AI-based Ecological Monitoring of Handwriting to Early Detect Cognitive Decline

Simone Toffoli
Electronics, Information and Bioengineering Department
Politecnico di Milano
Milano, Italy
simone.toffoli@polimi.it

Francesca Lunardini
Center for Clinical Neuroscience
Hospital los Madroños
Madrid, Spain
francesca.lunardini@inntegra.es

Carmen Galán de Isla
Fundación para la Formación e Investigación de los Profesionales de la Salud de Extremadura
Mérida, Spain
carmen.galan@fundesalud.es

Simona Ferrante
Electronics, Information and Bioengineering Department
Politecnico di Milano
Milano, Italy
simona.ferrante@polimi.it

***Abstract*— The early detection of Mild Cognitive Impairment (MCI) is fundamental to initiate treatments for delaying the onset of dementia. Currently, the Mini Mental State Examination (MMSE) is one of the most common clinical scales used by geriatricians to assess cognitive function. A deviation of 1 to 3 points from the maximum score (30) is considered as sign of relevant cognitive decline. However, objective and affordable tools are needed to complement the screening process. The quantitative analysis of handwriting represents a suitable solution, as the gesture is significantly impaired in MCI subjects in terms of time, speed, fluency and applied pressure. This works presents the development and testing of classification models able to separate subjects at risk of cognitive decline (MMSE <= 28) from controls (MMSE > 28), starting from free-content handwriting data acquired with a smart ink pen, used on paper, from which 36 indicators were computed. Data were collected in 2 phases. The former involved 45 subjects and served for models training. In the latter, data were acquired from 23 subjects in a domestic longitudinal framework and were partially used for model refinement, but mainly for testing. Three different algorithms were tried (support vector machine, random forest and Catboost) The best test performances on the longitudinal data were obtained by a Catboost classifier, achieving accuracy 93.33%, precision 88.89%, recall 100% and f1 score 94.12%. The results support the use of computerized handwriting analysis as screening tool for cognitive decline detection.**

Keywords*— *Cognitive Decline, Handwriting, Smart Ink Pen, Machine Learning, Feature Engineering

I. INTRODUCTION

The ageing phenomenon is responsible for the increasing prevalence of neurological chronic pathologies [1]. Among these, dementia holds a huge burden, for both the patients and the healthcare system, as no disease-modifying treatments are currently available [2]. The Mild Cognitive Impairment (MCI) is considered a precursor of dementia [1], [2]. Subjects with MCI can live independently but exhibit a higher degree of cognitive decline compared to healthy age-matched individuals [2]. The prevalence of MCI in the population aged over 60 years old is 17.3%, while the rate of conversion to dementia is around 10-15% patients per year [3]. In this context, the identification of the first signs of cognitive decline plays a fundamental role. The early detection would indeed allow starting a treatment plan able to delay the onset of dementia [1]–[3]. To date, the screening procedure for MCI is performed by general practitioners and geriatricians through clinical scales, the Mini Mental State Examination (MMSE) being one of the most used [4]. It allows the quick evaluation of the subject's cognitive status in different domains, including language, attention and memory [4]. The score ranges from 0 to 30, which indicates cognitive integrity. In the healthy population aged over 60 years old, scoring 30 in the MMSE, a relevant cognitive decline has been demonstrated for reductions of 1 to 3 points [5]. However, to foster the early screening, the clinical scale should be complemented by objective assessment methods [2]–[4]. In recent years, increasing attention has been devoted to the monitoring of one's condition in the home setting through sensorized daily-use objects, able to collect data with high frequency [6]. For the MCI early screening, the quantitative analysis of the handwriting process has been proposed as a valid support tool [7]. Indeed, the brain alterations causing the disorder have a direct effect on the organizational and motor processes involved in the handwriting generation [7]. Lack of coordination, reduced smoothness and velocity were found in MCI during sentence dictation and copying [8]. In [9], subjects with MCI exhibited significantly greater time both with the pen on sheet and in air, coupled with lower applied pressure while performing copying tasks. The importance of pressure and kinematic features for distinguishing between controls and MCI subjects was furtherly confirmed in [10], where classification models sensitivity and specificity ranged from 73.9 to 100%, according to the considered task. These studies gathered the data through digitizers, which record the 2D coordinates on the pen on the screen and the applied pressure, while the subjects performed a set of standardized writing tasks under the operator' supervision. For home monitoring purposes, however, the digitizer represents a technological barrier for elders and does not allow an ecological data acquisition, hindering its introduction in such a scenario. This work aims at overcoming these limitations, proposing a novel approach for the early detection of cognitive decline in the home setting from writing activity performed using a smart ink pen [12].

Firstly, already available data, related to free handwriting tasks, were used to build classification models able to identify the presence of cognitive decline, following the classic machine

Research supported by the H2020 grant N.101016112 ESSENCE: Empathic platform to personally monitor, Stimulate, enrich, and aSsist Elders aNd Children in their Environment.

learning. paradigm (train-test split, cross validation, testing). Secondly, a longitudinal data acquisition campaign was conducted in the context of the European project ESSENCE. Users were provided with a smart ink pen to be used at home, without any supervision, for 10 weeks. A small portion of the longitudinal data was used to retrain the developed models. Then, the models were applied on the remaining data, to evaluate the user's cognitive status.

II. Materials and Methods

The handwriting data were collected with the smart ink presented in [11], which is used on paper as a normal pen, guaranteeing a more user-friendly execution with respect to the digitizer. The acceleration and angular velocity signals are recorded by a 6-axis inertial measurement unit, while the force exerted on the writing surface by a sensor connected to the pen tip. Both sensors operate with a sample frequency of 50Hz.

Participants recruitment was divided in two phases. The first phase was devoted to the collection of the data necessary for the training of the classification algorithm (this data acquisition phase is labelled "DC1" in the following). To this aim, Politecnico di Milano (Italy) and Fundación para la Formación e Investigación de los Profesionales de la Salud de Extremadura (FS, Spain) recruited participants according to the following inclusion criteria: i) being older than 60 years old; ii) showing sufficient cognitive ability, quantified by a Mini Mental State Examination (MMSE) score ≥ 24. Subjects were asked to use the smart ink pen on paper in tasks of writing production, under the supervision of an operator. They had to write content free texts and/or a grocery lists in their native language. The protocol was approved by the ethical committee of Politecnico di Milano (opinion n. 10/2018) and by the Comité de Ética de la Investigación de Badajoz (29/06/2021). In the second phase, 50 users were enrolled in Spain by FS and provided with a smart ink pen to be used at home for approximately 10 weeks. The subjects had to be older than 65 years old and be non-frail or pre-frail. The users were instructed to perform a battery of controlled tests and, in addition, to use the smart ink pen to write freely whenever they felt like, in their native language. The data acquisition of unconstrained handwriting activity was ecological and completely transparent to the subjects. The smart ink pen started the data acquisition when moved, stored the data on its on-board memory at the end of the activity and transmitted the data to a cloud server every night. Thus, despite using the pen, no user's interaction was required. The data coming from the unconstrained activity (named "DC2") were used in this study, mostly for the classification algorithms testing. The Comité de Ética de la Investigación de Badajoz (29/06/2021) approved the second data collection phase. In both recruitment phases, all participants were evaluated with the MMSE. The MMSE score was the selected outcome for the cognitive decline classification problem. According to [5], participants were divided in Controls (C), if MMSE score > 28, and at Risk (R) if MMSE score <= 28.

Given the findings related to the handwriting in MCI, the handwriting data were processed in MATLAB® 2021b to extract 36 indicators in the temporal, kinematic, pressure and smoothness domains. The indicators are presented hereafter.

Time: Number of strokes (continuous segments in which the pen is in contact with the paper) normalized by the execution time (*RelStrokeNum* [#/s]). Mean time spent on sheet (*meanOnSheet* [s]) and its coefficient of variation (*Onsheet_CV* [dimensionless]), percentage of time spent on sheet (*OnSheetRatio* [dimensionless]). Mean time spent in air (*meanInair* [s]) and its coefficient of variation (*Inair_CV* [dimensionless]). Mean time spent in air (*meanInair_nopause* [dimensionless]) and its coefficient of variation (*Inair_nopause_CV* [dimensionless]) excluding pauses (in air moments longer than 2 seconds). Ratio between mean time spent in air and mean time spent on sheet with (*AirSheet_Ratio_pause* [dimensionless]) and without (*AirSheet_Ratio_NOpause* [dimensionless]) pauses. Absolute number of pauses (*PauseNum* [#]) and number of pauses normalized by the execution time (*PauseNum_Rel* [#/s]), mean pause duration (*meanPause* [s]) and its coefficient of variation (*PauseCV* [dimensionless]).

Kinematics: Number of inversions per stroke in acceleration (*NCA_REL* [#/s]) and angular velocity (*NCG_REL* [#/s]), normalized by stroke duration. Mean of the absolute difference between consecutive peaks in angular velocity (*CPDG_AVG* [degrees/s]) and its coefficient of variation (*CPDG_CV* [dimensionless]). Pen inclination in the horizontal plane (*TILT* [degrees]), its coefficient of variation (*TILT_CV* [dimensionless]) and variance (*TILT_VAR* [(degrees(s)^2]).

Pressure: Mean pressure exerted while the pen is on sheet (*meanPonsheet* [arbitrary]), computed as the mean of all non-zero pressure values, and difference between the maximum and median exerted pressure (*P_OVS* [arbitrary]). Mean pressure computed as the mean of each stroke mean pressure (*meanP* [arbitrary]) and its coefficient of variation (*P_CV* [dimensionless]). Mean of the absolute difference between consecutive peaks in pressure (*CPDP_AVG* [arbitrary]) and its coefficient of variation (*CPDP_CV* [dimensionless]). Number of inversions per stroke pressure (*NCP_REL* [#/s]) normalized by stroke duration.

Smoothness: Logarithmic dimensionless jerk of acceleration (*LDLJ_A_median* [dimensionless]) and angular velocity (*LDLJ_A_median* [dimensionless]). Spectral arc length of angular velocity (*Sx_median* [dimensionless]), computed with 6 different thresholds for noise removal (x represent one of the possible thresholds among 10, 20, 30, 40, 45, 50%).

The computed indicators were exploited to build classification algorithms able to separate C from R. Three different algorithms were trained using Python® 3.8.10, namely Support Vector Machines (SVM), Random Forest (RF) and Catboost. Different training strategies were adopted. Two different indicator sets were considered: the first included all the 36 computed indicators (*ALL*), while the second removed the temporal features which include the pauses in their computation (*NP*): *PauseNum*, *PauseNum_Rel*, *meanPause*, *PauseCV*, *meanInair*, *Inair_nopause_CV* and *AirSheetRatio_pause*. This choice was made to investigate the contribution of pauses in the discrimination task, as one could expect a higher pause occurrence in people with reduced cognitive capabilities. Then, three different datasets sets were considered, starting from the one including DC1 only, named DS0. This dataset included all the available samples from the DC1 subjects. The remaining two sets progressively added data from DC2 to DC1: DS1) DC1 and

at maximum one sample from each DC2 subjects; DS2) DC1 and at maximum two samples from each DC2 subjects. The rationale was to simulate the workflow of a real monitoring application where, starting from a model built on already available data, the same model is updated with the newly acquired data to add information related to the subjects under examination. Given the longitudinal data availability, depicted in section III, a maximum of two sample per subjects was added, to avoid performing the model testing on too few data. Independently of the chosen algorithm, indicators and dataset, the following pipeline was adopted. The dataset was divided into 80% for the model training and 20% for its testing. During the training phase, the performances (accuracy, precision, recall, f1 score) were evaluated with a 5-fold cross-validation. To account for the bias due to the train-test split, the models were trained and evaluated on four different seeds. Thus, each combination of algorithm, indicators and dataset yielded four separate classification models. Mean performances on train and test were then retained. The models characterized by the best average performance were then applied for the classification of all the DC2 samples not included in the dataset, to test their generalization capability on completely unsupervised handwriting samples, not seen during training. This procedure was applied to evaluate a real-life scenario, where an already developed model is used on unsupervised, newly acquired data as a screening support tool. This step included only DC2 subjects who produced at least 3 handwriting samples. For a given combination, each DC2 sample was classified by the four models, the final prediction assigned by majority voting. Then, the overall prediction (i.e., C or R) for each subject was again assigned by majority voting, according to the single sample predictions for the subject, and compared with the real MMSE score. The Shapley Additive Explanation (SHAP) technique [12] was applied to the models used on the unseen data, to gain insight about the most important indicators in the prediction.

III. Results

The DC1 group included 45 subjects (age median = 75 years, age iqr = 13.5 years; MMSE median = 29, MMSE iqr = 3; 27 female (F), 18 male (M); 23 C, 22 R), while 23 subjects in DC2 produced at least one free handwriting sample (age median = 69 years, age iqr = 4.75 years; MMSE median = 29, MMSE iqr = 2; 23 F, 1 M; 15 C; 8 R), and were included in DS1 and DS2. No significant differences in MMSE (p=0.11) emerged between the groups (Mann-Whitney U test, 5% significance level). In DC2, a minimum of 3 handwriting samples was available for 15 users (age median = 68 years, age iqr = 3.75 years; MMSE median = 29, MMSE iqr = 2.75; 14 F, 1 M; 9 C, 6 R), who were considered for the model application in the real scenario. Table I summarizes the available samples for the three datasets.

The best obtained models are reported in Table II. No results for SVM are reported, as the performances were poorer. For each model, the algorithm, indicator set and dataset are shown. The metrics, highlighted in bold for the best performers, are presented as mean ± standard deviation on the four seeds, for both train and test. The RF model trained with the combination ALL + DS0 emerged as the best during training. When evaluating the test performances, the emphasis was put on the metrics standard deviation rather than the mean, as it reflects the variability associated to the different train-test seeds. In this sense, the above-mentioned model was again the best in the recall. As for the other metrics, the Catboost algorithm built on NP + DS2 showed the greatest consistency across the seeds. The two models were used to classify the 15 DC2 users in the real-life application. The resulting confusion matrices are reported in table III and IV, respectively. For the RF model, the achieved metrics were accuracy 86.67%, precision 77.78%, recall 100% and f1 score 87.50%. The relevant indicators for the RF model prediction, according to the SHAP analysis on the four seeds, mainly belonged to the temporal domain. The following indicators exhibited higher values for subjects at Risk compared to Controls: *meanInair*, *PauseNum*, *PauseNum_Rel*, and *meanPause*. On the other hand, lower values were observed in subjects at Risk for *RelStrokeNum* and *OnSheetRatio*. The only relevant kinematic indicator was *NCG_REL*, showing lower values in the R group. Lastly, *NCP_REL* in the force domain was found to be lower in subjects at Risk. The Catboost model misclassified only one subject, obtaining accuracy 93.33%, precision 88.89%, recall 100% and f1 score 94.12%. The SHAP analysis confirmed the results obtained by the RF model for the indicators which do not depend on pauses.

IV. Discussion

Objective screening tools for the early detection of MCI constitute an urgent need. Currently, this intermediate stage represents the unique window where intervention to delay dementia onset is possible. The literature established the potential validity of computerized handwriting analysis for MCI screening purposes. From these studies, a novel approach was proposed in the current work. The focus was shifted from standardized tests to unconstrained handwriting activities, to monitor the gesture with high frequency in the home scenario, rather than in the clinical setting. To this aim, the smart ink pen replaced the digitizer for data acquisition. Being used like a normal pen, it allows the ecological and transparent recording of quantitative data. Despite the different protocol, the obtained results were in line with the examined literature. The SHAP analysis revealed trends in the temporal domain which confirmed the findings in [8], [9], especially for the “in air” patterns. Indeed, at Risk subject produced not only a low number of strokes per second (i.e., they wrote slowly), but exhibited a great time spent with the pen in air. This behavior could be related to the cerebral deterioration underlying the cognitive decline, which causes impairment in both movement planning and execution [7]. Importantly, The SHAP analysis adds useful information: knowing the most relevant indicators in the model predictions allows understanding the reasons

TABLE I. Datasets composition

DS	Total Samples	Control Samples (C)	At Risk Samples (R)
DS0	108	50	58
DS1	127	60	67
DS2	142	69	73

TABLE II. PERFORMANCES OF THE BEST DEVELOPED MODELS

Algorithm	Indicators	DS	Accuracy [%]		Precision [%]		Recall [%]		f1 score [%]	
			Train	Test	Train	Test	Train	Test	Train	Test
RF	ALL	DS0	**88.13±2.98**	80.68±7.76	**91.07±4.13**	84.15±11.8	**85.63±2.39**	**81.25±4.16**	**87.86±2.84**	82.36±6.33
RF	NP	DS1	81.97±2.72	83.65±10.1	85.41±3.47	84.71±14.1	79.50±1.96	87.50±8.99	81.37±2.14	85.51±8.36
Catboost	NP	DS2	80.00±2.50	**80.77±0.00**	83.67±4.62	**77.98±3.94**	77.00±3.46	86.54±7.36	79.07±2.70	**81.75±1.31**

behind the outputs produced by the model. The results interpretability could foster the model adoption by the clinical staff. The sensitivity of the proposed approach complies with the requirements (>80%) for being used as a screening tool [4]. With respect to the traditional methods, the ecological, domestic handwriting data acquisition guarantees higher assessment frequency and time savings for the healthcare system, while assessing relevant cognitive domains for MCI, like attention and planning. The classification performances were comparable to the ones in [10] where, however, the sample size was small (17 controls and 12 MCI) and a test set was absent. Indeed, the added value of the current work is found in the training datasets and in the double testing procedure. The former contained unconstrained samples written by subjects from two different countries, allowing the model to infer the intrinsic handwriting characteristics associated to cognitive decline, regardless the product content and the employed language. This supports the robustness of obtained results. The testing phase in the real scenario, on the other hand, demonstrated the suitability of the approach for the home monitoring of handwriting. The obtained results were strongly promising, as both tested models were able to correctly identify all the subjects in the risk group. High sensitivity is indeed the desired characteristics of a screening instrument. This study had some limitations. None of the recruited subjects had a clinical diagnosis of MCI and the MMSE was the sole considered criterion for the separation between groups, while the real-life testing included a limited number of users. Lastly, the approach does not consider some cognitive domains (orientation, decision making) which are typically assessed by traditional tools. Future research should consider other assessment methods as the outcome for the classification problem, test the models on a bigger sample size and consider the development of subject-specific models. Furthermore, the relationship between handwriting indicators and MMSE score could also be studied in a regression framework. This way, it would be possible to predict the extent of cognitive decline, if any.

To sum up, this work demonstrated the feasibility of home-based, unconstrained handwriting monitoring for the detection of the first signs of cognitive decline. The approach could be a valuable screening tool, prompting a thorough clinical examination in case of classification in the risk group.

TABLE III. CONFUSION MATRIX OF MODEL RF ALL+DS0

	Prediction R	Prediction C
True R	6	0
True C	2	7

TABLE IV. CONFUSION MATRIX OF MODEL CATBOOST NP+DS2

	Prediction R	Prediction C
True R	6	0
True C	1	8

REFERENCES

[1] K. Ritchie, "Mild cognitive impairment: An epidemiological perspective," *Dialogues in Clinical Neuroscience*, vol. 6, no. 4. pp. 401–408, 2004.

[2] N. D. Anderson, "State of the science on mild cognitive impairment (MCI)," *CNS Spectrums*, vol. 24, no. 1. Cambridge University Press, pp. 78–87, Feb. 01, 2019.

[3] R. M. P. Pessoa, A. J. L. Bomfim, B. L. C. Ferreira, and M. H. N. Chagas, "Diagnostic criteria and prevalence of mild cognitive impairment in older adults living in the community: A systematic review and meta-analysis," *Revista de Psiquiatria Clinica*, vol. 46, no. 3. Universidade de Sao Paulo, pp. 72–79, May 01, 2019.

[4] L. Zhuang, Y. Yang, and J. Gao, "Cognitive assessment tools for mild cognitive impairment screening," *Journal of Neurology*, vol. 268, no. 5. Springer Science and Business Media Deutschland GmbH, pp. 1615–1622, May 01, 2021.

[5] F. Salis, D. Costaggiu, and A. Mandas, "Mini-Mental State Examination: Optimal Cut-Off Levels for Mild and Severe Cognitive Impairment," *Geriatrics (Switzerland)*, vol. 8, no. 1, Feb. 2023.

[6] A. Chkeir, J. L. Novella, M. Dramé, D. Bera, M. Collart, and J. Duchêne, "In-home physical frailty monitoring: Relevance with respect to clinical tests," *BMC Geriatr*, vol. 19, no. 1, pp. 1–9, 2019.

[7] C. De Stefano, F. Fontanella, D. Impedovo, G. Pirlo, and A. Scotto di Freca, "Handwriting analysis to support neurodegenerative diseases diagnosis: A review," *Pattern Recognit Lett*, vol. 121, pp. 37–45, 2019.

[8] J. Kawa, A. Bednorz, P. Stępień, J. Derejczyk, and M. Bugdol, "Spatial and dynamical handwriting analysis in mild cognitive impairment," *Comput Biol Med*, vol. 82, pp. 21–28, Mar. 2017.

[9] 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. 228–236, 2006.

[10] J. Garre-Olmo, M. Faúndez-Zanuy, K. López-de-Ipiña, L. Calvó-Perxas, and O. Turró-Garriga, "Kinematic and Pressure Features of Handwriting and Drawing: Preliminary Results Between Patients with Mild Cognitive Impairment, Alzheimer Disease and Healthy Controls," *Curr Alzheimer Res*, vol. 14, no. 9, Mar. 2017.

[11] F. Lunardini *et al.*, "A Smart Ink Pen for the Ecological Assessment of Age-Related Changes in Writing and Tremor Features," *IEEE Trans Instrum Meas*, vol. 70, 2021.

[12] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, 2020.