Age Group Discrimination via Free Handwriting Indicators

Eugenio Lomurno, Simone Toffoli, Davide Di Febbo, Matteo Matteucci, Francesca Lunardini, Simona Ferrante

***Abstract*— Ageing is a physiological phenomenon associated with cognitive and functional decline which, in the long term, could hamper the execution of daily life activities and threaten both social and independent life. The onset of chronic diseases can intensify this process, increasing the risk of hospitalisation and admission to long term care. This represents a significant burden on public health and reduces the quality of life of those affected. Early detection of unhealthy decline is therefore key, but the similarity to normal ageing hinders its prompt screening. This study presents a first step towards the early screening of unhealthy ageing, based on an innovative instrumented ink pen to ecologically assess handwriting performance in different age groups: 40-59 (Group 1), 60-69 (Group 2) and 70+ (Group 3) years old. Raw handwriting data were collected from 60 healthy subjects and used to extract fourteen indicators related to gesture and tremor. The indicators were then used to discriminate between subjects of different age groups in three binary classification tasks, using a selection of machine learning algorithms. This approach produced remarkable results, particularly in the task of greatest interest, identifying subjects at the very beginning of the ageing process (Group 2) from elderly subjects (Group 3), achieving an accuracy of 97.5%, an F1 score of 97.44% and a ROC-AUC of 95%. Explainability of the model, facilitated by the analysis of the Shapley values of the learned indicators, revealed age-dependent sensitivity of handwriting and tremor-related indicators. The proposed method represents a promising solution for the early detection of abnormal signs of ageing, and is designed for the remote, non-invasive, unsupervised home monitoring, to improve the care of older adults.**

***Index Terms*—Ageing, Handwriting, Ecological Home Monitoring, Smart Ink Pen, Machine Learning**

I. Introduction

Healthy ageing is a physiological phenomenon associated with a progressive structural, cognitive and functional decline. A consistent concept in the literature is that ageing is a complex and dynamic process. Complex because it involves both physical and cognitive systems, and dynamic because individuals tend to continuously progress in such a path [1]. Deterioration begins with low-level structural changes in various systems, such as a reduction in the volume of voluntary muscle tissue, asynchronous firing of motor units and slower transmission of electrical information [2]. Instead, changes in the region of the brain and in the sensory pathways alter the ability to process sensory inputs [3]. At the cognitive level, executive functions such as decision-making and planning are also affected [4]. Overall, the ability to carry out everyday activities is impaired as the whole system becomes less able to respond to internal and external stimuli [3]: maintaining posture, walking and manipulating objects become a challenge [2]. Reduced speed and accuracy in performing daily tasks has a dramatic long-term impact on quality of life, threatening both social and independent living [3].

In addition, the onset of various age-related pathological conditions (e.g. frailty or neurodegenerative disorders) can accelerate this decline, leading to an increased risk of hospitalisation and admission to long-term care, with a significant impact on public care systems [5], [6] and a poor quality of life for those directly affected and those around them [7], [8].

Early detection of unhealthy decline is therefore key to slowing and preventing its development until it reaches the severe, irreversible stage of pre-death [9].

Consequently, attention should be paid to people aged 65 years and older who are at risk of starting an unhealthy ageing process [10]. However, the scarcity of medical resources and the similarity with normal ageing often hinders a prompt screening of such a condition. To avoid this risk, an emerging solution consists of remote monitoring technologies used to continuously track the health status of community-dwelling seniors [11]. To detect early signs of decline, particular attention has been paid to the monitoring of daily activities [12]. Indeed, in older adults, any variation in the performance of daily tasks may conceal meaningful information about decline [13]. Among daily tasks, handwriting may be an optimal candidate for remote monitoring because it is a high-level skill that involves several cerebral and motor districts [14]: as a consequence, it undergoes significant variations with physiological or pathological age-related decline [15]. Indeed, the quantitative analysis of handwriting has been observed to be sensitive to several neuro-motor disorders, including Parkinson's disease [16], dystonia [17], Huntington's disease [18] and essential tremor [19].

The limitations of home-based handwriting monitoring lie in the devices currently available for data collection. Most studies in the literature have used commercially available tablets and digitising surfaces to study writing activities; however, the diffuse technological illiteracy of older adults makes their

This work was supported by the European projects MOVECARE (Grant Agreement: 732158) and ESSENCE (Grant Agreement: 101016112).

Eugenio Lomurno (corresponding author, eugenio.lomurno@polimi.it), Simone Toffoli, Davide Di Febbo, Matteo Matteucci, Francesca Lunardini and Simona Ferrante are with Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy.

everyday use rather intrusive [15], [20], [21]. Furthermore, previous research has mostly analysed handwriting in controlled settings, i.e., using a standardised writing protocol [19], although the home environment represents an uncontrolled context in which the results of standard tests cannot be assumed to be valid without supervision [22].

In a recent work of our research group [23], we presented an instrumented ink pen for the automatic acquisition and quantitative analysis of handwriting to allow the ecological home monitoring of the writing activity [24]. The tool can be used for everyday writing on paper without any further interaction by the user, therefore meeting the requirements of ecological validity. We have previously investigated the reliability of handwriting and tremor indicators in healthy subjects of different ages and demonstrated their ability to discriminate age groups in semi-uncontrolled (i.e., the acquisitions were supervised by an operator, while the content was left free to the subjects) conditions using paper-and-pen, content-free writing tasks [23].

In this work, we studied the ability of the handwriting indicators in [23] in the classification of three age groups of healthy subjects, performing two types of unconstrained writing tasks: assigning a subject to their corresponding age group through free handwriting analysis can be a powerful, ecological screening tool for the early detection of abnormalities associated with ageing [25]. Particularly for subjects in the first years of ageing (i.e., in the sixth decade of life), a potential affinity of their writing parameters with those generally observed in a category of older individuals could be a sign of a more pronounced age-related decline and be interpreted as a prompt for further investigation. The paper is structured as follows: Section II presents the instrument, experimental protocols, data processing and classification algorithms used in this work. Section III presents the results and Section IV discusses them. Finally, Section V summarises the main findings and suggests new avenues for research.

II. Methods

A. The smart ink pen

We used the smart ink pen, shown in Figure 1, developed in the European project MoveCare [12], [26], to collect handwriting data. The device consists of a commercial ink pen equipped with an inertial measurement unit (IMU) to record movement and a miniaturised load cell connected to the tip [26]. These sensors enable the acquisition of useful signals for characterising the handwriting gesture. In particular, eight time series are recorded during handwriting: timestamps, 3-axis linear acceleration, 3-axis angular velocities and the normal force applied on the pen tip. All signals are sampled at 50 Hz. Such specifications allow successfully capturing the highest frequency associated with human hand motion (12-15 Hz) while providing sufficient frequency resolution for tremor analysis (0.1 Hz with the proposed methodology). The main advantage of this device is that it acquires quantitative information while giving the typical feel of writing on paper.

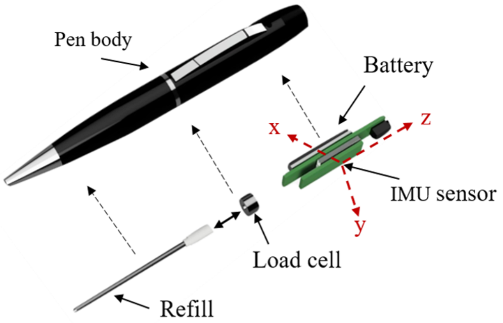

Fig. 1. A digital rendering of the smart ink pen with its internal components and the IMU reference frame orientation.

The pen is designed to automatically record signals when it is moved to write; the stored data can then be accessed via Bluetooth connection. All electronic components and the data storage mechanism are hidden from the user to ensure transparent use. This feature is particularly important when interacting with older adults who may be reluctant to use new technologies [21].

B. Participants and protocol

In order to be included in the study, participants had to be over 40 years old and healthy. Thus, any diagnosis of neurological, vascular or musculoskeletal disorders of the upper limbs was an exclusion criterion. Subjects over 65 years of age were included after verification of a Mini-Mental State Examination (MMSE) [27] score greater than 25, to ensure their cognition was preserved. During the recruitment phase, subjects were enrolled according to the following criteria: i) definition of three age groups to be studied, namely Group 1 (40-59 years old) representing adult subjects who haven't started the age decline yet, Group 2 (60-69 years old) representing subjects who are starting the age decline process and who would benefit the most from the early screening of abnormal ageing, and Group 3 (70+ years old), representing subjects for whom the ageing process is in place since several years; ii) a sample size of 20 subjects per group was defined in the protocol approved in [23] by the Ethics Committee of the Politecnico di Milano (n. 10/2018). One of the aim of the protocol was to assess the reliability of the handwriting and tremor indicators and such a sample was enough to demonstrate it in the different age groups. Thus, a sample size of 20 subjects per group was considered in the current work.

All subjects wrote a free text (*Text*, up to 7 lines) and a shopping list (*List*, up to 8 words) using the smart ink pen. The tasks had no specific constraints on the content to make them very similar to everyday writing.

C. Calculation of handwriting and tremor indicators

A set of 14 indicators related to handwriting kinematics and dynamics and to tremor were extracted from the raw data collected during each of the two writing tasks. These indicators were chosen because they have already demonstrated their statistical reliability and sensitivity in discriminating between subjects aged $<$ 40, between 60 and 69, and over 70 years old in a previous study using a similar acquisition protocol involving a reduced number of participants [23]. The calculation was implemented in Matlab® R2020b (Mathworks®, Natick, MA USA). The following indicators were calculated (for a detailed description, please refer to Appendix B):

- *Temporal handwriting measures*. Starting from the writing force signal, we divided handwriting into strokes, defined as the writing segments where the pen tip was in contact with the paper surface (non-zero force tracts). We then considered the averaged stroke duration in seconds within a writing task as the mean on-sheet time ($OnSheet$). Similarly, we kept the averaged duration of the non-writing segments (zero-force tracts) as the mean in-air time ($InAir$). The in-air time intervals longer than 2 seconds were excluded as we treated them as pauses. The ratio of the latter to the former was defined as the air-sheet time ratio ($AirSheetR$). These temporal parameters have been shown to grow with subjects' age [28].
- *Pen Tilt*. The tilt angle of the pen was calculated using the sensor fusion algorithm described in [23]. We retained the mean ($TiltMean$), coefficient of variation ($TiltCV$) and variance ($TiltVar$) of the tilt angle signal during writing (pauses excluded). We considered an angle of 90° when the pen was held in vertical position. Previous studies have also included pen tilt to characterise handwriting in different conditions [29], [30].
- *Writing Force*. Mean writing force ($Force$) was calculated by averaging the force signal over all strokes recorded during the writing task. The mean number of force changes (NCF), calculated as the average number of local maxima and minima within a stroke, was also retained as a measure of force variability. Force and force variability have been shown to change with age in handwriting [31].
- *Writing Smoothness*. We calculated the number of acceleration changes (NCA) as the average number of local minima and maxima in the 3D acceleration signal over all strokes. This quantity was observed to decrease with age [16].

To extract tremor, the 3D linear acceleration signal from the first to the last nonzero force value (i.e., from the beginning to the end of the execution) was considered. Such a signal was divided into nonoverlapping segments of 500 samples, thus allowing a frequency resolution of 0.1Hz in the spectral analysis [32]. We computed the power spectrum for each segment using the Hilbert-Huang transform (HHT) [33], which has been preferred in the literature for the study of tremor during voluntary activities over the standard Fourier transform [34], and we considered the first intrinsic mode function as the tremor component. The following tremor indicators were then calculated:

- *Tremor frequency*. We obtained the mean modal frequency ($Fmodal$) by averaging the frequencies of the highest peak in the power spectrum over all the segments [35].
- *Tremor Amplitude*. We calculated the root mean square (RMS) of the tremor signal in each segment and averaged it to retain the mean RMS.
- *Tremor entropy*. We considered the approximate entropy measure ($ApEn$), as in our previous study [23]. The entropy value (between 0 and 2) measures the unpredictability of the acceleration signal, which is influenced by the regularity of the tremor components. Entropy has been measured to decrease with age and pathology [36]. The entropy measure was computed on each segment and then averaged.
- *Nonlinear characteristics of tremor*. We applied the recurrence quantification analysis (RQA) to the 500 sample tremor segments. As in [32], we retained the recurrence ratio (RR) to measure the tendency of the tremor dynamics to express repeated patterns in time and the percentage of determinism (DET) to estimate the predictability of the oscillations during handwriting, averaging them over the number of 500 sample segments.

D. Classification tasks

From the collected data, three different datasets were created: Text and List, each including the 14 handwriting and tremor indicators, and Text + List, combining the two single datasets. We collected data from healthy participants, evenly divided into three age groups:

- Group 1: Subjects under the age of 60.
- Group 2: Subjects between 60 and 69 years old.
- Group 3: Subjects over 70 years of age.

The indicators were exploited to build algorithms able to perform binary age classifications between group pairs. Group 1 served as a baseline, consisting of adults with fully developed handwriting skills and theoretically no signs of age-related decline. Group 2 was selected for its central age of 65 years, which typically marks the beginning of the elder age. Early detection of abnormal ageing in this group would allow for personalized intervention plans to slow the decline. Group 3 represents individuals whose general conditions are already affected by the ageing process. Thus, if individuals in Group 2 exhibit handwriting characteristics similar to those of Group 3, they might be advised to undergo a comprehensive clinical evaluation to check their physical and cognitive health. Our primary interest was indeed the distinction between Group 2 and Group 3. However, testing the indicators sensitivity to the first manifestations of ageing was also valuable. Therefore, this classification challengethe classification between Group 1 and 2 was explored as well. Demonstrating that the handwriting indicators can discriminate between the more distant groups (Group 1 and Group 3) was also critical to prove that they effectively capture the essential features of the ageing process. This setup led to the definition of three binary classification scenarios across all possible group combinations. For each comparison, we labelled subjects from the younger age group as 0 and those from the older group as 1. To address the complexity and nonlinear dynamics of these analyses, we employed several machine learning classification techniques.

Our analysis began with Logistic Regression, a simple yet powerful linear classifier, which served as a crucial benchmark for performance due to its simplicity and effectiveness in binary classification tasks [37]. Support Vector Machines were included in the analysis as they excel in high-dimensional spaces by identifying the optimal hyperplane to separate different classes, providing a more refined approach to classification in scenarios with clear class boundaries [38]. We included Random Forest, an ensemble method that uses multiple decision trees to improve accuracy and robustness, effectively managing the classification task without succumbing to the overfitting common to individual decision trees [39]. AdaBoost was also added to the algorithms investigated for group age classification, due to its ability to focus on difficult-to-classify instances and its iterative adaptation to improve the accuracy of weak learners [40]. Finally, we included Catboost, which stands out as a state-of-the-art boosting algorithm not only for its adeptness in handling categorical data, but also for its superior performance and efficiency compared to other machine learning algorithms. Its gradient boosting framework also effectively mitigates overfitting, making it a leading choice for tackling a wide range of machine learning challenges [41]. Incorporating these algorithms into our analysis not only exploited their unique strengths, but also provided a comprehensive framework capable of moving from simple linear separations to complex, high-dimensional data landscapes.

E. Pipeline

The entire analysis pipeline is shown in Figure 2. The first phase called *Data Preparation* includes the collection of raw handwriting data and their elaboration, leading to the extraction of indicators. Subsequently, the phase called *Best Model Selection* aims to identify the best model for solving the task under analysis, through an iterative procedure including rigorous evaluation and extensive optimisation. In details, to ensure robustness even with a limited number of samples, the models were evaluated using the Leave One Out (LOO) cross-validation technique, with the aim of achieving an estimate with minimal bias. Each model incorporates the early stopping mechanism, with a patience set at 20 epochs with respect to the validation F1 Score. We chose this metric because it successfully captures and balances the presence of both false positive and false negative predictions. In addition, preprocessing was used to facilitate the learning process of the classifiers. It included a normalisation that adjusts all indicators to the [0,1] interval, regardless of their original value range. In this way, the models used in the analysis are not subject to initial biases associated with the value scales of individual indicators. The normalisation was applied coherently with the validation technique. To this end, at each iteration of the LOO cross-validation loop, normalisation values were computed with respect to the training set for that particular iteration. These values were then used to normalise both the training set and the validation sample. The performance thus obtained is used to search for the optimal parameters by means of a tuning process. This process, based on the Tree-structured Parzen Estimator sampling algorithm [42], allows the exploration of complex search spaces composed of the combinations of the various hyper-parameters of the different classification algorithms. For more information, please refer to Appendix A. Once the next configuration has been identified, a new model is built and evaluation via LOO cross-validation can be performed. For each model and each classification task, we let the parameters optimiser perform 50 iterations. After this step, the best model is thus identified and its weights and performance are stored. The final phase, called *Final Training and Explainability*, aims to train the model with the best parameters over the entire dataset. The model thus obtained is used to calculate the ranking of the indicators by means of explainability techniques, which are detailed in the following sub-section.

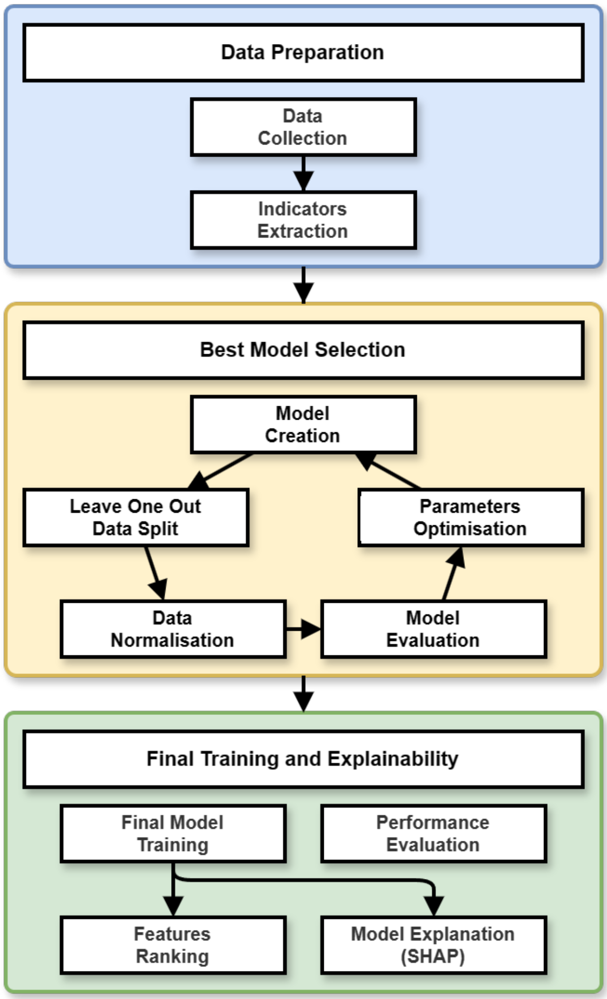

Fig. 2. The proposed experimental pipeline. Raw handwriting data are collected and processed to extract indicators. Then, they are used to build binary classification algorithms. After the best performer is found through a leave-one-out cross-validation and optimisation loop, its classification metrics are computed and the key features are investigated via SHAP analysis.

We used SHAP [43], [44], a model explanation technique based on game theory that computes the Shapley values [45] of the features according to their impact on the model predictions. In a binary classification task, SHAP first computes the baseline prediction value, i.e., the mean value predicted by the model given the observed samples, and then assigns a real number to weight each feature according to its average

TABLE I

THE RESULTS OF THE EXPERIMENTS CARRIED OUT ON THE CLASSIFICATION OF GROUP 1 AND GROUP 2. FOR EACH VARIANT OF THE DATASETS, I.E. TEXT, LIST AND TEXT + LIST, THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN LIGHT BLUE, YELLOW AND GREEN RESPECTIVELY. FINALLY, THE BEST OVERALL RESULT FOR EACH METRIC IS SHOWN IN BOLD.

Model	**Accuracy**	**Precision**	**Sensitivity**	**F1 Score**	**Specificity**	**ROC-AUC**
Logistic Regression (Text)	70.00	70.00	70.00	70.00	70.00	77.25
Support Vector Machines (Text)	82.50	78.26	90.00	83.72	75.00	76.75
Random Forest (Text)	67.50	64.00	80.00	71.11	55.00	65.25
Adaboost (Text)	70.00	72.22	65.00	68.42	75.00	70.00
Catboost (Text)	**95.00**	95.00	95.00	**95.00**	**95.00**	92.13
Logistic Regression (List)	62.50	61.90	65.00	63.41	60.00	60.25
Support Vector Machines (List)	80.00	71.43	100.0	83.33	60.00	83.13
Random Forest (List)	62.50	60.87	70.00	65.12	55.00	64.50
Adaboost (List)	65.00	63.64	70.00	66.67	60.00	64.25
Catboost (List)	**95.00**	**100.0**	90.00	94.74	**100.0**	**94.25**
Logistic Regression (Text + List)	62.50	61.90	65.00	63.41	60.00	63.75
Support Vector Machines (Text + List)	67.50	64.00	80.00	71.11	55.00	71.75
Random Forest (Text + List)	70.00	66.67	80.00	72.73	60.00	66.25
Adaboost (Text + List)	72.50	71.43	75.00	73.17	70.00	72.00
Catboost (Text + List)	**95.00**	95.00	**95.00**	**95.00**	95.00	92.75

TABLE II

THE RESULTS OF THE EXPERIMENTS CARRIED OUT ON THE CLASSIFICATION OF GROUP 2 AND GROUP 3. FOR EACH VARIANT OF THE DATASETS, I.E. TEXT, LIST AND TEXT + LIST, THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN LIGHT BLUE, YELLOW AND GREEN RESPECTIVELY. FINALLY, THE BEST OVERALL RESULT FOR EACH METRIC IS SHOWN IN BOLD.

Model	**Accuracy**	**Precision**	**Sensitivity**	**F1 Score**	**Specificity**	**ROC-AUC**
Logistic Regression (Text)	80.00	77.27	85.00	80.95	75.00	82.25
Support Vector Machines (Text)	82.50	78.26	90.00	83.72	75.00	84.25
Random Forest (Text)	77.50	73.91	85.00	79.07	70.00	83.25
Adaboost (Text)	82.50	84.21	80.00	82.05	85.00	82.50
Catboost (Text)	95.00	**100.0**	90.00	94.74	**100.0**	90.50
Logistic Regression (List)	80.00	83.33	75.00	78.95	85.00	83.25
Support Vector Machines (List)	82.50	84.21	80.00	82.05	85.00	83.50
Random Forest (List)	82.50	84.21	80.00	82.05	85.00	81.75
Adaboost (List)	80.00	80.00	80.00	80.00	80.00	73.25
Catboost (List)	**97.50**	**100.0**	**95.00**	**97.44**	**100.0**	95.00
Logistic Regression (Text + List)	75.00	72.73	80.00	76.19	70.00	83.25
Support Vector Machines (Text + List)	80.00	77.27	85.00	80.95	75.00	82.00
Random Forest (Text + List)	85.00	85.00	85.00	85.00	85.00	86.13
Adaboost (Text + List)	80.00	77.27	85.00	80.95	75.00	79.25
Catboost (Text + List)	95.00	95.00	95.00	95.00	95.00	**95.50**

TABLE III

THE RESULTS OF THE EXPERIMENTS CARRIED OUT ON THE CLASSIFICATION OF GROUP 1 AND GROUP 3. FOR EACH VARIANT OF THE DATASETS, I.E. TEXT, LIST AND TEXT + LIST, THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN LIGHT BLUE, YELLOW AND GREEN RESPECTIVELY. FINALLY, THE BEST OVERALL RESULT FOR EACH METRIC IS SHOWN IN BOLD.

Model	**Accuracy**	**Precision**	**Sensitivity**	**F1 Score**	**Specificity**	**ROC-AUC**
Logistic Regression (Text)	87.50	85.71	90.00	87.80	85.00	88.00
Support Vector Machines (Text)	90.00	90.00	90.00	90.00	90.00	88.00
Random Forest (Text)	90.00	90.00	90.00	90.00	90.00	90.75
Adaboost (Text)	87.50	85.71	90.00	87.80	85.00	89.00
Catboost (Text)	97.50	**100.0**	95.00	97.44	**100.0**	96.00
Logistic Regression (List)	90.00	90.00	90.00	90.00	90.00	95.25
Support Vector Machines (List)	85.00	88.89	80.00	84.21	90.00	88.50
Random Forest (List)	82.50	80.95	85.00	82.93	80.00	89.25
Adaboost (List)	82.50	84.21	80.00	82.05	85.00	83.13
Catboost (List)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Logistic Regression (Text + List)	85.00	85.00	85.00	85.00	85.00	90.75
Support Vector Machines (Text + List)	90.00	100.0	80.00	88.89	100.0	93.50
Random Forest (Text + List)	90.00	90.00	90.00	90.00	90.00	94.25
Adaboost (Text + List)	85.00	85.00	85.00	85.00	85.00	92.00
Catboost (Text + List)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

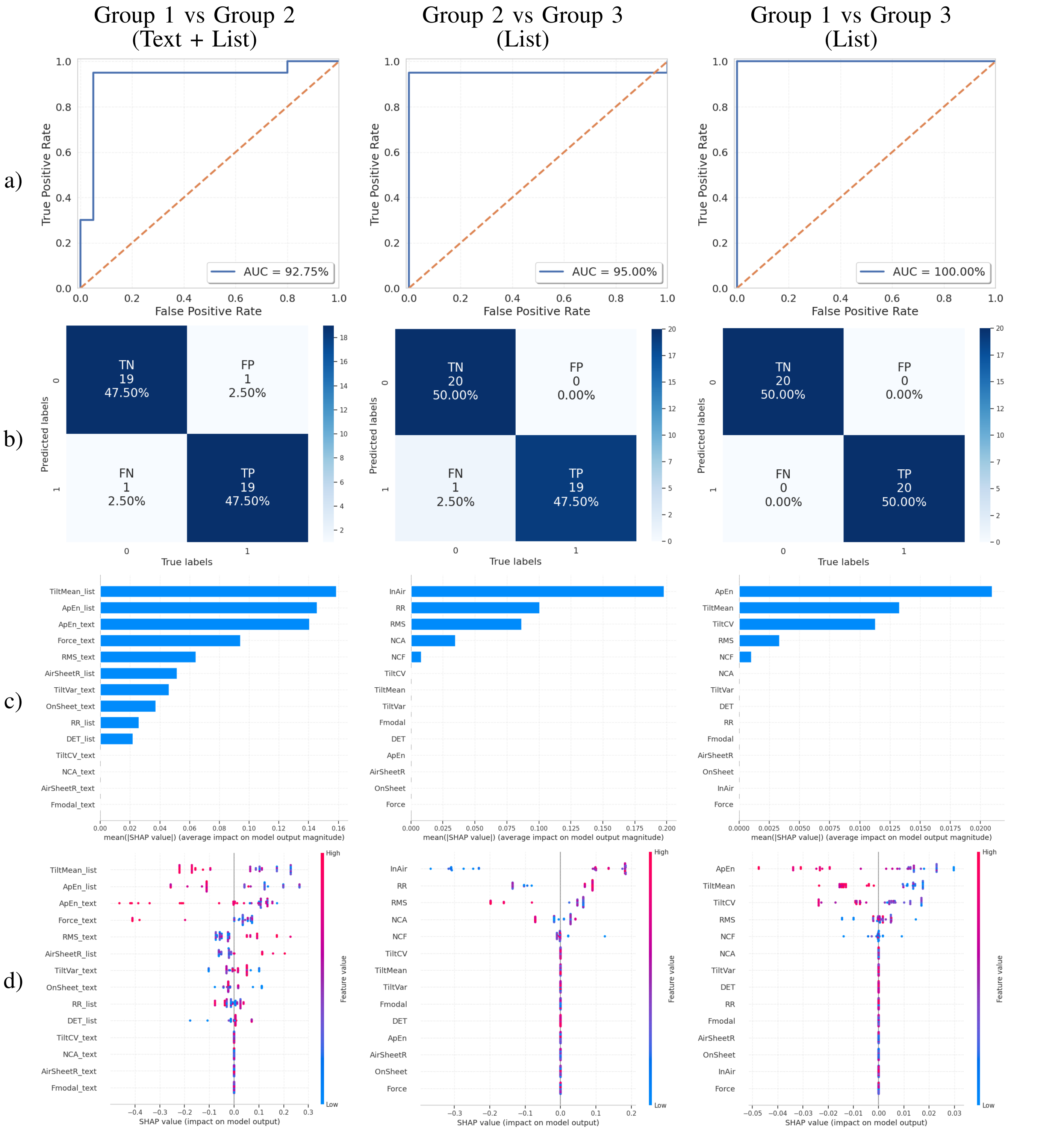

Fig. 3. Classification performance and model explanation plots for the Group 1 vs Group 2, Group 2 vs Group 3 and Group 1 vs Group 3 tasks. The ROC curves of the best models are shown in row a); the corresponding confusion matrices are shown in row b); rows c) and d) show the absolute average Shapley values and the Shapley values of the indicators for each sample, respectively.

contribution in feature coalitions, i.e., its Shapley value. It is then possible to explore the role of each feature in the classification of individual samples, independent of the fact that the model has learned them during the training step. The sample prediction represents the sum of the feature contribution starting from the baseline. If a feature has a positive influence, it influences the prediction in favour of class 1 and vice versa. This step was useful to understand, for each sample and age group, how much each indicator leads the model to predict class 0 (younger group) or 1 (older group).

III. Results

In this study, 60 healthy participants were recruited, evenly divided into three age groups: Group 1, comprising 12 males and 8 females with an average age of 57.70 ± 6.28 years; Group 2, including 10 males and 10 females with a mean age of 65.45 ± 2.20 years; Group 3, 6 males and 14 females with an average age of 80.2 ± 7.00 years. Thus, in each classification task, the List and Text datasets were made up of 20 samples per group. The Text + List dataset included 40 samples per group. Subjects were compliant with the instructions they were provided with in terms of numbers of written lines in the text (Group 1 = 7 ± 0, Group 2 = 6.80 ± 1.08, Group 3 = 6.43 ± 0.79) and items in written in the list task (Group 1 = 8 ± 0, Group 2 = 7.87 ± 0.35, Group 3 = 7.86 ± 0.38). In the list task, the global trend was that of writing a single word per item (Group 1 = 1.01 ± 0.04, Group 2 = 1.06 ± 0.19 , Group 3 = 1 ± 0). The adopted writing style was either uppercase or cursive. The cursive allograph was adopted in 13 (65%), 16 (80%) and 15 (75%) of the cases in the text task by Group 1, Group 2 and Group 3, respectively. As for the list, the occurrence of cursive was 5 (25%) for Group 1, 12 (60%) for Group 2 and 13 (65%) for Group 3.

The performance metrics of each classification model employed for Group 1 vs Group 2, Group 2 vs Group 3, and Group 1 vs Group 3 are detailed in Table I, Table II, and Table III, respectively.

Within each table, accuracy, precision, sensitivity, F1 score, specificity and ROC-AUC are reported for the best model of each classification algorithm according to the evaluation metrics (F1 score), considering each dataset. Figure 3 details the overall best performing model for each classification task, which always resulted based on the Catboost algorithm, regardless of the employed dataset. The first column shows the results for the best model obtained in the Group 1 vs Group 2 classification task, while the second and third columns show the best models for the Group 2 vs Group 3 and Group 1 vs Group 3 tasks, respectively. Row (a) shows plots comparing the True Positive Rate and the False Positive Rate, with the results calculated as ROC-AUC. Row (b) shows the associated confusion matrices to give another perspective and a deeper understanding of the predictive ability of the models with respect to unseen data. Row (c) shows the indicator ranking of the final models trained on the full available dataset and tuned via LOO cross-validation. This ranking is designed to show only the magnitude of the indicator influence, which is therefore shown in absolute value. Finally, row (d) expands the results of the previous row, by explaining how the learned samples were predicted based on their Shapley value. Specifically, each point represents the Shapley value of the indicator for a given sample. The blue-red colour scale indicates the value of the indicator (low to high): negative Shapley values pushed the prediction towards class 0 (the younger group), while positive values favoured the classification of the subject in class 1 (the older group).

IV. Discussion

In this paper we have demonstrated the capability of the quantitative analysis of handwriting to discriminate between healthy subjects of different age groups. We used a novel smart ink pen to collect handwriting data during tasks that mimicked everyday writing. In fact, participants were asked to write a short free text and a shopping list without any constraints on content or writing modality. This particular setting was chosen to maximise ecological validity, with the ultimate goal of using our findings to develop, in the future, home-based solutions dedicated to the early detection of unhealthy decline in seniors. Therefore, particular attention was paid to the classification of the individual's age: in the real use of the developed classifiers, the association of one's handwriting characteristics with those of an older age group could be interpreted as a clinically relevant anomaly [46].

Five different machine learning algorithms, namely Logistic Regression, Support Vector Machine, Random Forest, Adaboost and Catboost, were used to carry on three different binary classification tasks, based on the set of handwriting indicators we computed from the raw free text and shopping list data. The indicators described the execution from different points of view, including the temporal characteristics, the pen inclination in the horizontal plane, the force exerted on the writing surface, and the movement smoothness. Tremor was analysed as well, considering its amplitude and frequency properties, together with its regularity and nonlinear characteristics.

In the analysis, we considered all the possible binary classification problems (i.e., Group 1 vs Group 2, Group 2 vs Group 3 and Group 1 vs Group 3). The aim of the first two tasks was to test the sensitivity of the models to the variations in handwriting performance that might be expected between healthy individuals with small age differences [47]. Excellent performance was obtained with the Catboost classifier in both tasks (accuracy between 95.00% and 97.50%, precision from 95.00% to 100%, recall from 90.00% to 95.00%, F1 score from 94.74% to 97.44%, specificity from 95.00% to 100% and ROC-AUC from 90.50% to 95.50%), considering all three datasets composed by the indicators calculated from the Text, List and combined Text + List data. The Catboost model outperformed the other algorithms in the classifications between adjacent age groups, in all considered datasets. Indeed, the performance obtained by the other models ranged from 62.50% to 85.00% accuracy, 60.87% to 85.00% precision, 65.00% to 100% sensitivity, 63.41% to 85.00% F1 score, 55.00% to 85.00% specificity and 60.25% to 86.13% ROC-AUC.. Therefore, we could expect a high sensitivity to the changes in handwriting indicators due to an abnormal or pathological ageing decline [47].

In the third task, we looked instead at the classification between nonadjacent age groups. As expected, scores were higher in this task, as the differences in handwriting were likely more pronounced. Once again, the Catboost classifier produced the best overall result, achieving 100% in all metrics when considering the List and the Text+List dataset. This result, that we want to stress is obtained using the Leave One Out cross-validation method, suggests, on the one hand, the existence of non-linear separating hyperplanes that are able to correctly classify individuals into two different groups. On the other hand, it shows that the information provided by the indicators is sufficient for the Catboost algorithm to effectively approximate these separating hyperplanes. As the two age groups are more distant from each other, and therefore have more different writing features, such behaviour was expected, or at least conceivable. When the Text dataset was used, all the metrics were above 95.00%, with precision and specificity equal to 100%. Interestingly, the differences between Catboost classifiers and the other models were less pronounced with respect to the first two classification tasks. Considering all models and all datasets, all metrics resulted greater than 80.00%. Overall, these findings suggest that the extracted handwriting indicators are able to successfully capture the intrinsic and evident differences in the gesture between adult and elderly healthy individuals, thus making the classification problem almost trivial.

When looking at the discriminative power of the different datasets for the Catboost model, the List one turned out to be the best overall. For the classification between Groups 1 and 2, the List dataset was the one on which the model performed best in terms of precision, recall, and ROC-AUC. The trend was even more evident in the classification between Group 2 and 3, where the List obtained the best results in all metrics but ROC-AUC, which however reached 95.00%. Lastly, List and Text + List did not make any misclassifications when examining Group 1 vs Group 3, as already mentioned. As a whole, we would have predicted a greater discriminative power associated with the Text dataset. Indeed, writing a free text requires a higher cognitive demand, which should have a greater impact on the handwriting characteristics as age increases. Firstly, the instructions per se made the text exercise longer than the list. Secondly, both the semantic and the orthographic content are more complex in a text: the written product should convey a meaningful message (semantic), which requires the sapient combination of different grammatical components (nouns, verbs, articles and prepositions). In a shopping list, semantic remains important but is restricted to a narrower range, while the written items are likely to be nouns in most of the cases. A possible reason behind the obtained results could lie in the writing dynamics of the two exercises. Indeed, each item in the list was written in a new line, thus making the execution more segmented with respect to the text, with possible greater effects on older subjects' execution. The adopted allograph could also have had an impact, particularly in the classification tasks involving Group 1: in writing the list, these subjects preferred the uppercase style, while cursive was predominant in the other two groups. Given the relatively small number of samples, chance factors could not be excluded and further research on this topic is warranted. Nevertheless, it is worth noting that the differences in the results were small, suggesting that both data collection methods are valid and contain intrinsic age-related information.

We further analysed our experiments using the SHAP model explanation technique to understand the impact and the behavior of each handwriting indicator in the different classification tasks. In this paper we detail the results and analysis of three classifications tasks, always considering Catboost based models. In particular, the dataset showing the best F1 score is represented, as it was the optimization metric during the algorithm training.

In the first classification task, the groups aged 40 to 59 and 60 to 69 years old (Group 1 and Group 2) were considered. These two groups represent, respectively, a population of healthy subjects in which the effects of age decline should be absent, and a population in which a decline in physical or cognitive functionality may be at an early stage [10], [48]. As shown in Figure 3 first column, row (b), the Catboost model trained on the Text+List dataset was able to correctly classify 38 subjects over 40, with only 1 misclassified subject per group. Our results confirm previous findings in the literature, where it has been observed that handwriting varies significantly in middle and older adults [49]. The model explanation (Figure 3, lines c and d) showed that the tilt of the pen ($TiltMean\_list$) was the most important indicator in this classification with Group 2 demonstrating lower values (i.e., the pen was held more horizontally). According to Marzinotto et al. [49], a greater pen inclination is typical in middle-aged adults (Group 2). The approximate entropy ($ApEn\_list$ and $ApEn\_text$) also played a significant role, indicating a lower predictability of the handwriting time series of the younger class. This result is in line with the findings of our previous work [23], where, using similar experimental settings, significant differences between age groups were found. The trend of decreasing entropy with age was consistent with previous literature studying resting and postural tremor in younger and older adults [35], [36], [50]. As for the other nonlinear characteristics of tremor, higher values of the percentage of determinism (DET_list) were correctly associated with older ages [32]. The more predictable tremor patterns in Group 2 were also coupled with increased tremor amplitude (RMS_text). Although its variation was not statistically significant between different age groups in Lunardini et al. [26], in the current study writing force ($Force\_text$) emerged as the fourth most predictive indicator. The predictions were shifted towards the older group (Group 2) when the force values were lower. This was in line with the study by Engel Yeger et al. [3] in 2012, Caligiuri [51] in 2014 and Marzinotto et al. [49] in 2016.

The effect of the temporal indicators was revealed only by 2 indicators ($AirSheetR\_list$ and $OnSheet\_text$). The former confirmed the tendency of the older class to have more prolonged non-writing moments with respect to the on-sheet time, found in our previous work [23], and others [28],

[52]. As a whole, these results seem to suggest that the first signs of ageing are subtle and hidden in characteristics which are difficult to detect by the naked eye [1]. These insights were also useful to understand why the two misclassifications happened. The false positive (i.e., a subject belonging to Group 1 and classified as Group 2), exhibited low values in $TiltMean\_list$ and $Force\_text$ indicators, which are associated by the model with the older age group. On the other hand, the false negative subject showed high values of approximate entropy (both in List and Text dataset) and reduced tremor amplitude (RMS_text), thus being separated from Group 2 trend.

The second classification, between the groups aged 60 to 69 and 70+ years old (Groups 2 and 3), was the most relevant for investigating the suitability of our approach in the scenario of early detection of decline. In the normal ageing process, physical or cognitive decline is expected to be more consistent in the older group [10], [48]. Therefore, whenever an individual in the younger group is associated with the older one, it could be interpreted as a sign of abnormal decline. Our results showed that the classifier trained on the List dataset may be suitable for the monitoring of decline due to its high F1 score of 97.44%. Only 1 subject out of 40 was wrongly classified as younger, while the false negatives were 0 (Figure 3, second column, row (b)). The model explanation (Figure 3, second column, row (c) and (d)) showed that the in-air time parameter ($InAir$) was much more influential in the classification than all the others (twice as much as the second most important indicator in absolute value): higher $InAir$ were associated with individuals of the older class, highlighting a slower execution, as in [4]. Such an outcome has a two-fold interpretation. The oldest subjects can be facing decline both from the motor output perspective, slowing down movement execution, and the cognitive perspective, affecting the high-level executive functions associated with memory, motor planning and attention. Recurrence rate was the second indicator of importance and confirmed the trend of increasing tremor regularity with age emerged in the first classification task. On the other hand, RMS exhibited an opposite behaviour, with lower values associated with older participants. However, the reduced amplitude of the oscillations could also be indicative of decreased movement speed in general. Lastly, gesture fluency, measured by NCA, did not reveal a clear trend, as Group 2 subjects were evenly split between very high and very low values. Thus, in this classification task, the SHAP analysis depicts older subjects' handwriting performance as inefficient in terms of speed and standardised in terms of tremor manifestation. This was not true for the misclassified subject, a 73 year old man whose List performance was characterised by reduced time spent with the pen in the air ($InAir$), stochastic (RR) and strong (RMS) involuntary oscillations.

The third classification was between Group 1 and Group 3, with handwriting performance expected to be markedly different. As a consequence, the ability of the model in discriminating between these classes increased. When training the Catboost algorithm with the List dataset, all the subjects were correctly classified. The model explanation (Figure 3, third column, rows (c) and (d)) showed that the two most important indicators were the same, although reversed in order, found in the first classification task (Group 1 vs Group 2). Indeed, tremor approximate entropy ($ApEn$) revealed the highest differentiation power, with clearly distinguishable values between the groups: high for the youngest and low for the oldest. The angle of the pen with respect to the normal to the writing surface ($TiltMean$) dropped from the first position in the classification Group 1 vs Group 2 to the second position in the third classification in the impact ranking, while still keeping the same behaviour, revealing a higher degree of inclination in older subjects. Additionally, Group 3 was characterised by a lower variability in pen angle ($TiltCV$), possibly indicating a rigid manipulation of the writing tool. Besides these three indicators, only RMS and NCF had a meaningful impact on the model prediction. No clear trends were revealed, although it is worth noting that their contribution was one order of magnitude lower compared to the first three indicators. These results confirmed the findings of the classification between Group 1 and 2: with respect to people undergoing age-related modifications, either in early or advanced phases, adult subjects exhibit more stochastic tremor patterns [35] and hold the pen more vertically while writing [49].

Interestingly, the combination of List and Text datasets in the classification between Group 1 and 2 revealed ten relevant indicators for the model prediction, equally distributed between the two handwriting exercises. Among these, only the approximate entropy was present for both datasets. On the other hand, when considering the List alone (second and third classification tasks in Figure 3), only five indicators happened to have a marked impact on the model reasoning. This finding suggests the utility of both List and Text for age characterization: when considered together, they can enhance the understanding of age-related handwriting differences, providing non overlapped details on the various inspected domains.

The model explanation revealed that the impact of the handwriting indicators was task dependent, i.e., it changed according to the age ranges we considered in the classifications. These differences in the indicators importance highlighted the complexity of the age-driven decline in handwriting. However, the behaviour of the indicators in the different age intervals was consistent with the previous findings in literature in populations of healthy subjects. This result helped in the interpretability of the models, giving the possibility to understand their decisions, as they relied on known handwriting quantities. Such an aspect is critical for the potential adoption of the proposed models in real practice and would allow to understand the domain where handwriting impairments due to ageing occur, if any.

It is possible to point out certain limitations of the present work. The limited number of participants recruited in the study implied the use of Leave One Out cross-validation without the addition of a test set of sufficient size to provide a consistent estimate. This provided optimal validation results, but an additional dataset could have been used to assess further aspects of the models' performance. No distinctions were made in terms of the allograph used. Future research should deepen this topic by providing more specific insights into possible allograph-dependent writing disorders that occur with ageing.

V. Conclusions

In conclusion, this work showed the quantitative analysis of handwriting to classify healthy individuals belonging to different age groups. The developed classifiers may offer a novel and non-invasive instrument for the domestic monitoring of handwriting in elderly individuals. The interest in our findings is enhanced by the innovative tool employed to collect the subject's writing data, allowing the ecological acquisition of daily-life handwriting.

The results support the use of handwriting to detect age-related anomalies: an abnormal decline could be highlighted in individuals classified as belonging to an older age group, thus prompting a thorough clinical investigation of their conditions by the general practitioner or a specialist. Moreover, precise information about the nature of the decline could be achieved by investigating pathological handwriting changes, as Parkinson's disease and dementia, in diagnosed patients and developing illness-specific classifiers.

References

[1] D. E. Vaillancourt and K. M. Newell, "Changing complexity in human behavior and physiology through aging and disease," *Neurobiology of aging*, vol. 23, no. 1, pp. 1–11, 2002.

[2] S. Morrison, K. Newell *et al.*, "Aging, neuromuscular decline, and the change in physiological and behavioral complexity of upper-limb movement dynamics," *Journal of aging research*, vol. 2012, 2012.

[3] B. Engel-Yeger, S. Hus, and S. Rosenblum, "Age effects on sensory-processing abilities and their impact on handwriting," *Canadian journal of occupational therapy. Revue canadienne d'ergothérapie*, vol. 79, pp. 264–74, 12 2012.

[4] S. Rosenblum, B. Engel-Yeger, and Y. Fogel, "Age-related changes in executive control and their relationships with activity performance in handwriting," *Human movement science*, vol. 32, no. 2, pp. 363–376, 2013.

[5] M. C. Polidori, S. Maggi, F. Mattace Raso, and A. Pilotto, "The unavoidable costs of frailty: a geriatric perspective in the time of covid-19," *Geriatric Care*, vol. 6, 03 2020.

[6] K. Ensrud, A. Kats, J. Schousboe, B. Taylor, P. Cawthon, T. Hillier, K. Yaffe, S. Cummings, J. Cauley, L. Langsetmo, and S. Fractures, "Frailty phenotype and healthcare costs and utilization in older women," *Journal of the American Geriatrics Society*, vol. 66, 03 2018.

[7] G. Kojima, S. Iliffe, S. Jivraj, and K. Walters, "Association between frailty and quality of life among community-dwelling older people: A systematic review and meta-analysis," *Journal of Epidemiology and Community Health*, vol. 70, pp. jech–2015, 01 2016.

[8] E. Hoogendijk, B. Suanet, E. Dent, D. Deeg, and M. Aartsen, "Adverse effects of frailty on social functioning in older adults: Results from the longitudinal aging study amsterdam," *Maturitas*, 09 2015.

[9] M. Puts, S. Toubasi, M. Andrew, M. Ashe, J. Ploeg, E. Atkinson, A. P. Ayala, A. Roy, M. Rodriguez-Monforte, H. Bergman, K. McGilton, and K. McGilton, "Interventions to prevent or reduce the level of frailty in community-dwelling older adults: A scoping review of the literature and international policies," *Age and ageing*, vol. 46, 01 2017.

[10] C. Trevisan, N. Veronese, S. Maggi, G. Baggio, E. Toffanello, S. Zambon, L. Sartori, E. Musacchio, E. Perissinotto, G. Crepaldi, E. Manzato, and G. Sergi, "Factors influencing transitions between frailty states in elderly adults: The progetto veneto anziani longitudinal study," *Journal of the American Geriatrics Society*, vol. 65, 11 2016.

[11] 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 Geriatrics*, vol. 19, 02 2019.

[12] F. Lunardini, M. Luperto, M. Romeo, J. Renoux, N. Basilico, A. Krpič, N. A. Borghese, and S. Ferrante, "The movecare project: home-based monitoring of frailty," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[13] A. Buchman, R. Wilson, L. yu, B. James, P. Boyle, and D. Bennett, "Total daily activity declines more rapidly with increasing age in older adults," *Archives of gerontology and geriatrics*, vol. 58, 08 2013.

[14] R. Plamondon, C. O'Reilly, C. Remi, and T. Duval, "The lognormal handwriter: Learning, performing, and declining," *Frontiers in psychology*, p. 945, 12 2013.

[15] S. Rosenblum, P. Werner, T. Dekel, I. Gurevitz, and J. Heinik, "Handwriting process variables among elderly people with mild major depressive disorder: A preliminary study," *Aging clinical and experimental research*, vol. 22, pp. 141–7, 04 2010.

[16] J. Walton, "Handwriting changes due to aging and parkinson's syndrome," *Forensic science international*, vol. 88, pp. 197–214, 09 1997.

[17] K. Zeuner, M. Peller, A. Knutzen, I. Holler, A. Münchau, M. Hallett, G. Deuschl, and H. Siebner, "How to assess motor impairment in writer's cramp," *Movement disorders : official journal of the Movement Disorder Society*, vol. 22, pp. 1102–9, 06 2007.

[18] M. Caligiuri, C. Snell, S. Park, and J. Corey-Bloom, "Handwriting movement abnormalities in symptomatic and premanifest huntington's disease," *Movement Disorders Clinical Practice*, vol. 6, 08 2019.

[19] J. Alty, J. Cosgrove, D. Thorpe, and P. Kempster, "How to use pen and paper tasks to aid tremor diagnosis in the clinic," *Practical Neurology*, vol. 17, pp. practneurol–2017, 08 2017.

[20] R. Camicioli, S. Mizrahi, J. Spagnoli, C. Büla, J. Demonet, F. Vingerhoets, A. Gunten, and B. Santos-Eggimann, "Handwriting and pre-frailty in the lausanne cohort 65+ (lc65+) study," *Archives of Gerontology and Geriatrics*, vol. 61, 01 2015.

[21] E. Brooks, C. Turvey, and E. Augusterfer, "Provider barriers to telemental health: Obstacles overcome, obstacles remaining," *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, vol. 19, 04 2013.

[22] M. Schmuckler, "What is ecological validity? a dimensional analysis," *Infancy*, vol. 2, 10 2001.

[23] F. Lunardini, D. Di Febbo, and et al, "A smart ink pen for the ecological assessment of age-related changes in writing and tremor features," *IEEE Transactions on Instrumentation and Measurement*, 2020.

[24] S. Toffoli, E. Lomurno, F. Lunardini, G. Carmen, C. Pilar, F. Stefania, M. Malavolti, M. Matteucci, S. Ferrante *et al.*, "Activity and age classification from handwritten samples acquired with a smart ink pen," in *BHI 2022, International Conference on Biomedical and Health Informatics*, 2022.

[25] B. Jin, T. Thu, E. Baek, S. Sakong, J. Xiao, T. Mondal, and M. Deen, "Walking-age analyzer for healthcare applications," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, p. 1034, 01 2014.

[26] D. Di Febbo, F. Lunardini, M. Malavolti, A. Pedrocchi, N. A. Borghese, and S. Ferrante, "Iot ink pen for ecological monitoring of daily life handwriting," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5749–5752.

[27] A. Larner, "Mini-mental state examination: diagnostic test accuracy study in primary care referrals," *Neurodegenerative Disease Management*, vol. 8, 09 2018.

[28] S. Rosenblum, B. Engel-Yeger, and Y. Fogel, "Reprint of 'age-related changes in executive control and their relationships with activity performance in handwriting'," *Human movement science*, pp. 1056–69, 10 2013.

[29] T. Asselborn, P. Dillenbourg, and M. Chapatte, "Extending the spectrum of dysgraphia: A data driven strategy to estimate handwriting quality," *Scientific Reports*, vol. 10, 02 2020.

[30] J. Garre-Olmo, M. Faundez-Zanuy, K. Lopez-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," *Current Alzheimer research*, vol. 14, 05 2016.

[31] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, and M. Faundez-Zanuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease," *Artificial intelligence in medicine*, vol. 67, 01 2016.

[32] A. Meigal, S. Rissanen, M. Tarvainen, S. Georgiadis, P. Karjalainen, O. Airaksinen, and M. Kankaanpää, "Linear and nonlinear tremor acceleration characteristics in patients with parkinson's disease," *Physiological measurement*, vol. 33, pp. 395–412, 03 2012.

[33] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C.-C. Tung, and H. Liu, "The empirical mode decomposition and the hilbert

spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. A*, vol. 454, pp. 679–699, 01 1998.

[34] J. Z. Zhang, B. T. Price, R. D. Adams, K. Burbank, and T. J. Knaga, "Detection of involuntary human hand motions using empirical mode decomposition and hilbert-huang transform," in *2008 51st Midwest Symposium on Circuits and Systems*. IEEE, 2008, pp. 157–160.

[35] S. L. Hong, E. James, and K. Newell, "Coupling and irregularity in the aging motor system: Tremor and movement," *Neuroscience letters*, vol. 433, pp. 119–24, 04 2008.

[36] D. Vaillancourt, A. Slifkin, and K. Newell, "Regularity of force tremor in parkinson's disease," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 112, pp. 1594–603, 10 2001.

[37] J. Berkson, "Application of the logistic function to bio-assay," *Journal of the American statistical association*, vol. 39, no. 227, pp. 357–365, 1944.

[38] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[39] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[40] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[41] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[42] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[43] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[44] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[45] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[46] G. Marzinotto, J. Rosales, M. El Yacoubi, S. Garcia-Salicetti, C. Kahindo, H. Kerhervé, V. Cristancho-Lacroix, and A.-S. Rigaud, "Uncovering major age-related handwriting changes by unsupervised learning," vol. GNB2020 - Seventh National Congress of Bioengineering - Proceedings 2020, 01 2016.

[47] P. Spaan, J. Raaijmakers, and C. Jonker, "Alzheimer's disease versus normal ageing: A review of the efficiency of clinical and experimental memory measures," *Journal of clinical and experimental neuropsychology*, vol. 25, pp. 216–33, 05 2003.

[48] S. Gale, D. Acar, and K. Daffner, "Dementia," *The American Journal of Medicine*, vol. 131, 02 2018.

[49] G. Marzinotto, J. Rosales, M. El Yacoubi, S. Garcia-Salicetti, C. Kahindo, H. Kerhervé, V. Cristancho-Lacroix, and A.-S. Rigaud, "Age-related evolution patterns in online handwriting," *Computational and Mathematical Methods in Medicine*, vol. 2016, 08 2016.

[50] M. Sturman, D. Vaillancourt, and D. Corcos, *Journal of neurophysiology*, 2005.

[51] M. Caligiuri, C. Kim, and K. Landy, "Kinematics of signature writing in healthy aging," *Journal of forensic sciences*, vol. 59, 02 2014.

[52] S. Rosenblum, P. Werner, T. Dekel, I. Gurevitz, and J. Heinik, "Handwriting process variables among elderly people with mild major depressive disorder: A preliminary study," *Aging clinical and experimental research*, vol. 22, pp. 141–7, 04 2010.

[53] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[54] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018.

Appendix A

This section provides an overview of the search spaces used to tune the machine learning models used in the experimental part of this study. All values are summarised in Table IV. The optimisation framework used is Optuna [53], which incorporates a Tree-structured Parzen Estimator sampling method [42] alongside a Hyperband pruning mechanism [54]. For each optimised model, a series of 50 search iterations were performed with the objective of maximising F1 score in order to identify its optimal parameters.

TABLE IV
SEARCH SPACES FOR THE PROPOSED ALGORITHMS

Algorithm	**Search Space**
Logistic Regression	C: loguniform(1e-3, 1e3) Penalty: ['l1', 'l2', 'none'] Solver: ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'] Max Iter: 1000
Support Vector Machines	C: loguniform(1e-3, 1e3) Kernel: ['linear', 'poly', 'rbf', 'sigmoid'] Max Iter: 1000 Probability: true Degree: integer(2, 5) Gamma: loguniform(1e-4, 1e1) Coef0: uniform(-1, 1)
Random Forest	N estimators: integer(10, 200) Max Depth: integer(2, 8) Max Features: ['sqrt', 'log2', None] Criterion: ['gini', 'entropy']
AdaBoost	N estimators: integer(10, 200) Learning Rate: float(0.01, 1.0) Max Depth: integer(2, 8) Max Features: ['sqrt', 'log2', None] Algorithm: ['SAMME', 'SAMME.R'] Criterion: ['gini', 'entropy']
CatBoost	Iterations: 1000 Depth: integer(2, 8) Learning Rate: float(0.01, 0.3) L2 Leaf Reg: loguniform(1, 10) Bagging Temperature: float(0, 10) Border Count: integer(128, 256) Loss Function: 'CrossEntropy' Evaluation Metric: 'F1' Bootstrap Type: 'Bayesian'

Appendix B

This section explains how the indicators used in the paper are computed in MATLAB R2020b. We define:

- f_s the sampling frequency of the smart ink pen IMU, load cell and timestamp;
- t the timestamp time series measured in seconds;
- F the force signal time series measured in arbitrary unit;
- A_x, A_y and A_z $[\text{mm}/s^2]$ the linear acceleration time series along the x, y and z axis respectively;
- AV_x, AV_y and AV_z $[\text{deg}/s]$ the angular velocity time series along the x, y and z axis respectively;

Stroke Segmentation

Strokes are defined as segments where the pen tip is in contact with the paper. The single acquisition is characterized by an arbitrary number of strokes (defined as N_{strokes}) depending on the written content and the subject's personal handwriting style. The instant of beginning and end of each stroke were identified as follows:

$$t_{\text{start}} = t_i | F_{t_{i-1}} = 0 \wedge F_{t_i} > 0;$$
$$t_{\text{stop}} = t_i | F_{t_i} > 0 \wedge F_{t_{i+1}} = 0$$

The signals associated to the i-th stroke are defined as:

$$t^{\text{i-th}}_{\text{stroke}} = t(t^{\text{i-th}}_{\text{start}}, t^{\text{i-th}}_{\text{stop}});$$
$$F^{\text{i-th}}_{\text{stroke}} = F(t^{\text{i-th}}_{\text{start}}, t^{\text{i-th}}_{\text{stop}});$$
$$A^{\text{i-th}}_{xyz\ \text{stroke}} = A_{xyz}\ (t^{\text{i-th}}_{\text{start}}, t^{\text{i-th}}_{\text{stop}});$$
$$AV^{\text{i-th}}_{xyz\ \text{stroke}} = AV_{xyz}\ (t^{\text{i-th}}_{\text{start}}, t^{\text{i-th}}_{\text{stop}});$$

Temporal handwriting measures

Temporal handwriting indicators are computed as:

$$\text{OnSheet (s)} = \frac{\sum_{i=1}^{N_{\text{strokes}}} (t^i_{\text{stop}} - t^i_{\text{start}})}{N_{\text{strokes}}};$$
$$\text{InAir [s]} = \frac{\sum_{i=1}^{N_{\text{strokes}}-1} (t^{i+1}_{\text{start}} - t^i_{\text{stop}})}{N_{\text{strokes}}};$$
$$\text{AirSheetR [ ]} = \frac{\text{InAir}}{\text{OnSheet}};$$

Pen Tilt

For Tilt computation the following steps were followed for each stroke:

- Az (the linear acceleration component directed as the longitudinal axis of the pen) was low pass filtered with cutoff frequency set to 10Hz.
- A first approximation was computed as: $\theta_{\text{approx}} = \sin^{-1}\left(\frac{Az^i_{\text{stroke}}}{9.81}\right)$
- Defined constant $k_1 = 1.5$ and $k_2 = 0.4$ the Tilt was updated using the angular velocity tilt estimation as:

$$\theta(t) = \theta(t-1) + k_1\left(\theta_{\text{approx}}(t-1) - \theta(t-1)\right) + k_2\left(AVx^i_{\text{stroke}}(t-1) + AVy^i_{\text{stroke}}(t-1)\right)$$

- Finally the following Tilt features were extracted:

$$\text{TiltMean[deg]} = \frac{\sum_{i=1}^{i=N_{\text{strokes}}} \theta_i}{N_{\text{strokes}}};$$
$$\text{TiltCV[]} = \frac{\text{std}(\theta_i)}{\text{TiltMean}};$$
$$\text{TiltVar[]} = \text{var}(\theta_i);$$

Writing Force

The mean force applied in each stroke was computed as:

$$\text{Force [arbitrary]} = \frac{\sum_{i=1}^{N_{\text{strokes}}} \frac{\left(\int_{t^i_{\text{start}}}^{t^i_{\text{stop}}} F^i_{\text{stroke}}\, dt\right)}{t^{\text{i-th}}_{start} - t^{\text{i-th}}_{stop}}}{N_{\text{strokes}}}$$

For the NCF computation we firstly summed the number of F signal maxima and minima within each stroke to obtain NCF_i then we extracted:

$$NCF[\#] = \frac{\sum_{i=1}^{N_{\text{strokes}}} \frac{NCF_i}{t^{\text{i-th}}_{start} - t^{\text{i-th}}_{stop}}}{N_{\text{strokes}}}$$

Writing Smoothness

The 3D acceleration was computed. The number of A_{3D} maxima and minima was summed up within each stroke to obtain NCA_i to then average it over N_{strokes}:

$$A_{3D} = \sqrt{Ax^2 + Ay^2 + Az^2};$$

$$NCA[\#] = \frac{\sum_{i=1}^{N_{\text{strokes}}} \frac{NCA_i}{t^{\text{i-th}}_{start} - t^{\text{i-th}}_{stop}}}{N_{\text{strokes}}}.$$

Tremor Indicators

Firstly the A_{3D} signal was divided into windows (W_i) of 500 samples each without distinguishing between on-sheet and in-air tracts: $W_i = A_{3D}(t_i, t_i + 500)$. We define N_{windows} the number of windows found in the raw data. The empirical mode decomposition was applied to each W_i to extract the window intrinsic mode functions (imf_i). This was done through the MATLAB® routine `emd`: $imf_i = \texttt{emd}(W_i)$. The first imf (imf_{i1}) was considered as tremor. The Hilbert spectra of imf_i were obtained by applying the MATLAB® routine `hht`: $hs_{i,f(t)} = \texttt{hht}(imf_i, fs)$.

Tremor Frequency

In each W_i, we defined $FrPeak_{i1}$ the frequency (f) of imf_{i1} for which:

$$hs_{i1}[FrPeak_{i1}] \geq hs_{i1}[f] \quad \forall f$$

Finally, we obtained:

$$F_{\text{modal}}[Hz] = \frac{\sum_{i=1}^{N_{\text{windows}}} FrPeak_{i1}}{N_{\text{windows}}}$$

Tremor Amplitude

In each W_i, we computed the RMS (RMS_{i1}) and then we averaged them over the number of windows.

$$RMS_{i1} = \texttt{rms}(imf_{i1}); \quad RMS[mm/s^2] = \frac{\sum_{i=1}^{N_{\text{windows}}} RMS_{i1}}{N_{\text{windows}}}$$

Tremor Entropy

The tremor approximate entropy was computed on each W_i using the MATLAB® routine `approximateEntropy`: $\text{ApEn}_i = \texttt{approximateEntropy}(W_i, \text{'Dimension'}, 2, \text{'Radius'}, 0.2 * \texttt{std}(W_i))$. Finally, we obtained:

$$ApEn[] = \frac{\sum_{i=1}^{N_{\text{windows}}} \text{ApEn}_i}{N_{\text{windows}}}$$

Nonlinear Characteristics of Tremor

We applied RQA analysis on each W_i following these steps:

1) Delay (D_i) estimation using the mutual information algorithm in MATLAB® (link):

$$D_i = \texttt{MutualInformation}(imf_{i1}, imf_{i1})$$

2) Embedding dimension (ED_i) estimation with the false nearest neighbor chaotic algorithm in MATLAB® (link):

$$ED_i = \texttt{embeddingDIM}(imf_{i1}, D_i)$$

3) RQA analysis in MATLAB® (link):

$$RP_{imf_i} = \texttt{MathworksRPplot}(imf_{i1}, ED_i, D_i);$$

$$RR_i, DET_i = \texttt{MathworksRPplot}(RP_{imf_i}, ED_i, D_i)$$

Finally, we obtained the overall recurrence rate ($RR\%$) and determinism ($DET\%$) as follows:

$$RR\% = \frac{\sum_{i=1}^{N_{\text{windows}}} RR_i}{N_{\text{windows}}}; \quad DET\% = \frac{\sum_{i=1}^{N_{\text{windows}}} DET_i}{N_{\text{windows}}}.$$