scientific reports

Check for updates

OPEN

Explainability of CNN-based Alzheimer's disease detection from online handwriting

Jana Sweidan[1], Mounim A. El-Yacoubi[1]✉ & Anne-Sophie Rigaud[2,3]

With over 55 million people globally affected by dementia and nearly 10 million new cases reported annually, Alzheimer's disease is a prevalent and challenging neurodegenerative disorder. Despite significant advancements in machine learning techniques for Alzheimer's disease detection, the widespread adoption of deep learning models raises concerns about their explainability. The lack of explainability in deep learning models for online handwriting analysis is a critical gap in the literature in the context of Alzheimer's disease detection. This paper addresses this challenge by interpreting predictions from a Convolutional Neural Network applied to multivariate time series data, generated by online handwriting data associated with continuous loop series handwritten on a graphical tablet. Our explainability methods reveal distinct motor behavior characteristics for healthy individuals and those diagnosed with Alzheimer's. Healthy subjects exhibited consistent, smooth movements, while Alzheimer's patients demonstrated erratic patterns marked by abrupt stops and direction changes. This emphasizes the critical role of explainability in translating complex models into clinically relevant insights. Our research contributes to the enhancement of early diagnosis, providing significant and reliable insights to stakeholders involved in patient care and intervention strategies. Our work bridges the gap between machine learning predictions and clinical insights, fostering a more effective and understandable application of advanced models for Alzheimer's disease assessment.

Keywords Alzheimer's disease, Online handwriting, 1D-CNN, Explainability

Nowadays, over 55 million people worldwide suffer from dementia. Annually, almost 10 million new cases are reported. Dementia arises from diverse diseases and injuries impacting the brain, with Alzheimer's disease (AD) accounting for approximately 60–70% of cases[1]. AD, a prevalent and progressive neurodegenerative disorder, exacts a heavy toll on those it afflicts, slowly eroding cherished memories and the very essence of one's life. AD represents a formidable health challenge due to its elusive onset and devastating impact on cognitive function. Detecting AD at its early stages, therefore, is paramount for effective intervention and treatment.

In the early stages of AD, individuals may experience subtle but notable changes in their motor skills[2]. These changes can manifest as difficulties in tasks requiring fine motor control, such as writing or buttoning a shirt. As Handwriting (HW) is a complex psychomotor skill that requires fine motor control, specific neuromuscular coordination, and visuospatial functions[3], HW analysis has been investigated to detect AD at early stage[4–6]. Early detection of AD gives hope to patients holding onto their life's narrative, preserving invaluable moments that define who they are.

Existing research has leveraged machine learning (ML) and deep learning (DL) techniques to analyze HW dynamics[7–10], yet the critical issue of model explainability remains unexplored. These models are often considered "black box" systems, lacking transparency in decision-making processes. This limitation impedes their clinical applicability and trustworthiness, particularly in healthcare settings where interpretability is paramount. In this paper, we take a further step from ML/DL helping in detecting AD early based on HW, as we aim to give healthcare experts an explanation of DL decisions.

Proposed Work: In this research, we build upon the 1DCNN model used in[7] to achieve high-accuracy classifications. Our focus is on applying explainability methods to gain insights into the decision-making process. We employed three explanation methods: DeepSHAP[11], 2-step TSR[12], and CoMTE[13], and analyzed their outcomes to understand the model's decision-making and gain insights on the behaviors of the different classes, namely Early-stage Alzheimer's disease and Healthy controls. Overall, our contributions are as follows:

[1]Samovar/Télécom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France. [2]AP-HP, Groupe Hospitalier Cochin Paris Centre, Hôpital Broca, Pôle Gérontologie, 75005 Paris, France. [3]Université Paris Descartes, 75005 Paris, France. ✉email: mounim.el_yacoubi@telecom-sudparis.eu

- We conduct comprehensive evaluations of the classification model at both loop and subject levels, utilizing diverse metrics.
- We propose an in-depth exploration of model explainability using sate-of-the-art techniques.
- We provide detailed analysis of the insights gained from the explainability methods, shedding light on the model's decision-making processes, which is the first of its kind in the context of Alzheimer's disease assessment from online handwriting on tablet.

The rest of the paper is organized as follows: section "Literature review" overviews the related work. Section "Materials and methods" describes the dataset and classification model, and provides an in-depth exploration of the three explainability methods considered in this study. Section "Results" presents a thorough analysis of the obtained results, comparing the outcomes of the three methods in terms of explainability. Section "Discussion" engages in a discussion of the findings, drawing connections to existing literature and highlighting the significance of identified patterns. Section "Conclusion" concludes the paper.

Literature review

Neurodegenerative disorders profoundly impact fine motor movements, emphasizing the significance of HW patterns as crucial biomarkers. Previous research has delved into diseases such as Alzheimer's, Parkinson's, and Huntington's, focusing keenly on the kinematic intricacies of HW movements[14–16]. Dynamic (online) HW acquisition[17,18], capturing real-time temporal data, has emerged as pivotal, as it offers a comprehensive view compared to paper-based (offline) acquisition, which lacks nuanced temporal patterns[19–22]. Previous studies often rely on statistical tests that assess global kinematic parameters, each separately, assuming a single behavioral pattern for AD[5,6]. This approach, however, overlooks valuable temporal nuances embedded within the time series data. Recent research advocated for exploring the complete dynamics of raw data, highlighting the drawbacks of relying solely on global kinematic parameters[7,17,23]. El-Yacoubi et al.[17] utilized advanced techniques based on temporal clustering with k-Medoids and Dynamic Time Warping (DTW), achieving 74% in classification accuracy. Another pioneering study adopted a 1D-Convolutional Neural Network (1DCNN) model for early-stage AD classification using HW dynamics as time series data, achieving an impressive 85% accuracy without data augmentation[7]. CNNs have been demonstrated to achieve state-of-the-art results in time series classification[24]. Their power lies in their ability to autonomously learn meaningful patterns and representations from the data, effectively eliminating the need for manual feature engineering. In alignment with these advancements, our work builds upon the same 1DCNN model architecture employed by[7] and leverages its efficacy in unraveling intricate HW patterns associated with AD.

Despite their remarkable performance, however, neural networks are often referred to as "black box" models, due to the inherent challenge of explaining how and why they make particular decisions. The lack of explainability in these models presents a significant drawback, as it hinders our ability to understand the underlying reasons for model's behavior. This has been a focal point of research, often referred to as the accuracy versus interpretability dilemma[11], due to the trade-off between the remarkable accuracy achieved by DL models and the challenge of understanding their complex abstractions. This emphasis on interpretability is also crucial for establishing trust in models as a model that lacks trustworthiness is unlikely to find practical use, especially in health. Understanding the decisions made by ML models is crucial for building trust among healthcare professionals and patients[25].

In the realm of AD detection from online HW, several ML and DL methods have been proposed[7,8,17,23]. However, a critical unexplored aspect is the explainability of these models' decisions. Despite the advancements in accuracy achieved by DL models, the lack of attention to explainability poses a significant limitation in translating these models to practical and trustworthy tools for healthcare experts. To the best of our knowledge, existing literature does not address the explainability problem in the context of AD detection from online HW data.

Model explainability can be provided by either global or local methods: A global method interprets globally the model by seeking to understand the overall structure of how a model makes decisions. A local method, by contrast, seeks understanding how the model made a decision for a single instance. Within local methods, we distinguish sample-based explanations that provide different samples as explanations, and feature-based explanations that indicate the features impacting the decision the most[13,26]. One of the most prominent explainability methods, SHapley Additive exPlanations(SHAP), stands out for offering both global and local interpretations[11]. SHAP, a method derived from coalitional game theory, distributes fairly the "payout" (prediction outcome) across the inputs features. Recent research has underscored SHAP as an effective tool even for time series data, as it demonstrates superior performance compared to other methods not specifically tailored for time series data[12,13,26]. Furthermore, state-of-the-art research on the explainability of multivariate time series have converged on the efficacy of two specialized local methods: the Counterfactual explanation-based method (Comte) and Two-Step Temporal Saliency Rescaling (TSR)[12,13,26,27]. These methods, purpose-built for multivariate time series data, have showcased exceptional results and have their implementations readily accessible in the TSInterpret library[27]. In our work, we focus on three distinct methods: DeepSHAP[11], offering both local and global explanations, CoMTE[13], a sample-based local explainability method providing counterfactual explanations, and 2-Step TSR[12], a feature-based local explanation method. Detailed discussions on each of these methods is given in the forthcoming explainability section.

Materials and methods

In this section, we begin by describing the dataset, including the participants and data collection process (section "Dataset"). This is followed by a discussion on the derivation and significance of velocity features from the raw data (section "Velocity features"). Next, we outline the architecture and configuration of our classification

model, including the training process and performance evaluation metrics (sections "Classification model" and "Evaluation of the classification model"). Finally, we detail the explainability methods applied to interpret our model's predictions (section "Explainability").

Dataset

The dataset utilized in this study is the same as the one reported in[23]. It was collected at Broca Hospital in Paris and consists of 54 participants, 27 Early-Stage Alzheimer's Disease (ES-AD) patients and 27 Healthy Control (HC) subjects, with a mean age of 79.7 ± 6.4 and 73.2 ± 5.7 respectively. All participants freely signed an informed consent form after receiving information on the study's aim and content, and all methods were carried out in accordance with relevant guidelines and regulations. Furthermore, all experimental protocols were approved by the French Advisory Committee on the Processing of Health Research Information (CCTIRS : Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé). The inclusion criteria for ES-AD patients followed the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria[28], with a requirement for a Mini Mental State Examination(MMSE) score above 20. HC subjects underwent neuropsychological tests to confirm normal cognitive profiles, and individuals with medical conditions such as stroke and other neurodegenerative diseases were excluded from the study.

The HW data used in our work corresponds to the cursive-l dataset, where participants were instructed to produce a series of four sets of four $'\ell'$ letters, specifically $'\ell\ell\ell\ell'$, by writing them on a tablet to create the pattern illustrated in Fig. 1a. These handwritten patterns were collected using a WACOM Intuos Pro Large Tablet, operating at a sampling rate of 125 Hz. The tablet systematically records several variables, including the x-coordinate (X), y-coordinate (Y), pen pressure (P), pen azimuth (Az), and the pen's altitude during its movement slightly above the tablet's surface (Al). Notably, the tablet enables real-time data acquisition by also registering the time-stamp for each point during the data collection process, as illustrated in Figure 1 of the Appendix A.

Loops series can serve as a valuable indicator of behavioral patterns associated with individuals' health status, by allowing us to uncover meaningful trends while filtering out fluctuations caused by variations in words or characters[17]. Indeed, when individuals create a series of loops, particularly when drawing or writing, their motor skills, coordination, and overall cognitive functioning are put to test. These loops represent a unique and dynamic aspect of their handwriting. By focusing on these loops and disregarding variations introduced by changes in the actual words or characters being written, researchers can gain deeper insights into the underlying behavioral trends related to the subjects' health conditions and not the changes between subjects due to the different words they write. By isolating and analyzing the loop patterns independently of the text content, it becomes possible to discern subtle but meaningful alterations in motor function or cognitive abilities, which could serve as early indicators of health conditions.

Considering our dataset's limited size, comprising just 54 subjects, there is a crucial need to increase the sample count, especially for training DL models. Drawing inspiration from[23], a repetitive pattern within the $'\ell'$ letter loops is identified. These loops could be separated into unique training instances, resulting in substantial data expansion. Specifically, for each subject, where there are typically 16 $'\ell'$ letter loops, we strategically split the samples into strokes. We then retain only the loops, by discarding ligatures between them, as depicted in Fig. 1a. This approach significantly boosts our training dataset, providing approximately 16 times more data. It is important to note that while

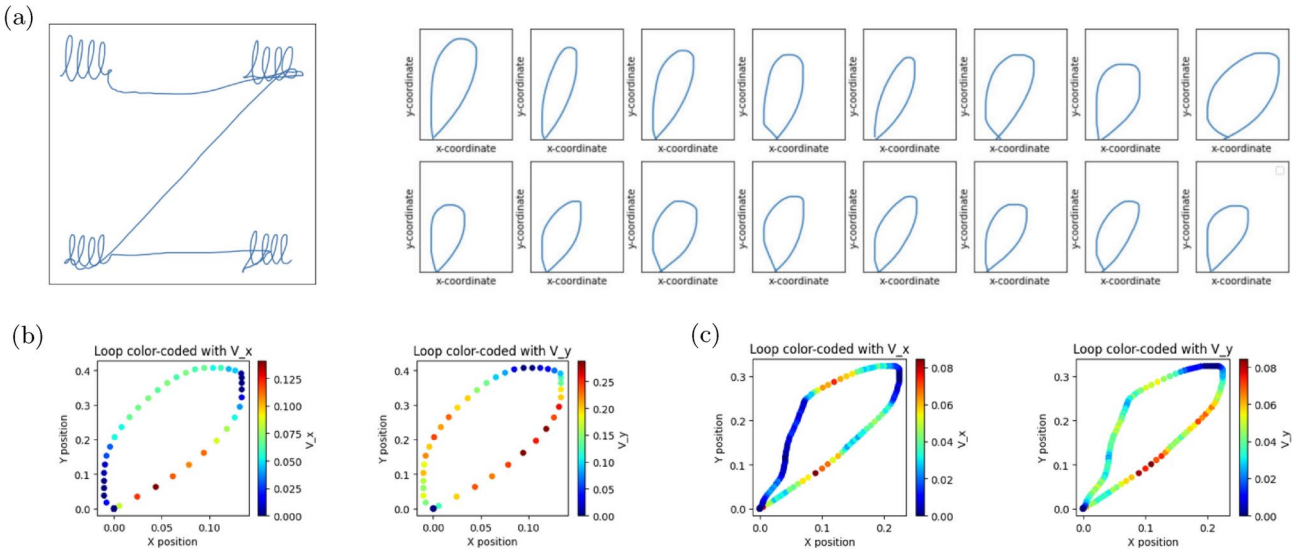


Fig. 1. Extracted loops and their velocity color mappings. (**a**) Example of cursive ℓ loops drawing data for one subject and the corresponding extracted loops[7]. (**b**) Example of HC subject's loop color coded with its velocity magnitude (**c**) Example of an AD patient's loop color coded with its velocity magnitude.

the typical count of loops is 16 for most subjects, there may be variations where some patients produce fewer or more loops.

Velocity features

In addition to the raw features provided by the recording tablet, new features can be encoded from HW dynamics to better represent the differences in task performance between HC and AD subjects. Previous research[23] has demonstrated the effectiveness of velocity features, calculated from the coordinates and timestamps, in improving classification performance. Velocity represents the rate of change of position w.r.t time. It provides crucial insights into the speed and direction of movement. After synchronizing the timestamps of the loops, we calculate velocity features, namely $V_x(n)$ (velocity in the x-direction) and $V_y(n)$ (velocity in the y-direction), from the synchronized positions, as defined in Eqs. (1) and (2) below:

$$V_x(n) = \frac{x(n+1) - x(n-1)}{t(n+1) - t(n-1)} \tag{1}$$

$$V_y(n) = \frac{y(n+1) - y(n-1)}{t(n+1) - t(n-1)} \tag{2}$$

By incorporating the additional velocity features, the total number of time series channels expands to eight, encompassing timestamps, X, Y, P, Az, Al, V_x and V_y. However, the findings of[7] indicated that the best performance was achieved solely with the two velocity features. Introducing extra features did not yield improved accuracy. Therefore, we selected the vertical and horizontal velocities as our features, where the trajectory of each loop is now captured by its individual point-wise velocities, $V_x(n)$ and $V_y(n)$, illustrating the pen's movement path. Figure 1b, c show examples of extracted loops, belonging to the AD and HC classes, color-coded with their feature values V_x and V_y.

Classification model

Instead of doing feature engineering to extract parameters from the time series data, the CNN itself performs feature extraction automatically by learning meaningful patterns from the raw input data. The convolutional layers of the CNN act as feature detectors, capturing local patterns or motifs within the time series. The CNN learns features directly from the data without relying, therefore, on predefined feature engineering, influenced by human biases or assumptions, and not always aligning with the intricate patterns present in the data. These biases could inadvertently shape the features in a way that reflects the engineer's perspective rather than the objective patterns in the data. The inherent objectivity of the CNN allows the model to uncover patterns that might be overlooked or misinterpreted by human-designed features. As a result, CNNs not only offers us a more automated and efficient approach but also provides an unbiased perspective, enhancing thereby the model's ability to capture diverse and subtle data patterns to discriminate ES-AD from HC.

Our 1DCNN model architecture, inspired from[7], is shown in Figure 2 of the Appendix A . It comprises two 1D convolutional layers, each sequentially followed by Rectified Linear Unit (ReLU) activation. The first convolutional layer comprises 128 1D filters, while the second layer utilizes 64 filters, all with dimensions of 4×1. Both convolutional layers are followed by Max Pooling layers with filter sizes of 2×1. After each Max Pooling operation, dropout with a rate of 0.2 is applied, based on[7]. The model's output is then flattened and passed through a fully connected layer (FC) before being subjected to a sigmoid activation function that provides the probability of ES-AD given the HW data. The learning rate is fixed and set to 0.001, also follows the settings described in[7].

Evaluation of the classification model

Our results are assessed using three pivotal metrics: accuracy, sensitivity, and specificity as defined in Eqs. (3), (4), and (5) respectively:

- Accuracy evaluates the overall capability of the model to make accurate classifications

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

- Sensitivity measures the model's ability to correctly identify AD

$$\text{Sensitivity} = \frac{tp}{tp + fn} \tag{4}$$

- Specificity measures the model's ability to accurately classify HC

$$\text{Specificity} = \frac{tn}{tn + fp} \tag{5}$$

These metrics are essential for comprehensively evaluating the model's performance and ensuring reliable comparison with existing literature. To assess the performance and generalization capabilities of the proposed model, we conducted comprehensive evaluations, taking into account both loop-based and subject-based assessments. We employed a leave-one-out cross-validation (LOOCV) approach. In each iteration of LOOCV, we held out one subject along with their corresponding loops for testing, while training the model on the remaining subjects. This process was repeated iteratively, systematically excluding each patient in turn. We have in total, therefore, 54 folds corresponding to the 54 subjects. In each fold, the model undergoes training for a specific number of epochs determined by early stopping, with a patience of 10 that monitors the validation loss to prevent overfitting. The training process stops when the model no longer shows improvement on the validation set after 10 consecutive iterations.

Loop based classification

In the loop-based evaluation, we treat each of the 866 individual loops as a separate sample, enabling us to examine the model's performance at the loop level, without considering the broader subject context leveraging all their loops. The results presented in Table 1 show the different metrics with the input features normalized and without normalization. With normalized features, we observe a slight increase in accuracy (86.76%) compared to non-normalized features (85.43%). Similarly, sensitivity improves from 85.99% to 88.26% with normalization, indicating refined detection of Alzheimer's-related patterns. Specificity also shows a slight enhancement from 84.86% to 85.27%. These findings underscore the importance of feature normalization, showcasing subtle yet valuable improvements at the individual loop level.

In light of the observed performance enhancement resulting from the normalization of velocity features, the subsequent experiments were conducted exclusively with normalized features, as the improvement in the metrics is essential for guaranteeing the reliability of our subsequent explainability methods. A robust and highly accurate model is fundamental to producing insightful and dependable explanations. With these enhancements, we are ensuring the groundwork for rigorous and reliable model interpretations.

Subject based classification

For subject-based evaluation, we adopt two distinct approaches to evaluate the model. Firstly, we employ a hard voting mechanism (majority voting) to determine the final classification for each subject, by aggregating the predictions of all the loops of a particular subject and assigning the subject to the class receiving the majority of votes. This method showcases the model's ability to classify subjects based on their individual loops' classifications.

Secondly, we utilize a soft voting strategy for subject-based evaluation, which aggregates class probabilities assigned to each loop, reflecting the model's confidence in its predictions. The final output is determined by calculating the average probability for each class, assigning the subject to the class with the highest average probability. This method allows us to discern subtle distinctions in the classification process. Voting methods are illustrated in Figure 3 of the Appendix A.

Table 1 provides the different performance metrics for both soft and hard voting methods. Notably, the soft voting approach exhibits a remarkable accuracy of 94.44%, surpassing the accuracy achieved through hard voting, which stands at 90.74%. Moreover, we observe that the model's sensitivity remains high for both methods, with soft voting achieving 92.86% and hard voting 89.29%. Additionally, the model exhibits exceptional specificity, further affirming its ability to accurately classify healthy subjects, with soft voting achieving 96.15% and hard voting 92.31%. These results demonstrate the robustness of our model in capturing subtle patterns and underscore the potential of employing soft voting for enhanced predictive performance and model reliability.

Explainability

For interpretability, we apply three different state-of-the-art explainability methods that proved to work best for multivariate time series data. In this section, we introduce these methods to understand the behavior of the different subjects. A comprehensive study is done on examples from the data to have an overall view of the methods, how they agree or disagree, and draw conclusions about the behavior of the two classes (ES-AD vs. HC) and the efficiency of the explainability methods.

Metric	Accuracy (%)	Sensitivity (%)	Specificity (%)
Individual loops			
With normalized features	86.76	88.26	85.27
Without normalization	85.43	85.99	84.86
Subject-based evaluations			
Soft voting	94.44	92.86	96.15
Hard voting	90.74	89.29	92.31

Table 1. Model performance.

Counterfactual explanations for machine learning time series (CoMTE)
Counterfactual (CF) explanations aim to provide insights into why a model made a particular prediction by showing what would have happened if the input data had been different in some way. This idea for ML explainability was first introduced by[29]. With CoMTE[13], the objective is to offer CF explanations specifically tailored to ML models designed for multivariate time series data. For a black-box ML model that accepts multivariate time series as input and produce class probabilities as output, the explanations aim to highlight which time series components require modification, and precisely how those modifications should occur in order to achieve the desired alteration in the classification outcome for a given sample. The method is specifically designed to identify the smallest number of time series substitutions from the chosen distractor instance X_{dist}. A distractor is a sample selected from the training dataset that belongs to the counterfactual class (opposite to the predicted class), and that can result in a change in the prediction. A two-step process is used involving the selection of suitable distractors from the CF class and the application of the Sequential Greedy Approach. The latter is an iterative method that replaces the time series modalities, i.e. V_x and V_y, in the test sample with those from chosen distractors, to maximize prediction probability until it surpasses a predefined threshold (0.95 in CoMTE).

In CoMTE, if V_x alone is sufficient for a CF, it returns only V_x's counterfactual, excluding V_y. In cases where both modalities are crucial, both are provided. This aligns with human preference for concise explanations, highlighting most significant factors rather than extensive lists of potential causes, especially in multivariate data sequences where each time series corresponds to a distinct metric[13].

Two-step temporal saliency rescaling (TSR)
Among the frequently employed explainability techniques of black-box classifiers, saliency methods stand out as a popular choice. The concept of "saliency" has its roots in explaining image models, where it entails identifying the most crucial pixels responsible for a classifier's output, typically depicted in a saliency map. This concept is not limited to image models and can be extended to explain Time Series Classification (TSC)[30]. Common methods for obtaining saliency-based explanations in TSC include two main approaches, gradient backpropagation-based and perturbation-based methods[31]. The authors in[12] conducted a comparison of well-known saliency methods that assess the importance of input features at specific time steps. They observed that classical saliency techniques do not yield satisfactory interpretations when used for *multivariate* time series data. To tackle this issue, they introduced the Two-Step Temporal Saliency Rescaling (TSR) approach, a novel technique designed to enhance the adaptability of any existing saliency method for time series data. In summary, this approach functions as follows:

- Initially, a time-relevance score is computed for each time step by determining the cumulative alteration in saliency values when that specific time step is masked.
- Subsequently, within time steps where the time-relevance score surpasses a predefined threshold, a feature-relevance score is computed for each individual feature. This is achieved by quantifying the collective change in saliency values upon masking a particular feature.
- The ultimate importance score for a given (time, feature) pair is then determined as the product of the corresponding time and feature relevance scores.

Remarkably, this approach enhances the quality of saliency maps generated by various methods when applied to time series. In particular, the authors found that TSR combined with GRAD[32] (the gradient of the output w.r.t the input) outperformed other saliency methods. Thus, we used TSR coupled with GRAD in deriving our model's explanations.

SHAP DeepExplainer
SHAP (SHapley Additive exPlanations) is one of the most used techniques of explaining a ML model and understanding how the features of the data are related to the model's output. It is a method derived from coalitional game theory to provide a way to distribute the "payout" (prediction outcome) across the features fairly. One of the biggest advantages of SHAP Values is that they provide both global and local explainability. DeepLIFT[33], another method developed for interpreting DL predictions, calculates the influence of changing inputs from their original values to reference values. The authors in[11] recognized a connection between DeepLIFT and Shapley values and opened the door to a novel approach, Deep SHAP. Deep SHAP leverages both the SHAP framework and the DeepLIFT method to provide explanations for DL models. While Deep SHAP does not inherently accommodate the direct utilization of time series data, it is the responsibility of the user to bridge the gap between the explanations and the framework[13].

Results
This section discusses the results of the interpretability methods used to analyze the model's predictions. First we present the results obtained with CoMTE (section "Counterfactual explanations for machine learning time series (CoMTE)"), then 2-step TSR (section "Two-step temporal saliency rescaling (TSR)"), followed by DeepSHAP (section "SHAP DeepExplainer"). Finally, we analyze the results to understand how velocity features influence model decisions (section "Results analysis").

Counterfactual explanations for machine learning time series (CoMTE)
To implement the CoMTE method, we utilized the TSInterpret library[27], which offers a comprehensive and user-friendly framework for generating explanations for time series data. Figure 2 shows several examples of CoMTE output for loops of different subjects. The blue curve displays the feature values of the input instance

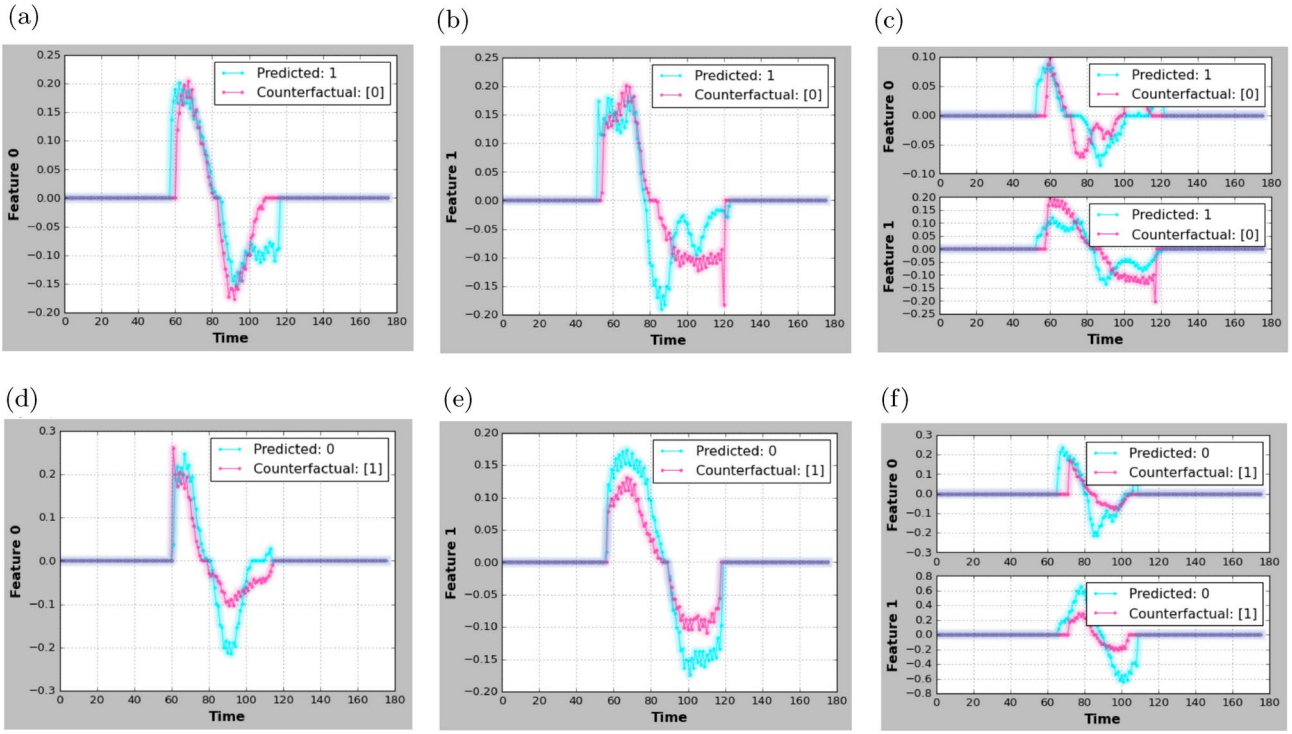


Fig. 2. Counterfactual explanations for an AD patient (**a**–**c**) and HC subject (**d**–**f**). (**a**) Only V_x modality (feature 0) is returned. (**b**) Only V_y modality (feature 1) is returned. (**c**) Both V_x and V_y modalities are returned. (**d**) Only V_x modality (feature 0) is returned. (**e**) Only V_y modality (feature 1) is returned. (**f**) Both V_x and V_y modalities are returned.

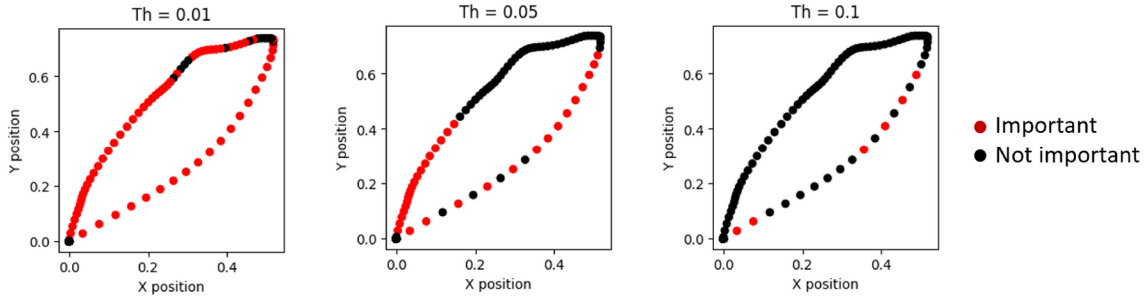


Fig. 3. Sample loop that shows the effect of the threshold on the visualization. (**a**) 2-step TSR relevance scores for HC subject loop. (**b**) 2-step TSR relevance scores for AD patient loop.

V_x or V_y corresponding to a loop, while the pink curve represents its CF from the training data, i.e. the closest curve from the training data to the input instance that would flip the predictions if replaced with it. In all figures, class 1 refers to AD, and class 0 refers to HC. Feature 0 refers to V_x modality, and feature 1 refers to V_y. Note that the horizontal line where the two curves overlap at the beginning and end of each graph correspond to the zero padding done in the preprocessing step. Let us take for example Fig. 2a. Suppose a medical expert wants to know how the original instance is predicted as AD (blue) instead of HC. In such a case, the expert applies CoMTE to generate a CF in the class direction of HC, resulting in a different velocity of V_x during certain time steps.

To better visualize the important regions of an input loop that need to be changed, and also to compare CoMTE results with the other explainability methods, we create new plots for this method. To find the regions that need to be changed, we need to take the point-wise difference between the original input instance and the produced CF. Unfortunately, since CoMTE produces the CF as a sample from the training data, and not the minimal time-steps that should be changed, there will always be a small difference between the two curves over all the time series steps as they cannot be identical. This brings the need to have a threshold to be able to plot the really important regions. By experimenting, we choose the threshold to be $Th = 0.05$, as a smaller value would make most loop parts important, and a higher value might lead to missing important regions. Figure 3 shows this approach for a loop, where red points form the important regions for the V_x modality.

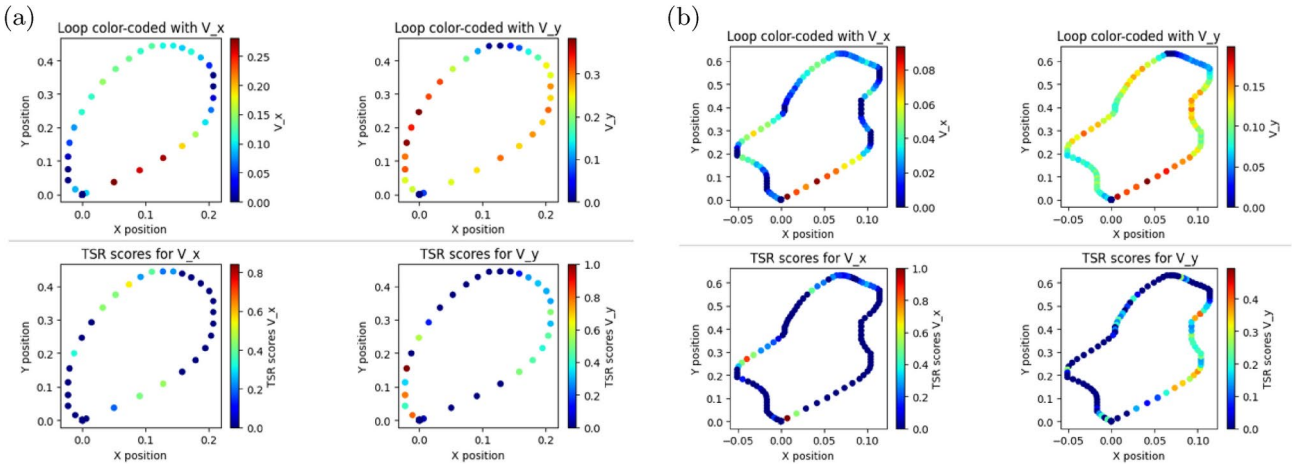


Fig. 4. 2-step TSR relevance scores for HC and AD subjects.

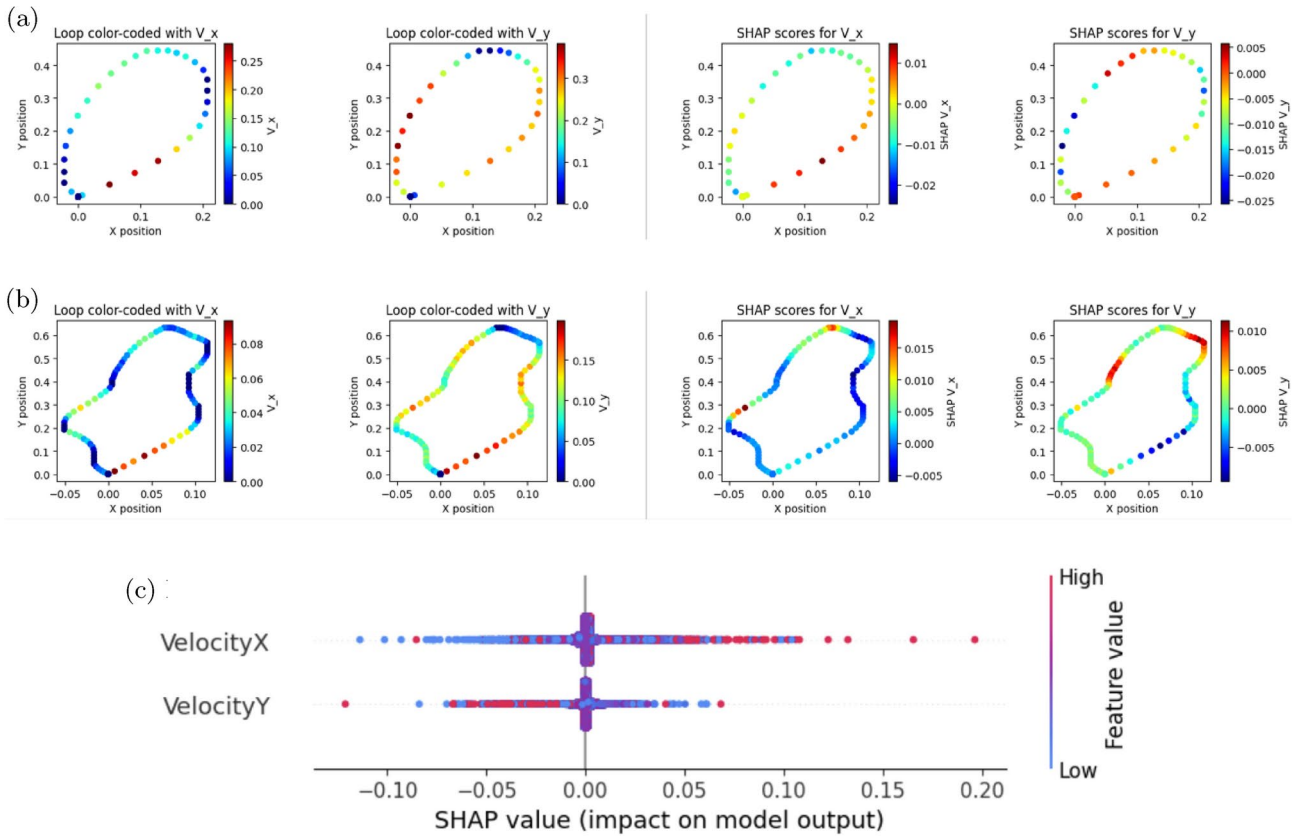


Fig. 5. DeepSHAP local and global interpretations. (**a**) Right: DeepSHAP importance values for HC subject loop. Since it belongs to HC class, negative values (blue color) are the relevant points. Left: loop color coded with its velocity features. (**b**) Right: DeepSHAP importance values for AD patient loop. Since it belongs to AD class, positive values (red color) are the relevant points. Left: loop color coded with its velocity features. (**c**) DeepShap global interpretation

Two-step temporal saliency rescaling (TSR)

We used the TSInterpret library[27] to implement the 2-step TSR method. Given an input instance, TSR returns normalized time slices and feature importance scores in range [0, 1]. We then color-code the loops with these relevance scores and display them alongside the loops color-coded with velocity features. This allows for a detailed analysis of the distinct regions in the loops, considering both the velocity values and importance scores. Figure 4 shows two examples of TSR output for loops of subjects belonging to different classes. In Fig. 4a, belonging to

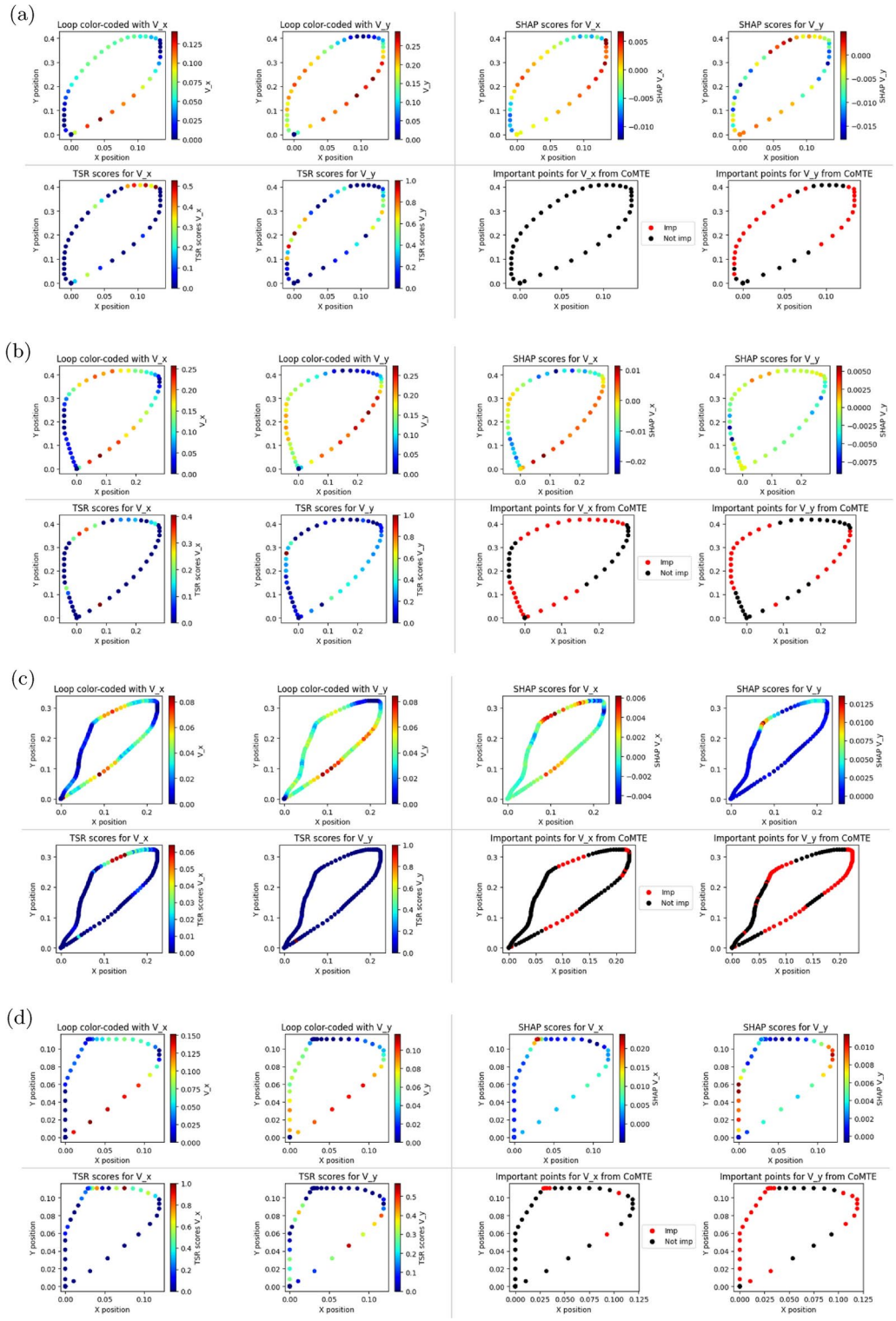


Fig. 6. Local explanations for different healthy subjects and Alzheimer's patients. (**a**) Healthy subject. (**b**) Healthy subject. (**c**) Alzheimer's patient. (**d**) Alzheimer's patient.

HC, high TSR values for specific points mean that these points contributed to the prediction of a HC. We observe higher TSR scores during the ascending and descending phases of the loop with high velocities V_x and V_y. On the other hand, in Fig. 4b, belonging to an AD patient, it is harder to relate the velocity values to the TSR scores and uncover a pattern. In this case, high TSR scores for certain points means that these points contributed to the prediction of AD patient. For V_x, we observe some scattered points that have high importance; as for V_y, TSR sheds importance especially on the ascending phase, during which there are sudden changes of direction.

SHAP DeepExplainer

Given an input instance, unlike TSR, DeepSHAP produces values attributing a directionality to the contribution. The magnitude of these values is the measure of how strong the effect is. Positive SHAP values positively impact (increases the probability of) the prediction of the AD class, while negative values have a negative impact on the AD class. After obtaining the shap values, we color-code the loop figures with these values and compare them with the loops color-coded with the velocity features to get insights about the important regions and their effect on the prediction.

Figure 5a, b show two examples for loops of subjects belonging to different classes. In Fig. 5a, belonging to a HC subject, negative shap values for specific points mean that these points contributed positively to the prediction of HC class. The bigger the magnitude (dark blue), the more the effect on the prediction. We see that we have negative values during the start of the descending phase of the loop where the velocity V_x is moderate, and at the end of the descending phase where V_x is slow. As for V_y, negative shap values are observed during the descending and ascending phases when V_y is high. On the other hand, in Fig. 5b, belonging to an AD patient, the focus is on the positive shap values that that highlight the points contributing positively to the prediction of AD class; the bigger the magnitude (dark red), the more the effect on the prediction. We observe positive values during the start of the descending phase of the loop where the velocity V_x is very slow and then gets faster, and at the end of the descending phase where V_x is high and an unusual direction change occurs. As for V_y, positive shap values are observed during the ascending and descending phases when V_y is low and sudden direction changes take place.

It is worth noting that DeepSHAP stands out as the only method out of the three offering both local and global interpretations. DeepSHAP allows us to understand the model's behavior across the entire dataset. This broader perspective is valuable, helping us grasp overall patterns in our time series data. To produce global interpretations, we aggregate the DeepSHAP shapley values from all folds into a unified plot as shown in Fig. 5c. Our classifier outputs probabilities between 0 and 1, where an output above 0.5 is assigned to the AD class. In the visualization, negative SHAP values correspond to HC while positive values indicate AD. Red-colored points signify features with high values, whereas blue represents low feature values. Notably, horizontal velocity feature, V_x, demonstrates greater overall importance than V_y in influencing the model's predictions.

The global view of SHAP illuminates notable distinctions between the loops of AD patients and HC subjects. Key areas of AD patients' loops often display a tendency towards slower V_y and faster V_x, whereas HC subjects exhibit the opposite trend. Note that while this pattern is dominant in our visualization, it is not absolute; there are regions of scattered red and blue colors, indicating variability and a less definitive pattern. It is crucial to consider that discriminating AD from HC subjects do not rely solely on velocity features. Factors such as the loop's region, whether it is in the ascending or descending phase, slant, and other characteristics, may also play a significant role. Therefore, the global view might not reflect a definitive trend in some cases. Instance-based explanations are essential for a deeper understanding of varied factors influencing predictions and uncovering nuances in the classification process.

Results analysis

In this subsection, we delve into a comprehensive comparison of the local explainability results obtained through the three methods: DeepSHAP, 2-step TSR, and CoMTE. To facilitate a comprehensive analysis, these results are visually presented in a unified figure alongside the loop color-coded with the velocity features V_x and V_y . This setup allows us to spot similarities and differences in how these methods explain the model's decisions. Our aim is to uncover common ground, disparities, and potential limitations in these explanation techniques. This comparison helps us understand how each method contributes to our model's interpretations, and gain insights into how well we can interpret our model's classifications. We showcase explanations for distinct loops belonging to different HC subjects and AD patients in Fig. 6. More examples can be found in Figures 4 and 5 of the appendix A. Note that we selected the loops that were well classified with probability greater than 90%, to make sure the interpretations represent correctly their respective classes.

Analyzing HC loops' explanations

Observations from the instance-based explanations of HC subjects reveal several consistent patterns. First, there is a notable emphasis on the importance of vertical velocity V_y during both the loops' ascending and descending phases. Second, there is a distinct focus on V_x at the transition between the ascending and descending phases, particularly at the initiation of the descending phase. Lastly, there is a noteworthy emphasis on slow V_x at the end of the descending phase, as indicated by both CoMTE and DeepSHAP explanations. These observations collectively depict a characteristic profile for HC subjects' loops, characterized by fast vertical velocities during both the ascending and descending phases, a transitional phase marked by medium to fast V_x and slow V_y, and slow V_x at the end of the descending phase. Importantly, these observations align with the global explanations provided by SHAP, where regions of high importance for V_x exhibit slow to medium velocities, while regions of high importance for V_y correspond to high velocities. This consistency reinforces the robustness and reliability of the DeepSHAP method.

Analyzing ES-AD loops' explanations

Upon detailed analysis of the instance-based explanations for loops belonging to ES-AD patients, several consistent patterns and irregularities emerge, providing valuable insights into their movement dynamics. For instance, opposite to HC behavior, high importance is given to fast V_x near the end of the descending phase, often followed suddenly by very slow V_x, indicating a tendency towards abrupt stops in their movement. Additionally, importance to fluctuations in both V_y and V_x during ascending and descending phases highlight the lack of smoothness and rhythm in their loops. These fluctuations often lead to sudden changes in both speed

and direction, further underscoring the erratic nature of their movements. In some cases, high importance is given to V_y at the end of the ascending phase, specifically when the change of direction occurs early when the y-position is still very small (< 0.1 in Fig. 6d for example). On the contrary, in HC loops, the y-position is mostly > 0.4, indicating larger loop sizes. Notably, these sudden and sharp changes of directions were captured mostly by DeepSHAP and CoMTE, while TSR struggles to highlight such critical areas, emphasizing the limitations of this method in extracting meaningful explanations where subjects fail to sustain a regular rhythm in their movements. Overall, DeepSHAP gives better insights on the importance of the features in specific regions of the loops. If a medical expert wants to know why the model predicted a normal instance instead of AD, conclusions can be drawn from the counterfactual approach with its initial visualization as depicted in section "Counterfactual explanations for machine learning time series (CoMTE)". This is because the choice of the threshold might not be optimal, and might be the reason for missing on important information.

Discussion

The regions highlighted as significant, especially by DeepSHAP, remarkably match the conclusions drawn in the previous work by[17]. Similar characteristics were identified for HC and AD loops although they used a completely different approach consisting of a two-stage clustering of loops based on velocity trajectories, and then uncovering patterns in the clusters. As documented in[17], for AD loops, DeepSHAP highlights the areas where ES-AD subjects "write faster at the onset of the ascending or descending phase" but then "fail to maintain the rhythm." These highlighted regions precisely correspond to the "loss of fluidity", characterized by a "sudden change of loop velocity or slant". Similarly, in the case of HC loops, DeepSHAP emphasizes regions with "highly fluid loops" and "medium to high velocity during their ascending and descending phases." These specific correlations not only validate the efficacy of our explainability methods, but also affirm these highlighted regions as pivotal indicators of unique movement styles, and, in particular, highlight in a fine way how ES-AD patients fail to maintain a fluid handwriting.

Ultimately, the consistent patterns revealed in healthy subjects' and Alzheimer's patients' loops provide valuable insights into the underlying movement dynamics. The emphasis on specific velocity features at different phases of the loops sheds light on the distinctive characteristics of healthy and Alzheimer's movements. Recent research has found that AD patients' handwriting shows alterations in spatial organization and poor movement control. Several studies have also identified common anomalies such as micrographia, slower movements, jerkiness, and loss of fluidity[34,35]. For instance, our findings of smaller loop sizes for AD patients align with the characteristic of micrographia. Additionally, the abrupt stops and erratic changes in direction observed in AD patients' loops indicate irregularities in their motor patterns, aligning with the clinical understanding of Alzheimer's disease as a condition affecting motor control.

Employing three distinct explanation methods -CoMTE, TSR, and DeepSHAP- strengthens our diagnostic approach. While these methods may sometimes yield diverse results, we view this variety as an asset rather than a limitation. In the complex landscape of medical diagnosis, having multiple perspectives enhances our understanding. It equips healthcare professionals with a range of insights, allowing them to select the method that best aligns with the specificities of a given case. It is important to note that the methods employed in our research are generic and can be applied across various deep learning models, ensuring broad applicability beyond the specific models utilized in this study.

Conclusion

Our study delved deep into the intricate handwriting patterns of healthy individuals and Alzheimer's patients, by employing advanced interpretability methods. Through this exploration, distinctive motor behaviors emerged. Healthy subjects exhibited consistent, smooth movements, while Alzheimer's patients demonstrated erratic patterns marked by abrupt stops and direction changes. Indeed, our dataset's limited size and diversity might not cover all possible variations in motor patterns. Future research could explore larger and more diverse datasets to validate and expand upon our findings. These findings not only enhance our understanding of disease-related motor behavior, but also pave the way for targeted interventions and therapies. Clinically, these insights could aid in the early diagnosis and monitoring of patients. Moreover, they provide a foundation for developing assistive technologies and rehabilitation strategies tailored to the specific needs of individuals with Alzheimer's. By focusing on the identified patterns, interventions can be designed to enhance smoothness and regularity in movements, potentially improving the quality of life for affected individuals. Conversely, the deep learning and interpretability techniques developed in this study could be leveraged for rehabilitation to monitor how a patient is improving their fine motor skills through therapy. In essence, our study acts as a stepping stone, bridging the gap between machine learning predictions and clinical understanding, ultimately striving for better outcomes and improved patient care.

Data availability

The datasets analysed during the current study are not publicly available due to the consent rules signed by the participants. However, the data may be made available from the corresponding author upon reasonable request.

Received: 3 April 2024; Accepted: 9 September 2024
Published online: 27 September 2024

References

1. World Health Organization. *Dementia Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/dementia (2023).

2. Buchman, A. S. & Bennett, D. A. Loss of motor function in preclinical Alzheimer's disease. *Expert Rev. Neurother.* **11**(5), 665–676 (2011).
3. Albert, M. S. *et al.* The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7**(3), 270–279 (2011).
4. Hayashi, A. *et al.* Neural substrates for writing impairments in Japanese patients with mild Alzheimer's disease: A SPECT study. *Neuropsychologia*. **49**(7), 1962–1968. https://doi.org/10.1016/j.neuropsychologia.2011.03.024 (2011).
5. Yan, J. H., Rountree, S., Massman, P., Doody, R. S. & Li, H. Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. *J. Psychiatr. Res.* **42**(14), 1203–1212. https://doi.org/10.1016/j.jpsychires.2008.01.006 (2008).
6. Yu, N. Y. & Chang, S. H. Kinematic analyses of graphomotor functions in individuals with Alzheimer's disease and amnestic mild cognitive impairment. *J. Med. Biol. Eng.* **36**(3), 334–343 (2016).
7. Dao, Q., El-Yacoubi, M. A. & Rigaud, A.-S. Detection of Alzheimer disease on online handwriting using 1d convolutional neural network. *IEEE Access* **11**, 2148–2155. https://doi.org/10.1109/ACCESS.2022.3232396 (2023).
8. Erdogmus, P. & Kabakus, A. T. The promise of convolutional neural networks for the early diagnosis of the Alzheimer's disease. *Eng. Appl. Artif. Intell.* **123**, 106254. https://doi.org/10.1016/j.engappai.2023.106254 (2023).
9. Mitra, U. & Rehman, S. U. Ml-powered handwriting analysis for early detection of Alzheimer's disease. *IEEE Access* **12**, 69031–69050. https://doi.org/10.1109/ACCESS.2024.3401104 (2024).
10. Hakan, Ö. A novel approach to detection of Alzheimer's disease from handwriting: Triple ensemble learning model. *Gazi Univ. J. Sci. Part C Des. Technol.* 1–1 (2024).
11. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (eds. Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.) vol. 30, 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (2017).
12. Ismail, A. A., Gunady, M., Bravo, H. C. & Feizi, S. Benchmarking deep learning interpretability in time series predictions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20* (Curran Associates Inc., 2020).
13. . Ates, E., Aksar, B., Leung, V. J. & Coskun, A. K. Counterfactual explanations for multivariate time series. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. https://doi.org/10.1109/icapai49758.2021.9462056. https://doi.org/10.11091109%2Ficapai49758.2021.9462056 (2021).
14. Werner, P., Rosenblum, S., Bar-On, G., Heinik, J. & Korczyn, A. Handwriting process variables discriminating mild Alzheimer's disease and mild cognitive impairment. *J. Gerontol. Psychol. Sci.* **61**(4), 228–236 (2006).
15. Teulings, H.-L. & Stelmach, G. E. Control of stroke size, peak acceleration, and stroke duration in parkinsonian handwriting. *Hum. Mov. Sci.* **10**(2–3), 315–334 (1991).
16. Slavin, M. J., Phillips, J. G., Bradshaw, J. L., Hall, K. A. & Presnell, I. Consistency of handwriting movements in dementia of the Alzheimer's type: A comparison with Huntington's and Parkinson's diseases. *J. Int. Neuropsychol. Soc.* **5**(1), 20–25. https://doi.org/10.1017/s135561779951103x (1999).
17. El-Yacoubi, M. A., Garcia-Salicetti, S., Kahindo, C., Rigaud, A.-S. & Cristancho-Lacroix, V. From aging to early-stage Alzheimer's: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. *Pattern Recognit.* **86**, 112–133 (2019).
18. Mwamsojo, N. *et al.* Reservoir computing for early stage Alzheimer's disease detection. *IEEE Access* **10**, 59821–59831. https://doi.org/10.1109/access.2022.3180045 (2022).
19. Impedovo, D. & Pirlo, G. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: A pattern recognition perspective. *IEEE Rev. Biomed. Eng.* **12**, 209–220. https://doi.org/10.1109/RBME.2018.2840679 (2019).
20. Almendra Freitas, C.O., El Yacoubi, A., Bortolozzi, F. & Sabourin, R. Brazilian bank check handwritten legal amount recognition. In *Proceedings 13th Brazilian Symposium on Computer Graphics and Image Processing (Cat. No.PR00878). SIBGRA-00*. https://doi.org/10.1109/sibgra.2000.883901 (IEEE Comput. Soc).
21. El-Yacoubi, A., Sabourin, R., Gilloux, M. & Suen, C. Y. Off-line handwritten word recognition using hidden markovmodels. In *Knowledge-based intelligent techniques in character recognition* (eds. Jain L.C. & Lazzerini B.) 191–229 (CRC Press, 1999).
22. El-Yacoubi, A., Sabourin, R., Gilloux, M. & Suen, C.Y. Improved model architecture and training phase in an off-line hmm-based word recognition system. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)* vol. 2, 1521–15252. https://doi.org/10.1109/ICPR.1998.711997 (1998).
23. Kahindo, C., El Yacoubi, M., Garcia-Salicetti, S., Rigaud, A.-S. & Cristancho-Lacroix, V. Characterizing early-stage Alzheimer through spatiotemporal dynamics of handwriting. *IEEE Signal Process. Lett.* https://doi.org/10.1109/LSP.2018.2794500 (2018).
24. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **33**(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1 (2019).
25. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. arXiv:1905.05134 (2019).
26. Rojat, T. *et al.* Explainable artificial intelligence (XAI) on timeseries data: A survey (2021).
27. Höllig, J., Kulbach, C. & Thoma, S. TSInterpret: A unified framework for time series interpretability. https://doi.org/10.48550/arXiv.2208.05280 (2022).
28. American Psychiatric Association. DSM-5 Task Force: Diagnostic and Statistical Manual of Mental Disorders: DSM-5™ 5th edn. https://doi.org/10.1176/appi.books.9780890425596 (2013).
29. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. J. Law Technol.* https://doi.org/10.2139/ssrn.3063289 (2018).
30. Meng, H., Wagner, C. & Triguero, I. Explaining time series classifiers through meaningful perturbation and optimisation. *Inf. Sci.* **645**, 119334. https://doi.org/10.1016/j.ins.2023.119334 (2023).
31. Jin, W., Li, X., Fatehi, M. & Hamarneh, G. Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. *MethodsX* **10**, 102009. https://doi.org/10.1016/j.mex.2023.102009 (2023).
32. Baehrens, D. *et al.* How to explain individual classification decisions. *J. Mach. Learn. Res.* **11**, 1803–1831 (2010).
33. Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not just a black box: Learning important features through propagating activation differences (2017).
34. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G. & Scotto di Freca, A. Handwriting analysis to support neurodegenerative diseases diagnosis: A review. *Pattern Recognition Letters* **121**, 37–45. https://doi.org/10.1016/j.patrec.2018.05.013 (2019) (**Graphonomics for e-citizens: e-health, e-society, e-education**).
35. Fernandes, C. P., Montalvo, G., Caligiuri, M., Pertsinakis, M. & Guimarães, J. Handwriting changes in Alzheimer's disease: A systematic review. *J. Alzheimers Dis.* **96**(1), 1–11. https://doi.org/10.3233/JAD-230438 (2023).

Author contributions

JS: models development, manuscript writing; M.A.E.: methodical guidance - models, manuscript writing; A-S.R.: subject assessment; dataset annotation.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72650-2.

Correspondence and requests for materials should be addressed to M.A.E.-Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024