As reported in a very recent survey on handwriting analysis during neuropsychological assessments, the automatic characterisation of visual or procedural biomarkers of brain health is a very interesting field deserving more attention (Moetesum, Diaz, Masroor, Siddiqi, & Vessio, 2022). The methodologies used to analyse a graphomotor response of a neuropsychological test can be grouped in two categories: visual analysis techniques, evaluating drawings/handwriting only statically after completion, and procedural analysis techniques, focusing on the dynamic evaluation of gesture production (Moetesum et al., 2022).

Among visual analysis techniques, early works tried to automate the search of the ROCF patterns location by identifying all the suitable basic geometric shapes in the drawing (Canham, Smith, & Tyrrell, 2000, 2005; Crevier & Lepage, 1997; Fairhurst & Smith, 1991). However, given the high level of distortion of the shapes that could be found in the figures, they accurately identified only 6 patterns including triangles, rectangles, prisms and simple lines, by representing the connectivity of their collinear lines, using the attributed relational graph algorithm (Messmer & Bunke, 1995). Then, they rated the quality of its representation according to basic spatial rules, grouped into categories of position, orientation and size, using fuzzy logic (Zadeh, 1965). The automated rating of the 75% of the patterns, compared with the scores given by 6 independent human raters, showed a discrepancy less than 5%.

Visual analysis techniques used to evaluate the ROCF showed huge challenges in localising and segmenting the specific patterns used to score the test. Indeed, free handwriting is characterised by imprecision, ambiguity, distortion that are even amplified when drawing such a complex figure. More recently, with the growing interest of the AI, the visual techniques were more directed towards global-level analysis of the whole figure based on black box deep learning approaches. As an example, Youn and collaborators adopted deep learning to predict cognitive impairment starting from the ROCF tests (Youn et al., 2021). They classified healthy, MCI and severely cognitive impaired

individuals obtaining accuracy performance of about 70% in the 3-class classification and between 80 and 90% for the binary classification problems. Their work was however restricted to the prediction of the diagnosis and additionally they were not reporting the most difficult binary classification problem (i.e. MCI vs healthy).

Vogt et al. proposed a cascade deep neural network algorithm to estimate both the 18 pattern scores and the overall patient performance starting from 303 ROCF drawings (Vogt et al., 2019). A good performance in the overall accuracy was found despite it was not strictly equivalent to the human ratings scores collected. The focus of such a work was to develop an automated scoring method for the raw image of the test, but they did not provide any additional information to support the clinician in the evaluation of the results. Moreover, Vogt et al. (2019) presented only preliminary results, without the explanation of the full details of their approach.

Among procedural approaches to evaluate the ROCF test, a very recent study, Petilli, Daini, Saibene, and Rabuffetti (2021), proposed a novel tablet-based ROCF automatic evaluation system aimed at differentiating the overall performance in its three main cognitive sub-domains: constructional, organisational and motor abilities. Data collected on healthy subjects during the whole execution of the ROCF copy was used to extract 12 indicators of interest. Using principal component analysis on the indicators 3 components explained the 80% of the variance and thus the 3 composite scores were computed accordingly. A subgroup of 35 healthy subjects was evaluated with an additional battery of tests specifically provided to assess the 3 cognitive domains considered (constructional, organisational and motor abilities). The composite scores resulted by the tablet-based evaluation system moderately to highly correlated (r : ranges from 0.41 to 0.85) with the 3 sub-domains of interest. However, such results should be clinically validated on patients and are still far from becoming a clinical standard.

To conclude, previous works have shown the difficulties in developing an automated evaluation system for the ROCF test able to extract scores equivalent to human raters and, at the same time, provide useful information about the specific cognitive components required to obtain an accurate drawing.

3. Materials

3.1. Data collection

We acquired retrospective data from subjects and neurological patients who underwent a neuro-psychological examination at the Istituto Palazzolo, Fondazione Don Carlo Gnocchi in Milan (Italy), within a time period from Jan 2017 to Dec 2018. During the examination, they performed the ROCF tests. We collected digital versions of the images \hat{F} drawn by the users during the copy variant of the ROCF test, where the subject has the availability of the template image F . Drawn copies are digitalised by scanning the paper records filed in the annual clinical registers. The figures were scanned with a resolution of 300dpi and saved as portable graphics format (PNG) in RGB colours. Personal data (age, years of education), the neurological assessment outcome (healthy, MCI or dementia) and other information such as the date of the visit and the MMSE score were also retained. Both patients and their caregivers provided written informed consent to participation in retrospective studies as this one.

A total of 57 ROCF samples were acquired for healthy subjects, 131 for the MCI and 62 for patients affected by dementia, for a total of 250 ROCF samples. The healthy individual's reported a median age of 75 (IQR 70–81), median years of education (y.o.e.) of 8 years (IQR 8–13) and median MMSE of 29 (IQR 28–29.5); the MCI patients had a median age of 79 (IQR 75–83), median y.o.e. of 8 years (IQR 8–11) and median MMSE of 27 (IQR 26–29); while the dementia patients' median age was 82 (IQR 79–85), the median y.o.e. was 8 (IQR 5–10) and the median MMSE was 22 (IQR 22–24).

3.2. DataSet

To get the ground truth for the pattern evaluation, each of the 18 patterns of the image \hat{F} was inspected by an expert and labelled with a score. More precisely, the outcome of the tests is the total score computed as the sum of the single scores the clinician assigned to each one of the patterns as drawn by the subject. To evaluate the accuracy of the drawing, a standard protocol is the Osterrieth system (Osterrieth, 1994). A continuous numerical score (between 0 and 2) is given to each pattern according to the quality of its representation. The guidelines suggest 0 if indistinguishable or absent, 0.5 if deformed and misplaced, 1 if correct and misplaced or just deformed, and 2 if correct and well placed in the figure. In this study, we considered a simplified pattern scoring system based on four qualitative scores, each one representing a single error type category¹:

- 0 (*omitted*), if the pattern was not represented nor recognisable in the figure;
- 1 (*distorted*), if a distorted,² yet recognisable, version of a pattern was represented in the figure;
- 2 (*misplaced*) if the pattern was not distorted but placed differently from the expected location;
- 3 (*correct*) if all the previous conditions did not applied.

The sharp categorisation of the Osterrieth system's pattern scores in error types simplified the manual inspection and labelling of the dataset and allowed us to configure the pattern evaluation problem as 4-classes classification instead of a regression between 0–2, as it would be the case of the Osterrieth system.

In the final dataset, each ROCF sample collected was associated to the 18 scores assigned manually by a clinician to each pattern (each score ranged from 0 to 3 as explained above) and an overall label correspondent to the clinical outcome of the neuropsychological visit (the label can be dementia, MCI, healthy). The composition of the dataset, including the number of occurrences for each one of the patterns, divided for each label, is reported in Table 1.

4. Methods

4.1. Workflow

The analysis and automated evaluation of images \hat{F} consists of a set of steps executed in sequence, which are explained in the following sections. As images \hat{F} had evaluation marks made by clinicians, see Fig. 1, the first step of the system consisted of the *image pre-processing* in which the ROCF samples were cleaned from marks and standardised (Section 4.2). Then, the analysis of a sample ROCF image was implemented in two main stages: the *pattern evaluation* (Sections 4.3 and 4.4) and the *diagnosis formulation* (Section 4.5). Pattern evaluation consists of an initial *detection* step, in which a pattern is searched in the figure to determine its presence or absence, and an *evaluation* step in which a label between 0,1,2 or 3 was assigned to the pattern. A schematic representation of the method workflow is reported in Fig. 2. The patterns of ROCF were divided into two categories according to their complexity, simple patterns and complex ones, as shown in Fig. 3. During their evaluation, the detection from \hat{F} of those patterns was dealt with two different methods; simple patterns are detected using computer vision (Section 4.3); complex patterns are detected using deep learning (Section 4.4).

In the diagnosis formulation stage, the labels assigned to each pattern by the tool were used as features for the classification of the

¹ The labels were treated as categorical. Yet they ranked from 0 to 3 reflecting the correctness of the drawing

² The patterns which were not topological equivalent to the template were labelled as *distorted*.

Table 1
The count of the patterns, divided for each label.

Pattern number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Omitted	15	26	55	170	43	67	33	34	70	55	79	36	73	18	24	39	34	34
Distorted	104	139	76	16	107	0	0	153	30	55	23	101	65	55	46	11	94	112
Misplaced	42	0	43	0	8	12	56	0	36	10	8	0	8	15	24	38	0	0
Correct	89	85	76	64	92	171	161	63	114	130	140	113	104	162	156	162	122	104

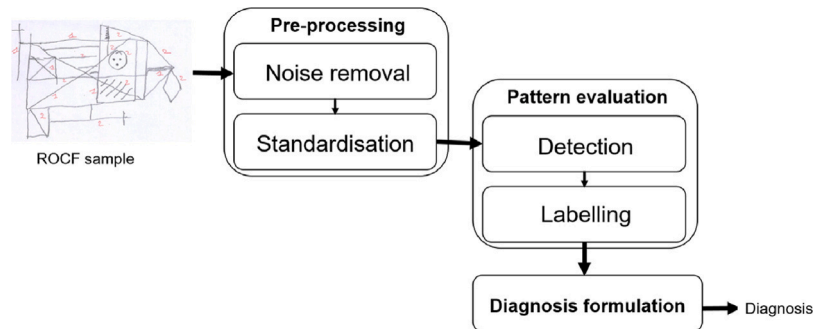


Fig. 2. The workflow of the tool.

ROCF copy \hat{F} in one of the 3 clinical outcomes, thus determining the subject's most probable diagnosis. Furthermore, we then analyse which patterns were most important into the classification process to provide such data as a diagnosis support mechanism to the clinician.

4.2. Image pre-processing

The pre-processing stage of images \hat{F} started with the noise removal, which was mostly characterised by the clinician's annotations during the ROCF copy evaluation. The clinician's signs consisted in the numerical score (marks) they sketched upon the drawing to track the patterns they already evaluated. In all the collected images, scores were written using red and green ink pen, while the ROCF lines were drawn with a common pencil. This difference between the clinician's and patient's signs allowed the noise removal through the use of colour filtering techniques. The images have been converted in the hue-saturation-value (HSV) colour space (Levkowitz & Herman, 1993) to easily identify the colour shades to remove. We found the HSV triplets [0 30 10], for red ink, and [160 30 10], for green ink, as the optimal values to identify the noisy pixels (clinician's marks) in the image. Then, we replaced the colour of those pixels with the most frequent pixel colour in \hat{F} , which we assumed to be the background colour of the paper sheet. Other spurious noise component were attenuated using low pass and median filtering with 3×3 kernels.

After noise removal, images have been binarized. We used unimodal thresholding (Rosin, 2001) to separate the pixels belonging to the drawing (set to 0 grey-scale intensity, *black*) from the ones belonging to the page (set to 255 in grey-scale, *white*). However, the handmade drawings \hat{F} were still imprecise. For examples they presented strokes with irregular thickness and non-perfectly closed shapes. To attenuate the effect of these additive disturbances, we first used image erosion with a 9×9 kernel to cover all the gaps between close tracts. Then we applied a Skeletonization algorithm (Abuain, Abdullah, Bataineh, Abu-Ain, & Omar, 2013) to the negative-binary image to obtain one pixel wide lines. At last, we dilated the drawing with a 3×3 kernel to enhance the objects in the image.

All images were very different in shapes, dimensions and proportion. Therefore a figure standardisation was required before further analysis. This required the support of an expert user, who manually selected, with a GUI embedded within our method, five reference point in the image. Those points, indicated in Fig. 4(a), identify the four vertices of the pattern 2 of the ROCF plus the rightmost point of the figure, which coincided with a vertex of the right triangle. The 5

reference points were used to perform an image homography (Szeliski, 2011) to match the reference points to their respective locations in the template model, preserving the structure of the patient's drawing. The template model image was a binary representation of the original ROCF centred in a 428×733 pixels binding box. An example of the figure standardisation is shown in Fig. 4 (b). Noise removal, image processing and the homography were implemented using the image processing Python libraries Open CV 4.5.1.48 (Bradski, 2000). Noise removal was applied using the Median Blur operator with a 3×3 kernel. The proposed standardisation method should reduce localisation and segmentation challenges in the following steps of analysis.

4.3. Simple patterns evaluation via computer vision

As each pattern was located into a different characteristic part of the image, the initial coordinates of the area in the image in which a pattern was searched, i.e. its region of interest (ROI), was chosen from the template model as shown by the red shapes in Fig. 3. The initial ROI of a pattern could be adjusted during the analysis, using the information of those previously detected. For the detection of some patterns, additional rotated variants of the initial ROI were considered (the parameters m_{-clk} and m_{-cjk} determined the clockwise or counter-clockwise inclination respectively), as shown in Fig. 5. The pattern were iteratively searched by moving the ROIs in the horizontal and in the vertical direction. In every iteration, A ROI was shifted by m_{sh} pixels, until it reached the image limits or additional pattern-specific bounds. In the case of the simple patterns (those characterised by an easy and regular geometrical structure, as a line, a regular polygon or multiple similar elements) we applied CV algorithm for their automatic detection and evaluation. The presence or absence of a simple pattern was determined using *line detection* algorithm, for patterns number 1, 4, 6, 7, 8, 11, 13 and 16, which were mainly composed by lines, and *shape detection* techniques, for the numbers 2, 9 and 14, which consisted in shapes. The identified pattern were then evaluated using *topological analysis*, as described in Section 4.3.3.

4.3.1. Line detection

Line detection is performed using the probabilistic Hough transform algorithm implementation available in the OpenCV Python library (Matas, Galambos, & Kittler, 2000). The algorithm's parameters radius resolution r and the angle θ have been fixed to 1 pixel and to $180/\pi$ respectively. The other parameters, namely the threshold for the minimum number of edges to detect a tract (th_e), the minimum line

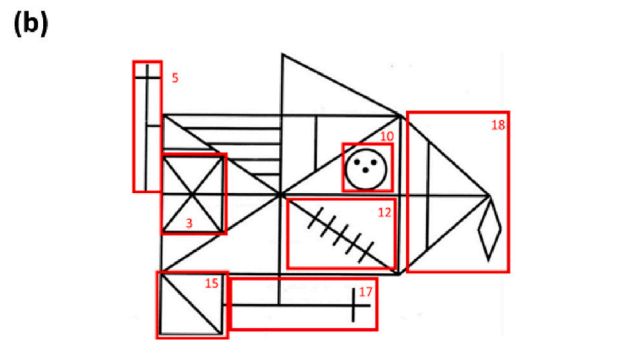
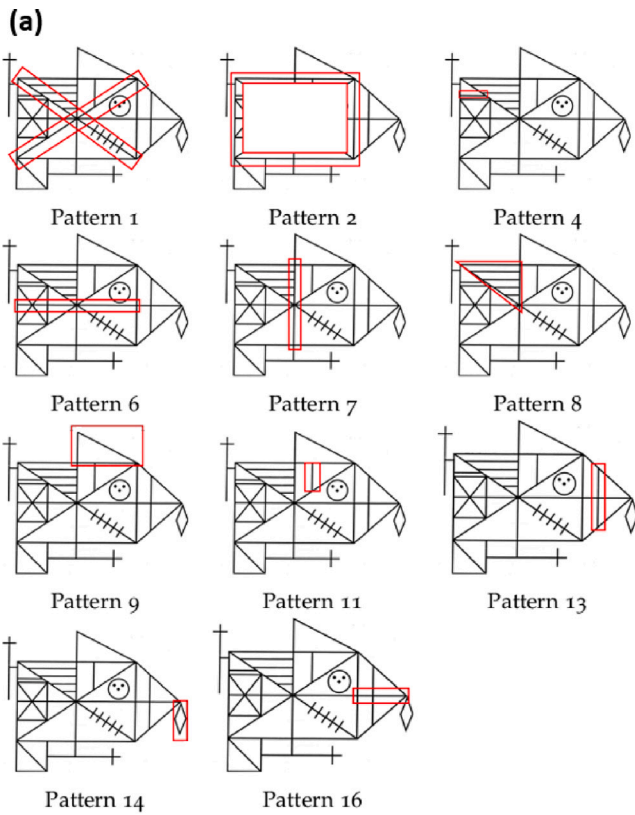


Fig. 3. The original ROCF figure F is composed of 18 patterns. We divide them in two categories, simple and complex patterns, according to the difficulty to automatically identify them in copied figures \hat{F} . Panel (a) shows the simple patterns and panel (b) shows the complex patterns. The red squares indicate the region of interest of each pattern.

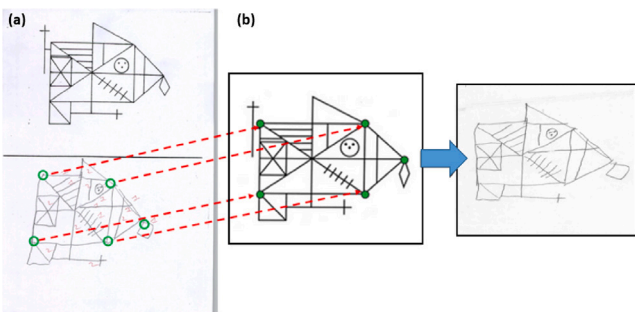


Fig. 4. Example of the standardisation of a ROCF sample, Panel (a) represent the scan the of paper sheet where the template (in the upper part) and the patient's replica (lower part) of the ROCF are figured. The green dots indicates the reference points selected by the examiner and the red lines associate the drawing reference points to the ones of the template model. Panel (b) shows the ROCF sample after the homography transformation.

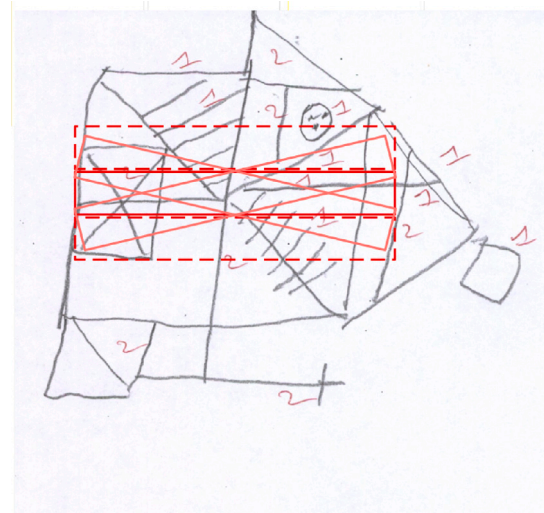


Fig. 5. An example of rotating and shifting the ROI for a pattern (Pattern 6 in this case). The original one has a continue red line. Other shapes are the variants.

segment length (l_{min}) and the maximum number of gap between points allowed in a line (max_g), were set differently for each pattern. The algorithm returned the coordinates of all the detected line edges. The lines belonging to the adjacent patterns and the other noise components were filtered out by setting a specific threshold (t_a) on their inclination w.r.t. the horizontal direction. Then, a set of lines was selected by counting the percentage of the pixels coinciding with those of the original pattern overlapping the sample ROCF and the template model. By defining a threshold on such percentage (A_c) we decided how much a collection of detected segments had to cover the template pattern to be identified as a pattern. From the chosen set of segments, we approximated a single line using linear regression with the edges of all the lines in the set. In Pattern 8, composed by more than one line, the procedure was repeated for each line separately.

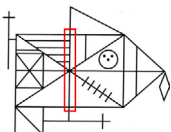
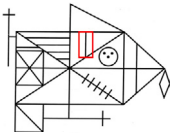
4.3.2. Shape detection

For shapes recognition, the contour detection method in OpenCV (Suzuki & be, 1985) was used to find a set of geometrical objects in a ROI. To approximate the obtained contours to a regular figures, the OpenCV implementation of the Douglas–Peucker’s iterative end-point fit algorithm (Douglas & Peucker, 1973) was used. Given a curve composed of line segments, a similar closed curve with fewer points was returned. Then a single shape was selected applying the following criteria: (i) the convex shapes with the same number of vertices of the pattern to search was checked, evaluating the convex hull (Schmidtman, Jennings, & Kingdom, 2016); (ii) if more than one shape met the requirements of (i), a threshold P_{min} was set to discard all the shapes whose perimeter was inferior to such value; (iii) if there was still more than one eligible shape, a hierarchical representation of contours was computed and the innermost node (i.e. the outermost shape) was retained.

4.3.3. Topological analysis of simple patterns

The label was assigned to a pattern in the sample ROCF according to the following rules: (i) if no line or shape was detected in a ROI, the score *omitted* was assigned to that pattern; (ii) if some shape was found, but none with the right number of vertices, the score *distorted* was assigned to the convex shape with a perimeter longer than P_{min} ; (iii) otherwise, the detected object, defined by the set of its edges coordinates, was topologically analysed using the Python library Shapely (Gillies et al., 2007). The object was inspected to check if it satisfied a set of structural properties (as continuity, intersections and

Table 2
Topological analysis rules for the evaluation of the simple patterns.

Pattern 2	<i>Main rectangle.</i> The main rectangle was identified by 4 of the reference points selected by the user, transformed by the homography. The topological analysis returned: <i>correct</i> , if rectangular shape was found; <i>distorted</i> , if no rectangular shape was found.	
Pattern 1	<i>Main diagonals.</i> The initial ROI was split in four part, each of them was dedicated to the detection of one half of a diagonal. The topological analysis returned: <i>distorted</i> , if at only one half per diagonal has been detected or at least two halves of the same diagonal did not intersect; <i>misplaced</i> , if none of the previous conditions applied and at least one vertex did not intersect the corresponding vertex of pattern 2; <i>correct</i> , if none of the previous conditions applied.	
Pattern 4	<i>Short horizontal line.</i> The topological analysis returned: <i>distorted</i> , if more than one lines was detected; <i>correct</i> , if a single line was detected.	
Pattern 6	<i>Horizontal line.</i> If pattern 2 was previously found, the ROI ⁶ ₀ was centred with its horizontal axis of symmetry. The topological analysis returned: <i>distorted</i> , if a line was detected and at least one of its vertices did not intersect an edge of pattern 2; <i>misplaced</i> , if none of the previous conditions applied and the line did not intersect pattern 1 in a single point; <i>correct</i> , if none of the previous conditions applied.	
Pattern 7	<i>Vertical line.</i> If pattern 2 was previously found, the initial ROI was centred with vertical axis of symmetry. The topological analysis returned: <i>distorted</i> , if a line was detected and at least one of its vertices did not intersect an edge of pattern 2; <i>misplaced</i> , if none of the previous conditions applied and the line did not intersect pattern 1 in a single point; <i>correct</i> , if none of the previous conditions applied.	
Pattern 8	<i>Parallel lines.</i> Pattern 8 was composed by four parallel segments so the initial ROI was divided in four sub-parts. The procedure was repeated iteratively by searching a segment in each sub-ROI. When a line was detected, the respective sub-ROI was excluded in the next iteration. The topological analysis returned: <i>distorted</i> , if less or more than four lines were detected and they did not intersect each other; <i>correct</i> , if none of the previous conditions applied.	
Pattern 9	<i>Topmost triangle.</i> Pattern 9 was composed by four parallel segments so ROI ⁹ ₀ was divided in four sub-parts. A segment was iteratively searched in each sub-ROI. When a line was detected, the respective sub-ROI was excluded in the next iteration. The topological analysis returned: <i>distorted</i> , if a convex shape with more than 3 edges was found; <i>misplaced</i> , if none of the previous conditions applied and the shape had no intersections with other lines of the figure; <i>correct</i> , if none of the previous conditions applied.	
Pattern 11	<i>Short vertical line.</i> The topological analysis returned: <i>distorted</i> , if more than one line was detected or one line was detected but it did not intersect either patterns 1 and 2; <i>misplaced</i> , one line which intersect either patterns 1 and 2 is detected and it intersect the two diagonals in the same point; <i>correct</i> , if none of the previous conditions applied.	
Pattern 16	<i>Right horizontal line.</i> The topological analysis returned: <i>distorted</i> , if more than one line was detected or one line was detected but it did not intersect patterns 2 and 6; <i>misplaced</i> , if none of the previous conditions applied and the line did not intersect the manual rightmost point; <i>correct</i> , if none of the previous conditions applied.	
Pattern 13	<i>Right vertical line.</i> The topological analysis returned: <i>distorted</i> , if more than one line was detected or one line was detected but it did not intersect at least one inclined edges of the delimiting triangle; <i>misplaced</i> , if none of the previous conditions applied and the line intersected the pattern 16 in its rightmost half; <i>correct</i> , if none of the previous conditions applied.	
Pattern 14	<i>Rhombus.</i> The topological analysis returned: <i>distorted</i> , if a convex shape with more or less than 4 vertex was detected; <i>misplaced</i> , if none of the previous conditions applied and the shape did not intersect the rightmost manual point. <i>correct</i> , if none of the previous conditions applied.	

gaps) to be considered a *correct* drawn pattern. If the properties were partially satisfied the pattern could be labelled whether *misplaced* or *distorted*.

Empirically, several straight lines were in fact drawn as a set of sequential segments. Similarly, several edges common between different

segments (forming, e.g., a 90° angle) present, due to small inaccuracies, small ‘gaps’ that could led to miss the detection of the edge itself. To cope with this, the boundary margins were considered to account for these types of natural inaccuracy of hand-made drawings. Therefore, the thickness of the lines and shapes was increased according to two

Table 3
Parameters setting of the computer vision algorithms for the detection and evaluation of the simple patterns.

Pattern n°	Definition of the ROI variants			Line/shape detection						Topological analysis	
	m_{sh} [pixel]	m_{clk}^- [°]	m_{clk}^+ [°]	th_e	l_{min} [pixel]	max_g	t_a [°]	A_{cov} [%]	P_{min} [pixel]	d_v [pixel]	d_l [pixel]
1	20	30	30	50	40	20	20–60	60	–	15	1.5
2	–	–	–	50	40	20	–	–	500	30	–
4	10	5	5	40	20	5	<10	60	–	20	1.5
6	15	10	10	75	40	20	<10	80	–	15	1.5
7	15	15	25	75	40	20	>70	80	–	15	1.5
8	5	5	12	30	10	5	<15	30	–	15	1.5
9	–	–	–	–	–	–	–	–	200	25	3
11	10	10	30	50	20	10	>60	60	–	15	1.5
13	10	10	30	50	20	10	>80	50	–	15	2
14	–	–	–	–	–	–	–	–	100	20	1.5
16	10	10	30	45	60	50	<10	80	–	15	3

parameters, d_v for each vertex and d_l for each segment, as margin error, in pixels, resulting ultimately in thickening the drawing and closing ‘gaps’ in straight lines. The pattern-specific properties to be satisfied in the topological analysis are listed in Table 2, in the same order as the research in the algorithm occurs. The particular parameter choice for the CV algorithms used for the detection of simple patterns is reported in Table 3, for each of the patterns.

4.4. Complex patterns evaluation via deep learning

The detection of complex patterns is performed by a DL algorithm that finds the regions in the ROCF sample image \hat{F} most similar to the template F . The measure of similarity has been defined as the euclidean distance (L_2) between two images, calculated by mapping them into a 1024-dimensional embedding space (Schroff, Kalenichenko, & Philbin, 2015). Two images are the more similar the shorter the L_2 distance between their vectors in the embedding space.

The embedded representation of the images was obtained using a modified ResNet50V2 neural network architecture (He, Zhang, Ren, & Sun, 2016). The fully-connected layer on top of the network was removed, the output of the residual part flattened, and a 1024-rectified linear units (ReLU) fully-connected layer was added. A triplet loss, reported in Eq. (1), was chosen as cost function, which has been shown to be efficient for this type of tasks (Schroff et al., 2015). The triplet loss, Eq. (1), encourages the images of the same pattern to be projected onto very close points in the embedding space and it also enforces the margin between images of different objects by considering triplets of vectors.

$$L = \max(m + D(\xi_a, \xi_p) - D(\xi_a, \xi_n), 0). \tag{1}$$

In the equation of the loss, ξ_a is the vector of a reference image (the anchor) in the embedding space, ξ_p is the vector of an image of the same object of the anchor (the positive) and ξ_n is the vector of an image of a different object (the negative); $D(\xi_i, \xi_j)$ is the squared euclidean distance between the vectors of i and j , and m is the margin between positive and negative pairs. The loss minimisation must also satisfy the constraint in Eq. (2),

$$D(\xi_a, \xi_p) + m < D(\xi_a, \xi_n), \tag{2}$$

therefore, $D(\xi_a, \xi_p)$ was pushed to zero and $D(\xi_a, \xi_n)$ to be greater than the former plus m .

We trained a different network for the recognition of each complex pattern. Pattern-specific datasets were created by manually cropping the pattern representations from the ROCF of all the subjects. For the similarity measurement, correct and distorted patterns only were considered. The scarce amount of samples retrieved was incremented ten times applying the following data augmentation techniques. The Python library ImgAug (Jung et al., 2020) was used to perform the following image transformations in random order:

- Gaussian blur with variance ranging from 0 to 0.5,

- Aspect ratio preserving scaling with a factor ranging from 0.85 to 1.15,
- Rotation by -10 to 10 degrees,
- Shear mapping by -15 to 15 degrees,
- Translation by -40 to 40% on x -axis and y -axis independently.

Datasets were strongly unbalanced as the portion of wrongly drawn patterns was consistently lower. In particular the percentage of correctly drawn patterns in images \hat{F} was 66%, 77.4%, 86.7%, 70.8%, 88.8%, 72.5% and 70.2% for patterns 3, 5, 10, 12, 15, 17 and 18 respectively. Therefore, heavier forms of image augmentation to 100 correct samples to obtain other 1000 ‘distorted’ versions were performed, with the aim of re-balancing the datasets. The new data augmentation parameters were manually set to better resemble the actual distorted samples.

To train a network, a batch of 32 pattern images was randomly extracted from a dataset and each of them was paired with the same pattern of the template model. A positive pair was generated if the image was labelled as correct, or a negative pair if it was labelled distorted. The network computed the L_2 distance between the images of each pair and the triplet loss was computed using the batch-hard strategy (Pereira & Campos, 2020), i.e. by selecting the hardest positive pair (with the maximum L_2) and the hardest negative pair (with the minimum L_2) only. The training consisted in 50 epochs in which the dataset was split in training and validation/test set by randomly picking the 30% of the samples from both the correct and distorted class. The Adam optimiser (Kingma & Ba, 2014) was used with a learning rate of 0.0001 and the model with the lowest validation loss value among each epoch was retained.

In the complex pattern detection, the set of its initial ROIs was moved in the vertical and horizontal direction for a maximum distance of 50 pixels. For each pattern, the ROI with the maximum value of similarity with respect to the template was retained. Then, the assignment of the label according to the following procedure:

- *Omitted*. A linear support vector machine (SVM) classifier was trained and used to discriminate between the presence or the absence of the pattern counting the portion of non-white pixels contained in the ROI. Pattern-specific dataset were used, including the cropped samples of all the omitted and the non-omitted patterns. The balanced accuracy of the classification was estimated with the leave-one-out cross-validation (LooCV); The label *omitted* was assigned when the SVM returned the absence for the pattern.
- *Misplaced*. If a pattern was detected, topological analysis was applied to examine its correct location by checking all expected intersections with the surrounding patterns;
- *Distorted* or *Correct*. If the pattern was neither omitted nor misplaced, a second linear SVM classifier was trained and applied to discriminate between a distorted and a correct representation, using the similarity measure above defined. Pattern-specific dataset, with correct and distorted patterns only, were used as training data and the L_2 from the template pattern was considered as single input feature. The balanced accuracy of the classification was estimated with LooCV.

4.5. Diagnosis formulation

The diagnosis formulation stage was aimed at associating the most probable subject’s diagnosis (between *healthy*, *MCI* and *dementia*) to each ROCF sample, using the 18 labels assigned to its patterns as predictors. The ability of the system to discriminate between groups was investigated by setting various classification tasks. Four classifications were considered: *healthy vs MCI*, *healthy vs dementia* and *MCI vs dementia*; and a multi-class classification task including the 3 groups with all the samples. Since the classes were unbalanced, the four classification tasks were also performed with new datasets created by randomly sampling (without replacement) 50 elements per class and by averaging the outcomes of 50 iterations, for a more robust estimate. A state-of-art boosting algorithm, Catboost (Dorogush, Ershov, & Gulin, 2018), was trained to solve the classification tasks by choosing a weighted cross entropy loss function (Phan & Yamamoto, 2020) and setting a number of 500 iterations. The performance were evaluated by estimating the Accuracy, F1, Precision and Recall scores with the Leave-one-out cross-validation (LooCV). The normalised numerical labels of the patterns (0, 1, 2, 3) were normalised (between 0 and 1) and used as input features. The classification tasks were implemented with the Python library sciKit-learn.

4.5.1. Model explanation

The binary classification tasks were further analysed by applying the model explanation technique SHAP (Lundberg & Lee, 2017). Here, 3 randomly sampled datasets with 50 samples per class were used to avoid the stronger influence of the more numerous group in the classification. SHAP uses game theory to rank the features (i.e. the patterns) importance and to assess the contribution of each feature in the binary classification of a ROCF sample in the trained models. The single feature contribution in the classification of each sample was quantified by a weight (the Shapely value) which moved its prediction towards a class or the other, if negative or positive, by an amount proportional to its magnitude. The features rank was then obtained by considering the average absolute weight of each feature for each sample. With this technique, one could better interpret the effect of each single pattern in the binary classification tasks and thus appreciate the importance of the patterns in discriminating between healthy and pathological individuals and between different levels of cognitive decline.

5. Results

5.1. Simple and complex patterns evaluation

For each simple pattern, the 4-class evaluation Accuracy was calculated as the percentage of the correctly labelled patterns over the total number of patterns (with the labels 0, 1, 2 and 3). The Accuracy scores are reported in Table 4, for the simple patterns, and in Table 5, for the complex ones. Table 5 shows the balanced Accuracy of the SVM models (the first predicting the absence/presence and the second predicting the correct/distorted representation of the pattern) and the Accuracy for each of complex patterns. The Accuracy score corresponded to a 4-element labelling task and lower accuracy scores were observed for the patterns 3, 5 and 9 (61.9%, 61.4% and 61.9% respectively). Patterns 6, 7, 16 and 17 achieved the highest Accuracy scores (78.9%, 77.6%, 74.9% and 77.1% respectively) and the rest of the patterns gained an average Accuracy score of 68.8%.

5.2. Diagnosis classification

The results of the 4 classification tasks (healthy vs MCI, healthy vs dementia, MCI vs dementia and 3-class task) in terms of accuracy F1 precision and recall are reported in Table 6. On the left are listed the outcomes of the classifications made using all the samples of each group. On the right, the average outcomes (with standard deviation) are

Table 4

Simple pattern scoring accuracy.	
Simple patterns	Acc [%]
1	65.5
2	66.8
4	72.2
6	78.9
7	77.6
8	68.2
9	61.9
11	72.2
13	71.3
14	68.2
16	74.9

Table 5

Complex pattern scoring accuracy.			
Complex patterns	Omitted/others [balanced acc.]	Correct/distorted [balanced acc.]	Acc [%]
3	0.94	0.81	61.4
5	0.81	0.76	61.9
10	0.91	0.77	70.0
12	0.83	0.86	68.2
15	0.89	0.89	65.9
17	0.71	0.82	77.1
18	0.65	0.82	68.2

shown for the tasks performed with random sampled datasets (N = 50 subjects per group). High to excellent performance were obtained in the binary classification tasks with all the samples, with Accuracy scores ranging from 87% (healthy vs dementia) to 92% (healthy vs dementia); F1 between 79% (MCI vs dementia) and 93% (healthy vs MCI); Precision between 85% (MCI vs dementia) and 100% (healthy vs dementia); and Recall between 74% (MCI vs dementia) and 90% (healthy vs MCI). Similar performance were achieved mediating the outcomes of the classifications with the random sampled dataset. The lower performance were obtained in the more complex multi-classification tasks: scores from 73% to 76% were obtained for the Accuracy and scores from 73% to 74% for the other metrics. The confusion matrices and the receiving-operator curves (ROC) for the 4 classifications in which all the samples were retained are shown in Fig. 6.

The explainable artificial intelligence SHAP tool was used to better interpret the models decision by estimating the single contribution of the patterns in the determination of a subject’s diagnosis. Furthermore, it supplied insights on the patterns importance to different levels of severity of the cognitive decline. The SHAP tool allowed the automatic analysis of the classification results giving the Shapely values as output, which in turn were used to estimate the pattern importance and impact in the determination of the samples diagnosis. The Shapely values indicated how much a pattern score pushed the prediction towards a class: negative values favoured the healthy class (or the less cognitive impaired class between the two groups considered), while positive values favoured the dementia class (or the more cognitive impaired one). The average absolute Shapely values of each pattern are displayed in Fig. 7, column (a), for the 3 tasks. The SHAP analysis revealed that patterns 9, 11, 3, 6 and 4 were the most important in the discrimination between healthy subjects and MCI patients (as their average absolute Shapely value was greater). Alternately, in the classification between MCI and dementia patients the most informative resulted in the patterns number 10 (predominantly), 13, 3, 18 and 17. In the remaining task (healthy-dementia), a mix of the pattern found in the previous cases was found as the most important, and additionally pattern 1. The plots in Fig. 7, column (b), show the single Shapely values of the patterns for the classification of each sample. Negative Shapely values pushed the classification of a sample towards the ‘less severe diagnosis’ (i.e healthy in the first two tasks and MCI in the third), while

Table 6
Results of the diagnosis classification tasks.

	All samples (unbalanced)				Random sampling (balanced)			
	Healthy vs MCI	Healthy vs dementia	MCI vs dementia	3-class	Healthy vs MCI	Healthy vs dementia	MCI vs dementia	3-class
Accuracy	0.89	0.92	0.87	0.76	0.85 ± 0.03	0.91 ± 0.02	0.83 ± 0.03	0.73 ± 0.04
F1	0.93	0.91	0.79	0.74	0.85 ± 0.04	0.90 ± 0.03	0.84 ± 0.03	0.73 ± 0.03
Precision	0.95	1.00	0.85	0.74	0.89 ± 0.04	0.96 ± 0.02	0.84 ± 0.04	0.74 ± 0.04
Recall	0.90	0.84	0.74	0.74	0.81 ± 0.05	0.85 ± 0.04	0.89 ± 0.04	0.73 ± 0.04

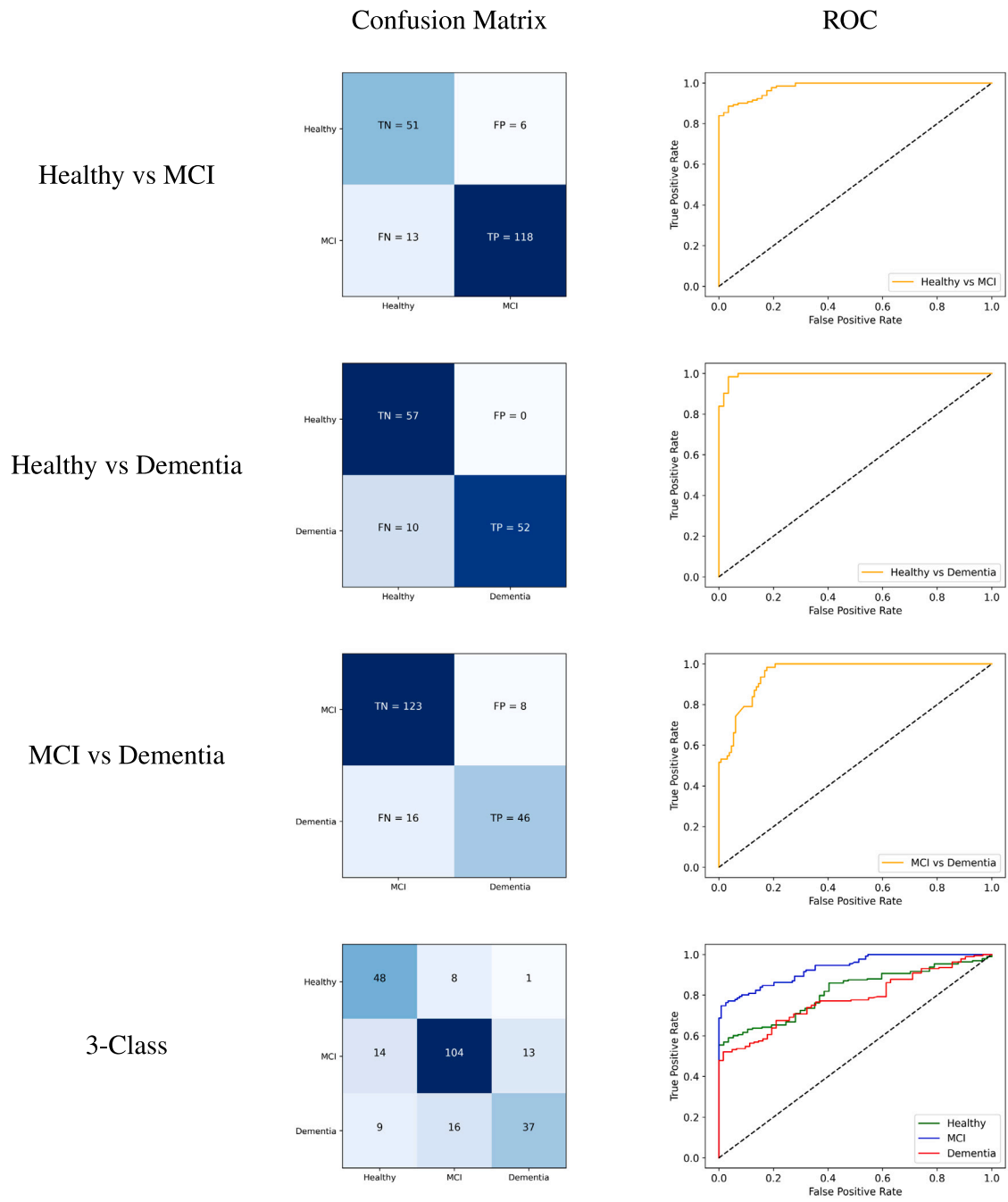


Fig. 6. Performances for the 3 binary and the multi-class classification tasks performed with all the samples. The confusion matrices are reported in the left column and the ROC curves in the right column. For the 3-class classification, the ROC curves are figured in the same plot in different colours: green for the healthy class, blue for the MCI class and red for the dementia class.

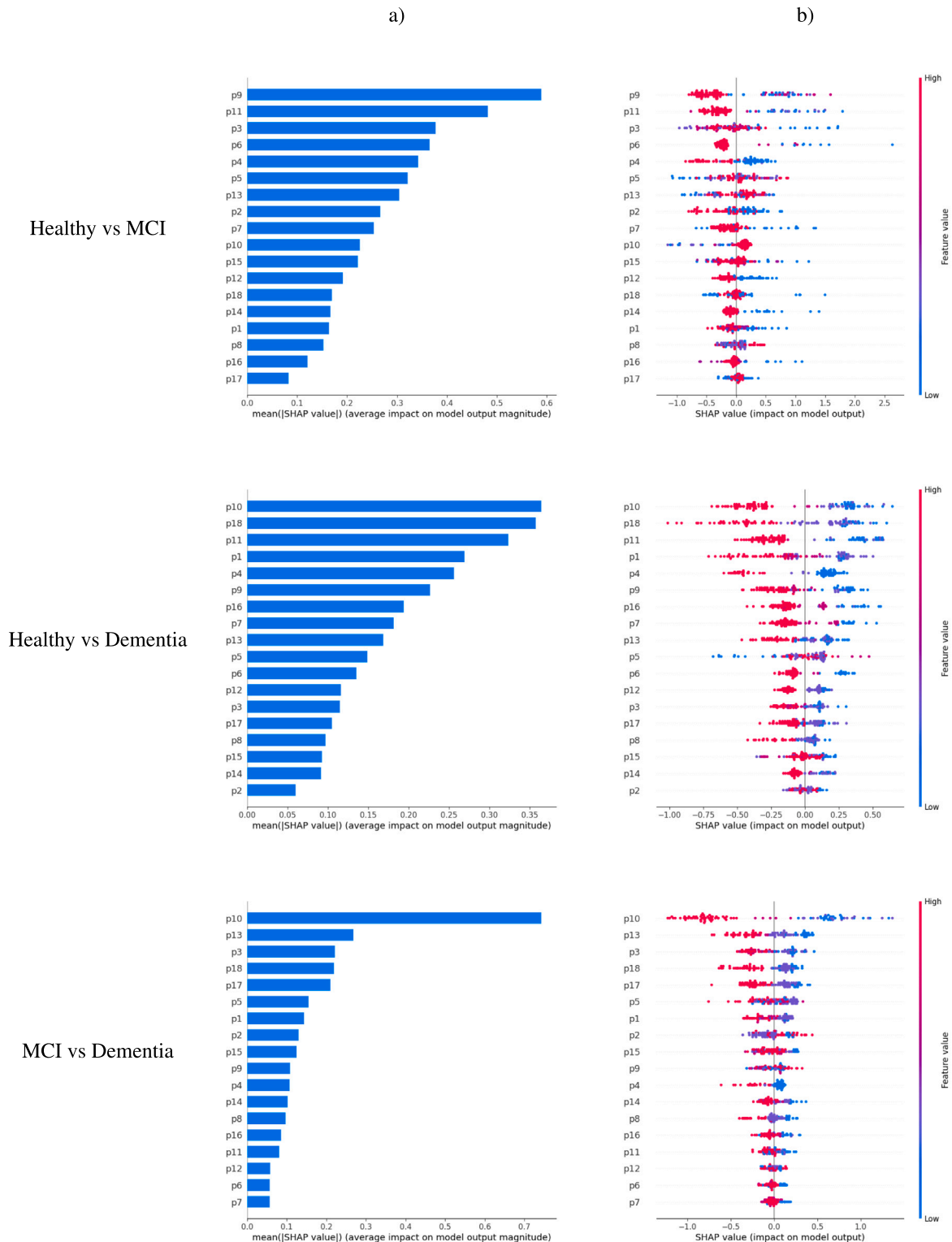


Fig. 7. Shapely values for the binary classifications.

positive values moved the prediction towards the other diagnosis. The blue-red colour-map of the dots indicated the quality of the pattern representation, as a continuous interpolation from the minimum score of 0 (in blue), corresponding to the label *omitted*, to the maximum score

of 3 (in red), which corresponded to the label *correct*. For example, the difference in the quality representation of pattern 9 in the classification healthy-MCI resulted very sharp between the two classes (with higher scores for the healthy one). The same behaviour is visible for the more

significant patterns in the other classifications (e.g. numbers 10, 18 and 11). Other patterns (such as the numbers 5 and 7) instead did not show a clear difference in the scoring between the classes in all the classifications. It suggested that individuals of both classes were likely to make similar mistakes in the drawing of these elements.

6. Discussion

6.1. System performance

In this work, a method for the analysis of the ROCF copy test has been proposed as a DSS to support clinicians in the evaluation of the ROCF copy test and in the prediction of the patient's diagnosis. The system performs a semi-automatic analysis of a scanned ROCF image, returning a qualitative score for each of the 18 patterns used in the standard scoring method (4 levels) and the most probable diagnosis for the subject between healthy, MCI or dementia.

The simple patterns were evaluated using predefined expert-based rules and low to medium Accuracy scores were obtained in their labelling. The main reason behind the choice of CV methods for the evaluation of the simple patterns consisted in the small amount of samples available for each label category. In some cases indeed, the rules were adjusted on limited examples of differently labelled patterns. For example, pattern 9 resulted the hardest to be evaluated (with an Accuracy of 61.9%), since only a small portion was labelled as *distorted* and a very diverse group of representations fell into that category (Table 1). The variety of the patterns replicas represented the main issue to the formulation of general rules for the exact labelling, although some positive Accuracy scores were obtained in the evaluation of the patterns 6, 7 and 16.

For the evaluation of the complex patterns, the articulated shapes represented a further barrier to the use of explicit labelling rules. Therefore, deep learning models were applied in their evaluation. The Accuracy scores (in Table 5) were similar to those achieved with the simple patterns. Overall, the main limitations in the single pattern evaluation are the unbalance of the labels available and the small sample size. Thus, as further development, we envisage the collection of a larger amount of retrospective data to increase the accuracy in the evaluation of the 18 patterns and make the system more robust. Also the complex pattern evaluation might improve increasing the sample size.

Very good to excellent classification metrics were obtained in the discrimination between all the healthy and MCI individuals. Despite of the class unbalance, high Precision and Recall were achieved (95% and 90% respectively). The ROC analysis revealed that the threshold of the classifier could be adjusted to calibrate the true positive rate to the maximum of 100%, tolerating a false positive rate of 30%. Even higher performances were achieved in the classification between healthy and dementia individuals, with the 100% of Precision score and a Recall equal to 84%. Such behaviour was quite expected since an increased level of impairment of cognitive functions is present in patients with dementia. Slightly lower performance was obtained in the MCI vs dementia classification with all the samples. However, the initial Recall of 74% could be improved to 100% by selecting a different threshold with a Precision equal to 80%. The fourth classification, with the 3 groups, presented the lower performances (still good with 76% for the accuracy and 74% for the other metrics) as the task complexity increased. The same tasks were repeated using random sampled balanced datasets and mediating the classification metrics of 50 iterations. The outcomes converged to high performances for the 3 binary tasks, yet lower than the case in which all the samples were included. A higher Precision was obtained in the tasks healthy vs MCI and healthy vs dementia, while Recall resulted higher in the MCI vs dementia classification. Similar performances were kept in the 3-class task. Our results were analogous in terms of accuracy to those reached by a recent work by Youn et al. (2021). In the study, they used a total number of 980

ROCF copy images to train convolutional neural networks as screening tool to classify between healthy, MCI and severely cognitive impaired individuals. Their accuracy performance were comparable to ours in the classification between healthy and severely impaired patients (90%) and in the 3-group classification task (71%). However, our system was able to achieve excellent accuracy (85%) in the classification between healthy and mild cognitive impaired individuals, which is the most relevant classification problem addressing early screening. Such classification problem was not reported in the previous study by the previous study (Youn et al., 2021). Furthermore, the proposed DSS was also capable of evaluating individually the 18 test patterns thus enriching the information provided to clinicians.

6.2. System impact

The DSS presented in this work is a prototype that can be used in the study and development of medical expert systems applied in the diagnosis of cognitive decline. It was developed as a part of a retrospective study in which past data samples of the ROCF copy test were collected to build the system's knowledge base. It represents an advantage with respect to the tablet-based solutions as it allows to expand the dataset by simply scan existing ROCF copy images, instead of asking new participants to execute the test. The amount of past ROCF examples constitutes the information level of the expert system, therefore it is likely to increase as the number of data samples grows. A higher level of information enhance the accuracy of the tool in the pattern evaluation and the confidence in the formulated diagnosis. Currently, samples data consist of experts evaluated ROCF copies, combined with the patient's true diagnosis. However, larger amount of samples might also allow the exploitation of unsupervised techniques for the image analysis, removing any residual subjective component in the evaluation of the test.

The results of the SHAP analysis revealed that some ROCF patterns were the most important in discriminating between healthy people and MCI patients or between MCI and dementia patients. As an example it could be seen from Fig. 7 how Pattern 11, which is a small vertical line that could be easily neglected (see Fig. 3(b)), is particularly important to distinguish between healthy and MCI patients. This finding, which could have direct clinical utility and applicability to support diagnosis, could be well regarded as a first step in the discovery of a new clinical sign. Indeed, the outcome of the neuropsychological assessment is based not only on the patient's performance score on the test, but also on whether certain signs of cognitive dysfunction appear. In general, a pathological sign is a complex but discrete, fairly invariable and recognisable pattern of responses or behaviours given by the patient during the performance of the test (Abbate et al., 2019). The experienced clinician may recognise some typical pathological signs in many neuropsychological tests. The clinical value of a sign is sometimes more relevant than the value of a low score returned on the test, because the sign directly suggests a particular cognitive dysfunction and corresponding brain damage. For example, in the case of the ROCF, the closing-in phenomenon, where the user draws the image \hat{F} close to or on the image F or on the margin of the drawing area, is a sign highly suggestive of environmental dependence syndrome, which corresponds to frontal-parietal brain damage. However, it is difficult to define objectively such underlying clinical signs, and inter-rater agreement on their assessment is often low. In this context, the application of an automated decision support system software to the ROCF, by providing further and objective information, could have the additional benefit of helping to more objectively define a neuropsychological sign at the test, contributing to a more reliable assessment. Furthermore, the proposed DSS tried to estimate also the same labels (omitted, distorted, misplaced, correct) used by clinicians in the standard scoring to evaluate the 18 patterns. This allowed the DSS to track and provide the expert with information about the type of the patient's mistakes, which is not included in the overall numerical score. The expert clinician then

could use this information to associate the patient's performance with other cognitive dysfunctions (Trojano & Gainotti, 2016). For example, a performance characterised mainly by omissions might be related primarily to deficits in visual attention and/or prefrontal executive monitoring. Instead when dislocation errors prevail, it could be associated primarily to a deficit in visuospatial skills, and in a minority of cases also to a particular cognitive disorder known as simultanagnosia. Finally, distortion errors made on small elements of the figure (figure content) might be related to deficits in grapho-motor skills and/or visual perception; instead distortion error made on frame elements of the figures might be associated to deficits in prefrontal executive planning. Unfortunately, in the standard correction system the different types of mistakes even observed and used concurrently to compute the overall score, are not exploited individually and not saved. So, in the actual clinical practice, the analysis of the qualitative aspects of the execution and consequently a possible analysis of the supposed different cognitive abilities involved is not directly available. The proposed decision support system is designed to support the clinician in retrieving, cataloguing, and leveraging also this valuable information regarding the type of error made by the patient. In the future an extensive use of our DSS in clinical practice is envisaged to evaluate its full potentialities and usefulness.

7. Conclusions

This work presented a decision support system for the analysis of the ROCF copy test, able to qualitatively evaluate the 18 patterns of the figure and to formulate the subject's most probable diagnosis with high accuracy. The system was tested using retrospective data and it achieved a very high accuracy performance of about 85%, 91%, and 83% in discriminating healthy subjects from MCI, healthy subjects from dementia and MCI from dementia respectively. Those results are even more relevant considering that the ROCF copy test represents just partial information in the clinical process of the cognitive assessment. The DSS integrates also explainable AI methods to better interpret the results of the diagnosis prediction. In particular, the DSS can identify which are the main patterns responsible for the prediction. Focusing the clinician attention on such a specific information might enrich his evaluation and can be a first step towards the discovery of specific signs possibly associated to specific cognitive dysfunctions. To conclude, the proposed DSS can enrich the standard evaluation of the ROCF copy test. Therefore, an extensive use of the system in clinical practice might pave the way towards a more robust use of the ROCF test in the process of cognitive evaluation.

CRedit authorship contribution statement

Davide Di Febbo: Methodology, Data curation, Writing – original draft, Writing – review & editing. **Simona Ferrante:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Marco Baratta:** Methodology, Data curation, Writing – review & editing. **Matteo Luperto:** Conceptualization, Writing – review & editing. **Carlo Abbate:** Resources, Writing – review & editing. **Pietro Davide Trimarchi:** Resources, Writing – review & editing. **Fabrizio Giunco:** Resources, Writing – review & editing. **Matteo Matteucci:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the project ESSENCE funded by the European Union under the call SC1-PHE-CORONAVIRUS-2020-2B (Horizon 2020, GA ID: 101016112).

References

- Abbate, C., Trimarchi, P. D., Inglese, S., Tomasini, E., Bagarolo, R., Giunco, F., et al. (2019). Signs and symptoms method in neuropsychology: A preliminary investigation of a standardized clinical interview for assessment of cognitive decline in dementia. *Applied Neuropsychology: Adult*, 28, 1–15. <http://dx.doi.org/10.1080/23279095.2019.1630626>.
- Abuain, W., Abdullah, S., Bataineh, B., Abu-Ain, T., & Omar, K. (2013). Skeletonization algorithm for binary images. *Procedia Technology*, 11, 704–709. <http://dx.doi.org/10.1016/j.protcy.2013.12.248>.
- Bertolani, L., Renzi, E., & Faglioni, P. (1993). Test di memoria non verbale di impiego diagnostico in clinica: Taratura su soggetti normali. *Archivio di Psicologia, Neurologia e Psichiatria*, 54, 477–486.
- Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, URL: <https://opencv.org/>.
- Broderick, M., Van Gemmert, A., Shill, H., & Stelmach, G. (2009). Hypometria and bradykinesia during drawing movements in individuals with Parkinson's disease. *Experimental Brain Research*, 197, 223–233. <http://dx.doi.org/10.1007/s00221-009-1925-z>.
- Canham, R., Smith, S., & Tyrrell, A. (2000). Automated scoring of a neuropsychological test: the rey osterrieth complex figure. 2. In *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future* (pp. 406–413 vol.2). <http://dx.doi.org/10.1109/EURMIC.2000.874519>.
- Canham, R. O., Smith, S. L., & Tyrrell, A. (2005). Location of structural sections from within a highly distorted complex line drawing. In *IEE proceedings: vision, image and signal processing* (pp. 741–749). <http://dx.doi.org/10.1049/ip-vis:20045166>.
- Crevier, D., & Lepage, R. (1997). Knowledge-based image understanding systems: A survey. *Computer Vision and Image Understanding*, 67, 161–185. <http://dx.doi.org/10.1006/cviu.1996.0520>.
- Donders, J. (2019). The incremental value of neuropsychological assessment: A critical review. *The Clinical Neuropsychologist*, 34, 56–87. <http://dx.doi.org/10.1080/13854046.2019.1575471>.
- Dorogush, A., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *NeurIPS*, 6639–6649. <http://dx.doi.org/10.48550/arXiv.1810.11363>.
- Douglas, D., & Peucker, T. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10, 112–122. <http://dx.doi.org/10.3138/FM57-6770-U75U-7727>.
- Elderkin-Thompson, V., Boone, K., Kumar, A., & Mintz, J. (2004). Validity of the boston qualitative scoring system for the rey-osterrieth complex figure among depressed elderly patients. *Journal of Clinical and Experimental Neuropsychology*, 26, 598–607. <http://dx.doi.org/10.1080/13803390409609784>.
- Fairhurst, M., & Smith, S. (1991). Application of image analysis to neurological screening through figure-copying tasks. *International Journal of Bio-Medical Computing*, 28, 269–287. [http://dx.doi.org/10.1016/0020-7101\(91\)90081-O](http://dx.doi.org/10.1016/0020-7101(91)90081-O).
- Gerth, S., Klassert, A., Dolk, T., Brenner-Fliesser, M., Fischer, M., Nottbusch, G., et al. (2016). Is handwriting performance affected by the writing surface? Comparing preschoolers', second graders', and adults' writing performance on a tablet vs. Paper. *Frontiers in Psychology*, 7, <http://dx.doi.org/10.3389/fpsyg.2016.01308>.
- Gillies, S., et al. (2007). Shapely: manipulation and analysis of geometric objects. URL: <https://github.com/Toblerity/Shapely>.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 630–645). Cham: Springer International Publishing.
- Jung, A. B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., et al. (2020). imgaug. Online; <https://github.com/aleju/imgaug>. (Accessed 01-Feb-2020).
- Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations*. <http://dx.doi.org/10.48550/arXiv.1412.6980>.
- Levkowitz, H., & Herman, G. (1993). GLHS: A generalized lightness, hue, and saturation color model. *CVGIP: Graphical Models and Image Processing*, 55, 271–285. <http://dx.doi.org/10.1006/cgip.1993.1019>.
- Liberman, J., Stewart, W., Seines, O., & Gordon, B. (1994). Rater agreement for the rey-osterrieth complex figure test. *Journal of Clinical Psychology*, 50, 615–624. [http://dx.doi.org/10.1002/1097-4679\(199407\)50:4<615::AID-JCLP2270500419>3.0.CO;2-R](http://dx.doi.org/10.1002/1097-4679(199407)50:4<615::AID-JCLP2270500419>3.0.CO;2-R).
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). <http://dx.doi.org/10.48550/arXiv.1705.07874>.
- Matas, J., Galambos, C., & Kittler, J. (2000). Robust detection of lines using the progressive probabilistic hough transform. *Computer Vision and Image Understanding*, 78, 119–137. <http://dx.doi.org/10.1006/cviu.1999.0831>.

- Merten, T., & Blaskewitz, N. (2008). The Rey complex figure test and recognition trial in clinical neuropsychological assessment. *Neurologie und Rehabilitation*, *14*, 195–202.
- Messmer, B., & Bunke, H. (1995). Automatic learning and recognition of graphical symbols in engineering drawings. *Graphics Recognition Methods and Applications*, *1072*, 123–134. http://dx.doi.org/10.1007/3-540-61226-2_11.
- Moetesum, M., Diaz, M., Masroo, U., Siddiqi, I., & Vessio, G. (2022). A survey of visual and procedural handwriting analysis for neuropsychological assessment. *Neural Computing and Applications*, *34*(12), 9561–9578. <http://dx.doi.org/10.1007/s00521-022-07185-6>.
- Osterrieth, P. (1994). Filetest de copie d'une figure complex: Contribution a l'etude de la perception et de la memoire [the test of copying a complex figure: A contribution to the study of perception and memory]. *Archives de Psychologie*, *30*, 286–356.
- Pereira, T., & Campos, T. (2020). Domain adaptation for person re-identification on new unlabeled data. *Computer Vision and Pattern Recognition*, 695–703. <http://dx.doi.org/10.5220/0008973606950703>.
- Petilli, M., Daini, R., Saibene, F., & Rabuffetti, M. (2021). Automated scoring for a Tablet-based Rey Figure copy task differentiates constructional, organisational, and motor abilities. *Scientific Reports*, *11*, <http://dx.doi.org/10.1038/s41598-021-94247-9>.
- Phan, T., & Yamamoto, K. (2020). Resolving class imbalance in object detection with weighted cross entropy losses. *Computer Vision and Pattern Recognition*, <http://dx.doi.org/10.48550/arXiv.2006.01413>.
- Rasmussen, L., Larsen, K., Houx, P., Skovgaard, L., Hanning, C., & Moller, J. (2001). The international study of postoperative cognitive dysfunction: the assessment of postoperative cognitive function. *Acta Anaesthesiologica Scandinavica*, *45*, 275–289. <http://dx.doi.org/10.1034/j.1399-6576.2001.045003275.x>.
- Rey, A., & Osterrieth, P. (1993). Translations of excerpts from Andre Rey's 'psychological examination of traumatic encephalopathy' and osterrieth's 'the complex figure test'. *Psychological Examination of Traumatic Encephalopathy and P. A. Osterrieth's the Complex Figure Test*. *Clinical Neuropsychologist*, *7*, 2–21. <http://dx.doi.org/10.1080/13854049308401883>.
- Rosin, P. (2001). Unimodal thresholding. *Pattern Recognition*, *34*, 2083–2096. [http://dx.doi.org/10.1016/S0031-3203\(00\)00136-9](http://dx.doi.org/10.1016/S0031-3203(00)00136-9).
- Scarpina, F., Ambiel, E., Albani, G., Pradotto, L., & Mauro, A. (2016). Utility of boston qualitative scoring system for rey-osterrieth complex figure: evidence from a Parkinson's diseases sample. *Neurological Sciences*, *37*, 1603–1611. <http://dx.doi.org/10.1007/s10072-016-2631-9>.
- Schmidtman, G., Jennings, B., & Kingdom, F. (2016). Shape recognition: Convexities, concavities and things in between. *Scientific Reports*, *5*, <http://dx.doi.org/10.1038/srep17142>.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815–823). <http://dx.doi.org/10.1109/CVPR.2015.7298682>.
- Shin, M.-S., Park, S.-Y., Park, S.-R., Seol, S.-H., & Kwon, J. (2006). Clinical and empirical applications of the rey-osterrieth complex figure test. *Nature protocols*, *1*, 892–899. <http://dx.doi.org/10.1038/nprot.2006.115>.
- Strauss, E., Sherman, E., & Spreen, O. (2006). A compendium of neuropsychological tests: Administration, norms, and commentary. *Oxford University Press*, *14*, 62–63. <http://dx.doi.org/10.1080/09084280701280502>.
- Suzuki, S., & be, K. (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, *30*, 32–46. [http://dx.doi.org/10.1016/0734-189X\(85\)90016-7](http://dx.doi.org/10.1016/0734-189X(85)90016-7).
- Szeliski, R. (2011). *Computer vision: algorithms and applications* (vol. 5). Springer-Verlag London Limited, <http://dx.doi.org/10.1007/978-1-84882-935-0>.
- Trojano, L., & Gainotti, G. (2016). Drawing disorders in Alzheimer's disease and other forms of dementia. *Journal of Alzheimer's Disease*, *53*, 31–52. <http://dx.doi.org/10.3233/JAD-160009>.
- Tupler, L., Welsh, K., Asare-aboagye, Y., & Dawson, D. (1995). Reliability of the Rey-Osterrieth Complex Figure in use with memory-impaired patients. *Journal of Clinical and Experimental Neuropsychology*, *17*, 566–579. <http://dx.doi.org/10.1080/01688639508405146>.
- Vogt, J., Kloosterman, H., Vermeent, S., van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, *34*, <http://dx.doi.org/10.1093/arclin/acz035.04>, 836–836.
- Westin, J., Ghiamati, S., Memedi, M., Nyholm, D., Johansson, A., Dougherty, M., et al. (2010). A new computer method for assessing drawing impairment in Parkinson's disease. *Journal of Neuroscience Methods*, *190*, 143–148. <http://dx.doi.org/10.1016/j.jneumeth.2010.04.027>.
- Youn, Y. C., Pyun, J.-M., Ryu, N., Baek, M., Jang, J. W., young ho, P., et al. (2021). 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 and Therapy*, *13*, <http://dx.doi.org/10.1186/s13195-021-00821-8>.
- Zadeh, L. (1965). Fuzzy sets. *Information Control*, *8*, 338–353. [http://dx.doi.org/10.1016/S0019-9958\(65\)90241-X](http://dx.doi.org/10.1016/S0019-9958(65)90241-X).
- Zucchella, C., Federico, A., Martini, A., Tinazzi, M., Bartolo, M., & Tamburin, S. (2018). Neuropsychological testing. *Practical Neurology*, *18*, 227–237. <http://dx.doi.org/10.1136/practneurol-2017-001743>.