Machine Vision and Applications (2022) 33:49
https://doi.org/10.1007/s00138-022-01297-8

SPECIAL ISSUE PAPER

Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction

Nicole D. Cilia[1,2] · Tiziana D'Alessandro[1] · Claudio De Stefano[1] · Francesco Fontanella[1]

Received: 28 March 2021 / Revised: 28 January 2022 / Accepted: 7 March 2022 / Published online: 20 April 2022

Abstract
Neurodegenerative diseases, such as Alzheimer's Disease or Parkinson's disease, are unfortunately still incurable, although there are many therapies that can slow down the progression of the disease and improve patients' lives. An essential condition, however, is the early diagnosis of these disorders to begin therapies as soon as possible: In fact, when the signs of the disease become evident, damages may be already significant and irreversible. In this context, it is generally agreed that handwriting is one of the first skills affected by the onset of cognitive disorders. For this reason, in a preliminary study, we considered a database of handwriting and drawing specimens and proposed a method for selecting the most relevant information for diagnosing neurodegenerative disorders. The basic idea was to generate, for each handwriting sample, a color image to exploit the ability of convolutional neural network to automatically extract features from raw images. In the generated images, the color of each elementary trait encodes, in the three RGB channels, the dynamic information associated with that trait. Starting from the very encouraging obtained results, the aim of this study is twofold: On the one hand, we have tried to improve the feature extraction phase, associating further dynamic information with each handwritten trait. On the other hand, we have expanded the database of handwriting samples by adding specimen derived from more complex drawing tasks. Finally, we carried out a large set of experiments for comparing the results obtained by using standard online features with those obtained with our feature extraction approach.

Keywords Alzheimer's disease prediction · Feature extraction · Convolutional neural networks · Transfer learning · Classification methods

1 Introduction

Alzheimer's disease (AD) is an irreversible and progressive neurodegenerative disease. Close monitoring and an early diagnosis of AD are essential to prevent a rapid progression of the disease. Therefore, prediction of a person's cognitive performance of participants is an important research topic in the study of Alzheimer's disease.

Usually, in the machine learning field, EEG signals or MRI images are used for detecting cognitive impairment. However, in recent years, it has become increasingly evident that other signals can be useful predictors of cognitive status. For example, it is generally agreed that early signs of Alzheimer's disease produce alterations in handwriting [1–4], which is based on an ensemble of kinesthetic and motor-perceptive skills. Several steps forward have been made in this field, starting from the definition of some handwriting protocols, which specify the writing or drawing tests to be performed. However, it should be noted that there is no general agreement on the number and type of task to be adopted and that there are few standard databases available collecting this type of data, which generally refer to a very limited number of participants.

This aspect represents a further difficulty in the context of machine learning techniques, which typically require a huge amount of information. Furthermore, there is no general agreement on the types of features on which researchers should concentrate [5,6]. Indeed, the problem of detecting effective features that allow the system to distinguish the

This work was partially supported by MIUR (Minister for Education, University and Research, Law 232/216, Department of Excellence).

✉ Claudio De Stefano
destefano@unicas.it

[1] Department of Electrical and Information Engineering, University of Cassino and Southern Lazio, Cassino, FR, Italy

[2] Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands

natural handwriting alterations due to age from those caused by neurodegenerative disorders is still an open issue, which strongly influences the obtainable results and the practical applicability of early diagnosis support techniques. In the large majority of cases, the sets of features are typically selected by hand, generally considering the dynamics of the handwriting process in order to detect motor disorders closely related to AD. Features directly derived from handwriting generation models have also been used for AD diagnosis.

It should also be remarked that, to the best of our knowledge, the feature sets considered in the studies published in the literature do not include shape information of handwritten traces, which may be very helpful in many cases. (The only exception is the evaluation of *micrographia*[1] that is normally used in Parkinson's disease detection.) The presence, for instance, of irregular or fragmented handwriting, often associated with changes in the thickness of the strokes, can indicate difficulties in fine motor control and possibly the onset of neurodegenerative disorders. Furthermore, these studies do not investigate the correlation between changes in the shape of handwritten strokes and those in the dynamics of the writing process that produced those strokes.

Moving on from these considerations, in a preliminary study [7] we tried to verify if the combined use of both shape features and dynamic features allows a decision support system to improve performance for AD diagnosis. Starting from a database of online handwriting samples, in which the sequence of points acquired at a given frequency is recorded in terms of x-y coordinates and pressure value of each point, we generated a synthetic offline color image (hereafter denoted as RGB) for each of them.The color of each elementary trait encodes the dynamic information associated with that trait in the three RGB channels. According to our procedure, a synthetic offline color image is generated by drawing an elementary trait for each pair of consecutive points in the corresponding online handwriting sample: The end points of each elementary trait have the same x-y coordinates of the corresponding pair of points in the online handwriting sample, while the color of the trait is obtained by using as RGB values, velocity, jerk and pressure relative to these points. Thus, in the obtained images, the shape of each elementary trait is correlated with the information about the dynamics with which that trait has been produced and the pressure exerted when it was written.

Moreover, we exploited the capability of deep neural networks (DNNs) to automatically extract features from offline color images. More specifically, we used convolutional neural networks (CNNs), because they are particularly suitable for processing raw images [8]. In this way, the presence of significant differences among patients and healthy controls, regarding the shape of the traits or the way in which these traits were produced (speed, jerk, pressure), can be automatically derived through the learning process of CNN, which produces, for each handwriting sample, a feature vector representation. The results obtained employing two tasks of the whole protocol presented in [9] and different CNNs pre-trained on the public database ImageNet [10] and then fine-tuned using the synthetic images generated according to the above procedure, have been very promising. For comparison purposes, in the above study, we also considered the results obtained by extracting standard dynamic features from the same data.

Although the results obtained were very encouraging, reporting an increase in performance compared to those obtained by using standard dynamic features, the analysis of the experimental data also showed that in some cases the features obtained with the CNNs approach were not able to distinguish healthy people (HC) from people with AD (PT), especially in the initial phases of the disease: This is probably due to the simplicity of the considered drawing tasks, which do not allow the system to adequately capture the alterations in the writing performance.

Moving from these considerations, we extended the set of experiments presented in [7], by selecting other graphic tasks belonging to our protocol with higher level of difficulty. In particular, we considered tasks requiring a higher level of fine motor control, as well as tasks that involve a higher cognitive load and a greater complexity in spatial organization. Also in this study, we have chosen to only consider graphic tasks that require participants to produce handwritten graphic forms that are not as familiar to them as characters and words in their native language. Our rationale is that if people suffering from neurodegenerative disorders write regularly this could make less evident alterations in their handwriting, making it more similar to that of healthy people who do not write regularly. In other words, we have selected writing tasks that the participants are not familiar with, and therefore not very automated from the neuromotor control perspective. In this way, the differences between the writing characteristics of healthy participants and those affected by neurodegenerative disorders should emerge more clearly [5,9].

Furthermore, in order to verify the relevance of the dynamic information associated with each handwritten trait, we generalized the procedure for generating the offline synthetic images, adding a further channel, to the three previously considered ones. According to this new procedure, a multi-channel TIFF image (hereafter denoted as MC) is generated for each handwriting sample, where the values of first three channels encode the dynamic information used in our previous study, namely velocity, jerk and pressure, while the value of the fourth channel encodes the acceleration. Finally,

[1] A secondary motor symptom experienced by some people with Parkinson's disease, resulting in an abnormal small or cramped handwriting

we also exploited in this case the ability of CNN to automatically extract features on TIFF images.

In summary, for each task we generated three different datasets, namely the one based on standard dynamic features [11–14], the one based on the features provided by CNN applied to synthetic color images [7] and the one provided by CNN applied to the MC images. Moreover, for each task and for each dataset, the performance was evaluated using the same classification schemes, namely random forest, K-nearest neighbor, multilayer perceptron and support vector machines. This choice allowed us to easily compare the experimental results relative to the different feature vector representations and, therefore, the role played by the shape and by the combined use of both shape and dynamic information. Finally, a further comparison was made by considering the classification results directly provided by the fully connected layer of CNN. The main contributions of the paper can be summarized as follows:

- Assessing the contribution, in terms of performance of an AD diagnosis system, of dynamic information encoded as color values in the RGB channels of specifically generated images. We also compared the results achieved by using these images with those achieved by using multi-channel images, generated by encoding a further dynamic feature in the fourth channel;
- Assessing CNNs ability as an automatic feature extractor tool comparing their performance with that achieved with widely used handcrafted features;
- Evaluating the effectiveness of the method presented in [7] on more tasks; these new tasks allowed us to test participant long-term motor planning ability;
- Comparing two different classification approaches. In the first one, we classified the participants considering handcrafted features and applying well-known machine learning algorithms. In the second we classified the participants considering the features automatically extracted by CNNs, using both RGB and multi-channel images. For comparison purposes, we also considered the classification results provided by the fully connected layers of CNNs.

The remainder of the paper is organized as follows. Section 2 discusses the related work, whereas Sect. 3 presents the architecture of our system. In particular, this section details the data acquisition process (Sect. 3.1), the handcrafted feature extraction and the deep feature extraction (Sects. 3.2 and 3.3, respectively), and the classification step (Sect. 3.4). The experimental results are shown in Sect. 4, while discussion and future works are eventually left to Sect. 5.

2 Related work

As anticipated in the Introduction, to the best of our knowledge, this is the first attempt to exploit information about shape changes in handwritten traces, as well as the correlation between changes in the shape and changes in the dynamics of the handwriting process, by means of a deep transfer learning approach. In other words, the studies published in the literature do not address the specific problem of verifying if shape features extracted from handwriting through deep learning techniques are more suitable for characterizing cognitive impairment then handcrafted features. Indeed, deep learning techniques in the assessment of the AD are usually employed starting from MRI images as input signals [15–17] or from EEG signals [18,19].

Although a lot of handwriting studies are still conducted in the field of psychology, where standard statistical techniques are used, increasing attention from the machine learning field in the analysis of data from handwriting is visible, especially for the PD diagnosis [20–23].

Regarding the AD diagnosis from handwriting, in [24] the authors performed semi- or unsupervised learning to uncover homogeneous clusters of participants, and analyzed how much information these clusters carry on the cognitive profiles. Furthermore, they introduced a new temporal representation learning from handwriting trajectories that uncovered a rich set of features simultaneously, like full velocity profile, size and slant, fluidity and shakiness, revealing how these features jointly characterize the cognitive profiles.

In [6], the authors performed kinematic measures of the handwriting process to assess the importance of features for differentiating groups and for assessing the characteristics of the handwriting process across five different and functional tasks of copying. The results showed that the kinematic measures together with the MMSE score were able to distinguish effectively between the patients belonging to the different groups considered. As for the feature analysis, pressure and time-in-air obtained the best performances. Also in [25] the authors analyzed the stability of the offline handwritten word "mamma" (mum in Italian) to distinguish AD patients from healthy controls. The stability of the word was computed by splitting its image into elementary parts and measuring the similarity of the adjacent parts. As a classification algorithm the authors adopted the Yoshimura approach, based on the comparison of the stability features among the sample to be recognized and those of the training samples.

In [26], the authors presented a novel approach in which handwritten signatures were analyzed for the early diagnosis of AD. Patients' signatures were represented by using

the Plamondon's Sigma-Normal model, by means of twelve features.

Finally, the goal of the work reported in [27] was to distinguish participants belonging to three different groups (AD, MCI and control group) by comparing their handwriting kinematics. The authors used discriminant analysis as a classification algorithm and adopted a protocol consisting of seven tasks, which included copying and drawing tasks. In these experiments, the authors, for the same task, investigated which were the most discriminating features and the best distinguished groups. They found that: (i) discriminating features depended on the type of group to be discriminated; (ii) some tasks, e.g., the clock drawing test, allowed some groups, e.g., AD vs. MCI, to be well discriminated (100% of specificity and sensitivity).

Starting from the results of this research, which highlights the lack of a unique handwriting protocol and the limited number of participants involved, we started a research activity in collaboration with relevant hospitals to define an experimental protocol capable of capturing the most relevant aspects of the onset of neurodegenerative diseases, involving a large number of participants on the basis of rigorous recruitment criteria. A first result of this activity is the definition of an experimental handwriting protocol according to which we collected the handwriting samples of one hundred eighty participants, including both AD patients and healthy controls. In particular, in [9] the experimental handwriting protocol consisting of 25 tasks to record the dynamics of handwriting, when different motor skills are employed, is presented. Using a subset of the above tasks, in a first set of experiments, whose results are reported in [11–13], we tested 130 participants (both patients and healthy controls) employing two classification algorithms. The tasks considered in the above-mentioned studies were selected in order to evaluate the alterations on kinematic and pressure properties in repeating complex graphic gestures, which have a semantic meaning, such as letters and words of different lengths and with different spatial organizations. To improve the performance of these systems in [14], a genetic algorithm has been used which selects the best subset of tasks among those belonging to the above protocol.

3 The architecture of the system

The architecture of the whole system is reported in Fig. 1: The figure shows that the acquired data are processed for generating both the set of standard dynamic features (denoted as *handcrafted features*), and the two groups of RGB and MC images: Each group of images is then forwarded to the corresponding CNN to extract a new set of features; thus, at the end of this step, two further sets of features are generated, namely those obtained by the RGB images (denoted as *RGB-deep features*), and those obtained by the MC images

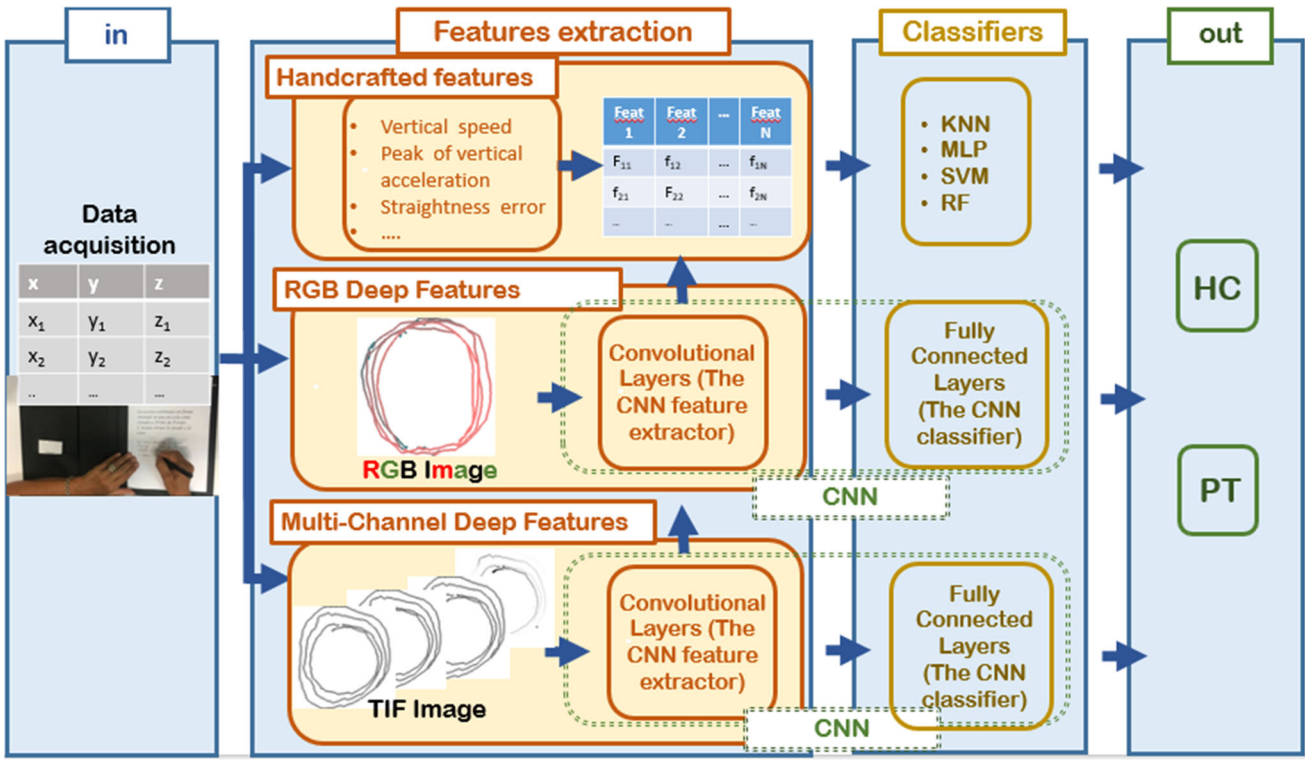


Fig. 1 Chain of the whole system

(denoted as *MC-deep features*). These three sets of features are individually used by each of the considered classification schemes to implement the classification stage. Note that this stage also includes the classification results directly provided by the *fully connected layer* of each CNN.

A detailed description of each stage of the proposed system will be provided in the following sections.

3.1 Data acquisition

The first step of our system, as shown in Fig. 1, is devoted to the acquisition and recording of participants' handwriting, produced by participants according to a given protocol, in terms of x-y-z coordinates of each point, acquired at a constant sampling rate equal to 200Hz. The first two coordinates are the point position in the two-dimensional space representing the surface on which the writing is produced, while the third is a measure of the pressure exerted by the person at that point. This measure assumes a positive value when the pen is resting on the sheet and a null value when the pen is detached, up to a maximum distance of 3 cm from the sheet, beyond which the system is not able to receive information. The application developed using the C programming language drives the graphic tablet and acquires the coordinates of the movements of the participants while they are writing on a A4 sheet fixed to the tablet surface. Furthermore, since writing skills can be influenced by age, education level and type of work, this information is also stored.

For the recruitment of participants involved in the study, with the support of the geriatric ward, Alzheimer unit, of the "Federico II" hospital in Naples, we used standard clinical tests, such as the Mini-Mental State Examination (MMSE), the Frontal Assessment Battery (FAB), the Montreal Cognitive Assessment (MoCA) to distribute the participants into two groups: healthy people (control group) and patients. All participants were right-handed and comfortably positioned approximately 70 cm from the sheet of paper.

Our database includes 180 participants (90 patients and 90 healthy control), each performing the 25 tasks defined in our protocol. As anticipated in the Introduction, in this study we only considered the handwriting samples relative to six graphic tasks produced by all the participants.

Starting from the x-y-z acquired coordinates, three different feature extraction approaches were implemented:

(i) Typical online features are extracted and used for implementing the classification schemes described in Sect. 3.2.
(ii) The dynamic information relative to each stroke is considered for generating the images used for deep learning classification experiments, as further detailed in Sect. 3.3.1.
(iii) Similarly, in the third approach, MC images are generated and used for deep learning classification experiments (see Sect. 3.3.1).

We analyzed participants' handwriting while drawing lines to predict their cognitive status. In particular, we asked the participants to perform six tasks, as detailed in the following.

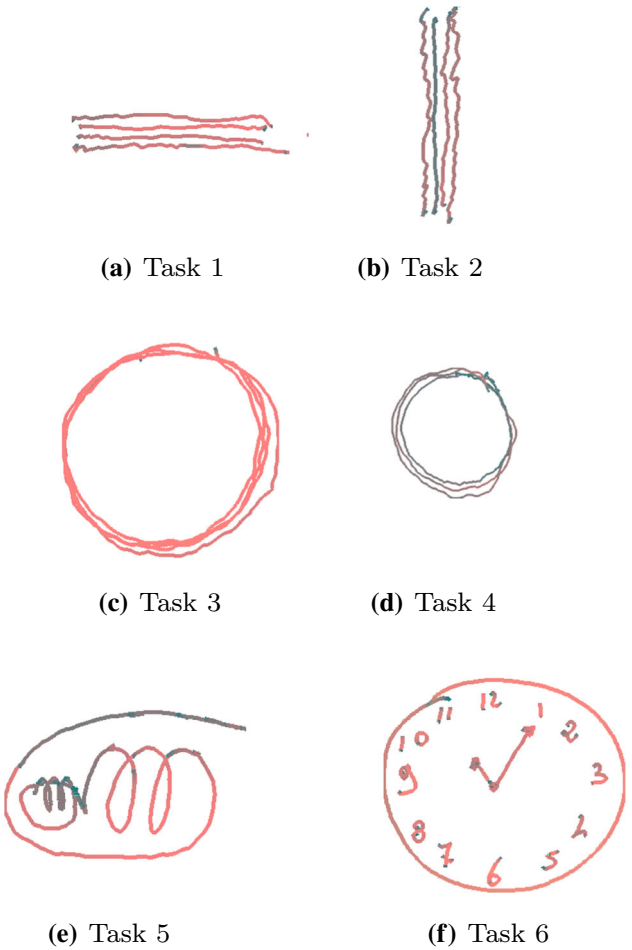
The first two tasks consisted of joining two points 5cm apart with a straight continuous horizontal (task 1) or vertical (task 2) line, continuously for four times. This kind of tasks investigates elementary motor functions [28]. Horizontal movements require movements of the arm, keeping fingers in a fixed position. Vertical movements require small finger and wrist movements. In addition, drawing a single continuous line four times requires the execution of long-term motor planning, which is a typically compromised function in individuals with cognitive impairments.

The third and fourth tasks consisted of retracing a 3cm (task 3) or 6cm (task 4) wide circle four times. These tasks highlight the continuity of the line by retracing, a circular shape of various dimensions. The continuity and distancing from the background figure to be traced are indicative of cognitive deterioration. These tasks make it possible to check the automaticity of the movements and the regularity and coordination of the sequence of movements [29].

The fifth task consisted of retracing a complex form specifically devised to test the participant's motor control skills. This task investigates the alteration of the handwritten traits independently of any letter, word or the related semantic usage. The handwriting movements needed to retrace the form require a constant motor re-modulation. The shape of the form consists of a continuous line that presents radii of different curvatures with the aim of testing both fine control and long-term motor motion planning [30,31].

Finally, the sixth task was the well-known clock drawing test: The participant is asked to draw a clock face, including the numbers, and then to draw the hands at five past eleven. The clock-drawing test (CDT) is used for the screening of cognitive impairments and dementia. It is also used to assess participants' spatial dysfunction and lack of attention. It was originally used to evaluate visuo-constructive abilities but it has been shown that abnormal clock drawing occurs in other cognitive impairments. The test requires verbal understanding, memory and spatially coded knowledge in addition to constructive skills [32]. Moreover, in [33] the authors found that CDT shows a high sensitivity for the diagnosis of mild Alzheimer's disease.

Examples of tasks are shown in Fig. 2.



(a) Task 1 **(b)** Task 2

(c) Task 3 **(d)** Task 4

(e) Task 5 **(f)** Task 6

Fig. 2 Examples of tasks performed by a participant involved in the experiments

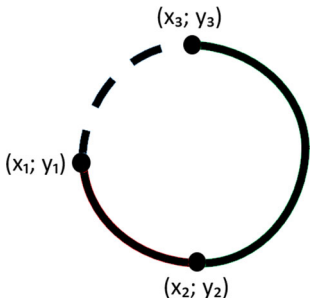


Fig. 3 Example of generated strokes

3.2 Handcrafted feature extraction

From the acquisition phase, the handwriting trajectories, in terms of **x, y** coordinates, are available. For each point acquired a third piece of information representing the pressure (**z** coordinate), is also provided. From these coordinates, we have calculated the handcrafted features used for the classification step as detailed below.

Each feature is computed referring to a stroke, defined as the sequence of traits produced between a pen down point and the following pen up or a change of direction on the y-axis (see Fig. 3). As shown in this figure, from the starting point (x_1, y_1) to the point (x_2, y_2) we detected a stroke since in (x_2, y_2) there is a change of direction along the y-axis. Similarly, from (x_2, y_2) to (x_3, y_3) we detected another stroke because in (x_3, y_3) a pen up occurs. We denoted such strokes as *on-paper*, since they are acquired by the system when the pen tip is touching the sheet. Moreover, from (x_3, y_3) to (x_1, y_1) the pen tip is lifted from the sheet but within the maximum distance that allows the system to receive information: Thus, we can detect a further stroke, traced between a pen up and the following pen down, but keeping the pen tip close to sheet. We denoted such strokes as *in-air*.

Many studies in the literature have shown that the analysis of in air traits can provide significant information for identifying neurodegenerative disorders: In-air movements, indeed, characterize the motor planning activities related to the positioning of the pen tip between two successive written traits. Moving from these considerations, in a previous study [7] we decided to extract the features from both in-air and on-paper strokes. On the other hand, in this study, we have decided to eliminate the features calculated from the in-air traits so that the comparison between the system with handcrafted features and the CNN system is more faithful. As we will see in Sect. 3.3, the CNN system uses input images that do not include in-air strokes.

For each stroke, we extracted twenty-two features, which can be grouped into two categories, namely static and dynamic, as detailed in Table 1. A feature vector is obtained for each task performed by each person by averaging the values for all the strokes relative to that task.

3.3 Deep feature extraction

In this step, the online handwriting samples, each represented as sequence of points acquired at a given frequency, in terms of x-y coordinates and pressure value of each point, are processed for generating two groups of images, namely RGB and MC images. The images of each group are forwarded to the corresponding CNN, which operate as feature extractor. To this aim, CNN are pre-trained on the public database ImageNet [10] and then fine-tuned using such images. The result, is the production of two feature sets, each representing the whole database of handwriting samples, but including the features extracted from the corresponding group of images. This process is detailed in the following sections.

3.3.1 Image generation

Starting from the same raw data used for handcrafted feature extraction, stored in terms of x-y coordinates and pressure

Table 1 Feature list

#	Name	Description	Type
1	Duration	Time interval between the first and the last points in a stroke	D
2	Start vertical position	Vertical start position relative to the lower edge of the active digitizer area	S
3	Vertical size	Difference between the highest and lowest y coordinates of the stroke	S
4	Peak vertical velocity	Maximum value of vertical velocity among the points of the stroke	D
5	Peak vertical acceleration	Maximum value of vertical acceleration among the points of the stroke	D
6	Start horizontal position	Horizontal start position relative to the lower edge of the active tablet area	S
7	Horizontal size	Difference between the highest (rightmost) and lowest (leftmost) x coordinates of the stroke	S
8	Straightness error	It is calculated estimating the length of the straight line, fitting the straight line, estimating the (perpendicular) distances of each point to the fitted line, estimating the standard deviation of the distances and dividing it by the length of the line between beginning and end	D
9	Slant	Direction from the beginning point to endpoint of the stroke, in radiant	S
10	Loop surface	Area of the loop enclosed by the previous and the present stroke	S
11	Relative initial slant	Departure of the direction during the first 80 ms to the slant of the entire stroke.	D
12	Relative time to peak vertical velocity	Ratio of the time duration at which the maximum peak velocity occurs (from the start time) to the total duration	D
13	Absolute size	Calculated from the vertical and horizontal sizes	S
14	Average absolute velocity	Average absolute velocity computed across all the samples of the stroke	D
15	Road length	length of a stroke from beginning to end, dimensionless	S
16	Absolute y jerk	The root-mean-square (RMS) value of the absolute jerk along the vertical direction, across all points of the stroke	D
17	Normalized y jerk	Dimensionless as it is normalized for stroke duration and size	D
18	Absolute jerk	The root-mean-square (RMS) value of the absolute jerk across all points of the stroke	D
19	Normalized jerk	Dimensionless as it is normalized for stroke duration and size	D
20	Number of peak acceleration points	Number of acceleration peaks both up-going and down-going in the stroke	S
21	Pen pressure	Average pen pressure computed over the points of the stroke	D
22	#strokes	Total number of strokes of the task	S

Feature types are: dynamic (D) and static (S)

of the points acquired for each online handwriting sample, we have generated two type of images to be submitted to the CNN networks.

The traits of both types of synthetic images are obtained by considering the points (x_i, y_i) as vertices of the polygonal that approximates the original curve. As regards the first type, we also used kinematic information encoded in the RGB channels. In particular, these synthetic images are obtained by considering: i) the triplet of values (z_i, v_i, j_i) assumed as RGB color components for the ith trait, delimited by the couple of points (x_i, y_i) and (x_{i+1}, y_{i+1}).

ii) the triplet of values are obtained as follows:

- z_i is the pressure value at point (x_i, y_i) and it is assumed constant along the ith trait;
- v_i is the velocity of the ith trait, computed as the ratio between the length of the ith trait and interval time of $5ms$ corresponding to the period of acquisition of the tablet;
- j_i is the jerk of the ith trait, defined as the second derivative of v_i.

The values of the triplets (z_i, v_i, j_i) have been normalized into the range [0, 255] in order to match the standard 0-255 color scale, by considering the minimum and the maximum value on the entire training set for these three quantities. An example of a trait generated from these images is reported in Fig. 4, where the color of the first trait corresponds to the

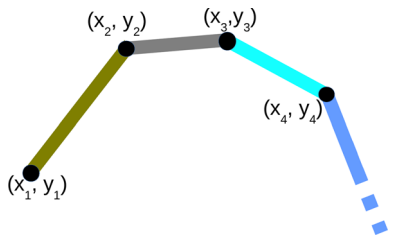


Fig. 4 Example of encoding for the trait generation in a RGB image

triplet (z=127, v=127, j=0), while that of the second one to the triplet (z=127, v=127, j=127).

As previously mentioned, in order to improve the dynamic information encoded, we also created multi-channel TIFF images, storing four representations (frames) of the same handwriting sample into a single image file. Each frame is a grayscale representation of the traits obtained according to a procedure similar to that previously described for RGB images. More specifically, considering the points (x_i, y_i) as vertices of the polygonal that approximates the original curve, pixel values in each frame are assigned according to the following criteria (see Fig. 5):

- The first frame implements the acceleration feature: The acceleration of the ith trait is defined as the derivative of v_i;
- The second frame implements the jerk feature: The jerk of the ith trait is defined as the second derivative of v_i;
- The third frame implements the velocity feature: The velocity of the ith trait is computed as the ratio between the length of the ith trait and interval time of $5ms$ corresponding to the period of acquisition of the tablet;
- The fourth frame implements the pressure feature that it is assumed constant along the ith trait.

As stated in Sect. 1 and better detailed in Sect. 3.3.2, we adopted four CNN models that accept input images that are automatically resized to 256x256 for VGG19, to 224x224 for ResNet50, to 299x299 for InceptionV3 and InceptionResNetV2, respectively. Taking into account these constraints for both type of images, the original **x, y** coordinates have been resized into the range [0, 299] for each image, in order to provide ex ante images of suitable size and minimize the loss of information related to possible zoom in/out.

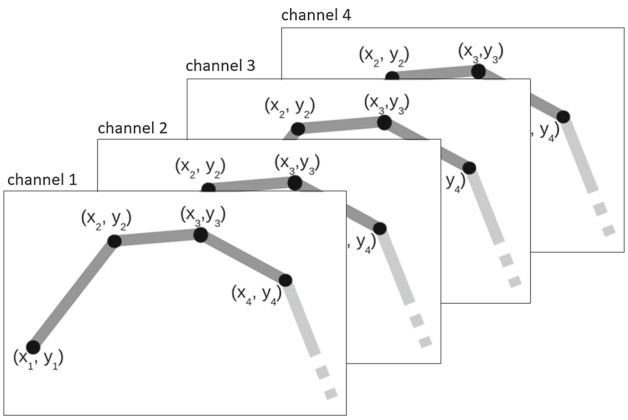


Fig. 5 Example of encoding for the trait generation in a MC image

3.3.2 Deep transfer learning for feature extraction

Deep transfer learning is gaining much popularity nowadays to solve image classification problems as it is possible to employ different CNNs trained on public datasets like ImageNet [10] reaching the highest classification performance in many application fields. In this paper, we adopted four different CNNs models: VGG19 [34], ResNet50 [35], InceptionV3 [36], InceptionResNetV2 [37]. These models differ one from another in several details, such as the introduction of new structural elements (inception, residual, dropout) or the number of layers and, consequently, the number of trainable parameters. The VGG19, in fact, is a model of tens of layers and twenty-five millions of parameters, while the deeper InceptionResNetV2 is made of hundreds of layers and consequently the number of parameters increases to sixty-two millions (see Table 2).

The adopted CNN are composed of a convolutional and a classification part. The first part is conceived for feature extraction (FE) from the images used to feed the network, whereas the second part is for the classification (C) (see Fig. 6).

The transfer learning (TL) step is followed by a retraining of the network using the fine tuning (FT) approach, which requires the retraining of both parts (FE and C) of the network. In order to apply FT, the parameters of the feature extraction are initialized with the weights obtained on ImageNet, whereas the classification part is initialized with the weights obtained during the previous TL step.

Table 2 Number of parameters and input/output size of the CNN used in the experiments

Model	Parameters	Input size	Output size (N)
VGG19	25M	256x256	512
ResNet50	32M	224x224	2048
InceptionV3	30M	299x299	2048
InceptionResNetV2	62M	299x299	1536

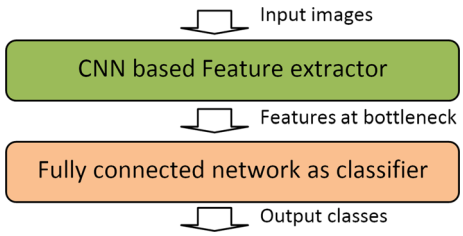


Fig. 6 The general structure of the networks used

The original classification part of the network has been replaced with a unique classifier for all the models, as described in the next section.

After the training phase, the CNN networks were used for both deep feature extraction, and classification with the final fully connected layers (the classifier section of the deep network). The output of the FE part of the network, for each input, consists in a vector of features, denoted also as bottleneck (i.e., the last activation map before the fully connected layers in the original model). This is a flattened vector of extracted features and its size depends on the architecture of the considered CNN (the number of features for each model is shown in Table 2).

Once the CNN architectures are chosen, it is necessary to assess them through an experimental phase with the aim of maximizing the mean accuracy of every model by selecting the following hyper-parameters and settings:

- Stochastic gradient descent (SGD) with learning rate 0.001, momentum 0.9: optimization method used to minimize the loss function.
- Categorical cross-entropy: is the adopted loss function.
- Batch size 16 and 20, respectively, for RGB and MC images: number of training set images considered in each iteration.
- Max epochs equal to 2, 000: One epoch is one pass on the entire training set and contains a number of iterations equal to $(trainingsetsize)/batchsize$.
- Patience 200: If the validation accuracy does not improve for a number of 200 epochs, the training is stopped.
- Accuracy as a measure of performance.

During the training phase, a fivefold cross-validation strategy was adopted, using a validation set to reduce or avoid the undesired over-fitting phenomenon. Following the standard cross-validation procedure, the data set was randomly partitioned into five equally sized folds. At each iteration, all the samples belonging to a single fold were used as test set, while the samples belonging to the other four folds were further divided into two subsets: a validation set obtained by randomly selecting 10% of these samples, and a training set consisting of the remaining ones. In practice, at each iteration, the training procedure exploits the validation set to stop the learning if the performance on such data begins to deteriorate, thus avoiding the over-fitting phenomenon. The cross-validation process is repeated five times, with each of the five folds used exactly once as a test set.

3.4 The classification step

The classification step was carried out considering nine different features sets, as already mentioned in the previous sections. Specifically, the first set consists of the twenty-two handcrafted features (see Sect. 3.2), while the remaining ones are extracted from both for RGB and MC images, through the FE part of each of the four CNNs described in Sect. 3.3.1. Summarizing, the first feature set includes standard dynamic features obtained from the online handwriting samples, four feature sets are provided by CNN applied to RGB images, and the remaining four feature sets are provided by CNN applied to MC images.

The classification was conducted following two different approaches. The first one involved the using of a standard classifier, which takes as input the features provided by the CNN feature extractor. The second approach, on the other hand, consists in using an unique classifier composed of fully connected layers for each CNN (see Fig. 6), properly modified for our purposes, i.e., classifying two classes (healthy control or patient) instead of the thousands as is the case of the ImageNet dataset.

Regarding the first approach, we decided to consider four well-known classification schemes among the most used: random forest (RF) [38], multilayer perceptron (MLP), support vector machines (SVM) [39], and the K-nearest neighbors (K-NN). Those classifiers have different characteristics and each one represents a different kind of model, more precisely RF is an ensemble of decision tree, MLP is a connectionist network, K-NN is an instance based nonparametric regression algorithm and SVM is kernel-based.

The second approach, instead, relies on the use of a classifier, where the input layer comprehends a number of neurons equal to the number of features reported in Table 2, while the output layer has two neurons, each one corresponding to the desired class (healthy controls and patients). There are two hidden layers between the input and the output one, with two thousand forty-eight neurons each and an intermediate dropout layer.

4 Experimental results

The experiments reported in this section have been executed on the following system architecture:

- CPU and RAM: Intel Core i7-7700 CPU @3.60GHz equipped with 32GB of RAM;
- Graphics card: GPU Titan Xp;
- Software: Keras 2.2.2 and TensorFlow 1.10.0.

For the four classification schemes mentioned in Sect. 3.4 (RF, MLP, K-NN, and SVM) we performed thirty runs and used the fivefold cross-validation strategy to evaluate the classification accuracy. The results reported in the following have been computed averaging the accuracy achieved on the thirty runs performed. However, since the use of the fully connected layer of CNN as classifier needs to retrain

Table 3 Training times, expressed in seconds

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
VGG19	6000	6000	6000	5100	1860	1680
ResNet50	8100	6000	4500	3900	1740	1980
InceptionV3	8100	8400	12000	7200	2160	2400
InceptionResNetV2	9600	10800	9900	9000	3240	2700

Table 4 Values of the classifier parameters used in the experiments

Classifier	Parameter	Value
RF	trees	100
K-NN	K	3
MLP	Learning rate	0.3
	Momentum	0.2
	Hidden neurons	(#features + #classes)/2
	Epochs	500
SVM	Kernel	RBF
	C	1.0
	γ	0.5

the whole network (see Fig. 6), thus involving a large amount of time, the FC results were computed averaging the accuracy achieved on the five test folders, as detailed in Sect. 3.3.2. Table 3 shows the time needed to extract the features, while the values of the parameters used in the experiments are shown in Table 4.

The feature extraction procedure detailed above was applied to the data of the six drawing tasks described in Sect. 3.1. Since the extraction of the deep features requires a training phase of the CNN, to avoid any bias with the fivefold cross-validation strategy, we selected for each sample the feature vector provided by the CNN, when that sample was in the test fold, i.e., it was non included in the folds used for training.

To assess the effectiveness of our system, we performed three sets of experiments. In the first one, we evaluated and compared the results achieved by the CNNs used on the six tasks previously mentioned. In the second one, we compared the classification performance achieved by RGB and MC images. Finally, in the third set we compared the classification performance of our approach with those achieved by using the handcrafted features described in Sect. 3.2. The results of these experiments are detailed in the following subsections.

4.1 Tasks comparison

In this set of experiments, we tried to answer to the following questions: Which is, if any, the best performing task among those considered? Among the four CNNs used with RGB and MC images, there is one that performs better than the others? Do the CNNs exhibit similarities or differences on the tasks?

With the aim of reporting a detailed description of the results achieved, Tables 5 and 6 show the results for each task provided by the five considered classifiers, when using both RGB-deep features and MC-deep features, extracted with different CNNs. From both tables, we can observe that, for each task, the performance of each classifier varies widely as the deep features used vary. Similarly, for each task, the performance obtained by using the features extracted with each CNN, varies widely changing the classifiers. Furthermore, for each classifier, the performance obtained by using the features extracted with each CNN, significantly varies changing the tasks.

To summarize the results shown in Tables 5 and 6, we plotted two vertical bar graphs for each feature type. The first graph displays the accuracy achieved by each classifier, whereas the second one that achieved by each CNN. The plots of Fig. 7 refer to RGB-deep features, while those of Fig. 8 to MC-deep features. In both figures, the left plot shows for each task the mean accuracy of each classifier, averaged on the results achieved with the features provided by the four CNNs, while the right plot shows for each task the mean accuracy of each CNN averaged on the results of the five classifiers. The aim is to see "at a glance" if there is a CNN or a classifier that performs better than the others. From the figures, we can observe that task 2 achieved the worst performance, both for RGB and MC features. This result is explainable considering that task 1 requires a greater motor load than task 2. Indeed, when joining the points vertically, it is easier to carry out the task without moving the arm with small movements of both fingers and wrist.

From the figures showing the classifier performances (left plots), we can observe that RF and SVM outperform the others in most cases. These results on one hand confirm the effectiveness of the ensemble-based strategy of RF, as well as that of the SVM kernel-based approach, specifically devised for two class problems. On the other hand, they confirm that the simple K-NN algorithm was not able to effectively estimate the probability distributions underlying our data. These results also highlight an important point: The effectiveness of RGB features extracted by the CNNs is independent of the classification algorithm used to implement the classification layer. Furthermore, RF and SVM performance is better than

Table 5 Classification results achieved using RGB features

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19												
RF	67.6	2.3	62.8	2.8	70.0	2.4	68.2	1.6	63.7	2.4	**73.2**	2.1
K-NN	64.3	2.2	56.7	2.7	61.0	2.0	61.6	2.1	64.3	2.0	72.8	1.9
SVM	61.7	2.4	55.4	2.0	67.6	1.4	66.9	1.3	60.1	2.2	70.8	1.2
MLP	63.5	3.3	56.6	2.8	58.1	3.4	58.2	3.4	60.5	2.8	72.6	1.8
FC	64.1	15.2	59.0	5.1	70.7	7.5	64.7	11.0	64.8	11.7	70.4	13.5
ResNet50												
RF	69.7	2.1	60.3	2.4	71.2	2.4	71.7	2.2	66.0	2.2	69.1	2.3
K-NN	66.8	2.1	57.0	2.4	66.3	2.6	57.9	1.4	59.6	2.0	62.5	2.0
SVM	72.0	1.7	58.9	1.9	67.3	2.2	72.6	1.6	65.1	1.4	67.9	1.8
MLP	53.8	2.1	50.6	3.4	60.3	8.7	52.9	3.5	53.0	2.7	56.2	4.2
FC	61.1	8.0	53.5	9.5	68.8	7.7	64.4	12.7	64.3	9.3	66.8	11.4
InceptionV3												
RF	68.6	2.1	54.5	3.8	70.5	2.5	71.7	1.9	66.13	2.93	67.69	3.09
K-NN	66.5	1.8	56.0	2.8	62.7	2.5	66.8	2.0	64.79	2.06	58.58	2.43
SVM	71.1	1.5	55.4	2.7	71.1	2.0	**73.1**	1.9	**69.18**	1.32	69.84	2.28
MLP	67.5	2.5	53.5	1.5	63.9	1.7	62.8	4.7	58.12	5.55	63.24	3.87
FC	64.2	9.0	57.1	4.8	68.0	6.6	65.7	11.1	58.86	11.15	62.41	7.00
Inc.ResNetV2												
RF	70.4	2.4	**65.4**	2.0	**73.0**	2.1	69.2	2.3	65.2	2.1	65.3	2.4
K-NN	68.4	1.8	62.8	2.4	65.0	2.0	64.6	2.0	64.5	2.0	64.1	2.3
SVM	71.1	2.2	61.2	2.4	71.8	1.6	67.5	1.4	68.7	1.9	58.6	1.7
MLP	**74.6**	2.3	63.3	1.5	70.4	2.1	52.6	2.1	57.3	4.9	51.6	2.0
FC	60.0	15.9	62.7	7.3	68.0	10.6	65.3	10.1	62.0	8.5	55.3	9.4

Bold values highlight the overall best performance achieved on each task.

Table 6 Classification results achieved using MC features

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
VGG19												
RF	64.0	2.6	60.9	2.5	68.6	1.5	64.3	2.4	66.4	2.4	69.0	2.2
K-NN	56.9	2.5	55.5	2.3	59.7	2.4	59.5	2.0	62.6	2.9	65.4	1.8
SVM	59.7	1.8	55.1	2.5	64.5	2.1	62.5	2.0	61.9	2.0	68.3	1.7
MLP	58.2	3.3	56.6	2.8	61.5	2.9	62.1	2.4	61.3	2.3	65.6	2.9
FC	66.2	10.7	**62.7**	10.0	69.5	10.5	64.2	11.1	64.7	7.4	64.5	12.3
ResNet50												
RF	60.1	2.6	56.3	3.0	72.0	2.4	66.8	2.7	65.0	2.2	66.0	2.7
K-NN	59.3	2.3	54.9	1.7	61.1	1.5	62.8	2.1	58.3	2.8	58.3	1.9
SVM	57.4	2.5	53.7	2.3	72.2	1.7	67.2	1.9	66.6	1.8	**69.9**	2.1
MLP	55.6	2.8	49.7	1.0	64.0	5.4	57.7	4.8	57.2	5.2	58.3	4.3
FC	61.7	12.1	59.0	8.8	**72.8**	9.0	67.7	13.7	65.6	10.3	68.6	8.3
InceptionV3												
RF	67.5	2.2	57.9	2.5	68.8	2.5	67.6	2.2	65.6	3.0	62.2	2.8
K-NN	63.5	1.9	54.1	2.2	57.1	1.7	61.4	2.0	59.6	2.5	60.6	2.3
SVM	**67.7**	2.2	61.8	2.0	68.3	2.3	68.7	2.0	66.5	2.4	64.5	2.8
MLP	66.3	2.0	56.8	3.2	62.7	4.0	63.7	4.0	55.4	3.5	61.8	2.2
FC	67.2	8.9	58.9	13.1	70.5	2.6	**71.2**	13.0	61.1	7.1	65.4	13.6

Table 6 continued

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
Inc.ResNetV2												
RF	65.2	2.3	61.4	2.7	67.6	2.1	59.8	2.6	67.6	2.2	66.4	2.2
K-NN	58.9	2.9	56.9	2.1	63.0	2.1	57.9	2.2	59.6	2.5	60.7	2.2
SVM	63.3	2.7	59.9	2.6	65.6	1.8	58.2	2.6	67.9	1.7	66.8	1.7
MLP	66.7	2.5	58.5	2.3	59.7	3.0	58.2	3.3	68.4	2.8	61.4	2.6
FC	61.6	11.8	58.7	10.4	67.8	11.3	64.8	10.9	**70.7**	3.3	55.0	17.1

Bold values highlight the overall best performance achieved on each task.

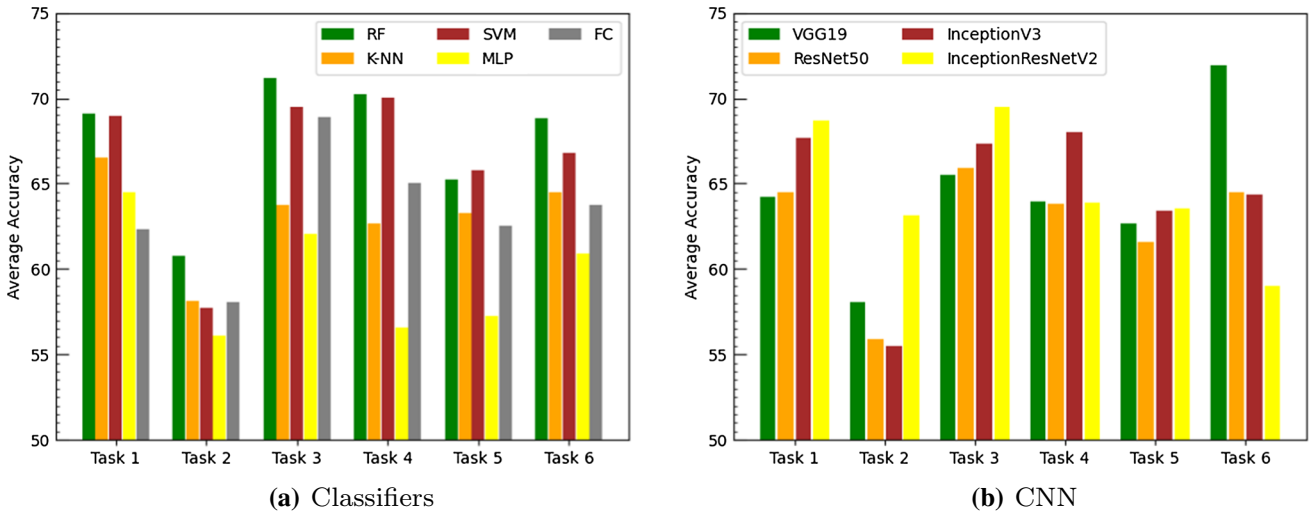


Fig. 7 Average accuracy achieved by the classifiers (**a**) and the CNNs (**b**) using RGB images

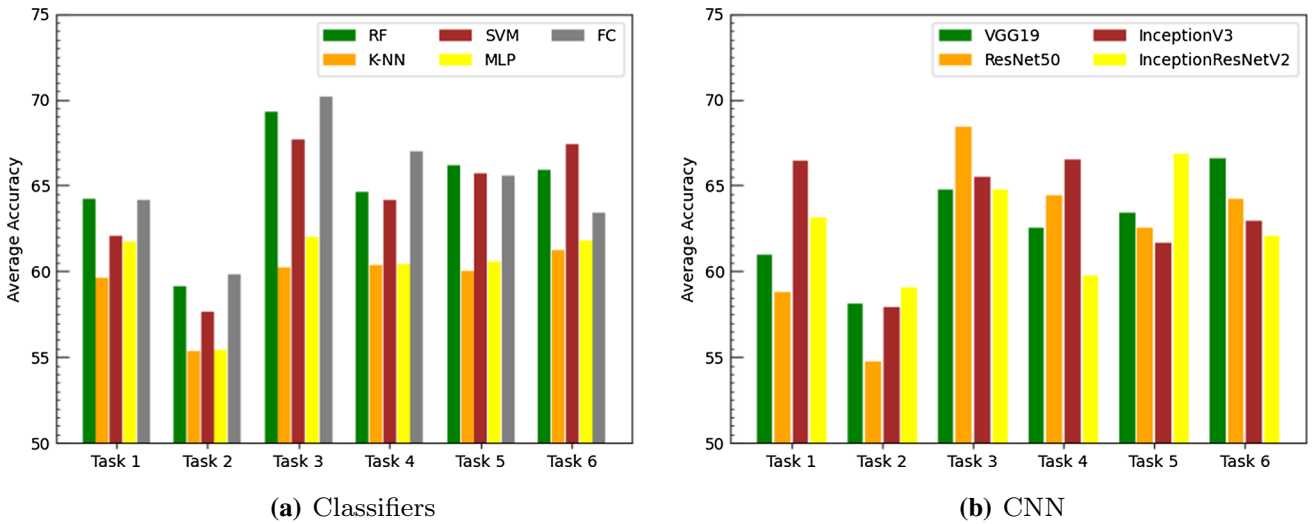


Fig. 8 Average accuracy achieved by the classifiers (**a**) and the CNNs (**b**) using MC images

that of the FC classifier trained during the process for feature extraction. The same does not occur for MC features.

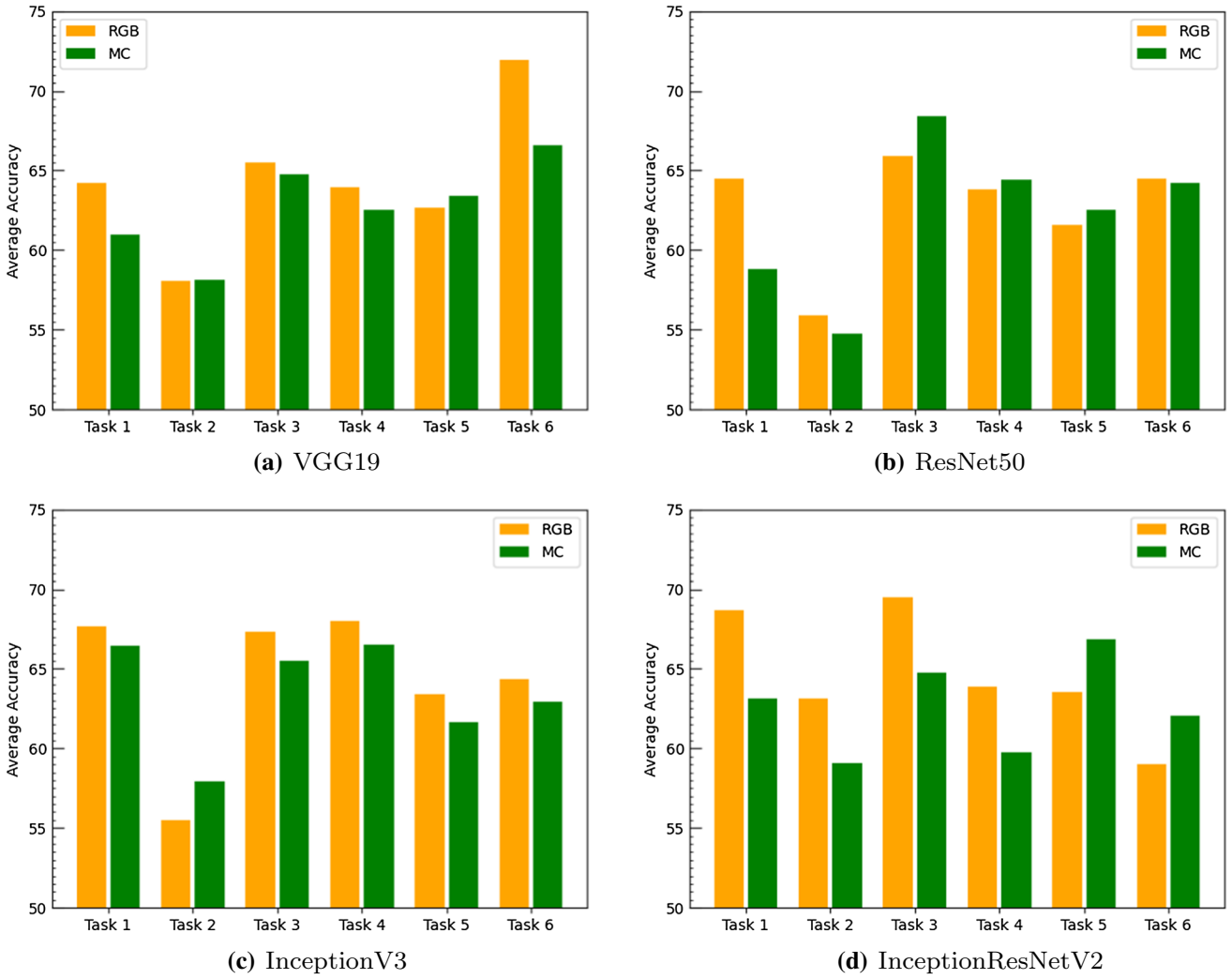
From Fig. 7b, showing the CNN performances using RGB features, we can observe that InceptionResNetV2 (60M parameters) achieved the best results on the first three tasks, whereas for the remaining tasks different cases occur. In particular, for tasks 4 and 6 the best performance was achieved by InceptionV3 (30M) and VGG19 (25M), respectively. Most probably, this is due to the higher complexity of these tasks when compared to the first three. Therefore, simpler CNNs allowed a more effective training on the available data, thus providing better results. On the other hand, on task 5 the CNNs achieve similar performances, confirming that in this case the number of parameters did not affect the training process.

Looking at Fig. 8b (CNN performance on MC features), we can see that, except for tasks 2 and 5, InceptionResNetV2 did not achieve the best performance. This result seems to suggest that the use of CNN with higher complexity does not allow the system to achieve better performance.

4.2 Comparison of MC and RGB features

In the second set of experiments, we compared the classification performance achieved by using RGB and MC features. The aim was to evaluate, in terms of performance, the contribution of the fourth channel in MC images.

Figure 9 shows the comparison between the classification performance achieved using MC and RGB features. To this aim, we averaged the accuracy achieved by the five classifiers used and plotted a vertical bar per task. From the graphs, we can observe that in most cases the performance achieved using RGB features is slightly better than (or comparable with) that achieved using MC features. This result confirms



(a) VGG19 **(b)** ResNet50

(c) InceptionV3 **(d)** InceptionResNetV2

Fig. 9 Accuracy for each task averaged over the results of five classifiers

Table 7 Classification results achieved by the FC classifier, using RGB and MC features

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	RGB	MC	RGB	MC	RGB	MC	RGB	MC	RGB	MC	RGB	MC
VGG19	64.0	66.2	59.0	62.7	70.7	69.5	64.7	64.2	64.8	64.7	70.4	64.5
ResNet50	61.1	61.7	53.5	59.0	68.8	72.8	64.4	67.7	64.3	65.6	66.8	68.6
InceptionV3	64.2	67.2	57.1	58.9	68.0	70.5	65.7	71.2	58.9	61.1	62.4	65.4
Inc.ResNetV2	60.0	61.6	62.7	58.7	68.0	67.8	65.3	64.8	62.0	70.7	55.2	55.0

Table 8 Results of classification with the handcrafted features

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD	ACC	SD
RF	**61.3**	2.5	**66.4**	1.7	53.0	3.2	**68.2**	1.5	**64.9**	2.9	**55.9**	2.4
K-NN	58.1	3.4	64.3	1.7	**57.9**	2.9	63.7	2.3	61.1	2.3	54.2	2.9
SVM	52.1	0.1	51.7	0.0	51.3	1.0	51.0	0.4	51.7	0.9	52.3	2.6
MLP	57.3	2.7	66.3	1.8	55.0	3.6	63.4	2.2	63.2	3.5	53.3	3.3
AVG	57.2		62.2		54.3		61.6		60.2		53.9	

Bold values highlight the overall best performance achieved on each task

that the information added by the fourth channel (see Sect. 3.3) did not allow our system to significantly improve its performance.

Moreover, as shown in Fig. 8a, in the case of MC features, the FC classifier achieved slightly better or comparable performance when compared with the other classifiers considered. This result confirms that during the training step, to deal with the higher complexity of the MC images, it was necessary to exploit the interaction between the feature extractor and the classification layer (see Sect. 3.3). Note that the results provided by the FC classifier show higher values for the standard deviation than those provided by the other classifiers (see Tables 5 and 6). This is probably due to the fact that, as previously mentioned, the FC results were computed by averaging the accuracy achieved over the five test folders, while the results of the other classifiers were averaged over 30 runs.

To highlight these aspects, we compared the performance of RGB and MC features achieved by using the FC classifier only (see Table 7).

4.3 Comparing deep and handcrafted features

In the last set of experiments we compared the classification performance achieved using RGB features with that achieved using the handcrafted features.

Table 8 shows the accuracy achieved using the handcrafted features. The last row of the table shows, for each task, the average accuracy computed over the four classifiers. From the table, we can observe that using these features the best performance is achieved by RF and K-NN. Thus, while confirming the effectiveness of the RF ensemble-based strategy, K-NN, in contrast to the deep features case, obtained satisfactory results. In this case, indeed, the K-NN algorithm was able to effectively estimate the probability distributions underlying our data represented through the handcrafted features. From the table, we can also observe that, in this case, task 2 allowed us to achieve good performance. These results suggest that for this task, in contrast to the deep features case, the information added by some of the handcrafted features allowed us to effectively distinguish between the handwriting of patients from that of the control group. On the contrary, tasks 3 and 6 achieved poor performance with these features. These results suggest that handcrafted features do not represent the shape and dynamics of handwriting in such a way to effectively distinguish handwriting samples of cognitively impaired people from those of the control group.

To summarize the comparison between deep and handcrafted features, we plotted a vertical bar graph showing the best overall accuracy achieved on each task (see Fig. 10). For each task, we plotted the best overall classification performance achieved by using deep-RGB features (bold values in Table 5) and handcrafted features (bold values in Table 8). From the plot, we can observe that our deep-based approach outperforms that based on the handcrafted features, except for task 2. These results confirm the effectiveness of our approach for combining shape and dynamic information. The slight performance difference in task 2 is probably due to the fact that the low complexity of this task does not allow the selection of discriminant features, as also confirmed by the poor classification results generally obtained by using both deep and handcrafted features.

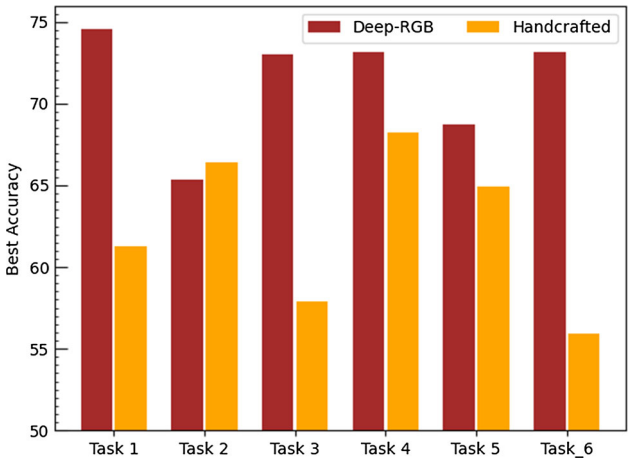


Fig. 10 Comparison results between deep-RGB and handcrafted features. The accuracy values shown are those highlighted in bold in Table 5

5 Discussion and conclusions

In this paper, we presented a deep transfer learning approach for feature selection applied to Alzheimer's disease diagnosis through handwriting analysis.

The rationale of our work was that of combining information derived from the shape of online handwritten traits with those related to the dynamics of the writing process used to produce such traits. To this aim, we generated synthetic offline multi-channel images, where each elementary trait in the handwritten trace is represented in each channel with a gray level encoding a single dynamic piece of information associated with that trait. Moreover, we exploited the capability of convolutional neural networks (CNNs) to automatically extract features from offline images.

In this study, we compared the results obtained by generating both three-channel (RGB) images and four-channel (TIFF) images. In the first case, the three channel encode for each elementary trait, velocity, jerk and pressure applied for producing that trait. In the second case, a fourth channel has been added to the previous ones to encode the acceleration. The experimental results obtained by exploiting the features extracted from these images were also compared with those obtained by using standard dynamic features directly derived from the original online handwriting samples. For the sake of comparison, the performance was evaluated using the same classifiers: This choice allowed us to easily analyze the experimental results relative to the different feature representations and, therefore, the role played by the shape and by the combined use of both shape and dynamic information. Finally, a further comparison was made by considering the classification results directly provided by the fully connected layer of CNN.

As a first consideration, we can observe that the deep features seem more promising than the handcrafted ones, reaching the best performance in terms of accuracy. Indeed, for each task and for each classification scheme, there is always a CNN model whose features allow us to obtain better results than those obtainable with the handcrafted ones. The only exception is task 2, where the best performance obtained by using handcrafted features was slightly better than that obtained with deep features.

Regarding the comparison between RGB and MC deep feature, the analysis of the results showed that the addition of a further channel in the generation of multi-channel images does not seem to allow better feature extraction: In fact, the classification results obtained with the RGB deep features are almost always better than those obtained with the MC deep ones. The only exception is the case of task 5, where the FC classifier, obtained by training Inc.ResNetV2 with MC deep features, produced slightly better results. However, it should be noted that these results were obtained using only handwriting samples related to graphic tasks: Thus, we intend to carry out, as a future work, a wider comparison using the data of all the writing tasks included in our protocol. Considering the whole set of tasks would also allow us to improve the overall performance of the diagnostic system, by combining for each participant the responses provided by the classifiers for each single task [40,41].

Finally, as a future development of our system, we would also like to include information related to in-air features: As mentioned above, we did not exploit this information because our aim was to evaluate the combined use of dynamic and morphological features. However, since in-air points are acquired from the tablet during the execution of the writing tasks, we could add these in-air traits when generating the synthetic images and evaluate their effects on the feature extraction process.

Declarations

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Conflict of interest The authors declare no conflict of interest.

References

1. Vessio, G.: Dynamic handwriting analysis for neurodegenerative disease assessment: a literary review. Appl. Sci. **9**(21), 4666 (2019)
2. Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Principles of Neural Science, 4th edn. McGraw-Hill Medical, New York (2000)
3. Lambert, J., Giffard, B., Nore, F., de la Sayette, V., Pasquier, F., Eustache, F.: Central and peripheral agraphia in alzheimer's disease: from the case of auguste d. to a cognitive neuropsychology approach. Cortex **43**(7), 935–951 (2007)
4. Neils-Strunjas, J., Groves-Wright, K., Mashima, P., Harnish, S.: Dysgraphia in Alzheimer's disease: a review for clinical and research purposes. J. Speech Lang. Hear. Res. **49**(6), 1313–30 (2006)
5. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., di Freca, A.S.: Handwriting analysis to support neurodegenerative diseases diagnosis: a review. Pattern Recognit. Lett. **121**, 37–45 (2018)
6. Werner, P., Rosenblum, S., Bar-On, G., Heinik, J., Korczyn, A.: Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. J. Gerontol. Psychol. Sci. **61**(4), 228–36 (2006)
7. Cilia, N.D., De Stefano, C., Marrocco, C., Fontanella, F., Molinara, M., di Freca, A.S.: Deep transfer learning for alzheimer's disease detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9904–9911 (2021)
8. LeCun, Y., Bengio, Y.: Convolutional Networks for Images, Speech, and Time-series. MIT Press, Cambridge (1995)
9. Cilia, N.D., De Stefano, C., Fontanella, F., Scotto di Freca, A.: An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. Proced. Comput. Sci. **141**, 466–471 (2018)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE Computer Society (2009)
11. Cilia, N.D., De Stefano, C., Fontanella, F., Molinara, M., Di Freca, A.S.: Handwriting analysis to support alzheimer's disease diagnosis: a preliminary study. In: Vento, M., Percannella, G. (eds.) Computer Analysis of Images and Patterns, pp. 143–151. Springer International Publishing, Cham (2019)
12. Cilia, N.D., De Stefano, C., Fontanella, F., Molinara, M., Di Freca, A.S.: Using handwriting features to characterize cognitive impairment. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) Image Analysis and Processing, pp. 683–693. Springer International Publishing, Cham (2019)
13. Cilia, N.D., De Stefano, C., Fontanella, F., di Freca, A.S.: How word choice affects cognitive impairment detection by handwriting analysis: A preliminary study. In: Cicirelli, F., Guerrieri, A., Pizzuti, C., Socievole, A., Spezzano, G., Vinci, A. (eds.) Artificial Life and Evolutionary Computation, pp. 113–123. Springer International Publishing, Cham (2020)
14. Cilia, N.D., De Stefano, C., Fontanella, F., Scotto di Freca, A.: Using genetic algorithms for the prediction of cognitive impairments. In: Castillo, P.A. et al. (eds) Applications of Evolutionary Computation. EvoApplications 2020. Lecture Notes in Computer Science, vol. 12104, pp. 479–493. Springer, Cham (2020)
15. Lei, B., Yang, M., Yang, P., Zhou, F., Hou, W., Zou, W., Li, X., Wang, T., Wang, S., Xiao, X.: Deep and joint learning of longitudinal data for alzheimer's disease prediction. Pattern Recognition **107247** (2020)
16. Cao, P., Liu, X., Yang, J., Zhao, D., Huang, M., Zaiane, O.: l2,1–l1 regularized nonlinear multi-task representation learning based cognitive performance prediction of alzheimer's disease. Pattern Recognit. **79**, 195–215 (2018)
17. Zhang, Y., Zhang, H., Chen, X., Liu, M., Zhu, X., Lee, S.W., Shen, D.: Strength and similarity guided group-level brain functional network construction for MCI diagnosis. Pattern Recognit. **88**, 421–430 (2018)
18. Bi, X., Wang, H.: Early alzheimer's disease diagnosis based on EEG spectral images using deep learning. Neural Netw. **114**, 119–135 (2019)
19. Fiscon, G., Weitschek, E., Cialini, A., Felici, G., Bertolazzi, P., De Salvo, S., Bramanti, A., Bramanti, P., De Cola, M.C.: Combining EEG signal processing with supervised methods for alzheimer's patients classification. BMC Med. Inform. Decis. Mak. **18**, 35 (2018)
20. Bevilacqua, V., Loconsole, C., Brunetti, A., Cascarano, G.D., Lattarulo, A., Losavio, G., Di Sciascio, E.: A model-free computer-assisted handwriting analysis exploiting optimal topology ANNs on biometric signals in parkinson's disease research. In: Huang, D.S. et al. (eds.) Intelligent Computing Theories and Application. ICIC 2018. Lecture Notes in Computer Science, vol. 10955, pp. 650–655. Springer, Cham (2018)
21. Loconsole, C., Cascarano, G.D., Brunetti, A., Trotta, G.F., Losavio, G., Bevilacqua, V., Di Sciascio, E.: A model-free technique based on computer vision and sEMG for classification in parkinson's disease by using computer-assisted handwriting analysis. Pattern Recognit. Lett. **121**, 28–36 (2019)
22. Diaz, M., Ferrer, M.A., Impedovo, D., Pirlo, G., Vessio, G.: Dynamically enhanced static handwriting representation for parkinson's disease detection. Pattern Recognit. Lett. **128**, 204–210 (2019)
23. Diaz, M., Moetesum, M., Siddiqi, I., Vessio, G.: Sequence-based dynamic handwriting analysis for parkinson's disease detection with one-dimensional convolutions and BiGRUs. Siddiqi **168**, 114405 (2021)
24. El-Yacoubi, M.A., Garcia-Salicetti, S., Kahindo, C., Rigaud, A.S., Cristancho-Lacroix, V.: From aging to early-stage alzheimer's: uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. Pattern Recognit. **86**, 112–133 (2019)
25. Impedovo, D., Pirlo, G., Mangini, F.M., Barbuzzi, D., Rollo, A., Balestrucci, A., Impedovo, S., Sarcinella, L., O'Reilly, C., Plamondon, R.: Writing generation model for health care neuromuscular system investigation. In: Proceedings of CIBB 2013, pp. 137–148. Springer (2014)
26. Pirlo, G., Cabrera, M.D., Ferrer-Ballester, M.A., Impedovo, D., Occhionero, F., Zurlo, U.: Early diagnosis of neurodegenerative diseases by handwritten signature analysis. In: ICIAP Workshops, pp. 290–297 (2015)
27. Garre-Olmo, J., Faundez-Zanuy, M., de Ipiña, K.L., Calvo-Perxas, L., Turro-Garriga, O.: Kinematic and pressure features of handwriting and drawing: Preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. Curr. Alzheimer Res. **14**, 1–9 (2017)
28. Yan, J.H., Rountree, S., Massman, P., Smith Doody, R., Li, H.: Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. J. Psychiatr. Res. **42**(14), 1203–1212 (2008)
29. Schröter, A., Mergl, R., Bürger, K., Hampel, H., Möller, H.J., Hegerl, U.: Kinematic analysis of handwriting movements in patients with alzheimer's disease, mild cognitive impairment, depression and healthy subjects. Dement. Geriatr. Cognit. Disord. **15**(3), 132–42 (2003)
30. Marcelli, A., Parziale, A., Santoro, A.: Modelling visual appearance of handwriting. In: Petrosino, A. (ed.) Image Analysis and Processing - ICIAP 2013. ICIAP 2013. Lecture Notes in Computer Science, vol. 8157, pp. 673–682. Springer, Berlin, Heidelberg (2013)
31. Marcelli, A., Parziale, A., Senatore, R.: Some observations on handwriting from a motor learning perspective. In: 2nd Interna-

tional Workshop on Automated Forensic Handwriting Analysis (2013)
32. Tseng, M.H., Cermak, S.A.: The influence of ergonomic factors and perceptual-motor abilities on handwriting performance. Am. J. Occup. Ther. **47**(10), 919–926 (1993)
33. Vyhnálek, M., Rubínová, E., Marková, H., Nikolai, T., Laczó, J., Andel, R., Hort, J.: Clock drawing test in screening for alzheimer's dementia and mild cognitive impairment in clinical practice. Int. J. Geriatr. Psychiatry **32**(9), 933–939 (2017)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Y. Bengio, Y. LeCun (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015). arxiv: abs/1409.1556
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2826 (2016)
37. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pp. 4278–4284. ACM (2016)
38. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
39. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1-27:27 (2011)
40. De Stefano, C., Fontanella, F., Marrocco, C., Di Freca, A.S.: A hybrid evolutionary algorithm for bayesian networks learning: An application to classifier combination. Appl. Evol. Comput. **6024**, 221–230 (2010)
41. De Stefano, C., Fontanella, F., Folino, G., Di Freca, A.S.: A bayesian approach for combining ensembles of GP classifiers. Mult. Classif. Syst. MCS **6713**, 26–35 (2011)

Nicole D. Cilia is currently an assistant professor at the Department of Computer Engineering, University of Enna. She participated in the PANOPTESEC EU project on Cyber Security in the Ontology for Reachability Matrix Computation package and in the national HAND project on Handwriting analysis against Neuromuscular Disease. She has been an invited speaker in international conferences and she has been awarded two starting grants regarding Eye-Tracking e Hand-Tracking tools to investigate analogical reasoning. She has been guest editor of two special issues on MDPI journals and co-chair of two workshops of the ICPR conference. She is a review editor for Frontiers in Robotics and AI and La mente e i sistemi cognitive, and she is a member of CVPL. In her research activity, she mainly deals with the methodological problems in Artificial Intelligence and with machine and deep learning techniques for the implementation of tools based on handwriting and eye-tracking analysis for the prediction of neurodegenerative diseases. She is the author of 40 scientific papers published in scientific journals and international conference proceedings.

Tiziana D'Alessandro received a Master's degree in Computer engineering from the University of Cassino and Southern Lazio (Italy) in 2019. She is currently a PhD student in Information Engineering at the University of Cassino and Southern Lazio. Her current research interests include artificial intelligence, deep learning, handwriting recognition, handwriting data analysis, historical document processing, and computer vision.

Claudio De Stefano received the Laurea Degree cum laude in Electronic Engineering and the PhD degree in Electronic and Computer Engineering both from the University of Naples "Federico II", Italy. He is currently Full Professor of Computer Science at the Department of Electrical and Information Engineering of the University of Cassino and Southern Lazio. He has been invited as visiting professor in many international research centers and, over the years, has participated, as a speaker, in relevant national and international conferences, and has authored over 160 publications in international journals and congresses. The results of his research activity have been published in relevant international journals. He joined the Program Committees of many important international conferences on image analysis, pattern recognition, handwriting analysis and recognition and was chair of international conferences. He also co-edited books as proceedings of international conferences and special issues of international journals. Since 1992, he has been a member of the International Association for Pattern Recognition (IAPR). Since 2017, he is the President of the International Graphonomics Society (IGS). He is Associate Editor of the journals Pattern Recognition Letters and Frontiers in Human Neuroscience (section Cognitive Neuroscience). The scientific interests of Claudio De Stefano include Artificial Vision, Image Processing, Pattern Recognition and Automatic Learning Systems. His current research interests include classifier combination paradigms and automatic learning methods based on the use of evolutionary algorithms and Neural Networks. Finally, handwriting analysis has recently been applied to the early detection of cognitive disorders.

Francesco Fontanella received the Laurea degree in Physics and the PhD degree in electronic and computer engineering both from the University of Naples "Federico II," Naples, Italy, in 2001 and 2005, respectively. He is an assistant professor at the Department of Electrical and Information Engineering of the University of Cassino and Southern Lazio. His research interests includes pattern recognition, evolutionary computation and bio-inspired computing. He has been guest editor of two special issues published on Pattern recognition Letters and co-chair of the International Workshop on Pattern Recognition of Cultural Heritage (PatReCH 2019, held in conjunction with ICIAP 2019) and technical co-chair of the 2018 IEEE International Conference on Metrology for Archaeology and Cultural Heritage (Metro Archeo 2018). He is also member of the board of SPECIES (Society for the Promotion of Evolutionary Computation in Europe and its Surroundings) and of the IEEE Computational Intelligence Society task force on evolutionary computer vision and image processing. He has authored over fifty scientific papers in journals and international conference proceedings.