Handwriting Task-Selection based on the Analysis of Patterns in Classification Results on Alzheimer Dataset

Vincenzo Gattulli[1], Donato Impedovo[1] , Giuseppe Pirlo[1] Gianfranco Semeraro[1,2]

[1] *Department of Computer Science, University of Studies of Bari "Aldo Moro", Via Edoardo Orabona 4, Bari, 70125, Italy*
[2] *University School for Advanced Studies IUSS Pavia, Piazza Della Vittoria 15, Pavia, 27100, Italy*

Abstract

Alzheimer's Disease (AD) is a major disease associated with Dementia and a new case of AD is discovered every three seconds. Research proved that handwriting can be used to assess the health status of a subject. Writing and drawing tasks with different level of complexity were developed and inspected so far. At the same time there is evidence that not all the tasks have similar importance to evaluate the health status of patients. In this work a handwriting task selection is proposed based on the analysis of patterns of difficult cases (users) to be classified. The handwriting task selection was performed on the dataset "DARWIN" which includes 25 on-line recorded handwriting tasks (represented by a common set of parameter features). A wide set of classification models have been used to identify common difficult users: Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Bayesian Networks (BN), Gaussian Naïve Bayes (GNB), Multilayer Perceptron (MP) and Learning Vector Quantization (LVQ). The approach also highlighted that the group of selected tasks results to be composed with different type of writing and drawing tasks, and different level of complexity, simpler and complex tasks. Hence, it is showed that in order to discriminate healthy subjects from subjects suffering Alzheimer Disease it is necessary to use different tasks considering the type of task and the level of complexity

Keywords:

Alzheimer Disease, Handwriting, Task Selection, Machine Learning

1. Introduction

Progressive brain cells destruction manifest itself affecting memory, thinking, behavioral and emotional skills. These symptoms are referred as Dementia [1], [2]. Largely Dementia evolves becoming more severe as Alzheimer's Disease (AD) or Parkinson's Disease (PD) or Lewy Body Dementia. About the 60-80% of peoples suffering Dementia develop AD [2]. Alzheimer's Disease causes a disruption of the brain cells and nerves by abnormal proteins. The number of people around the world that are suffering from Alzheimer disease is incrementing estimating about of a new case every three seconds, accordingly with the World Alzheimer Report [1].

The research is moving forward looking at application of new techniques to accurately detect Dementia, and disease related to Dementia as AD, PD etc., using machine learning techniques [3], [4]. Considering that Dementia severity is classified in three different stages (early, mild and late), the focus of research in Dementia detection is on the early stage, in order to assess the health status of the patient timely to help clinicians to prepare and assign appropriate medical treatment [3], [4].

The assessment of the health status of a patient can be performed by using a wide range of data: images of handwriting tasks as draw and write text (offline handwriting) [5], time-series related to hand movements while performing handwriting tasks (online handwriting) [3], [6], videos of gait [7], audio from breath and cough [8], audio from speech [9] etc.

IEEESDS'23: Data Science Techniques for Datasets on Mental and Neurodegenerative Disorders, June 22, 2023, Zürich, Switzerland
gianfranco.semeraro@iusspavia.it (G. Semeraro);
0000-0003-1666-8323 (G. Semeraro);

In this work, the focus is on the online handwriting [3], [6]. The analysis of the handwriting results to be crucial due the links between the performing of handwriting and the Dementia diseases [10]. Indeed, in literature several handwriting tasks are used in order to assess the patients' health status [11]. Writing and drawing on digital tablets are action here referred as handwriting tasks, tablets capture the actions during their realization process. In this way it is possible to use raw data as time-series of pen-tip x-y coordinates on the digital tablet's surface, pressure and other information as azimuth and altitude. Also, in-air movements are collected if the pen tip is at most at one centimeter from the digital tablet surface (the maximum distance depends on the specific device) [10], [11]. The previous described raw data can be enhanced using the so called "function features", extracted directly from the used device or computed as time function (as displacement) [10].

The handwriting tasks can be divided in three mainly categories: simple writing, simple drawing and complex tasks. The three categories differentiate in the amount of cognitive loads and the motors skills involved in the action [10]. In this way, several amount of cognitive loads (from the lightest as drawing straight lines to the heaviest as copying details of bank check) and the motors skills are evaluated and used to classify subjects' health status[10].

Handwriting tasks could be synthetized using features extracted from the time-series raw-data. The result of such features extraction is that each handwriting task is described a fixed-length vector, hence a fixed number of information are obtained (as the task duration, number of strokes etc.). These features are called "parameter features" [10].

An open issue in this field is the selection of the most profitable tasks to be used for classification aims. In this work, a classification results-based task selection is performed in order to evaluate handwriting tasks in classification performance. Specifically, the proposed methodology is based on identifying common patterns from the handwriting tasks analysis of classification results and analyzing the relationship of these patterns with the dataset. In this way, best tasks are selected.

The work is organized as follows: Section 2 sketches the related work. Section 3 explains the methods and pre-processing techniques. The dataset is discussed in Section 4. Experimental Set-Ups is explained in Section 5. Benchmarking results and discussion is presented in Section 6. Conclusion are explained in Section 7.

2. Related Works

Handwriting is used in a wide range of scenarios from the recognition of the handwriting, drawing or script to the health status assessment by handwriting analysis, including early diagnosis, and to the writer verification and identification [11].

Cilia et al. [6] proposed different machine learning algorithms comparing them each other in two different setups: firstly, training the algorithms on the entire dataset (all tasks) and secondly creating task-specific classifier and then combing task-specific classifier at decision level. For the last set up, authors proceed to combining the best task-specific classifier in order to fuse them to enhance classification performance. Authors present a dataset of neurodegenerative disease, called "DARWIN" [6]. Also in this work, the "DARWIN" dataset [6] was used in order to analyze tasks performed by subjects for comparison aims because in [6] it is also presented a kind of task selection based on the classification performance.

In particular, in [6] a first way to perform selection was a combination of the all task-specific classifier in a majority voting approach. The second way was to select the top best classifier for each task and then combining them in a majority voting approach. The third way was based on the analysis of the task-specific classifiers performance. Specifically, a Friedman test-based ranking was computed about the classification models. Then, the task-specific classifiers were grouped incrementally from the top to the bottom of the ranking and used with the majority voting approach.

Such approaches do not take into account the tasks, but the performance of machine learning algorithms, in other words, the selection is performance-based. Meanwhile, in this work an extensively analysis of the classification results is performed in order to highlight common pattern along different users and then to select tasks. For this reason, the results used in this work for a fair comparison are them from the experiments performed using all tasks.

In this work "task" is referencing the group of features related to an handwriting task performed by a person. The group of features is the same for each task performed by a person. Hence, "selection" is referencing the selection of "task" in order to enhance the performance of machine learning algorithms. The proposed "task selection" is performed basing it on the analysis of classification results of machine learning models in order to find pattern to enhance the selection, and the related classification performance, results. The proposed method is presented to overcome the "curse of dimensionality" problem present in the DARWIN dataset [6]. Hence, the comparison of the performance is done with the results in [6].

3. Methods

The work is based on the protocol represented in Figure 1. Several machine learning algorithms were considered in order to inspect their common behavior and to identify difficult users to be classified. Successively handwriting task

selection was performed. Hence, in this section the machine learning algorithms are firstly explained. Secondly, the statistical and similarity methods to perform task selection are explained.

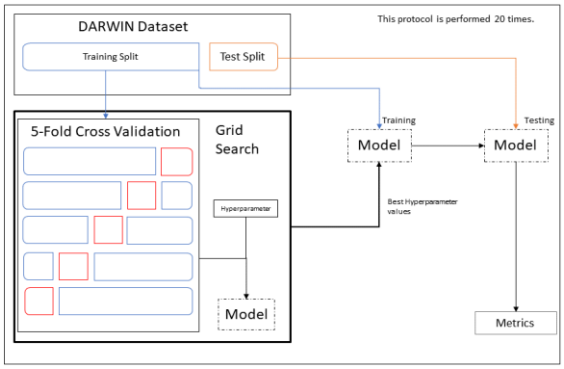

Figure 1 Experimental Protocol Representation.

3.1. Machine Learning Algorithms (Baselines models)

In this work a user is represented by a features vector belonging to a set of tasks. The following Machine Learning algorithms have been used to classify each user as a Healthy or a Patient:

- Decision-Tree (DT) is an algorithm based on the use of tree data-structure and it is suitable for classification task. Hence, the built tree is composed by nodes and leaf. Nodes represent features tests, meanwhile leaves represent class label. The result is a classifier able to predict the class label for the new input data.

- Random Forests (RF) is an ensemble algorithm. This means that the final model is composed of an ensemble of weak learners. In this case, the weak learners are decision trees. The classification is obtained by taking the class with the highest mean probability across the forest.
A different decision-tree is built introducing randomness while building the classifier, in order to decrease the variance of the forest classifier. The randomness is introduced using firstly the bagging technique and then the Random Subspaces technique.

- Logistic Regression (LR) is a classifier algorithm based on the use of the logistic function. The logistic function is used to model the class probabilities.

- K-Nearest Neighbors (KNN) is an algorithm used in both supervised and unsupervised methods. In this work, the KNN is used in a supervised manner in order to find the k data samples nearest to the new input. How much data are near to each other is computed using distance functions such as the Euclidean function. After the algorithm found the k nearest data to the input, the class is predicted using the majority voting principle: the class with a major occurrence in the neighborhood is the class predicted.

- Linear Discriminant Analysis (LDA) is an algorithm used for dimensionality reduction, data visualization, and classification. In this work, it was used as a classifier in order to build linear boundaries to classify data.

- Gaussian Naïve Bayes (GNB) is the implementation of the Naïve Bayes assuming that the likelihood is Gaussian. This algorithm is based on the Bayes Theorem whit the assumption that given a class, every feature pair is conditional independent.

- Support Vector Machine (SVM) is an algorithm based on the concept of decision planes. This algorithm aims to build hyperplanes in feature spaces, that can linearly separate two classes. The hyperplanes are computed by finding the maximum margins between the nearest training data and the hyperplane.

- Multi-Layer Perceptron (MLP) is an algorithm based on the use of stacked “neurons”. Specifically, neurons are the minimal computational element. Each element is composed by the application of, the so-called, activation function to the weighted combination of the input data features. The most famous activation function are: sigmoid, SoftMax, hyperbolic tangent, and ReLU.
The neurons can be stacked in different layers, and, for each layer, a different number of neurons could be used. The training is performed using the back-propagation algorithms that modify the weights of the combination.

- Learning Vector Quantization (LVQ) is an algorithm based on the use of prototypes. Specifically, the set of prototypes (called also neurons) is the solution for the classification. A prototype is an array of the same dimension of the data point. Hence, the prototypes are in the feature spaces. Each prototype belongs to one of the classes, this leads to a different number of prototypes for each class.
The algorithm consists of: for each data point, find the nearest prototype based on the distance function; then adjust the

position of that prototype accordingly with the class of the data point: moved near if the class is equal between data and nearest prototype, else moved away.

3.2. Task selection methodology

In general, the aim of task/features selection is aimed at identifying the most profitable tasks/features able to maximize performances. At the same time, it has been showed that there is not a common set of tasks/features able to be optimal for every users. Based on this consideration, in this work task selection is performed according to an analysis based on the most difficult users to be classified. In other words, for each classifier a list of misclassified users is identified:

$$M_c = \{u_i | o_i^c \neq l_i\},$$

With:

1. $i \in \{1,2,\dots,N\}$, being N the users within the dataset;
2. $c \in \{1,2,\dots,C\}$, being C the different classifiers considered (see the previous section);
3. $l_i = \begin{cases} HC \\ D \end{cases} being\ the\ label\ (correct\ class)\ for\ the\ i-th\ user$;
4. $o_i^c = \begin{cases} HC \\ D \end{cases} being\ the\ output\ provided\ by\ the\ c-th\ classifier\ for\ the\ i-th\ user.$

Then, all M_c subsets are intersected to identify the subset of users misclassified by all classifiers:

$$M \ = M_1 \cap M_2 \cap \dots \cap M_C,$$

This subset is here referred as "common misclassified individuals". Successively the similarity of features vectors of these individuals is evaluated to highlight if the subset is composed of similar individuals and thus to identify clusters. To the aim, the cosine similarity was adopted for the fast interpretation of the results. It is a similarity measure between two vectors of the same dimension: in our case a vector identifies a user. Considering two vectors the cosine similarity computes the cosine of the angular between the two input vectors. This measure ranges between [-1,1], where 1 indicates that the vectors are equal, -1 indicates the vectors are opposite and 0 indicates that vectors are orthogonal.

The refinement of the analysis was possible by looking for similarities between the "common misclassified individuals", referred as M, and the remaining subjects. In order to clearly define the remaining subjects, let use DD to identify the dataset composed by N subjects, then the remaining subjects are the subjects in the dataset DD but not in the set of "common misclassified individuals" M i.e., the remaining subjects are referred as $R = DD \backslash M$.

In order to compare similarities, once all the similarities were computed, the similarities were subject-based and class-based divided. This means that for each subject in M, the similarity with subjects in R are divided considering the correct class l_i. Mathematically, the notation are:

$$S_i^{HC} = \{cosine(u_i, u_k) \mid \forall\, u_k \in R\ and\ l_k = HC\ and\ u_i \in M\}$$

Indicating all similarities between the i-th subject in M and all subjects in R that have the correct class as HC, and:

$$S_i^{D} = \{cosine(u_i, u_k) \mid u_k \in R\ and\ l_k = D\ and\ \ u_i \in M\}$$

Denoting all similarities between the i-th subject in M and all subjects in R that have the correct class as D.

Since the high cardinality of S_i^{HC} and S_i^{D}, because of the amount of subjects in both classes, the analysis was carried out using a statistical method called One-Way ANOVA. In this way, the information carried out was about the statistical difference between the means of the two set of similarities, considering also the variance intra- and inter- the two sets. Such analysis was applied separately for each subject in M.

One-Way ANOVA assume as null hypothesis that the means of the two groups are statistically equal. Hence, the aim is to find the F-value to check if such value is higher or lower than the critical F-values defined assuming a significance level. Practically, the application of the analysis was carried out applying the following steps:

1. Computation of the overall mean: let $S_i^{TOT} = S_i^{HC} \cup S_i^{D}$, hence $mean_i^{TOT} = \frac{\sum_{cs_i \in S_i^{TOT}} cs_i}{|S_i^{TOT}|}$
2. Computation of the mean for each set: $mean_i^{HC} = \frac{\sum_{cs_i \in S_i^{HC}} cs_i}{|S_i^{HC}|}$ and $mean_i^{D} = \frac{\sum_{cs_i \in S_i^{D}} cs_i}{|S_i^{D}|}$
3. Computation of the sum of squared differences of mean between groups: $SS_{bg} = \sum_{l \in \{HC,D\}} \left(mean_i^l - mean_i^{TOT}\right)^2 * |S_i^l|$
4. Computation of the variance within groups: $SS_{within} = \sum_{l \in \{HC,D\}} \sum_{cs_i^l \in S_i^l} \left(cs_i^l - mean_i^l\right)^2$

5. Computation of between and within group degrees of freedom:
    - Between-group degree of freedom: $bg = N°\ of\ classes - 1$. In this case, $bg = 2 - 1$.
    - Within-group degree of freedom: $wg = Total\ number\ of\ subjects - N°\ of\ classes = \mathrm{N} - 2$
6. Computation of the mean square between and within the group: $MS_{between} = \frac{SS_{bg}}{bg}$ and $MS_{within} = \frac{SS_{within}}{wg}$
7. Computation of the F-Statistic : $F = \frac{MS_{between}}{MS_{within}}$
8. Finally, let consider the degree of freedom bg and wg, and the F-distribution. From $F(bg, wg)$, it was computed the critical F-values F_{crit} for a fixed significant level, in order to check if using the F-Statistic F the null hypothesis is rejected or not.
   The null hypothesis is rejected when $F > F_{crit}$.

Using one-way ANOVA, it was highlighted whether the similarity groups calculated for each of the subjects in M were different and, controlling for group means, which of the groups the subject was most similar to.

Considering that each subject is represented with a feature vector and that each feature vector is the concatenation of the same type of features computed for different writing tasks, the same statistical analysis, previously described, was repeated by dividing the feature vectors according to the reference task. In this way, it was possible to obtain, for each writing task and for each subject in M, the information regarding which class the subject was most similar to.

Using this information it was possible to perform a task-selection based on the similarity of subjects in M with the classes. Specifically, the tasks selected were those where at least one of the subjects in M was most similar to the correct class.

4. Dataset and experimental set-up

4.1. Dataset

The "DARWIN" dataset has been used [6]. "DARWIN" is the abbreviation for "Diagnosis AlzheimeR WIth handwriting". Hence, the dataset is composed by data related to handwriting tasks. In particular, a set of 174 subjects were examined. Such set is composed by 85 healthy peoples, that compose the Health Control group and were labelled with "H", and 89 subjects that suffer of Alzheimer, that were labelled with "P". The evaluation of the Alzheimer Disease severity was performed using clinical tests such as the Mini-Mental State Examination (MMSE), the Frontal Assessment Battery (FAB), and the Montreal Cognitive Assessment (MoCA). In order to build such dataset, subjects under medication or too compromised were excluded and a further selection of subjects was done in order to have two groups with similar statistical distribution of age, gender, type of work and educational level.

Each subject performed 25 handwriting tasks, drawing or writing. The performed tasks are showed in Table 1, and they was collected accordingly to the protocol described by Cilia N et al in [12] that consist of in performing each task on a paper sheet, an A4, placed on a digital tablet (a WACOM Bamboo Folio) in order to record the raw signal from the pen such as coordinates x and y, movements of the pen touching the paper sheet and movements near to the surface (in-air movements of the pen tip) and timestamps. Instructions were printed on the paper sheets and also provided by the instructor to ensure that the subjects performing the experiments understood the requirements.

Then, from each handwriting task, the same series of 18 features is extracted. The extracted features are related to the time spent to complete the task (total time); such time were also divided in on-paper time and in-air time; the average speed, acceleration and jerk computed separately for on-paper movements and in-air movements; mean and variance of the pressure; the Generalization of the Mean Relative Tremor (GMRT), which consists of the average of the sum of the differences between the i-th point and its d-th predecessor, firstly divided on-paper and in-air values, and the averaging the previous values; the maximal extension about x and y axis; and finally the Dispersion Index which consists of dividing the paper sheet in fixed-size boxes (e.g. 3x3) then counting how many boxes are covered by the handwriting traits and successively dividing that number with the total amount of the boxes; in this way, the coverage ratio of the paper sheet is computed.

For each subject, all the groups of features were concatenated obtaining vectors composed by 450 elements.

Table 1 Handwriting task performed by subjects in order to diagnose if they are suffering of Alzheimer or not. In the table there is a code to identify a task and its description.

Code	Task Description
1	Write the signature
2	Draw a horizontal line to join two points, continuously four times
3	Draw a vertical line to join two points, continuously four times

4	Redraw a circle with a diameter equal to 6 cm, for four times continuously
5	Redraw a circle with a diameter equal to 3 cm, for four times continuously
6	Copy letters such as the “l”, “m” and “p”
7	Copy the same letters of task 6 on an adjacent row
8	Write as a sequence for four times the letter “l” in cursive
9	Write as a sequence for four times the bigram “le” in cursive
10	Copy a specific word, “foglio” that means “sheet” in Italian
11	Copy a specific word, “foglio” that means “sheet” in Italian, above a line
12	Copy a specific word, “mamma” that means “mom” in Italian
13	Copy a specific word, “mamma” that means “mom” in Italian, above a line
14	Memorize and rewrite words “telefono”, “cane” and “negozio” (Italian version for “telephone”, “dog” and “store”)
15	Copying the word “bottiglia” (Italian translation of “bottle”) in reverse
16	Copying the word “casa” (Italian translation of “home”) in reverse
17	Copy six of regular, non-regular and non-words into the appropriate boxes. Specifically, “pane”, “mela” (respectively “bread” and “apple”) as regular words, “prosciutto”, “ciliegia” (respectively “ham” and “cherry”) as non-regular words, finally “taganaccio” and “lonfo” as non-words.
18	Write the object’s name shown in a picture (specifically, it was shown a chair)
19	Copying the fields of a postal order
20	Write the simple sentence dictated
21	Redrawing a complex shape
22	Copy a phone number
23	Write the dictated phone number
24	Draw a complete clock, with all hours, and put hands at 11:05 (Clock Drawing Test)
25	Copy a text paragraph. The paragraph has 110 characters accordingly with literature

4.2. Experimental Set-ups

The experimental pipeline is similar to the original one on the reference dataset for comparison aims [6]. More specifically, the dataset was randomly shuffled and divided into training and testing. This division was done 20 times. In this way, it is possible to alleviate the effects of the bias introduced by the randomness of the shuffling and the split. Each classifier was optimized on the training split to find the best hyperparameters using the grid search with a 5-fold cross-validation. The set of parameters to be optimized is equal to the set of parameters indicated in [6] and that set of parameters, along with the range of values in which to find the optimal solution, is represented in Table 2.

Table 2 For each classifier indicated the hyperparameters and the explored range of values.

Classifier	Parameter	Min Value	Max Value	Step
Random Forest	Max Depth	3	10	1
	N. Estimators	100	300	50
	Bootstrap	True	False	
	Min Samples Split	2		
	Min Samples Leave	1		
Logistic Regression	C	0.001	5	0.005
KNN	B Neighbors	5	15	1
Linear Discriminant Analysis	Solver	SVD		
Gaussian Naïve Bayes	Priors	2		
	Var. Smoothing	1e-9		

Support Vector Machine	Kernel	Radial Basis Function, Linear		
	C	0.5	1.5	0.1
	γ	0.5		
Decision Tree	Criterion	Gini, Entropy		
	Max Depth	2	10	1
	Min Samples Split	2	5	1
	Min Samples Leaf	2	20	2
	Max Leaf Node	2	20	2
Multi-Layer Perceptron	Activation	ReLu, Logistic, TanH		
	Hidden Layer Size	8	20	1
	Learning Rate Init.	0.05	0.4	0.05
	Max Iteration	1000		
	α	0.0001		
Learning Vector Quantization	Prototype for classes	1	50	5
	β	2	50	5
	Max Iteration	2500		

After that the best hyperparameters are found, the model is trained on the training split with the best hyperparameter, and then the classifier is tested. The metrics for each run of the experiments are averaged. The experimental set-setup explained is represented in Figure 1.

5. Results and Discussion

5.1. Preliminary results

The first step has been the evaluation of the different classifiers to the entire set of available tasks. This is the baseline for further analysis as well as the state-of-art on the DARWIN dataset [6]. The baseline values are reported in Table 3.

Table 3 Classification accuracy, specificity and sensitivity presented in [6] *This results are used as baseline.*

	RF	**LR**	**K-NN**	**LDA**	**GNB**	**SVM**	**DT**	**MLP**	**LVQ**
Accuracy [6]	0.8829	0.8186	0.7143	0.7214	0.85	0.79	0.7857	0.8314	0.7743
Specificity [6]	0.8618	0.7941	0.8941	0.7265	0.7912	0.8059	0.7441	0.8176	0.8735
Sensitivity [6]	0.9028	0.8417	0.5444	0.7167	0.9056	0.7750	0.8250	0.8444	0.6806

An in-depth analysis, as described in par 3.2, of the results showed that there is a subset of individuals who are misclassified at least once during the execution of the experiments from all the used classifiers. For the sake of clarity, detail how the subset of common misclassified individuals was identified is explained: by analyzing the predictions of each classifier for each of the 20 runs of the experiments, a subset of individuals misclassified by a specific classifier was identified. Then, all subsets of individuals misclassified by the specific classifier were intersected to identify the subset of common individuals misclassified by all classifiers. In this case $M_c = \{2,7,41,42,45,63,74\}$ is the set of common misclassified users. The range of cosine similarity values is [0.8175, 0.8946]. This result suggests that the individuals are remarkably similar, this also reflects the common behavior of the different classifiers seems to be justified. The values are represented in Figure 2.

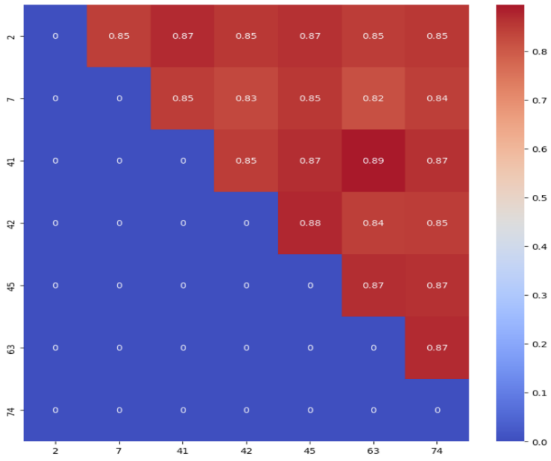

Figure 2 Cosine Similarity matrix. The similarity was calculated only for the upper triangle of the matrix and avoiding the computation for the diagonal since the diagonal should have all values equal to 1.00.

Furthermore, the subset of commonly misclassified individuals results to be composed only of patients. This means that these patients are (miss)-classified as healthy subjects by all the classifiers. This information leads to the hypothesis that the commonly misclassified individuals are patients more similar to the healthy class than the patients. In order to demonstrate that hypothesis, further analyses are performed comparing the cosine similarity between the commonly misclassified individuals and the patient's class and healthy class.

The healthy class results to have similarity in range [0.80 – 0.9275]. Meanwhile, the similarity range with all other patients is in the range [0.6764 – 0.9159]. The values are reported in

Table 4.

Table 4 Cosine Similarity between the common misclassified individuals and the healthy and patients' classes. In the table is also reported the number of individuals for both classes.

	N° Individuals	**Min**	**Max**
Patients	81	0.6764	0.9159
Healthy	85 – full class	0.80	0.9275

In order to refine such analysis, the One-Way ANOVA Test was performed as described in par. 3.2. In addition to the analysis of the distribution, a comparison between the means of both classes is performed to understand which class is more similar to each of the commonly misclassified individuals. The results are all synthesized in Table 5.

Table 5 One-Way Anova F Statistics, P Values, and means of similarity distribution separate are reported. As last row, there is the evaluation of such values. As it is noticeable, all the individuals are more similar to healthy class than the patients' class which is their real class

ID misclassified individuals	**2**	**7**	**41**	**42**	**45**	**63**	**74**
F Statistic	72.583	80.198	48.022	52.423	51.976	64.453	41.255

P Value		9.3063e-15	6.6378e-16	8.8757e-11	1.5867e-11	1.887e-11	1.7221e-13	1.3517e-09
Means of similarity with:	**Patients**	0.82357	0.81192	0.84661	0.82499	0.83452	0.83658	0.82622
	Healthy	**0.8599**	**0.8605**	**0.87797**	**0.85722**	**0.86502**	**0.87073**	**0.86859**

Also the statistical analysis of the distribution of the similarity confirms the results of the classifiers. Moreover, the statistical analysis confirms the hypothesis that the common misclassified individuals are patients more similar to healthy class than patients' class.

In order to perform tasks selection, the previous statistical analysis was extended to the task level. This means that a table similar to Table 5 is obtained for each task and based on these tables, a "task selection" was performed by choosing the task for which at least one of the commonly misclassified patients is most similar to the "patient" class. For the sake of readability, only the final selected tasks are here reported:

- task 2 (i.e., joint two points with a horizontal line, continuously four times);
- task 17 (i.e., copying six words in appropriate boxes);
- task 24 (i.e. drawing a clock with all hours and hands at 11:05);
- task 25 (i.e., copying a paragraph).

The selected tasks are classified, accordingly to [10]:

- task 2 is a simple drawing task, because it is suitable for the evaluation of simple motor performance;
- task 17 is simple writing tasks, because the used six words allows to capture information related to motor processes and planning skills, moreover the planning skill involves the identification of the proper target box. Another interesting issue is that words are composed with letters that contains similar movement patterns ("e", "l"), with letters that needs to lift the pen to complete the word ("t", "i"). Furthermore, using more than one words allows to capture the fatigue effects on handwriting;
- task 24 and task 25 as complex tasks: motor, cognitive and functional skills are involved, indeed the subject have to respectively remember the clock shape or read the text, have to search and find the right place to write and then the subject have to write the content.

Given the above, it is noticeable that the group of selected tasks is composed by different types of tasks, indeed there are drawing and writing tasks. Moreover, the tasks are of different complexity from the simpler, task 2, to the complex ones, task 24 and 25. Hence, based on the dataset and on the experiments, it is possible to highlight the following: firstly, there is not an unique task that allows to perfectly discriminate healthy subjects to subject that suffer of Alzheimer Disease; secondly, the selected tasks are of different type and complexity, hence to evaluate the healthy status of a patients a combination of different task is needed.

5.2. Results after Task-Selection

Finally, the experimental pipeline above described (see fig. 1) was applied using the information of the selected tasks: results are reported in Table 6.

Table 6 Accuracy, specificity, and sensitivity of the experiments performed on the selected tasks (2, 17, 24, and 25) and the accuracy, specificity, and sensitivity presented by the authors in [6]*. The bold results are better comparing the results presented in* [6].

	RF	**LR**	**K-NN**	**LDA**	**GNB**	**SVM**	**DT**	**MLP**	**LVQ**
Accuracy [6]	0.8829	0.8186	0.7143	0.7214	0.85	0.79	0.7857	0.8314	0.7743
Specificity [6]	0.8618	0.7941	0.8941	0.7265	0.7912	0.8059	0.7441	0.8176	0.8735
Sensitivity [6]	0.9028	0.8417	0.5447	0.7167	0.9056	0.7750	0.8250	0.8444	0.6806
Accuracy	0,8243	**0,8357**	**0,74 86**	**0,765 7**	0,77	**0,80 43**	0,7257	0,7129	**0,835 7**

Specificity	3	0,814	**0,8645**	**01**	**0,96**	**0**	**0,793**	**51**	**0,85**	**17**	**0,83**	3	0,709	5	0,722	5	0,851
Sensitivity	2	0,838	0,8153	**78**	**0,57**	**8**	**0,746**	5	0,70	**76**	**0,78**	1	0,741	8	0,722	**7**	**0,821**

It is noticeable that K-NN, Linear Discriminant Analysis and Support Vector Machine are the three algorithms that have major benefits from the task selection. Indeed, all the three metrics of accuracy, specificity and sensitivity are greater than the same metrics for the same classifier using all the set of available tasks [6].

In addition, the accuracy of Logistic Regression (LR) and Learning Vector Quantization (LVQ) has improved significantly. These improvements are accompanied by the improvement of specificity (an increase of about 7 percent) of LR and sensitivity (an increase of about 14 percent) of LVQ. Together, the sensitivity of LR and specificity of LVQ decreased by about 3% but remained above 80%.

Gaussian Naïve Bayes (GNB) specificity is the only improved metric, while the other metrics are decremented from the results presented in [6]. Considering that the task selection was performed by taking into account the commonly misclassified individuals and that the individuals are all patients, the specificity of GNB is improved accordingly with the proposed task selection.

In addition, for every model that improves on at least one metric, specificity is higher than sensitivity. This means that the models, using only selected tasks, are better able to detect whether a patient has Alzheimer's disease.

6. Conclusion

In this paper, a handwriting task selection methodology is proposed based the analysis of classification results. The analysis of classification results is based on statistical analysis and similarity analysis in order to find pattern in results.

The experiments were repeated 20 times to ensure classification results and to mitigate the effects of the randomness in the data due to the shuffling and split process. Indeed, at each repetition the dataset was randomly shuffled and the split in train and test subsets. The classifier models were optimized on the training set using a grid-search 5-fold cross validation, in order to evaluate the optimal hyperparameters values given a set of parameters and their range of values to evaluate. Furthermore, the experiments described above were performed on the entire dataset (using all the tasks and related features set) and on the group of selected tasks.

The analysis of the experimental results on the entire dataset had highlighted a subgroup of patients (subjects that suffer Alzheimer Disease) miss-classified at least once as healthy from all the classifiers (called commonly misclassified individuals). This is a very interesting results, in fact these could be early cases. Further analysis shown that this subgroup of patients results to be more similar to the healthy subject.

Hence, an analysis at task level was performed. From such analysis a subgroup of tasks was selected. The selected tasks are the tasks where at least one patients of the commonly misclassified individuals results to be more similar to the patients class.

The proposed methodology results in improved classification performance of several machine learning models. In particular, for the K-NN, Linear Discriminant Analysis, and Support Vector Machine models, a complete improvement was obtained on all performance metrics used. In addition, due to this methodology, the specificity of the previous models, together with Logistic Regression and Learning Vector Quantization. Hence, the models are more capable of identifying individuals with Alzheimer's disease using only a small number of “tasks”.

The selected tasks are of different levels of complexity, from simple task to complex task, and they are of different type, writing and drawing tasks. It is possible to conclude that it is necessary to use a group of different tasks in order to evaluate the healthy status of a subject.

Even if results have revealed that the approach is able to reveal early cases of patients miss-classified by the different classifiers, and even if performances are improved in many cases adopting the proposed schema, it is clear that more research is needed. Indeed, future improvements could be performed using the kinematic theory. Such approach is widely used in many domains as in fall detection [13], human activity recognition [14], and surely in handwriting [15]. Furthermore, considering the use of features instead of time-series, it could be possible build a client-server system in order to provide such application worldwide. Obviously, data should be protected following SoA techniques [16]–[21].

Acknowledgments

This paper and related research have been conducted during and with the support of the Italian national inter-university PhD course in Sustainable Development and Climate Change (link: www.phd-sdc.it)

References

[1] C. Patterson, *World Alzheimer Report 2018 - The state of the art of dementia research: New frontiers*. ALZHEIMER'S DISEASE INTERNATIONAL, 2018. Accessed: Jan. 31, 2023. [Online]. Available: https://apo.org.au/sites/default/files/resource-files/2018-09/apo-nid260056.pdf

[2] S. Gauthier, C. Webster, S. Servaes, J. A. Morais, and P. Rosa-Neto, *World Alzheimer Report 2022 – Life after diagnosis: Navigating treatment, care and support*. London, Englad: Alzheimer's Disease International, 2022. Accessed: Jan. 31, 2023. [Online]. Available: https://www.alzint.org/u/World-Alzheimer-Report-2022.pdf

[3] V. Gattulli, D. Impedovo, G. Pirlo, and G. Semeraro, "Early Dementia Identification: On the Use of Random Handwriting Strokes," *In: Carmona-Duarte, C., Diaz, M., Ferrer, M.A., Morales, A. (eds) Intertwining Graphonomics with Human Movements. IGS 2022. Lecture Notes in Computer Science. Springer, Cham.*, vol. 13424, pp. 285–300, Dec. 2022, doi: 10.1007/978-3-031-19745-1_21.

[4] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying Alzheimer's disease," *Simul Model Pract Theory*, vol. 99, p. 102023, Feb. 2020, doi: 10.1016/J.SIMPAT.2019.102023.

[5] V. Dentamaro, P. Giglio, D. Impedovo, and G. Pirlo, "Benchmarking of Shallow Learning and Deep Learning Techniques with Transfer Learning for Neurodegenerative Disease Assessment Through Handwriting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12917 LNCS, pp. 7–20, Sep. 2021, doi: 10.1007/978-3-030-86159-9_1.

[6] N. D. Cilia, G. de Gregorio, C. de Stefano, F. Fontanella, A. Marcelli, and A. Parziale, "Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking," *Eng Appl Artif Intell*, vol. 111, p. 104822, May 2022, doi: 10.1016/J.ENGAPPAI.2022.104822.

[7] V. Dentamaro, D. Impedovo, and G. Pirlo, "Gait Analysis for Early Neurodegenerative Diseases Classification through the Kinematic Theory of Rapid Human Movements," *IEEE Access*, vol. 8, pp. 193966–193980, 2020, doi: 10.1109/ACCESS.2020.3032202.

[8] V. Dentamaro, P. Giglio, D. Impedovo, L. Moretti, and G. Pirlo, "AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath," *Pattern Recognit*, vol. 127, p. 108656, Jul. 2022, doi: 10.1016/J.PATCOG.2022.108656.

[9] L. Ilias, D. Askounis, and J. Psarras, "Detecting Dementia from Speech and Transcripts using Transformers," *Comput Speech Lang*, vol. 79, Apr. 2023, doi: 10.1016/J.CSL.2023.101485.

[10] D. Impedovo and G. Pirlo, "Dynamic Handwriting Analysis for the Assessment of Neurodegenerative Diseases: A Pattern Recognition Perspective," *IEEE Rev Biomed Eng*, vol. 12, pp. 209–220, May 2018, doi: 10.1109/RBME.2018.2840679.

[11] M. Faundez-Zanuy, J. Mekyska, and D. Impedovo, "Online Handwriting, Signature and Touch Dynamics: Tasks and Potential Applications in the Field of Security and Health," *Cognit Comput*, vol. 13, no. 5, pp. 1406–1421, Sep. 2021, doi: 10.1007/S12559-021-09938-2/FIGURES/9.

[12] N. D. Cilia, C. de Stefano, F. Fontanella, and A. S. di Freca, "An Experimental Protocol to Support Cognitive Impairment Diagnosis by using Handwriting Analysis," *Procedia Comput Sci*, vol. 141, pp. 466–471, Jan. 2018, doi: 10.1016/J.PROCS.2018.10.141.

[13] V. Dentamaro, D. Impedovo, and G. Pirlo, "Fall detection by human pose estimation and kinematic theory," *Proceedings - International Conference on Pattern Recognition*, pp. 2328–2335, 2020, doi: 10.1109/ICPR48806.2021.9413331.

[14] V. Gattulli, D. Impedovo, G. Pirlo, and L. Sarcinella, "Human Activity Recognition for the Identification of Bullying and Cyberbullying Using Smartphone Sensors," *Electronics (Switzerland)*, vol. 12, no. 2, Jan. 2023, doi: 10.3390/ELECTRONICS12020261.

[15] V. Dentamaro, D. Impedovo, and G. Pirlo, "An Analysis of Tasks and Features for Neuro-Degenerative Disease Assessment by Handwriting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), vol. 12661 LNCS, pp. 536–545, 2021, doi: 10.1007/978-3-030-68763-2_41.

[16] A. Cannarile, V. Dentamaro, S. Galantucci, A. Iannacone, D. Impedovo, and G. Pirlo, "Comparing Deep Learning and Shallow Learning Techniques for API Calls Malware Prediction: A Study," *Applied Sciences 2022, Vol. 12, Page 1645*, vol. 12, no. 3, p. 1645, Feb. 2022, doi: 10.3390/APP12031645.

[17] D. Impedovo, V. Dentamaro, G. Abbattista, V. Gattulli, and G. Pirlo, "A comparative study of shallow learning and deep transfer learning techniques for accurate fingerprints vitality detection," *Pattern Recognit Lett*, vol. 151, pp. 11–18, Nov. 2021, doi: 10.1016/J.PATREC.2021.07.025.

[18] F. Castro, D. Impedovo, and G. Pirlo, "A Medical Image Encryption Scheme for Secure Fingerprint-Based Authenticated Transmission," *Applied Sciences (Switzerland)*, vol. 13, no. 10, May 2023, doi: 10.3390/APP13106099.

[19] D. Impedovo, A. Longo, T. Palmisano, L. Sarcinella, and D. Veneto, "An investigation on voice mimicry attacks to a speaker recognition system," *CEUR Workshop Proc*, vol. 3260, pp. 114–123, 2022.

[20] F. Carrera, V. Dentamaro, S. Galantucci, A. Iannacone, D. Impedovo, and G. Pirlo, "Combining Unsupervised Approaches for Near Real-Time Network Traffic Anomaly Detection," *Applied Sciences 2022, Vol. 12, Page 1759*, vol. 12, no. 3, p. 1759, Feb. 2022, doi: 10.3390/APP12031759.

[21] A. Coscia, V. Dentamaro, S. Galantucci, A. Maci, and G. Pirlo, "YAMME: a YAra-byte-signatures Metamorphic Mutation Engine," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2023, doi: 10.1109/TIFS.2023.3294059.