Offline handwriting image analysis to predict Alzheimer's disease via deep learning

Nicole Dalia Cilia*†, Tiziana D'Alessandro‡, Claudio De Stefano‡ and Francesco Fontanella‡§
*Department of Computer Engineering, University of Enna "Kore"
†Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands
‡Department of Electrical and Information Engineering, University of Cassino and Southern Lazio
§Corresponding author, email: fontanella@unicas.it

***Abstract*—In the framework of Alzheimer's disease prediction systems, it is widely agreed that handwriting seems to be one of the first skills to be influenced by the onset of such a disease. In the large majority of cases, the above systems consider information relating to the dynamics of the handwriting process, directly derived from online handwriting samples. This kind of features, however, are not able to capture the alterations in the shape, size and thickness of the handwritten traits, which may be produced by the alterations in motor control due to neurodegenerative disorders. Following this line of thought, in a previous study we combined shape and dynamic information by generating synthetic color images from online handwriting samples, where the color of each elementary trait encodes, in the three RGB channels, the dynamic information associated to that trait. Finally, we exploited the capability of Deep Neural Networks to automatically extract features from raw images, following the Transfer Learning approach. The results obtained with this approach did not show significant improvements compared to those obtained with the use of dynamic information only, probably because approximating the original traits with straight lines of predefined thickness results in a loss of information on their actual shape and thickness. Moving from these considerations, the purpose of our study is to verify whether automatically extracting features directly from offline handwriting images, thus considering the original shape of the handwritten trace, could provide better results. Again, we exploited the capability of Deep Neural Networks to automatically extract features from raw images. The preliminary experimental results confirmed the effectiveness of the proposed approach.**

I. Introduction

As reported in many recent studies, the incidence and the social and economic costs of neurodegenerative diseases are expected to grow considerably in the coming years. Among these, Alzheimer's disease (AD) is one of the most serious, with a very strong impact on both quality and life expectancy of those affected. This disease causes a progressive decline in cognitive abilities, often characterized by an initial deterioration of memory, which over time can affect people's behavior, speech, visuospatial orientation and motor skills. Furthermore, the lack of a truly effective treatment for this disease makes its early diagnosis fundamental: the treatments currently available, in fact, tend to delay the effects of the disease and therefore should be started as soon as possible.

In this context, it is now widely agreed that handwriting seems to be one of the first skills to be influenced by the onset of neurodegenerative diseases [1], [2], [3], [4]. Handwriting, in fact, involves both motor and cognitive functions and may be subject to changes due to loss of muscle control, confusion and oblivion, usually progressively worsening over time. For these reasons, the analysis of handwriting and the study of its alterations has become of great interest for the development of support tools for diagnosing and for controlling the progression of these diseases. It is worth noting, however, that although many studies have been published that use handwriting to detect the presence of neurodegenerative disorders, there is still no standard experimental protocol that describes the set of handwriting tasks to be submitted to potential patients. Furthermore, there is also no reference datasets large enough to allow an effective performance evaluation of diagnostic support systems [5], [6].

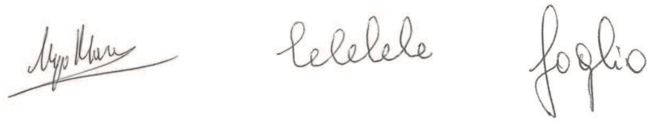
Another problem of great interest in this context is that of identifying the most significant information to be extracted from handwriting samples, which allow distinguishing the natural alterations of the handwriting due to age from those caused by neurodegenerative disorders. Also in this case, it should be noted that there is no universally accepted set of features, but the features that are considered most distinctive are those relating to the dynamics of the writing process. This is the main reason why it is preferred to use online handwriting samples acquired through tablets, which allow to detect information such as speed and acceleration, in addition to the pressure exerted by the writer during the writing process.

In this framework, we proposed an experimental protocol consisting of 25 tasks to analyze the impact of different motor skills on the performance of AD patients and healthy controls [7]. The protocol requires that the writing tasks be performed on tablets allowing subjects to write on normal paper sheets using a special pen that, in addition to producing the normal ink trace, also allows acquiring the handwritten trace with an assigned sampling frequency. The aim is to record the dynamics of the handwriting as well as to keep the handwritten trace on paper. The writing tasks defined in this protocol were administered to a group of approximately 180 subjects, including patients with an initial level of AD and a control group [7]. Both the AD patients and the control group were recruited with the support of the geriatrics department, Alzheimer's unit, of the "Federico II" hospital in Naples. The data obtained in this way allowed us to generate a quite large database, which was used to develop a system for the early diagnosis of AD based on the use of dynamic features (such as speed, jerk acceleration, etc.) [8], [9], [10], [11]. Although the experimental results obtained were encouraging, the analysis of the errors showed that by using only the information relating to the dynamics of the writing process, the proposed system was not able to distinguish in some cases patients with AD from the control group, especially in the early stages of the disease. This behavior is probably due to the fact that alterations in motor control due to neurodegenerative disorders can also produce significant alterations in the shape and size of the handwritten traits [12].

To solve this problem, in a previous study [13], we tried to combine the information related to the shape with that related to the dynamics of the handwriting process, by generating synthetic RGB images from online handwriting samples. In practice, for each online handwriting sample, a synthetic color image was generated picking up the points acquired by the tablet: for each pair of consecutive points, an elementary trait is produced (with a predefined thickness) whose color encodes, in the three RGB channels, the dynamic information associated with that trait. Finally, we exploited the ability of deep neural networks (DNN) to automatically extract features from raw images, following the Transfer Learning approach. The results obtained with this approach did not show significant improvements compared to those obtained with the use of dynamic information only, probably because the original traits produced by the writers were approximated with lines of predefined thickness, thus losing information on both their shape and their actual thickness.

Starting from these considerations, the purpose of our study is to verify whether the performance can be improved using the original offline images, obtained by digitizing the text written on paper sheets during the administration of our protocol. The rationale of such approach is to take into account shape, size and actual thickness of the handwriting samples. Again, we exploited the capability of Deep Neural Networks to automatically extract features from raw images, following the Transfer Learning approach. The analysis of the preliminary results, obtained by considering the data relating to 3 tasks of the protocol mentioned above, showed a significant performance improvement with respect to the use of feature extracted from RGB synthetic images, and performance comparable with those relative to the use of dynamic features. Finally, it is useful to underline that an important advantage that could derive from the use of this approach is the possibility to exploit for diagnostic purposes also parts of text previously written by the subjects, in order to verify if the initial signs of the disease were already present and to analyze its progression. The disadvantage is of course that of losing the dynamic information directly derived from the online data: however, it should be remarked that this effect does not seem to be relevant based on the preliminary results obtained.

The remainder of the paper is organized as follows: Section II describes the data acquisition process, Section III illustrates the feature extraction process, while Section IV presents the experimental results. Discussion and future works are eventually left to Section V.



(a) Task 1 (b) Task 2 (c) Task 3

Fig. 1. Example of offline images.

II. Data acquisition

As anticipated in the introduction, the data used for the experiments were obtained according to the experimental protocol presented in [7], which was administrated to 174 participants, 85 healthy controls and 89 patients (PT). The data were acquired at a frequency of 200Hz as a sequence of points stored in the form of x, y and z coordinates. The first two are the spatial coordinates of each point in the two-dimensional space representing the surface where the writing is produced, while the third is a measure of the pressure exerted in that point. Furthermore, since handwriting skills are influenced by age, gender, job type and education level, this information is also recorded and associated to the data of each handwriting sample.

The tasks considered in our protocol include the writing of groups of letters or words as well as the execution of graphic tasks, in order to evaluate motor control, memory and cognitive capacity of the subjects. For the purpose of this study, we considered only writing tasks: the participants were asked to write down their signature (Task 1), to write in cursive the bigram ”le” for four times continuously (Task 2) and to write the Italian word for sheet, ”foglio” (Task 3). The first task is very popular in the literature and it is typically performed with a highly automated gesture. The second task allows testing the fine motor control in repeating the same pattern in sequence but with a different size. Finally, the third is a copy task of a common word with an interesting graphic composition, as it includes ascending and descending traits. The choice of these tasks is because they allow checking the automatism, the regularity and the coordination of the motor sequences, as well as the spatial organization. Moreover, we have not included graphic tasks because one of the aim of this study is to verify if the features automatically extracted from offline images of handwriting allow an effective prediction of AD: under this hypothesis, earlier handwriting examples, which generally do not include graphic patterns, could also be exploited to check for previous signs of cognitive impairment, and to evaluate their progression.

A. Image generation

As discussed in the Introduction, in this study we considered two types of images: synthetic RGB images and offline images. The first were obtained from the online handwriting samples, starting from the points acquired by the tablet in terms of x, y and z coordinates. For each pair of consecutive points, the straight segment connecting them is generated with a pre-defined thickness. The effect is that the original curve is approximated by a sequence of straight segment, where the RGB color channels encode, for each segment, the dynamic information associated to the original movement, namely pressure, velocity and jerk. For further details we suggest checking out our recent publication [13].

As regards the second type of images, they were simply obtained by segmenting the original offline image produced for each task. Thus, in each image, the trace is exactly that produced by the participant while performing the task, and the pixel values correspond to the natural shades of gray left by the ink on the paper: such values depend on both the pressure exerted and the dynamics of the movements. Examples of offline images are shown in Fig. 1.

Finally, both RGB and offline images, were resized to adapt the original size of the images to the different input format required by the deep neural networks. The resizing was performed assuring that the trace was perfectly centered in order to minimize the loss of information related to possible zoom in/out.

III. Feature Extraction

Starting from the acquired data, two different features extraction approaches were adopted: handcrafted (HC) III-A and deep III-B. Regarding the first approach, the handcrafted features were extracted from the acquired coordinates by computing static and dynamic measures that characterize the handwriting process, considering also personal information. The second approach, instead, exploits the ability of Deep Neural Networks (DNNs) to extract features from the different types of images generated during the data acquisition phase. It is worth noting that online handwriting samples were used to extract both handcrafted and RGB deep features, while Offline deep features were obtained from the offline images of handwriting samples.

A. Handcrafted feature extraction

From the acquisition phase the trajectories of handwriting, in terms of **x, y** coordinates, are available. For each acquired point a third information representing the pressure (coordinate **z**), is also provided. As detailed in [14], from these coordinates, we computed the handcrafted features used for the classification step with standard machine learning (ML) algorithms.

We extracted the same set of features, 18 features, from all the tasks. The set included both static and dynamic features. The first ones were computed taking into account the shape or the position of the handwritten traits, whereas the second ones were related to the handwriting kinematics, like velocity and acceleration. Many studies in the literature have shown that the analysis of on air traits can provide significant information for identifying neurodegenerative disorders: on air movements, in fact, characterize the motor planning activities related to the positioning of the pen tip between two successive written traits. Moving from these considerations, we decided to extract some on air features calculated when the pen is not in contact with the paper sheet, but within a fixed distance (the value depends on the particular tablet used). Examples of dynamic features are *Time spent to perform the entire task*, *Velocity*, *Acceleration* and *Jerk*. Examples of static features are *Pendowns Number*, *X and Y Extension*. For further details see [14].

B. Deep feature extraction

We extracted the deep features from both RGB and offline images (see section II-A). These images were used to feed different models of Convolutional Neural Networks (CNNs): VGG19 [15], ResNet50 [16], InceptionV3 [17], InceptionResNetV2 [18]. These models have been improved over the years: first, new structural elements have been introduced; second, the number of layers and parameters has been increased. As previously mentioned, the input images (both RGB and offline) were resized to adhere the format required by the considered CNN: see the first column of table I.

CNNs architecture are composed of two parts: a feature extractor (FE) and a classifier (C). Both these parts have to be trained on a suitable dataset, to identify the weights that maximize classification performance. Preliminary experiments were performed to find effective values for the hyper-parameters. We selected the following values:

- 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: number of images from training set 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 $(trainingsetsize)/batch$.
- Patience 200: limit for epochs if the validation loss does not improve for a while.
- Accuracy as a measure of performance.

As regards the CNN training, we adopted a 5-fold cross validation strategy: the data set was randomly partitioned into 5 equally sized folds and at each iteration a different fold was used as test set. Moreover, to avoid the undesired over-fitting phenomenon, we used a validation set during the training phase. This implies that the samples belonging to the remaining 4 folds were divided into training and validation set. In particular, at each iteration, the percentage of samples used as test set is equal to 20%, while 70% of samples are used as training set and 10% as validation set.

As it is a good practice to initialize the parameters of the network, every CNN was pre-trained on ImageNet [19] and the training was done following two popular techniques: Transfer Learning (TL) and Fine Tuning (FT). During the TL step, all the parameters of the FE part were frozen, whereas the parameters of the classifier were randomly initialized and trained. During the FT step, both parts (FE and C) were involved in the training and all the parameters have been unfrozen. The FT assumes that the parameters of FE are

TABLE I
NUMBER OF DEEP FEATURES FOR EACH MODEL AT BOTTLENECK

Model	**Input size**	**Deep features size(N)**
VGG19	256x256	512
ResNet50	224x224	2048
InceptionV3	299x299	2048
InceptionResNetV2	299x299	1536

TABLE II
ML CLASSIFIERS AND THEIR HYPERPARAMETERS INVOLVED IN THE GRID SEARCH PROCESS

Classifier	Hyperparameters	constraits
XGB	min child weight	1, 5, 10
	gamma	0.5, 1, 1.5, 2
	subsample	0.6, 0.8, 1
	colsample bytree	0.6, 0.8, 1
	max depth	3, 4
RF	bootstrap	True, False
	max depth	10, 20, 50
	mas features	auto, sqrt
	min samples leaf	1, 2, 4
	min samples splir	2, 5, 10
	n estimators	100, 200
Tree	criterion	gini, entropy
	min samples split	2, 10
	max depth	2, 5, 10
	min samples leaf	1, 5, 10
	max leaf nodes	2, 5, 10
SVM	C	0.1, 1, 10, 100
	gamma	1, 0.1, 0.01, 0.001
	kernel	rbf
	class weight	balanced, None
MLP	hidden layer sizes	50, 100, 200
	activation	tanh, relu
	solver	lbfgs, sgd
	alpha	0.0001, 0.05
	learning rate	constant, adaptive

initialized with the weights obtained on ImageNet, while the C part is initialized with weights obtained during the previous TL step. After the whole training phase, the CNN networks have been used on one hand for deep feature extraction, and on the other hand for classification with the final fully connected classifier (FC). Note that the fully connected layer was specifically designed for our purposes. i.e. a two class classification problem (healthy control or patient). The same FC architecture was adopted for all the considered CNNs.

The output of the FE part of the network, for each input image, 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 Tab. I). As we processed two sets of images, RGB and offline, we obtained two sets of deep features, one for every image type.

IV. EXPERIMENTAL RESULTS

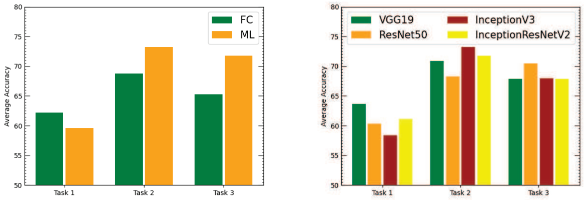
To test the effectiveness of the nine sets of features extracted (See Section III) we adopted a ML approach implementing five classification models: XGboost (XGB), random forest (RF), decision tree (DT), support vector machine (SVM), and multi layer perceptron (MLP). To allow each classifier to work in its best configuration, we performed a five-fold cross-validated grid search to select the best set of hyper-parameters for the classifier. In practice, we defined a set of values for each parameter to be tuned, and then exhaustively tested all parameter combinations. Table II shows the ranges of the hyper-parameters tested. Furthermore, to obtain statistically significant results, we performed 30 runs for each set of features. In each run, the dataset was randomly shuffled and split into a training and test set.

For each set of features, Table III shows the results of the top performing classifier in terms of accuracy (Acc). The table also shows the related Sensitivity (Sen), Specificity (Spe) and Precision (Pre). In the rows reporting ML results, the acronym in the Acc column denotes the best performing classifier in terms of accuracy for that set of features on that task. The table also shows the performance achieved by the fully connected (FC rows) layer of the CNNs. These classifiers consisted of two layers fully connected each made of 2048 neurons with a dropout layer between them. For each column, bold values highlight the best performance achieved.

For the sake of clarity, we report in the following the formal definition of the above evaluation metrics (the class Positive refers to the group of patients, whereas the class Negative refers to the group of healthy controls):

- Accuracy: fraction of correctly labelled examples among all predictions;
- Sensitivity (also denoted as True Positive Rate or Recall): rate of positive samples correctly classified;
- Specificity (also denoted as True Negative Rate): rate of negative samples erroneously classified;
- Precision: fraction of positives samples among those predicted as positives; it measures the precision of the learner when predicting positive samples;

From the table we can observe that the performance differs widely across the tasks and feature sets. From the table we can also see that the best accuracy was achieved for the first task by the RF classifier with the handcrafted features (65.86%), while for both the second and the third task by the XGB classifier with the offline features (74.7% and 74.4%, respectively). Overall, we can see that the first task shows the worst results, whereas the second task achieved the best



(a) Classifiers

(b) CNNs

Fig. 2. Average accuracy achieved by the classifiers (a) and the CNNs (b) using offline images.

TABLE III
PERFORMANCE COMPARISON.

	Task 1				Task 2				Task 3			
	Acc	**Sen**	**Spe**	**Pre**	**Acc**	**Sen**	**Spe**	**Pre**	**Acc**	**Sen**	**Spe**	**Pre**
RGB features												
VGG19												
ML	56.6 (RF)	61.6	51.6	59.3	68.6 (MLP)	67.1	64.7	68.7	66.5 (XGB)	67.1	66.5	67.7
FC	56.3	50.6	62.7	59.2	68.8	58.7	79.6	75.3	67.8	58.7	**77.2**	72.8
ResNet50												
ML	56.4 (XGB)	59.1	54.8	58.9	66.6 (RF)	69.9	64.6	68.6	65.3 (XGB)	67.9	64.1	66.8
FC	58.7	74.2	42.2	57.8	62.5	62.1	62.7	63.5	59.7	71.3	46.9	58.5
Inc.V3												
ML	59.97 (RF)	64.7	55.2	62.8	68.3 (XGB)	70.5	66.6	69.9	65.1 (RF)	69.9	61.6	66.8
FC	52.5	66.3	37.4	53.2	62.5	60.9	63.8	63.8	65.5	72.5	57.8	64.3
Inc.Res.V2												
ML	56.8 (RF)	60.8	54.7	59.5	69.6 (XGB)	69.3	66.1	69.5	63.3 (RF)	65.5	61.9	65.6
FC	55.1	58.4	51.8	56.5	68.8	60.9	**77.1**	73.6	60.7	**90.8**	28.9	57.3
Offline features												
VGG19												
ML	61.1 (XGB)	67.3	54.2	65.5	71.2 (XGB)	73.7	66.5	74.2	68.1 (MLP)	69.2	64.3	70.2
FC	66.3	**77.2**	53.8	65.3	70.4	73.8	66.6	71.4	67.6	69.3	65.3	69.3
ResNet50												
ML	61.2 (XGB)	67.3	54.8	64.4	70.6 (XGB)	73.3	68.0	74.9	**74.4** (XGB)	75.7	70.5	**77.8**
FC	59.3	57.9	60.2	62.1	65.9	88.6	41.1	62.9	58.9	82.9	32.1	57.9
Inc.V3												
ML	56.4 (XGB)	74.2	54.4	61.1	75.4 (XGB)	80.3	69.5	76.2	68.6 (RF)	80.3	51.7	66.5
FC	60.3	62.4	50.8	61.8	71.1	88.6	51.2	67.2	67.3	72.7	61.5	68.0
Inc.Res.V2												
ML	59.6 (XGB)	63.3	55.6	**66.4**	**75.7** (XGB)	78.5	72.8	**78.6**	68.4 (MLP)	73	62.8	71.5
FC	62.6	**77.2**	46.1	61.8	67.7	**92.1**	41.1	63.7	67.2	75	58.9	67.3
Handcrafted features												
ML	**65.8** (RF)	67.6	**63.8**	64.2	74.7 (RF)	78.8	70.8	72.2	69.2 (RF)	69.1	69.4	69.0

performance.

The first task requires a well known kinematic gesture, becoming almost an automatic graphic task, that doesn't require an important motor or cognitive attention. The second and third tasks, which include descending and ascending traits, require greater coordination and control skills and brings out the difference between patients and healthy controls.

From the table we can also observe that ML classifiers outperformed the FC ones, both for RGB and offline deep features, with the RF, XGBoost and MLP winning respectively seven, fourteen, and three times. This result demonstrates that: (i) ensemble-based architectures (RF and XGB) are more effective in capturing the differences between patients and healthy people; (ii) the grid-search procedure allowed us to optimize the performance of the ML classifiers.

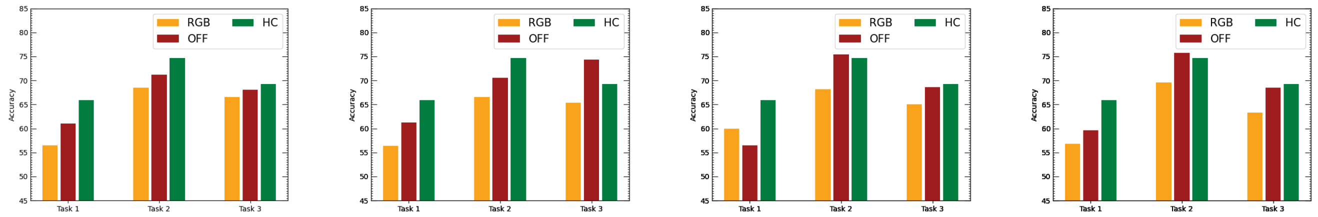
Furthermore, the accuracy values show that in most cases offline deep features outperformed the RGB deep features. These performance differences are larger for the ML classifiers, confirming that the latter were able to better exploit the information contained in the offline features. This is an interesting result, as these features were extracted from offline images containing the original traits of the participants' handwriting, without any information regarding the dynamics of the movements.

To further investigate the performance achieved using the offline features, we compared the performance of the FC and ML classifiers, as well as the those of the four CNNs used. In the first case, for each task we plotted the accuracy averaged over the CNNs. In the second case, for each task, we plotted the accuracy averaged over both the ML and FC classifiers. These plots are shown in Fig. 2. From the top plot we can observe that the ML classifiers outperformed the FC ones, confirming the effectiveness of the ML classifiers in characterizing the handwriting of people affected by AD. The bottom plot shows that the CNN performance varied across the tasks, and no CNN achieved best performance on all three tasks. These plots also confirmed that the second task allowed us to achieve the best performance in terms of accuracy.

To further compare the effectiveness of the three types of features extracted, for each task and for each CNN we plotted the average accuracy achieved by the ML classifiers in Fig. 3. Overall, the plots show that the HC features in most cases outperformed the others, while offline features always outperformed the RGB ones, except for the first task using InceptionV3. The offline and HC features show comparable performance for Task 2 and Task 3, which are more complex than Task 1: in fact, the sequence of ”le” bigrams and the word ”foglio”, require a higher motor control effort with respect to the signature, which is a highly automated gesture.

The fact that HC features generally provided better results is not surprising, since they have been widely used in the literature and include dynamic information related to the handwriting process: as previously mentioned the significance of these features to support the diagnosis of AD is well known in the field. The deep ones, instead, are automatically extracted from the CNN and their number far exceeds the number of handcrafted features.

Finally, to verify if combining the responses provided by



(a) VGG19

(b) ResNet50

(c) InceptionV3

(d) InceptionResNetV2

Fig. 3. Comparison between accuracy achieved by the ML classifiers from RGB, offline and Handcfrafted features.

the classifiers the overall performance could be improved, we performed a further set of experiments applying a majority vote rule: the results are summarized in Tab. IV. For each task, the table reports the accuracy obtained by combining the responses of the considered classifiers using the features extracted from both RGB and offline images: the first two rows refer to the results of the ML approach applied to RGB and offline features, respectively. Similarly, the third and fourth rows refer to the results of the FC layer of the CNNs using RGB and offline features, respectively. The table also reports the results obtained by combining the responses for all the tasks.

The results seem interesting, even if the performance increment is generally limited: this effect is probably due to the reduced number of classifier responses to be combined and to the simplicity of the combing rule. The best result was obtained by combining the responses of ML classifiers for offline features (accuracy equal to 75.94%). It is worth noting that the results obtained by combining all tasks are always lower than that obtained by combining the responses of Task 2: this is due to the fact that the accuracy obtained for the Task 2 is generally higher than that of the other tasks.

V. Discussion and Conclusions

The purpose of the study was to verify whether extracting features directly from original offline handwriting images, instead of synthetic RGB ones, thus considering the original shape of the handwritten trace, could provide better results with respect to those obtained with the handcrafted features. We exploited the capability of Deep Neural Networks to automatically extract features from raw images. The preliminary experimental results are very encouraging and confirm the effectiveness of the proposed approach.

As a first consideration, we can observe that offline deep features generally outperform the RGB deep features. This is an interesting result, because RGB images contain dynamic information, even if the shape of the reconstructed traits is an approximation of the original one. On the contrary offline images allow to capture all the shape details of the handwritten traits, which demonstrated to be very important for distinguishing patients from healthy controls. This result suggests that offline images can be used for diagnosing AD and that it may also be possible to consider previous examples of a person's handwriting to detect the presence of neurodegenerative disease and estimate its progression.

The second important results is that offline deep features show performance comparable with those obtained with handcrafted features. Even if in many experiments the accuracy provided by handcrafted features was higher, it should be remarked that this behavior is more evident for the data relative to Task1: the signature, in fact, represent an highly automated graphic gesture, whose shape is less influenced by the presence of AD. These results are in good accordance with those relative to the other tasks: in fact, in Task 2 and Task 3, which require the writer to produce more complex ink traces, shape changes are more relevant than changes in the dynamic features, thus allowing a better detection of AD patients.

It is worth noting that the sensitivity is generally higher in case of offline deep features: this results is very important in medical applications, where the cost of not identifying a subject with a given pathology is much higher than the cost of not correctly classifying a healthy subject.

Finally, although the performance obtained is still not adequate for the implementation of a real-world system, we believe that improving the feature extraction and selection process (e.g. selecting a specific set of features for each task) together with the adoption of more powerful combining rules, may significantly improve the overall classification accuracy.

TABLE IV
Combining results.

		Task 1	Task 2	Task 3	All Tasks
ML	**RGB**	59.05	70.25	67.46	69.35
	OFF	61.89	**75.94**	72.79	74.23
FC	**RGB**	60	72.34	69.5	68.02
	OFF	66.66	74.63	68.7	73.33

References

[1] G. Vessio, “Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review.” *Applied Sciences*, vol. 9(21):4666, 2019.

[2] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*, 4th ed. McGraw-Hill Medical, Jul. 2000.

[3] J. Lambert, B. Giffard, F. Nore, V. de la Sayette, F. Pasquier, and F. Eustache, “Central and peripheral agraphia in alzheimer's disease: From the case of auguste d. to a cognitive neuropsychology approach,” *Cortex*, vol. 43, no. 7, pp. 935–951, 2007.

[4] J. Neils-Strunjas, K. Groves-Wright, P. Mashima, and S. Harnish, “Dysgraphia in Alzheimer's disease: a review for clinical and research purposes,” *J Speech Lang Hear Res*, vol. 49, no. 6, pp. 1313–30, 2006.

[5] C. De Stefano, F. Fontanella, D. Impedovo, G. Pirlo, and A. Scotto di Freca A, “Handwriting analysis to support neurodegenerative diseases diagnosis: a review,” *Pattern Recognition Letters*, vol. 121, pp. 37–45, 2018.

[6] P. Werner, S. Rosenblum, G. Bar-On, J. Heinik, and A. Korczyn, “Handwriting process variables discriminating mild alzheimer’s disease and mild cognitive impairment,” *Journal of Gerontology: PSYCHOLOGICAL SCIENCES*, vol. 61, no. 4, pp. 228–36, 2006.

[7] N. Cilia, C. De Stefano, F. Fontanella, and A. Scotto di Freca, “An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis,” *Procedia Computer Science*, vol. 141, pp. 466 – 471, 2018.

[8] N. Cilia, C. D. Stefano, F. Fontanella, and A. Scotto di Freca, “Handwriting analysis for the diagnosis of alzheimers disease: A preliminary study,” in *LNCS - Computer Analysis of Images and Patterns*, vol. 11679. Springer, 2019, pp. 143–151.

[9] N. Cilia, C. De Stefano, F. Fontanella, and A. Scotto di Freca, “Using handwriting features to characterize cognitive impairment,” in *Lecture notes in computer science*. Springer, 2019.

[10] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto di Freca, “How word choice affects cognitive impairment detection by handwriting analysis: A preliminary study,” in *Artificial Life and Evolutionary Computation*, F. Cicirelli, A. Guerrieri, C. Pizzuti, A. Socievole, G. Spezzano, and A. Vinci, Eds. Cham: Springer International Publishing, 2020, pp. 113–123.

[11] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto di Freca, “Using genetic algorithms for the prediction of cognitive impairments,” in *Lecture notes in computer science*, vol. 12104. Springer, 2020, pp. 479–493.

[12] R. Inzelberg, M. Plotnik, N. K. Harpaz, and T. Flash, “Micrographia, much beyond the writer’s hand,” in *Parkinsonism Related Disorder*, vol. 26, 2016, pp. 1–9.

[13] N. Cilia, T. D’Alessandro, C. De Stefano, F. Fontanella, and M. Molinara, “From online handwriting to synthetic images for alzheimer’s disease detection using a deep transfer learning approach,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4243–4254, 2021.

[14] G. De Gregorio, D. Desiato, A. Marcelli, and G. Polese, “A multi classifier approach for supporting alzheimer’s diagnosis based on handwriting analysis.” in *ICPR 2020 Workshops (1)*, 2020, pp. 559–574.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[18] C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *AAAI*, 2016.

[19] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database.” in *CVPR*. IEEE Computer Society, 2009, pp. 248–255.