



OPEN HandPD37: advanced handwriting dataset for Parkinson's disease diagnosis using stack ensemble ML

Jungpil Shin¹✉, Abu Saleh Musa Miah^{1,6}, Akane Hashimoto¹, Koki Hirooka¹, Md. Al Mehedi Hasan², Najmul Hassan¹, Mitsuki Aihara³, Ryoma Takahashi⁴ & Shunsuke Kobayashi⁵✉

Parkinson's Disease (PD) is a neurodegenerative disorder that affects older individuals, especially in aging countries like Japan. A key symptom of PD is hand tremor, which impairs fine motor skills, including handwriting. Early detection is crucial for treatment and management. However, existing datasets often lack sufficient task variety for effective, computer-aided PD recognition. To address this, we propose the HandPD37 dataset, which includes handwriting samples from 78 PD patients and 54 healthy individuals across 37 tasks. These tasks assess both fine motor skills and linguistic complexity, capturing diverse handwriting characteristics indicative of PD. Data was recorded using tablet devices, providing detailed information like X–Y coordinates, pen pressure, and time stamps. Additional kinematic features, such as speed and acceleration, were extracted, totaling 171 features per sample. We applied a stack ensemble of 10 machine learning (ML) algorithms, with Logistic Regression (LR) for classification, and evaluated the tasks using a meta-ensemble approach. Tasks 18, 28, 33, and 34 were most effective for distinguishing PD patients from healthy controls. The results demonstrate that combining kinematic and pressure-based features with ML models improves PD recognition accuracy. The HandPD37 dataset and ensemble approach provide a promising solution for clinical decision-support for PD diagnosis. The 37 tasks, tailored for Japanese participants, enhance the dataset's applicability for more accurate and generalized PD detection.

Keywords Parkinson's disease, PaHaW dataset, Computer-aided disease recognition, Handwriting, Kinematic features, Dynamic movement, Decision support system, Feature extraction, Machine learning, Feature selection

Parkinson's disease (PD) is a complex neurodegenerative disorder that affects millions of people worldwide^{1–5}. Current estimates suggest a prevalence ranging from approximately 10 to 800 cases per 100,000 individuals, making PD one of the most common neurodegenerative diseases⁶. The condition primarily affects motor function and is typically associated with resting tremor, muscle rigidity, bradykinesia (slowness of movement), and impaired postural reflexes. These motor symptoms are closely related to deficits in motor planning, programming, sequencing, initiation, and execution of movements⁷. At present, there is no fully objective clinical test for diagnosing PD. Reliable diagnosis often requires months of observation to monitor symptom progression, and even then, misdiagnosis rates can reach approximately 25%⁸. Definitive confirmation is only possible through post-mortem pathological examination, underscoring the complexity and uncertainty of PD diagnosis. Consequently, there is a growing need for decision-support tools that can assist clinicians in early detection and treatment planning^{9,10}. In this context, the identification of reliable biomarkers has become a central goal in neurodegenerative disease research¹¹. One of the most prominent manifestations of PD is difficulty in performing learned motor skills, such as handwriting^{12,13}. Individuals with PD often struggle to smoothly coordinate components of movement sequences, resulting in fragmented or discontinuous actions. As a result, pauses and hesitations frequently appear between movement segments^{14,15}. Handwriting and other continuous motor tasks are typically performed more slowly than in healthy individuals. Numerous studies have

¹School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-0006, Japan.

²Research Institute for Electronic Science (RIES), Hokkaido University, Sapporo, Japan. ³CAC Corporation Applied Technology Research Department Research and Development Division, Tokyo, Japan. ⁴Department of Neurology, Dokkyo Medical University, Mibu, Japan. ⁵Department of Neurology, Teikyo University, Tokyo 173-8605, Japan.

⁶Present address: Dept. of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh. ✉ email: jpsin@u-aizu.ac.jp; skoba-ky@umin.net

demonstrated that handwriting analysis can serve as a valuable biomarker for PD diagnosis^{16,17}. One explanation for this phenomenon is that handwriting in PD patients may lose its automatic nature and instead rely more heavily on visually guided feedback, functioning as a closed-loop control process¹⁸.

To better analyze handwriting impairments and understand PD-related motor dysfunction, various handwriting tasks have been proposed in the literature. A widely used task for tremor assessment is drawing the Archimedean spiral, which has been extensively applied in evaluating motor performance in PD and other movement disorders^{12,19–21}. Another common task involves writing words containing repeated cursive letters, particularly the letter “l”^{14,22}. In addition to these established tasks, some studies have explored writing simple words and short sentences characterized by easy spelling and straightforward syntax. Beyond pen-tip trajectories, several studies have shown that additional handwriting modalities are critical for PD assessment. These include the absolute position of the pen, in-air movements when the pen is lifted but remains near the trajectory, and pressure applied to the writing surface^{16,17,23}. In particular, pen pressure has been identified as a key factor in characterizing PD-related motor impairment¹⁷.

Pereira et al. introduced the HandPD dataset, which includes two handwriting tasks: (1) Spiral HandPD and (2) Meander HandPD²⁴. This dataset consists of handwritten examinations collected from 92 participants at Botucatu Medical School in São Paulo, Brazil, including 18 healthy individuals and 74 PD patients. Participants completed four spiral and four meander drawings, which were subsequently cropped and stored as images. The dataset contains 736 labeled images and includes metadata such as patient ID, gender, handedness, and age, along with extracted features such as root mean square error, tremor metrics, and radius differences between templates and handwritten traces. To address dataset imbalance and enrich data modalities, the authors later proposed the NewHandPD dataset, comprising 66 individuals (35 healthy and 31 PD patients). This improved dataset includes 12 handwriting exams per participant, encompassing spirals, meanders, and circled movements. Data were recorded using a smart pen capable of capturing both static images and dynamic signals. The NewHandPD dataset provides 264 images and 792 signal recordings, offering a more balanced and informative resource for PD research. Shrivastava et al. proposed a handwriting dataset consisting of four tasks—spirals, waves, cubes, and triangles—and applied an XGBoost classifier, achieving an accuracy of 89.7%²⁵. However, relying on only four handwriting tasks limits the ability to capture tremor-specific and fine-grained motor features. Drotár et al. introduced the PaHaW dataset, which includes eight handwriting tasks: drawing an Archimedean spiral; writing the cursive letter “l”, the bigram “le”, and the trigram “les”; writing three Czech words (“lektorka,” “porovnat,” and “nepopadnout”); and writing a full sentence (“Tramvaj dnes už nepojede”)²⁶. These tasks were selected to allow continuous writing without lifting the pen. The PaHaW dataset includes detailed recordings of on-surface trajectories, in-air movements, and pen pressure, all of which have been shown to be important for PD classification. Using handcrafted features and SVM classifiers, Shin et al. reported near-perfect classification performance on this dataset²⁷. Despite these advances, extracting highly discriminative tremor-related features remains challenging when using a limited number of handwriting tasks, whether two²⁴, four²⁵, or even eight tasks²⁶. To the best of our knowledge, no existing study has explored handwriting datasets incorporating more than eight tasks to comprehensively capture tremor dynamics. To address this limitation, we propose a new handwriting dataset consisting of 37 distinct tasks, designed to better capture tremor-related and fine motor characteristics in PD patients.

The main contributions of this work are summarized as follows:

- Creation of a new PD handwriting HandPD37 dataset with 37 tasks: We created a novel HandPD37 dataset of PD handwriting that includes 37 distinct handwriting tasks. The dataset comprises handwriting samples from 132 individuals, including 78 PD patients (average age 69.7) from Teikyo University and Fukushima Medical University, and 54 healthy individuals (average age 61.09) from the University of Aizu. Each participant completed handwriting tasks, with the data recorded using a Wacom Cintiq Pro 16 LCD tablet. The dataset attributes include time stamps, pen pressure, X-Y coordinates, azimuth, and altitude. This dataset, with its diverse range of handwriting tasks, provides a rich resource for PD diagnosis.
- Kinematic and statistical feature extraction: To enhance the dataset's feature set, we extracted 21 kinematic features, such as speed, acceleration, jerk, and stroke angle, which capture the dynamic movement characteristics of handwriting. These features help to identify movement impairments commonly seen in PD patients. Additionally, we extracted eight statistical features for each kinematic measure, yielding a total of 171 features. The inclusion of these features enables a deeper understanding of handwriting disturbances in PD patients.
- Application of ensemble machine learning (ML) models for classification: We employed a stacking ensemble of 10 ML algorithms, utilizing LR as the meta-classifier for the final classification model. This ensemble approach combines the strengths of multiple algorithms, improving classification accuracy and robustness compared to individual models. The ensemble technique provides a more reliable and accurate classification of PD and healthy controls, demonstrating its potential for clinical decision-support deployment in diagnostic systems.
- Novelty and advantages of the proposed approach: This study overcomes the limitations of previous handwriting datasets by offering a larger and more comprehensive set of tasks (37 tasks compared to 2, 4, or 8 in previous works). The dataset's rich data, incorporating both kinematic and pressure-based features, captures subtle handwriting differences indicative of PD. By leveraging ensemble learning models and incorporating detailed handwriting features, we achieve higher classification accuracy, providing valuable insights for early, non-invasive PD diagnosis. This approach showcases the effectiveness of combining diverse features and algorithms to enhance the detection and understanding of PD.

The rest of this paper is organized as follows. “[Proposed PD handwriting dataset](#)” describes the dataset used in this study. “[Methodology](#)” presents the proposed methodology. The experimental results and performance

analysis are reported in “Experimental results”. Finally, “Conclusion” concludes the paper and outlines future research directions.

Proposed PD handwriting dataset

We created a handwriting dataset to aid in the diagnosis of PD, comprising 132 subjects, including 78 PD patients and 54 healthy controls matched by age. The dataset was collected in collaboration with Takeda General Hospital, Fukushima Medical University, and the University of Aizu. Subjects were seated comfortably in front of a desk and assigned handwriting tasks, which they completed following prepared templates on a Wacom Cintiq Pro 16 LCD tablet using the corresponding pen. Participants were allowed to retry tasks in case of mistakes. The dataset includes time-series data with time stamps, X–Y coordinates, pen pressure, azimuth, and altitude. Figure 1 illustrates the age distribution of the participants. The task templates are shown in Fig. 2, and data distribution is detailed in Figs. 3 and 4.

Subjects

Data were collected from 132 subjects (78 PD patients and 54 healthy controls). Although 132 subjects participated, not all 37 tasks were completed by each subject due to task complexity and the varying capacities of the patients. Table 1 shows the number of subjects for each task. PD patient data was acquired from inpatients at Takeda General Hospital and Fukushima Medical University, while healthy control data was collected from professors and staff at the University of Aizu. Figure 1 presents the age distribution of the subjects, with PD patients in Fig. 1a and healthy individuals in Fig. 1b. The PD patients range from 41 to 86 years, while healthy controls range from 50 to 79 years. Although the y-axis scales differ due to differences in participant counts per age group, this reflects the natural distribution of participants in the rehabilitation centers and university setting. Recruiting perfectly age-matched participants across both groups is challenging. Independent t-tests showed significant differences between the PD and healthy control groups in both age and years of education ($p < 0.05$), as reported in Table 2. Therefore, age and educational level may influence handwriting performance and classification outcomes and are considered potential confounding variables in this study. Summary statistics, including the mean and standard deviation of subject age, are shown in Table 2. MDS-UPDRS, years since diagnosis, and LEDD are shown in Table 3, while Table 4 provides detailed clinical and demographic information for all participants. All subjects were right-handed and Japanese.

Handwriting tasks

The tasks used in the study were designed to capture various aspects of handwriting behavior, particularly the tremors and motor control issues commonly observed in PD patients. These tasks involved writing repetitive characters, drawing geometrical shapes, and writing complex Japanese characters, each with a specific objective to assess different motor skills. Task 1 to 9: repetitive writing of ㇿ in various sizes: The first nine tasks focus on writing the Japanese character ㇿ repeatedly, with variations in writing size and form. These tasks are crucial for PD recognition as they help assess the fine motor control of patients. PD patients typically exhibit tremors and difficulty with consistent pressure during writing, especially for repetitive tasks. Task 1 involves writing the character ㇿ 10 times; Task 2 tests the ability to write the same character in different ways, challenging motor coordination. Task 3 involves varying sizes of the character, highlighting tremor or irregularity in writing patterns at different scales. Tasks 4 to 9 involve progressively larger sizes, testing not only the control over the pen but also the patients’ ability to maintain smooth handwriting despite tremors. The variation in task complexity also allows us to observe how tremor intensity and handwriting consistency change under different conditions.

Task 10 to 16: circle drawing tasks: These tasks focus on drawing circles, each with increasing complexity, to assess fluidity and tremor control. In Task 10, participants are instructed to draw three circles from the outside

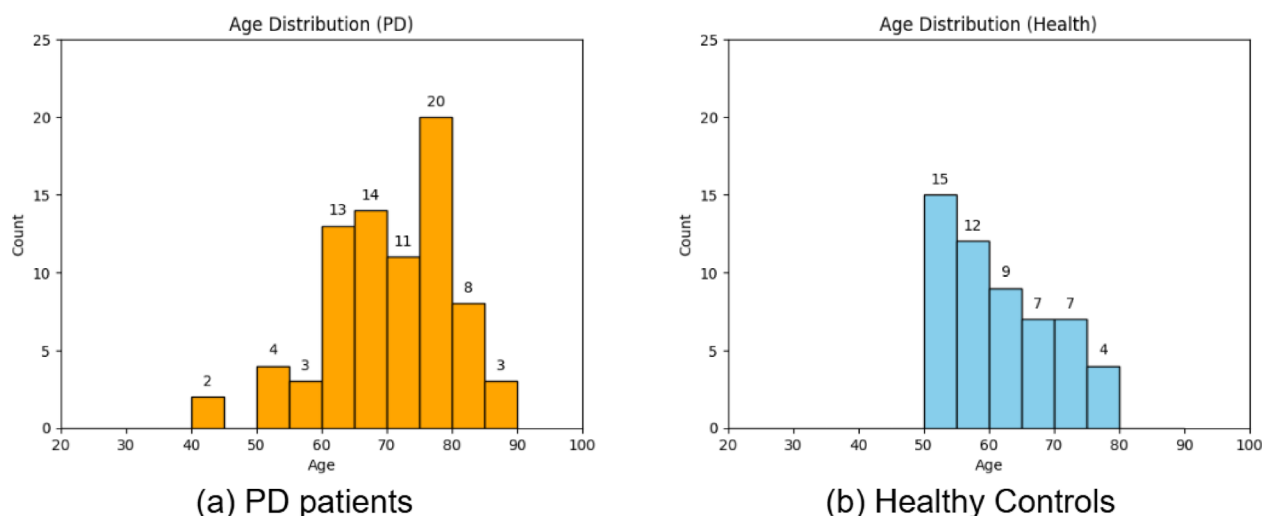


Fig. 1. Age distribution of the proposed dataset (a) PD patients (b) Healthy controls.

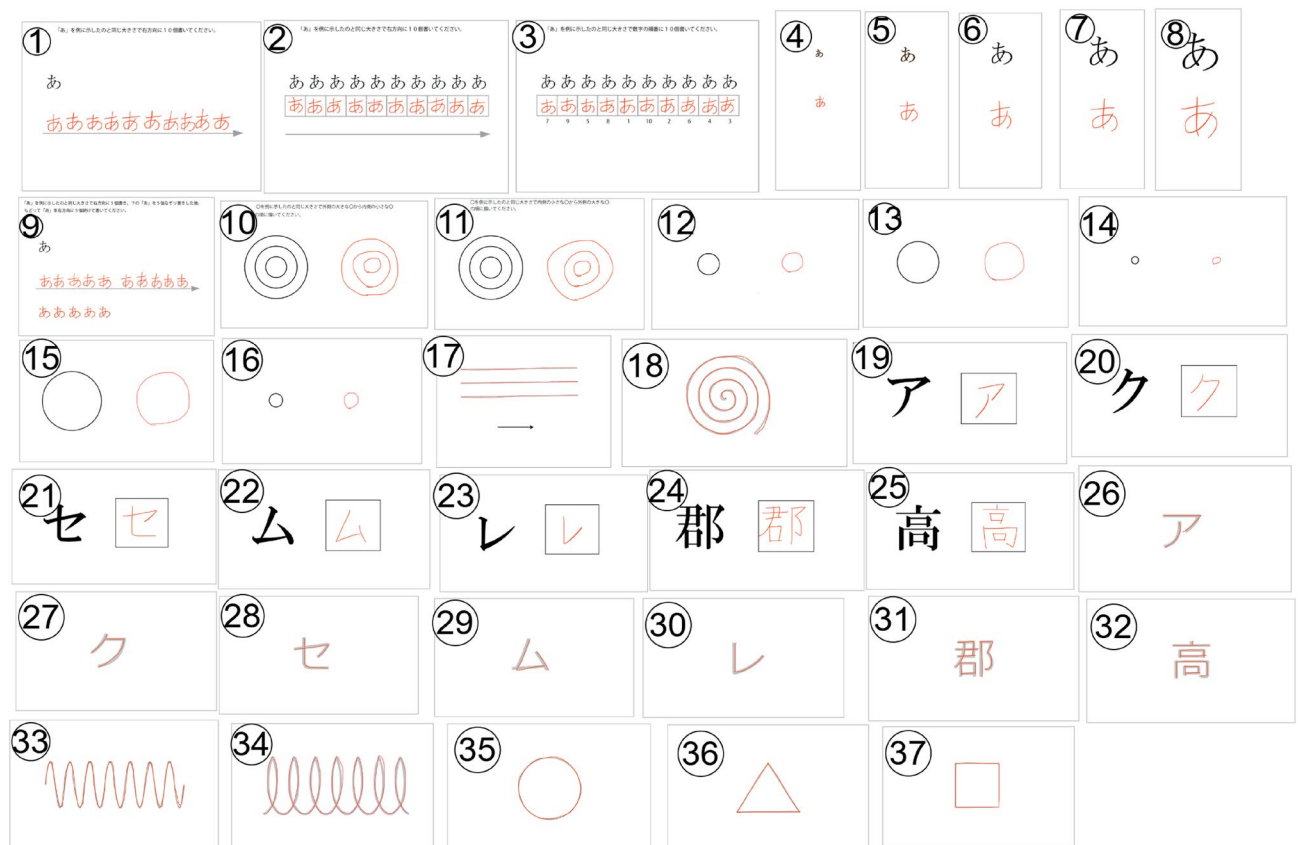


Fig. 2. Proposed 37 handwriting tasks.

in, and in Task 11, they are instructed to draw three circles from the inside out. Tasks 12 to 16 involve different sizes and types of circles, assessing how tremor affects the ability to control pen pressure and consistency. These tasks are vital because tremors in PD often manifest during circular or curved movements, providing insight into the patient's tremor amplitude and writing stability.

Task 17 to 18: parallel lines and spiral drawing: Task 17 involves drawing three parallel lines, which tests straight-line control—a fundamental motor skill. Parallel lines are particularly difficult for PD patients, as their tremors often cause deviations, providing a clear signature of motor dysfunction. Task 18 involves drawing a large spiral, which is also an effective test for tremors, as the spiral requires continuous motion and control. PD patients often struggle to maintain a steady hand, leading to irregularities in the spiral shape, making it an important task for tremor measurement.

Task 19 to 25: writing Japanese characters: These tasks assess the ability to write complex Japanese characters (Katakana and Kanji), which require intricate strokes and coordination. Task 19 tests the writing of the Katakana ア, while Task 20 assesses the character ク, and Tasks 21 to 23 require writing セ, ム, and レ. Tasks 24 and 25 test the writing of Kanji, 郡 and 高. These tasks involve multi-stroke characters that provide insight into motor skill degradation associated with PD. The varying complexity of each character allows for the assessment of fine motor control, tremor impact, and writing speed, which are key factors in PD diagnosis.

Task 26 to 32: Kanji and Katakana writing without template: Tasks 26 to 32 assess the ability to write Kanji and Katakana characters by tracing the template. Tracing tasks provide a standardized writing trajectory for all participants, thereby reducing inter-individual variability due to writing style and linguistic factors. Because tracing primarily reflects visuomotor control rather than cognitive or language processes, it enables a more specific assessment of PD-related motor impairments, such as bradykinesia, tremor, and movement irregularity. In addition, the presence of a reference path allows objective quantification of spatial and kinematic errors, which facilitates reliable feature extraction for ML analysis.

Task 33: wave drawing: Task 33 requires the subject to draw a wave-like shape seven times without seeing the template. Similar to spiral drawing, this task assesses the subject's ability to execute fluid, continuous motion, which is often disrupted by tremors in PD patients. The ability to complete this task without irregularities in motion can provide insights into the severity and frequency of tremors during fine motor tasks.

Task 34: Archimedean spiral and cursive 'l' writing: Task 34 involves drawing an Archimedean spiral and repetitively writing the cursive letter "l." These are ideal tasks for PD detection, as they require continuous motion and control. PD patients often exhibit "micrographia" (small, cramped writing) and tremors that manifest during repetitive or continuous movements. This task is particularly sensitive to those issues, providing valuable information on how tremors affect writing fluidity over extended periods.

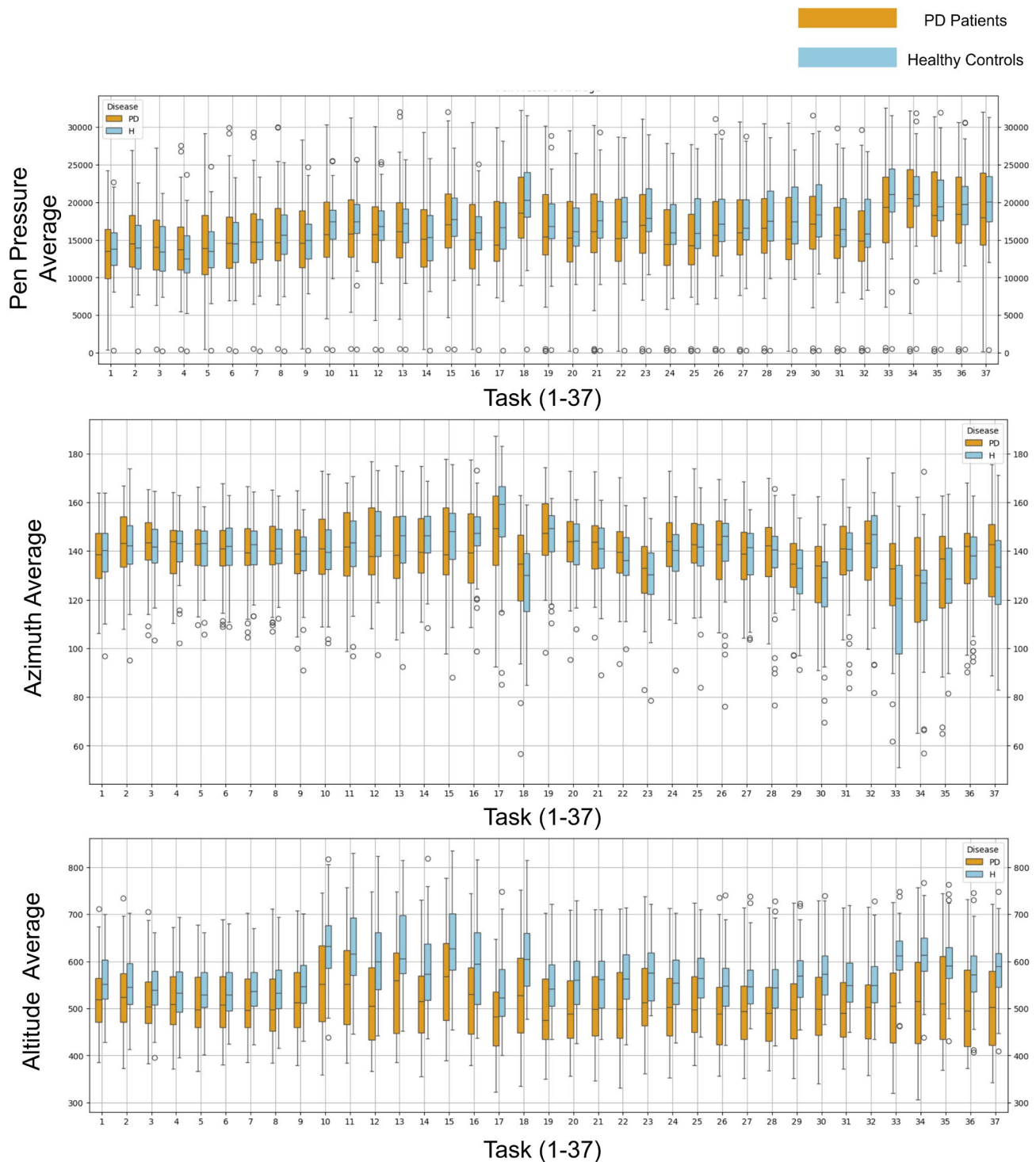


Fig. 3. Data attribute: average of pen pressure, altitude, azimuth distribution.

Task 35 to 37: geometrical drawing without template: Tasks 35, 36, and 37 involve drawing basic geometric shapes (circle, triangle, rectangle) without seeing the template. These tasks are designed to assess overall motor control and tremor severity in a non-complex, straightforward manner. The simple shapes are easy to replicate but require precise control. PD patients typically struggle with shapes that require straight lines or consistent curves, so these tasks help quantify tremor impact on geometric accuracy and overall writing legibility.

Overall task novelty and advantage for PD recognition These handwriting tasks are highly effective for PD recognition because they incorporate various elements of handwriting performance, such as repetition, size variation, complexity, and continuous motion. Tasks like circle and spiral drawing are sensitive to the tremors that characterize PD, while writing tasks using complex characters assess the patient's ability to maintain

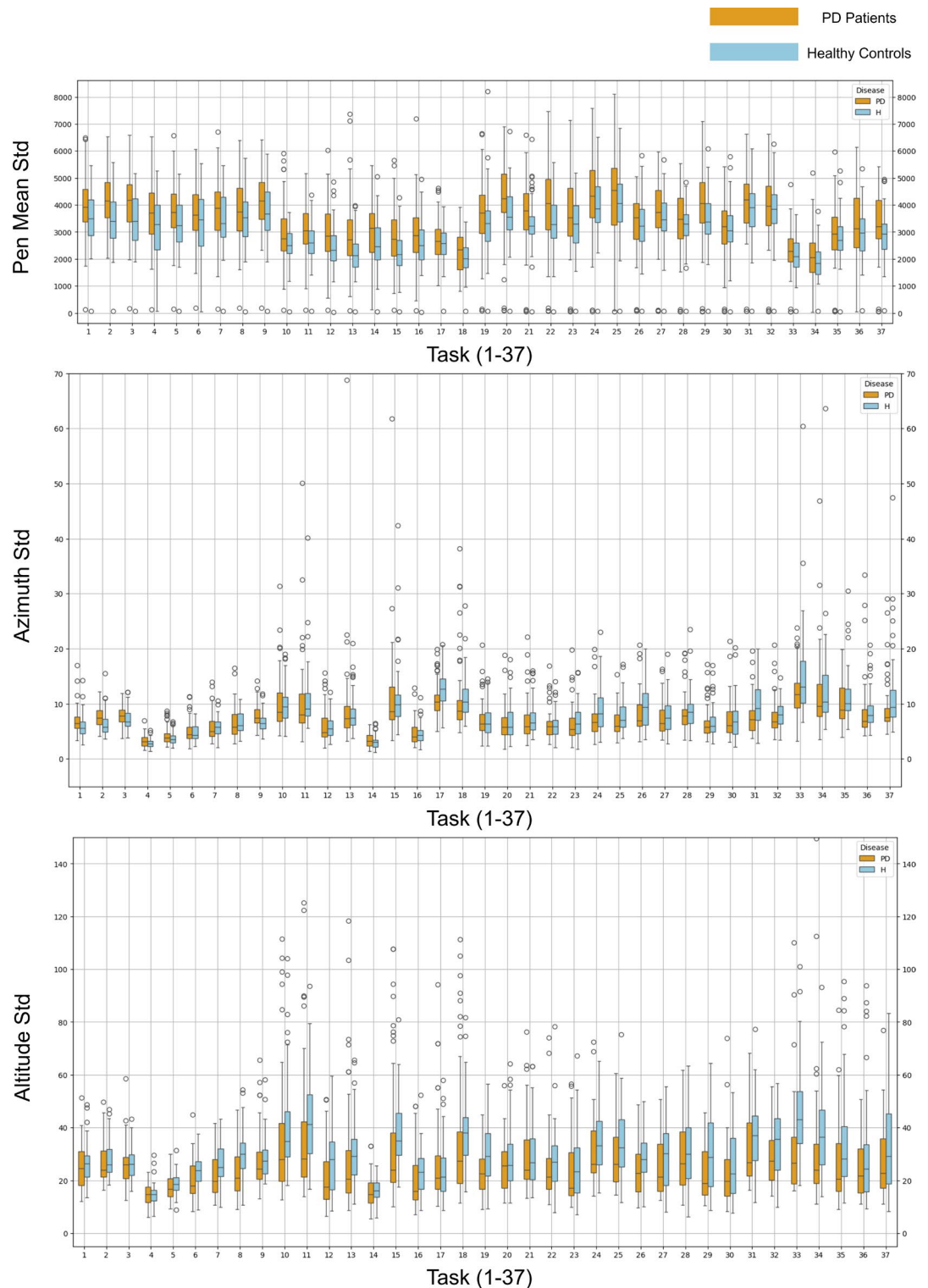


Fig. 4. Data attribute: standard deviation of pen pressure, altitude, azimuth distribution.

coordination and pressure. By capturing these diverse handwriting patterns, the dataset provides a detailed, non-invasive way to assess PD progression, particularly in the early stages. Tremors, micrographia, and rigidity all have distinct signatures in handwriting that can be identified using ML and deep learning models, making these tasks invaluable for automated PD diagnosis. Tremors, in particular, have distinct characteristics in the writing movements that can be effectively captured in these tasks, which ultimately aids in the early diagnosis and monitoring of PD.

Task no	Total	PD	H	Task no	Total	PD	H	Task no	Total	PD	H
1	130	76	54	13	128	74	54	25	103	49	54
2	128	74	54	14	128	74	54	26	104	50	54
3	128	74	54	15	129	75	54	27	104	50	54
4	130	76	54	16	129	75	54	28	103	49	54
5	129	75	54	17	125	71	54	29	103	49	54
6	129	75	54	18	124	70	54	30	103	49	54
7	129	75	54	19	104	50	54	31	103	49	54
8	128	74	54	20	104	50	54	32	103	49	54
9	129	75	54	21	104	50	54	33	102	48	54
10	131	77	54	22	104	50	54	34	101	47	54
11	128	74	54	23	103	49	54	35	102	48	54
12	129	75	54	24	104	50	54	36	103	49	54
								37	102	48	54

Table 1. Number of subjects for each task included here.

Variable	PD (Mean ± SD)	H (Mean ± SD)	t-value (PD vs H)	p-value (PD vs H)
Age	69.77 ± 9.84	61.09 ± 8.59	5.243	$6.22 \times 10^{-7} < 0.001$
Education (years)	12.3 ± 3.1	15.8 ± 2.7	-6.888	$2.57 \times 10^{-10} < 0.001$

Table 2. Demographic characteristics and statistical comparison of demographic variables between Parkinson’s disease (PD) and healthy control (H) groups.

Group	MDS-UPDRS Part III (mean ± Std)	Years since diagnosis (mean ± Std)	LEDD (mean ± Std)	Male/female
PD	33.80 ± 17.25	5.29 ± 3.71	395 ± 300	42/36
H	–	–	–	33/21

Table 3. Disease-related information for Parkinson’s disease (PD) participants.

Methodology

In this study, we propose a novel approach for diagnosing PD through a newly developed handwriting dataset, leveraging ML techniques as shown in Fig. 5. The dataset was created by collecting handwriting samples from 132 participants, including 78 PD patients and 54 healthy controls, across 37 carefully designed tasks. These tasks assess fine motor skills and linguistic complexity, with data captured using a Wacom Cintiq Pro 16 LCD tablet. The tablet recorded detailed time-series data, including time stamps, X–Y coordinates, pen pressure, azimuth, and altitude, providing precise and continuous information to detect subtle differences in handwriting indicative of PD. In the procedure, after processing the dataset, we extracted 21 kinematic features such as speed, acceleration, jerk, and stroke angle, which capture the dynamic aspects of handwriting. These kinematic features help reveal the motor impairments commonly seen in PD patients. Additionally, eight statistical features were calculated for each kinematic measure, resulting in a total of 171 features. The inclusion of these features allows for a more comprehensive understanding of handwriting disturbances in PD patients, providing valuable insights for classification. A stacking ensemble of 10 ML algorithms was employed, using LR as the meta-classifier to improve classification accuracy. This ensemble approach combines the strengths of multiple algorithms, enhancing the robustness of the model and making it more reliable compared to using a single algorithm. The models were trained on the dataset, and their performance was evaluated using cross-validation techniques to ensure generalization and prevent overfitting. Finally, the effectiveness of the tasks in distinguishing PD patients from healthy controls was assessed by ranking the tasks based on the results from the individual models and the ensemble approach. This ranking helps identify the most relevant handwriting tasks for PD diagnosis. By combining detailed handwriting features with ensemble learning models, this methodology demonstrates a promising approach for the early and non-invasive detection of PD. Furthermore, the biomarkers were detected with feature selection using a statistical method. In the following sections, we describe the data preprocessing procedure, kinematic feature extraction, statistical feature extraction, ensemble classification, and feature selection.

Data preprocessing

During the data preprocessing stage, the raw handwriting signals were inspected for missing values, duplicated timestamps, abnormal coordinate points, and incomplete traces. Coordinate points that differed significantly from the surrounding trajectory were identified based on sudden changes in inter-point displacement or

PD patients						Healthy controls		
ID	Sex	Age	LEDD	Length of PD	UPDRS III	ID	Sex	Age
1	M	41	0	0	32	79	M	50
2	M	41	0	0	32	80	F	50
3	F	52	489.4	4	7	81	F	50
4	F	52	489.4	4	7	82	F	50
5	F	52	489.4	4	7	83	M	51
6	F	54	465.5	15	5	84	M	52
7	F	55	769	11	3	85	F	52
8	F	57	541.02	1	39	86	M	53
9	F	57	541.02	1	39	87	M	53
10	M	60	818.7	11	50	88	M	53
11	M	60	818.7	11	50	89	F	53
12	F	60	629.55	7	4	90	F	53
13	F	60	629.55	7	4	91	F	53
14	M	63	0	2	20	92	M	54
15	M	63	0	6		93	F	54
16	F	63	1013.25	4	55	94	M	55
17	F	63	0	2		95	M	55
18	M	64	300.45	2	24	96	F	55
19	M	64	300.45	2	24	97	F	55
20	M	64	955	9	17	98	M	56
21	M	64	955	9	17	99	F	56
22	F	64	0	1	58	100	F	56
23	M	65	409.85	3	40	101	M	57
24	M	65	409.85	3	40	102	M	57
25	M	65	664	10	14	103	M	57
26	M	65	664	10	14	104	M	58
27	M	65	450	4		105	M	59
28	F	65	1027.84	5	65	106	M	60
29	F	65	1027.84	5	65	107	M	60
30	M	66	0	6		108	M	60
31	F	66	0	7	57	109	M	61
32	M	67	300	2	27	110	F	61
33	F	67	939.4	13	14	111	F	61
34	F	69	300	7	31	112	M	62
35	F	69	300	7	31	113	F	62
36	F	69	689.4	2	13	114	F	64
37	M	70	0	4		115	M	65
38	M	71	150.002	3	32	116	F	65
39	M	71	150.002	3	32	117	F	67
40	M	71	150.002	3	32	118	M	68
41	F	71	430.025	7	16	119	M	68
42	F	71	430.025	7	16	120	M	69
43	F	71	0	0	14	121	F	69
44	F	72	300	3	33	122	M	70
45	F	72	300	3	33	123	M	73
46	F	73	250	7		124	M	73
47	M	74	450	0	25	125	M	73
48	F	75	479.002	8	25	126	M	73
49	F	75	325	5	25	127	F	73
50	F	75	325	5	25	128	M	74
51	M	76	450	7		129	M	77
52	M	76	450	7		130	F	77
53	F	76	0	4	56	131	M	78
54	F	76	0	4	56	132	M	79
Continued								

PD patients						Healthy controls		
ID	Sex	Age	LEDD	Length of PD	UPDRS III	ID	Sex	Age
55	F	76	998.25	14	46			
56	M	77	450	4	52			
57	M	77	450	4	52			
58	M	77	0	0	50			
59	M	77	0	0	50			
60	M	77	450	4	52			
61	M	78	0	7				
62	M	78	450	5	29			
63	M	78	200	2	37			
64	M	79	0	8				
65	M	79	823	7	42			
66	M	79	823	7	42			
67	M	79	450	0	29			
68	M	80	200	2	57			
69	M	81	598.5	3	36			
70	M	82	505	6	36			
71	F	82	0	3				
72	F	83	359	4	41			
73	F	83	359	4	41			
74	F	83	240.1	15	64			
75	F	83	240.1	15	64			
76	M	85	450	7	43			
77	M	86	450	7	43			
78	M	86	300	8				

Table 4. Detailed clinical and demographic information about participants.

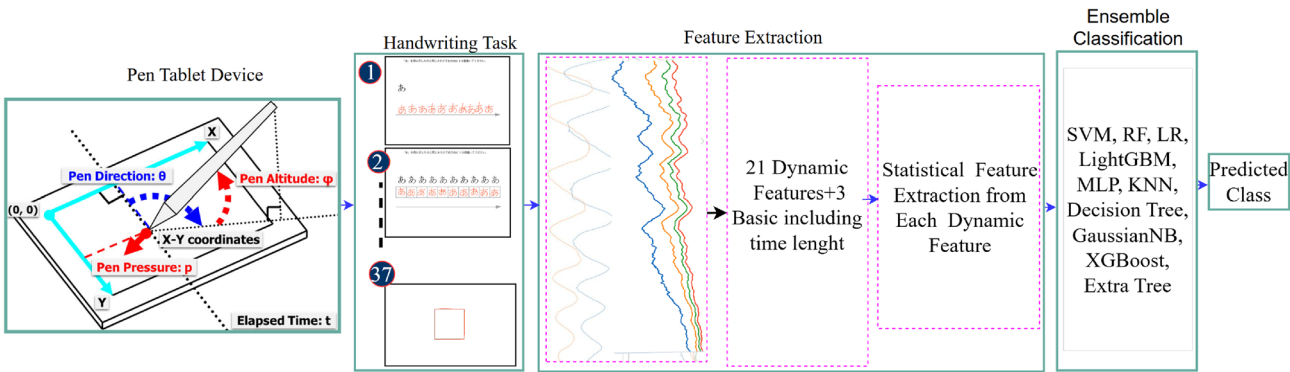


Fig. 5. Proposed framework architecture.

writing speed. Isolated abnormal points were corrected using interpolation from neighboring valid points, while samples with many abnormal points or extensive missing values were removed from the analysis. Missing values were filled using interpolation when the missing portion was small; otherwise, the corresponding sample was excluded. Finally, the data previously classified according to each patient were re-examined and reorganized according to task number and participant ID before feature extraction.

Kinematic feature extraction

In this study, we introduce the HandPD37 dataset, which includes several kinematic handwriting features. These features are crucial for analyzing the dynamics of handwriting and can provide valuable insights into various motor characteristics of writing. Below, we describe each feature, with its corresponding mathematical representation. In this section, the extraction of handwriting features from input data comprises statistical, kinematic, and angular aspects. Table 5 shows the list of kinematic features. The angle-in-stroke features were computed to quantify local trajectory curvature at different neighborhood window widths, rather than percentages of the stroke length.

Feature	Description
Time	Writing time
x-coordinate	A coordinate measured parallel to an x-axis to determine its value
y-coordinate	A coordinate measured parallel to a y-axis to determine its value
Azimuth	Pen orientation angle relative to the vertical plane of the tablet
Altitude	Pen tilt angle relative to the tablet writing plane
Writing acceleration	Rate of change of writing speed
Writing Jerk	Rate of change of writing acceleration
Writing Speed	Magnitude of velocity vector (from x and y coordinates)
Pressure	Pen pressure applied during writing
Altitude 1st Derivative	Rate of change of altitude with respect to time
Altitude 2nd Derivative	Rate of change of altitude 1st derivative
Angle in Stroke 10	Local stroke curvature computed using a neighborhood window width of 10
Angle in Stroke 30	Local stroke curvature computed using a neighborhood window width of 30
Angle in Stroke 50	Local stroke curvature computed using a neighborhood window width of 50
Angle in Stroke 100	Local stroke curvature computed using a neighborhood window width of 100
Angle in Stroke 300	Local stroke curvature computed using a neighborhood window width of 300
Angle in Stroke 500	Local stroke curvature computed using a neighborhood window width of 500
Azimuth 1st Derivative	Rate of change of azimuth with respect to time
Azimuth 2nd Derivative	Rate of change of azimuth 1st derivative
Pressure 1st Derivative	Rate of change of pressure with respect to time
Pressure 2nd Derivative	Rate of change of pressure 1st derivative

Table 5. Kinematic features and descriptions.

The kinematic features extracted from the HandPD37 dataset can be mathematically described as follows:

- x-coordinate: The x coordinate of the pen position is represented as:

$$x(t) \quad \text{for } t \in [t_0, t_n]$$

where t is the time variable and t_0, t_n represent the start and end times of the writing event.

- y-coordinate: Similarly, the y coordinate is represented as:

$$y(t) \quad \text{for } t \in [t_0, t_n]$$

- Writing speed: Writing speed $v(t)$ is the magnitude of the velocity vector derived from the x and y coordinates:

$$v(t) = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}$$

- Writing acceleration: Writing acceleration $a(t)$ is the rate of change of writing speed:

$$a(t) = \frac{dv(t)}{dt} = \frac{d}{dt} \left(\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \right)$$

- Writing jerk: Writing jerk $j(t)$ is the rate of change of writing acceleration:

$$j(t) = \frac{da(t)}{dt}$$

- Azimuth: The azimuth $\theta(t)$ is the angle of the pen with respect to the horizontal plane and can be defined as:

$$\theta(t) = \arctan\left(\frac{\Delta y}{\Delta x}\right)$$

- Altitude: The altitude $\phi(t)$ represents the pen's angle with respect to the vertical axis:

$$\phi(t) = \arccos \left(\frac{\Delta z}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \right)$$

- Derivatives of altitude: The first and second derivatives of altitude are computed as:

$$\frac{d\phi(t)}{dt} \quad \text{and} \quad \frac{d^2\phi(t)}{dt^2}$$

representing the rate of change of altitude and its acceleration, respectively.

- Pressure: The pressure applied during writing is denoted by $P(t)$. Its derivatives give the rate of change of pressure over time:

$$\frac{dP(t)}{dt} \quad \text{and} \quad \frac{d^2P(t)}{dt^2}$$

These kinematic features were selected because they describe clinically relevant handwriting movement characteristics, including movement speed, movement smoothness, pen-pressure variation, pen orientation, and trajectory irregularity. Such characteristics are closely related to motor symptoms of Parkinson's disease, including tremor, bradykinesia, rigidity, and impaired fine motor control. Therefore, the extracted kinematic features are essential for analyzing handwriting dynamics and provide quantitative indicators for PD-related handwriting abnormalities.

Statistical feature extraction from each kinematic feature

In the HandPD37 dataset, statistical features were extracted from each of the 21 kinematic features to analyze handwriting dynamics. The extracted statistical features provide essential insights into the distribution, variability, and central tendencies of each kinematic measurement. These features can be categorized into Central Tendency and Dispersion Features and Higher-order and Relationship Statistical Features. Here, we focus on the 8 core statistical features that provide a comprehensive description of handwriting dynamics^{28–30}:

These statistical features capture the central tendencies (mean, median) and the variability (variance, standard deviation) within the kinematic signals.

- Mean: The mean is the average value of the signal, indicating the typical magnitude of the signal across its entire duration. It is calculated as:

$$\text{Mean}(\mu) = \frac{1}{N} \sum_{i=1}^N x_i$$

- Median: The median represents the middle value of the sorted data, providing a robust measure against outliers. It is defined as:

$$\text{Median} = x_{\lceil N/2 \rceil} \quad (\text{for sorted data: } x_1, x_2, \dots, x_N)$$

- Variance: Variance measures the spread of the data, quantifying how much the values deviate from the mean. It is calculated as:

$$\text{Variance}(\sigma^2) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

- Standard deviation: The standard deviation is the square root of variance and represents the average deviation from the mean. It is given by:

$$\text{SD}(\sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

- Maximum: The maximum value identifies the peak or highest value in the signal, representing the extreme value in the data:

$$\text{Max} = \max(x_1, x_2, \dots, x_N)$$

- Minimum: The minimum value is the lowest value in the signal, capturing the trough or extreme low value:

$$\text{Min} = \min(x_1, x_2, \dots, x_N)$$

- 25th percentile: The 25th percentile represents the value below which 25% of the data points fall, highlighting the first quartile of the distribution.
- 75th percentile: The 75th percentile shows the value below which 75% of the data points fall, representing the third quartile of the distribution.

These statistical features are fundamental for capturing both the central tendencies and the variability in the handwriting dynamics of both healthy individuals and those affected by PD. In the HandPD37 dataset, we extracted a total of 171 features from the kinematic handwriting data, capturing various aspects of handwriting dynamics. These features include 21 kinematic features (such as writing speed) and 8 statistical features (mean, standard deviation, maximum, minimum, median, 25th percentile, 75th percentile, and variance) calculated for each kinematic feature, resulting in 168 statistical features. The remaining 3 features (width, height, and time length) complete the feature set, totaling 171 features.

Classification of PD and healthy control (HC)

To classify PD and Healthy Control (HC) groups in the HandPD37 dataset, an ensemble ML approach was employed, integrating 10 different classification algorithms. The process, as shown in Fig. 6, involved training individual models and combining their predictions to enhance classification accuracy. The models used in this ensemble approach included Random Forest (RF), LR, Support Vector Machine (SVM), Decision Tree (DT), Extra Trees (ET), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), LightGBM (LGBM), Multi-Layer Perceptron (MLP), and XGBoost (XGB). In addition, Optuna was used for parameter tuning when training each model.

Stacking ensemble learning with LR

To enhance the performance of our classification models, we employed a stacking ensemble learning approach, incorporating 10 distinct ML models followed by a meta-learning stage using LR. Similar to the training of base models, Optuna was used to tune the model's parameters. This ensemble methodology combines the outputs of multiple individual models to produce a more robust and accurate classification, capitalizing on the strengths of each base model.

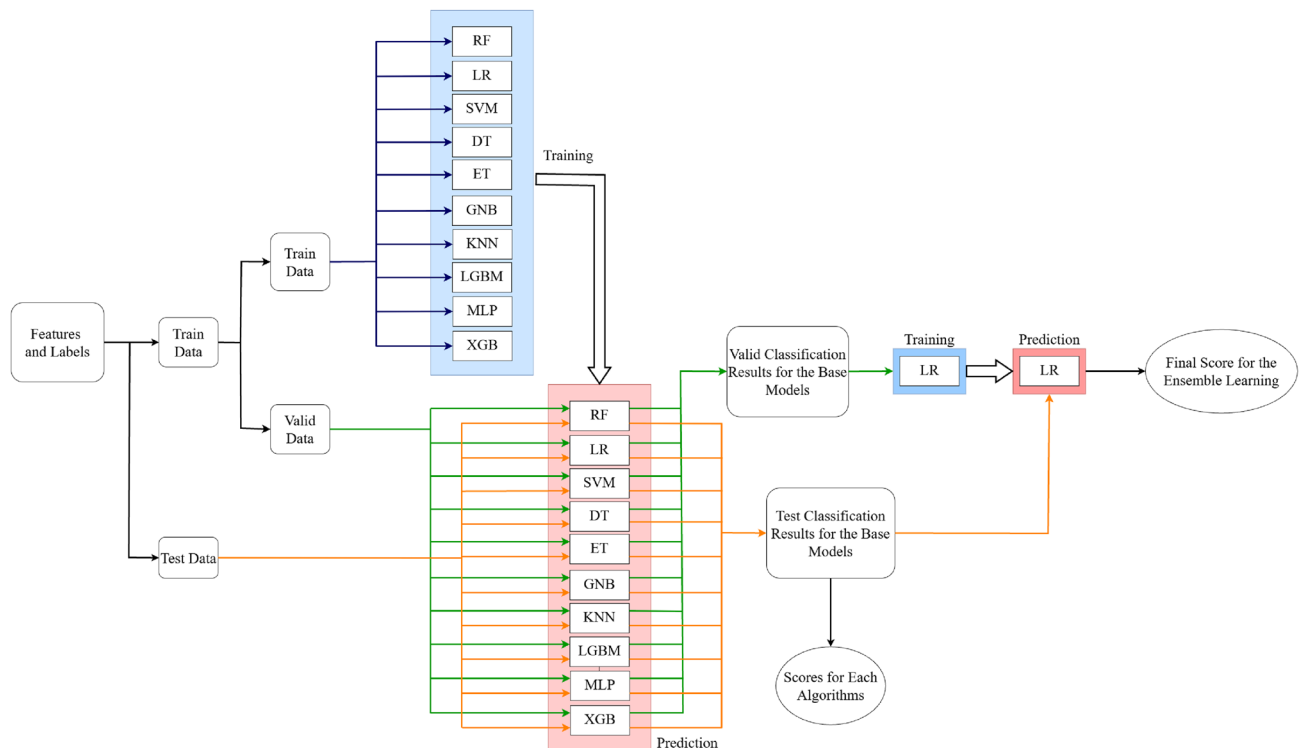


Fig. 6. Proposed stacking ensemble classification algorithm.

Training procedure for each model

For each of the ten algorithms, we implemented a custom training function that takes the features and ground-truth labels of the training data as input and returns the trained model. In the Optuna-based optimization, the following cross-validation and optimization settings were common to all algorithms: `cv=StratifiedGroupKFold(n_splits=3, shuffle=True, random_state=42)` and `n_trials=30`. In addition, the random seed values, such as `random_state`, were fixed to 42 for all models. All tasks were formulated as binary classification problems, and the performance of each cross-validation fold was evaluated using accuracy. The hyperparameter search spaces and fixed settings used for Optuna-based optimization are shown in Table 6. After selecting the optimal hyperparameters from the inner cross-validation, each base learner was retrained using all samples in the outer training set. The retrained models were then used to generate predicted probabilities for the corresponding outer test set. These probabilities were supplied to the trained meta-model to produce the final prediction. Logistic regression, which was used as the meta-model, was also optimized in the same manner as when it was used as a base model. In this procedure, only the training data were passed to the parameter optimization function. Therefore, the training, validation, and test data were not mixed at any stage of model training.

Base model training and out-of-fold prediction generation

First, a nested five-fold cross-validation framework was applied to the entire dataset. In the outer loop, the dataset was divided into five folds, where four folds were used for training and the remaining fold was reserved for testing. Within each outer training set, an inner five-fold cross-validation was performed. In each inner split, four folds were used to train the base models, while the remaining fold served as the validation set for generating out-of-fold (OOF) predictions. The overall nested subject-wise cross-validation and stacking workflow is illustrated in Fig. 7.

During both the outer and inner cross-validation procedures, `StratifiedGroupKFold(n_splits=n_splits, shuffle=True, random_state=42)` from `scikit-learn` was employed to preserve the class distribution while ensuring that all samples from the same participant were assigned to the same fold. Consequently, all data splitting was performed in a subject-wise rather than sample-wise manner. Since multiple handwriting samples were collected from each participant, this subject-wise splitting strategy prevented data leakage and avoided the model learning participant-specific characteristics. Therefore, samples from the same participant were never distributed across the training, validation, and test sets within the same cross-validation split. For each inner fold, hyperparameter optimization was performed exclusively on the corresponding inner training folds using Optuna. After selecting the optimal hyperparameters, the base model was trained using only those inner training folds, and predicted probabilities were generated for the corresponding validation fold. Since the validation fold was excluded from both hyperparameter optimization and model fitting, the resulting OOF predictions were obtained from previously unseen samples. This validation fold not used for optimization. By repeating this procedure across all five inner folds, OOF predictions were generated for every sample in the outer training set. These OOF predicted probabilities served as the input features for training the stacking meta-model. For probability estimation, the `predict_proba()` function in `scikit-learn` was used. Unlike `predict()`, which outputs only the predicted class labels, `predict_proba()` returns the estimated probability that each sample belongs to each class (Parkinson's disease or healthy control). For the SVM classifier, these probabilities were pseudo-probabilities obtained through sigmoid (Platt) calibration rather than direct probabilistic estimation. The predicted class label for each base model was determined by assigning the Parkinson's disease class when the predicted probability was greater than or equal to 0.5; otherwise, the healthy control class was assigned.

Meta-model training and test prediction

The out-of-fold predictions generated by the base models within the outer training set were used as the training data for the meta-model. After hyperparameter optimization, each base learner was retrained on the complete outer training set using the selected hyperparameters. The retrained base models were then used to generate predicted probabilities for the corresponding outer test fold, which served as the input features for the meta-model during testing. The meta-model was trained using the out-of-fold predicted probabilities generated by the base models from the inner cross-validation of the outer training set. It was then used to classify the test data based on the predicted probabilities obtained from the base models. This procedure was repeated across the five outer folds, and the final performance was evaluated based on the results obtained from all test folds. All experiments were conducted using a fixed random seed of 42. The evaluation was performed using five outer folds, rather than repeated runs with different random seeds. The selection of features was conducted using the Mann-Whitney U test and the correlation coefficient for each task, as illustrated in Fig. 8. Classification was executed in the same manner as with all 171 features, and the outcomes were subsequently compared. Features that could serve as potential biomarkers were identified based on a p -value < 0.10 across all tasks. The common features included mean writing speed, mean altitude, percentile 25 pressure second derivative, percentile 25 altitude, percentile 25 writing speed, percentile 25 altitude first derivative, median altitude, median writing speed, median angle in stroke 30, minimum altitude, maximum altitude, percentile 75 altitude, percentile 75 writing speed, and percentile 75 angle in stroke 10. In addition, classification was also conducted using only these common features to assess their predictive value.

Experimental results

In this study, we evaluated the performance of our proposed approach by applying it to 37 handwriting tasks collected from PD patients and healthy controls (HC). We first trained and tested each of the 10 individual ML models RF, LR, SVM, DT, ET, GNB, KNN, LGBM, MLP, and XGB on the dataset. After training, we ranked the

Algorithm	Hyperparameter	Value or Search space
RF	n_estimators	Int [50, 500]
	max_depth	Log-uniform Int [4, 32]
	min_samples_split	Int [2, 20]
	min_samples_leaf	Int [1, 20]
LR	C	Log-uniform [1e-3, 1e3]
	penalty	Categorical {l1, l2}
	solver	Fixed: liblinear
SVM	C	Log-uniform [1e-3, 1e3]
	kernel	Categorical {linear, rbf, poly, sigmoid}
	gamma	Log-uniform [1e-4, 1e1] (if kernel $\neq$ linear)
	degree	Int [2, 5] (if kernel = poly)
DT	criterion	Categorical {gini, entropy}
	max_depth	Int [2, 20]
	min_samples_leaf	Int [1, 15]
	min_samples_split	Int [2, 20]
	max_features	Categorical {sqrt, log2, None}
ET	n_estimators	Int [50, 500]
	max_depth	Log-uniform Int [2, 32]
	min_samples_split	Int [2, 20]
	min_samples_leaf	Int [1, 20]
	max_features	Float [0.1, 1.0]
	criterion	Categorical {gini, entropy}
GNB	var_smoothing	Log-uniform [1e-10, 1e-3]
KNN	n_components (PCA)	Dynamic range*
	n_neighbors	Int [2, 21]
	weights	Categorical {uniform, distance}
	metric	Categorical {euclidean, manhattan}
LGBM	n_estimators	Int [50, 500]
	learning_rate	Log-uniform [1e-3, 1e-1]
	num_leaves	Int [5, 30]
	max_depth	Int [2, 7]
	reg_alpha	Log-uniform [1e-8, 1e1]
	reg_lambda	Log-uniform [1e-8, 1e1]
	min_child_samples	Int [10, 30]
MLP	n_units	Int [10, 100]
	hidden_layer_sizes	Fixed: (n_units,) single layer
	activation	Categorical {relu, tanh}
	alpha	Log-uniform [1e-5, 1e-1]
	learning_rate_init	Log-uniform [1e-4, 1e-2]
	solver	Categorical {adam, lbfgs}
	max_iter	Fixed: 500
	early_stopping	Fixed: True
XGB	n_estimators	Int [100, 1000]
	max_depth	Int [2, 8]
	learning_rate	Log-uniform [1e-3, 1e-1]
	subsample	Float [0.5, 1.0]
	colsample_bytree	Float [0.4, 1.0]
	gamma	Log-uniform [1e-8, 1.0]
Continued		

Algorithm	Hyperparameter	Value or Search space
	reg_alpha	Log-uniform [1e-8, 1e1]
	reg_lambda	Log-uniform [1e-8, 1e1]
	objective	Fixed: binary:logistic
	eval_metric	Fixed: logloss
	use_label_encoder	Fixed: False

Table 6. Hyperparameter search spaces and fixed settings used for optuna-based optimization. *PCA was applied only to the KNN model. The upper bound of the number of principal components was dynamically set to avoid exceeding the minimum number of training samples in the inner stratified group cross-validation folds, while also considering the number of input features and an upper limit of 80.

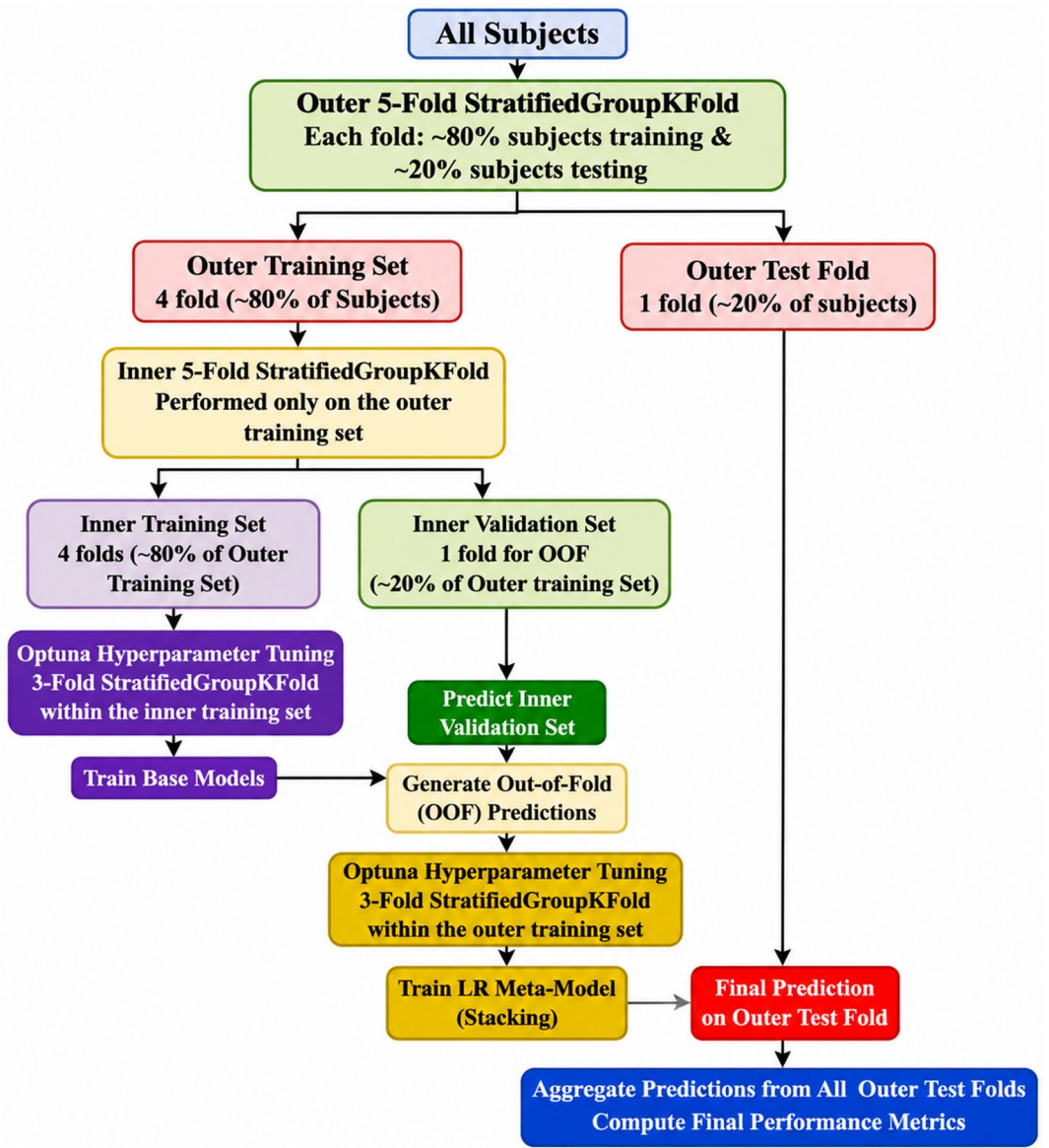


Fig. 7. Overview of the proposed nested subject-wise five-fold cross-validation framework with stacking ensemble learning.

tasks based on the performance of each individual model. These rankings were based on accuracy. The individual model performances allowed us to identify which tasks were more effectively distinguished between PD and HC samples for each algorithm. Subsequently, we applied a stacking ensemble approach to combine the predictions from all 10 base models. LR was used as the meta-classifier to generate the final classification score by integrating

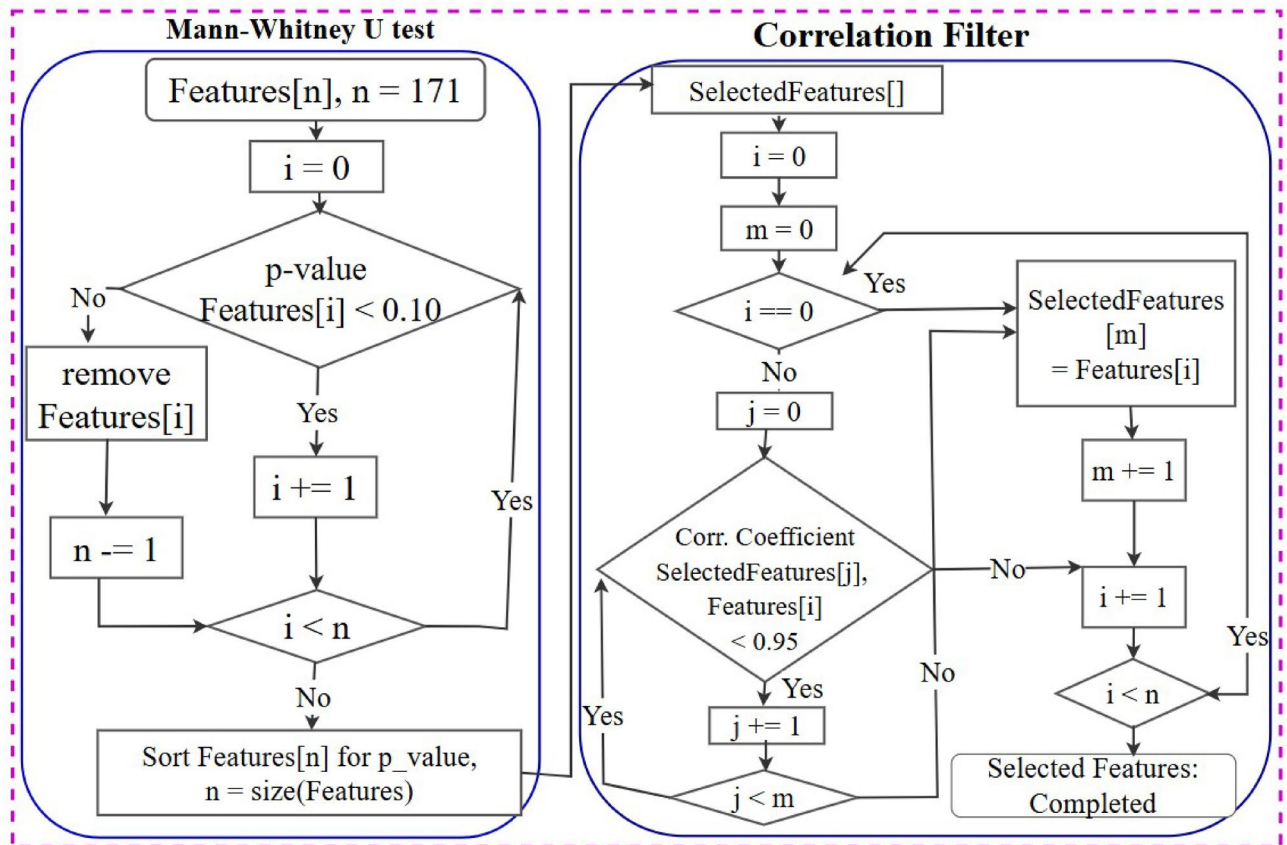


Fig. 8. Feature selection flowchart.

the outputs of the base models. We then ranked the tasks based on the performance of this ensemble model, providing a comprehensive ranking that reflected the contribution of each task to the classification accuracy. The ensemble model, which benefits from combining the strengths of multiple base models, provided improved classification results, enhancing the accuracy and robustness of the PD detection process. For all models and the ensemble method, a consistent nested five-fold cross-validation protocol was employed. In each outer fold, approximately 80% of the subjects were used for model training and 20% were reserved for testing. Within each outer training set, an inner five-fold cross-validation was performed for hyperparameter optimization and generation of out-of-fold predictions used to train the stacking meta-model. Subject-wise splitting was maintained throughout the entire procedure to prevent data leakage. The final performance was obtained by aggregating the predictions from all five outer test folds.

Performance results for each ML model

The experimental results presented in Table 7 provide an overview of the classification performance across various models for 37 distinct tasks. The table shows the accuracy of different models, including RF, LR, SVM, DT, ET, GNB, KNN, LGBM, MLP, and XGB, along with a final accuracy column that consolidates the results. The performance of the models varies across tasks. Overall, ET, LGBM, and XGB consistently perform well across multiple tasks, achieving high accuracy rates above 90%. To clarify whether the ensemble model effectively utilized the outputs of the base models, we compared the accuracy of the ensemble model with the best-performing single model for each task. The best accuracy for each task is shown in bold in Table 7. The ensemble model achieved the highest accuracy in 16 out of 37 tasks. Furthermore, the detailed performance metrics of the best-performing model for each handwriting task, including accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC, are summarized in Table 8 as the mean $\pm$ standard deviation obtained from five-fold cross-validation. In addition to the task-wise comparison, the ensemble model showed the highest minimum accuracy and mean accuracy across all tasks. This indicates that the ensemble model tended to improve the lower bound of classification performance. This interpretation was further supported by the independent t-test results shown in Table 9, which demonstrated that the ensemble model achieved significantly higher accuracy than eight of the ten individual algorithms ($p < 0.01$), with the exception of ET and LGBM. Thus, although the advantage of the ensemble model was not observed for every individual task, the statistical comparison and summary statistics suggest that it provided more stable overall performance than most individual algorithms. Overall, these findings suggest that ML and ensemble-based approaches have potential utility for classification tasks related to PD diagnosis based on handwriting analysis. However, the variation in model performance across tasks indicates that the effectiveness of these approaches is task-dependent and requires careful model selection

Task	RF	LR	SVM	DT	ET	GNB	KNN	LGBM	MLP	XGB	Final Result
1	90.72	96.92	96.92	93.88	96.97	95.49	91.62	96.20	93.81	96.18	97.66
2	90.73	93.24	94.62	88.23	96.95	99.23	95.52	98.52	93.97	94.44	99.29
3	90.36	93.69	97.50	81.90	94.52	90.83	95.12	97.62	96.90	92.26	95.36
4	93.01	91.41	87.34	90.68	96.89	83.31	90.10	93.01	86.54	92.93	94.47
5	93.09	91.72	88.47	89.90	95.34	89.38	86.24	95.43	88.57	94.60	95.43
6	89.73	92.88	91.51	89.79	96.79	91.50	84.53	93.59	90.72	91.12	96.76
7	91.89	96.85	92.29	89.58	96.76	91.40	82.27	96.76	89.11	92.63	98.33
8	90.89	96.70	92.91	83.09	93.97	98.32	90.68	94.88	95.30	92.43	95.93
9	89.58	92.20	93.83	92.01	95.99	92.20	92.20	96.83	94.57	93.59	96.89
10	90.13	97.72	95.44	92.42	94.67	89.92	90.72	96.15	97.69	96.15	98.46
11	95.20	96.83	94.48	96.06	98.43	79.66	85.94	96.09	94.48	95.29	98.43
12	92.76	97.03	93.66	97.10	99.20	86.02	92.19	100.00	93.76	99.23	100.00
13	90.66	94.45	93.89	91.89	96.83	80.06	90.85	97.69	93.95	95.99	99.17
14	94.46	94.56	92.95	94.59	95.85	89.06	89.86	96.07	94.56	93.60	94.46
15	94.32	96.30	96.30	94.39	94.39	80.80	86.41	94.49	94.86	94.39	95.39
16	94.36	93.76	96.86	91.16	96.03	83.71	88.48	96.83	94.53	96.03	96.89
17	94.41	95.89	95.15	91.00	96.72	94.35	95.06	97.46	95.89	95.89	95.89
18	98.21	99.23	96.79	95.15	99.29	91.39	96.14	98.25	97.62	99.05	100.00
19	96.34	90.44	89.39	91.44	93.36	86.23	83.68	97.95	89.39	98.08	99.13
20	97.03	96.03	96.21	90.26	97.89	93.23	89.44	96.16	96.34	95.97	95.16
21	97.39	95.52	90.49	96.26	97.21	92.86	88.70	96.34	93.55	98.26	96.34
22	95.29	92.86	91.68	88.02	95.29	67.28	88.94	94.23	89.76	93.55	93.36
23	91.26	85.90	90.09	87.52	91.13	82.51	79.35	95.02	91.09	92.22	92.00
24	96.16	96.16	94.29	94.84	97.03	96.16	91.42	96.21	94.29	97.03	94.29
25	93.12	93.22	93.22	87.95	94.09	95.04	90.18	91.98	93.30	91.11	94.17
26	97.03	95.97	95.29	91.13	98.08	90.44	93.23	98.95	92.18	97.21	98.08
27	95.29	95.34	95.47	92.13	98.08	89.52	90.49	96.84	93.55	98.08	98.95
28	98.09	97.14	95.19	95.14	98.09	92.23	89.32	99.00	94.14	99.09	100.00
29	94.19	95.23	94.10	91.32	97.14	86.46	87.42	96.14	94.28	96.05	97.10
30	97.00	95.14	97.05	95.05	96.00	93.19	88.28	98.00	95.10	98.09	98.05
31	96.00	97.14	94.05	93.28	97.05	88.28	92.41	99.00	95.05	97.09	99.00
32	96.00	97.05	96.05	91.28	97.00	91.28	93.28	92.05	97.05	94.05	96.05
33	95.99	99.05	97.14	94.99	100.00	93.04	95.04	97.99	100.00	94.99	98.95
34	99.05	93.98	93.98	97.04	100.00	87.32	93.98	95.99	95.04	99.05	99.05
35	95.34	95.90	93.72	93.30	96.21	88.29	88.32	96.39	94.72	96.52	99.13
36	96.05	95.32	95.19	93.27	99.09	91.28	90.23	99.09	95.19	98.18	99.05
37	93.42	91.68	93.51	87.51	96.98	96.11	90.68	96.31	93.51	92.47	96.20

Table 7. Classification results for different models and tasks. Bold values indicate the highest accuracy among the individual algorithms and the ensemble model for each task

and hyperparameter tuning. Hyperparameter optimization using Optuna was performed only within the inner training folds, which helps to avoid information leakage.

Task ranking for individual models

The tables present the top 10 tasks ranked by classification accuracy for each individual model, including RF, LR, SVM, DT, ET, GNB, KNN, LGBM, MLP, and XGB. Tables 10 and 11 show the top 10 ranked tasks for each algorithm. Figure 9 shows the number of times each task appeared among the top 5 tasks across the individual models. In addition, the images of the two most frequently appearing tasks in Fig. 9, together with their numbers of appearances, are shown in Fig. 10a. Task 18 stood out most prominently, appearing among the top 5 tasks in 7 out of 10 individual models. This suggests that Task 18 was a particularly informative task in this dataset for PD classification based on handwriting analysis. In addition, Task 33 appeared among the top 5 tasks in 5 out of 10 individual models, suggesting that this task may also contain useful handwriting-related information for PD classification.

Task ranking for the ensemble model

In this section, we present the ranking of tasks based on the performance of the ensemble model, which integrates the outputs of 10 individual ML models. Table 12 shows the final ranking of tasks based on classification accuracy in percentage format. In the ensemble model, Tasks 12, 18, and 28 achieved the highest classification accuracy

Task	Model	Accuracy	Sensitivity/recall	Specificity	Precision	F1-score	ROC-AUC
1	Ensemble	97.66 ± 1.92	97.42 ± 3.17	98.00 ± 4.00	98.67 ± 2.67	97.98 ± 1.65	100.00 ± 0
2	Ensemble	99.29 ± 1.43	98.89 ± 2.22	100.00 ± 0	100.00 ± 0	99.43 ± 1.14	100.00 ± 0
3	LGBM	97.62 ± 3.28	97.14 ± 5.71	98.18 ± 3.64	98.89 ± 2.22	97.89 ± 3.00	99.07 ± 1.27
4	ET	96.89 ± 1.60	95.98 ± 3.40	98.18 ± 3.64	98.75 ± 2.50	97.27 ± 1.45	99.38 ± 0.97
5	LGBM	95.43 ± 2.82	96.21 ± 3.13	94.70 ± 4.34	95.89 ± 3.37	96.01 ± 2.54	99.32 ± 0.83
	Ensemble	95.43 ± 2.82	96.21 ± 3.13	94.70 ± 4.34	95.89 ± 3.37	96.01 ± 2.54	99.02 ± 0.81
6	ET	96.79 ± 3.09	95.03 ± 7.09	98.00 ± 4.00	98.89 ± 2.22	96.74 ± 3.66	98.76 ± 2.19
7	Ensemble	98.33 ± 3.33	96.36 ± 7.27	100.00 ± 0	100.00 ± 0	98.00 ± 4.00	99.44 ± 1.12
8	GNB	98.32 ± 2.07	97.08 ± 3.63	100.00 ± 0	100.00 ± 0	98.49 ± 1.89	99.25 ± 1.00
9	Ensemble	96.89 ± 2.82	94.18 ± 5.21	100.00 ± 0	100.00 ± 0	96.93 ± 2.77	99.75 ± 0.31
10	Ensemble	98.46 ± 3.08	98.57 ± 2.86	98.33 ± 3.33	98.57 ± 2.86	98.57 ± 2.86	99.40 ± 1.19
11	ET	98.43 ± 1.92	98.57 ± 2.86	98.18 ± 3.64	98.75 ± 2.50	98.61 ± 1.70	99.74 ± 0.52
	Ensemble	98.43 ± 1.92	97.24 ± 3.39	100.00 ± 0	100.00 ± 0	98.57 ± 1.75	100.00 ± 0
12	LGBM	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
	Ensemble	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
13	Ensemble	99.17 ± 1.67	98.57 ± 2.86	100.00 ± 0	100.00 ± 0	99.26 ± 1.48	100.00 ± 0
14	LGBM	96.07 ± 2.36	95.68 ± 5.40	96.33 ± 4.52	97.42 ± 3.17	96.38 ± 2.19	98.20 ± 2.00
15	LR	96.30 ± 2.13	95.31 ± 4.11	98.33 ± 3.33	98.57 ± 2.86	96.81 ± 1.79	99.59 ± 0.57
	SVM	96.30 ± 2.13	96.64 ± 4.33	96.33 ± 4.52	97.32 ± 3.29	96.86 ± 1.79	99.16 ± 0.48
16	Ensemble	96.89 ± 1.59	96.16 ± 3.23	98.18 ± 3.64	98.46 ± 3.08	97.21 ± 1.48	99.62 ± 0.77
17	LGBM	97.46 ± 3.35	96.92 ± 6.15	98.18 ± 3.64	98.46 ± 3.08	97.53 ± 3.32	100.00 ± 0
18	Ensemble	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
19	Ensemble	99.13 ± 1.74	98.46 ± 3.08	100.00 ± 0	100.00 ± 0	99.20 ± 1.60	100.00 ± 0
20	ET	97.89 ± 4.21	94.29 ± 11.43	100.00 ± 0	100.00 ± 0	96.67 ± 6.67	100.00 ± 0
21	XGB	98.26 ± 2.13	100.00 ± 0	96.18 ± 4.69	97.03 ± 3.64	98.46 ± 1.89	99.85 ± 0.30
22	RF	95.29 ± 2.79	94.64 ± 6.59	94.18 ± 7.92	95.79 ± 5.45	94.90 ± 2.85	99.30 ± 0.62
	ET	95.29 ± 4.34	92.62 ± 11.08	96.18 ± 4.69	96.90 ± 3.81	94.26 ± 6.26	99.39 ± 0.88
23	LGBM	95.02 ± 5.49	95.96 ± 5.17	94.67 ± 6.86	93.33 ± 8.89	94.49 ± 6.49	98.34 ± 2.14
24	ET	97.03 ± 4.13	92.75 ± 11.07	100.00 ± 0	100.00 ± 0	95.87 ± 6.46	99.85 ± 0.31
	XGB	97.03 ± 4.13	92.75 ± 11.07	100.00 ± 0	100.00 ± 0	95.87 ± 6.46	99.45 ± 0.68
25	GNB	95.04 ± 3.02	88.82 ± 6.61	100.00 ± 0	100.00 ± 0	93.95 ± 3.66	98.93 ± 0.96
26	LGBM	98.95 ± 2.11	97.14 ± 5.71	100.00 ± 0	100.00 ± 0	98.46 ± 3.08	100.00 ± 0
27	Ensemble	98.95 ± 2.11	97.14 ± 5.71	100.00 ± 0	100.00 ± 0	98.46 ± 3.08	99.52 ± 0.95
28	Ensemble	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
29	ET	97.14 ± 2.34	93.96 ± 5.16	100.00 ± 0	100.00 ± 0	96.81 ± 2.73	99.58 ± 0.83
30	XGB	98.09 ± 2.34	98.46 ± 3.08	98.18 ± 3.64	98.00 ± 4.00	98.15 ± 2.30	100.00 ± 0
31	LGBM	99.00 ± 2.00	97.50 ± 5.00	100.00 ± 0	100.00 ± 0	98.67 ± 2.67	99.79 ± 0.42
	Ensemble	99.00 ± 2.00	97.50 ± 5.00	100.00 ± 0	100.00 ± 0	98.67 ± 2.67	100.00 ± 0
32	LR	97.05 ± 2.41	93.28 ± 5.55	100.00 ± 0	100.00 ± 0	96.44 ± 2.94	99.64 ± 0.73
	MLP	97.05 ± 2.41	93.28 ± 5.55	100.00 ± 0	100.00 ± 0	96.44 ± 2.94	99.02 ± 0.88
33	ET	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
	MLP	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
34	ET	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0	100.00 ± 0
35	Ensemble	99.13 ± 1.74	100.00 ± 0	98.00 ± 4.00	98.57 ± 2.86	99.26 ± 1.48	99.69 ± 0.62
36	ET	99.09 ± 1.82	98.46 ± 3.08	100.00 ± 0	100.00 ± 0	99.20 ± 1.60	100.00 ± 0
	LGBM	99.09 ± 1.82	98.46 ± 3.08	100.00 ± 0	100.00 ± 0	99.20 ± 1.60	100.00 ± 0
37	ET	96.98 ± 3.90	93.06 ± 9.04	100.00 ± 0	100.00 ± 0	96.17 ± 5.04	99.56 ± 0.54

Table 8. Performance metrics of the best-performing model for each handwriting task based on five-fold cross-validation.

Comparison	Mean difference (%)	t-value	p-value	Significant ($p < 0.01$)
Ensemble vs RF	2.93	5.26	1.40×10^{-6}	Yes
Ensemble vs LR	2.23	4.06	1.24×10^{-4}	Yes
Ensemble vs SVM	3.13	5.87	1.23×10^{-7}	Yes
Ensemble vs DT	5.36	7.99	1.58×10^{-11}	Yes
Ensemble vs ET	0.47	1.03	0.3088	No
Ensemble vs GNB	7.72	7.25	3.82×10^{-10}	Yes
Ensemble vs KNN	7.15	10.11	1.85×10^{-15}	Yes
Ensemble vs LGBM	0.63	1.35	0.1797	No
Ensemble vs MLP	3.20	5.66	2.88×10^{-7}	Yes
Ensemble vs XGB	1.62	3.10	0.0027	Yes

Table 9. Independent t-test results comparing ensemble accuracy with individual algorithms.

RF			LR			SVM			DT			ET		
Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy
1	34	99.05	1	18	99.23	1	3	97.50	1	12	97.10	1	33	100.00
2	18	98.21	2	33	99.05	2	33	97.14	2	34	97.04	1	34	100.00
3	28	98.09	3	10	97.72	3	30	97.05	3	21	96.26	3	18	99.29
4	21	97.39	4	28	97.14	4	1	96.92	4	11	96.06	4	12	99.20
5	20	97.03	4	31	97.14	5	16	96.86	5	18	95.15	5	36	99.09
6	26	97.03	6	32	97.05	6	18	96.79	6	28	95.14	6	11	98.43
7	30	97.00	7	12	97.03	7	15	96.30	7	30	95.05	7	28	98.09
8	19	96.34	8	1	96.92	8	20	96.21	8	33	94.99	8	26	98.08
9	24	96.16	9	7	96.85	9	32	96.05	9	24	94.84	8	27	98.08
10	36	96.05	10	11	96.83	10	27	95.47	10	14	94.59	10	20	97.89

Table 10. Top-10 ranked tasks for RF, LR, SVM, DT, and ET.

GNB			KNN			LGBM			MLP			XGB		
Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy	Rank	Task	Accuracy
1	2	99.23	1	18	96.14	1	12	100.00	1	33	100.00	1	12	99.23
2	8	98.32	2	2	95.52	2	36	99.09	2	10	97.69	2	28	99.09
3	24	96.16	3	3	95.12	3	28	99.00	3	18	97.62	3	18	99.05
4	37	96.11	4	17	95.06	3	31	99.00	4	32	97.05	3	34	99.05
5	1	95.49	5	33	95.04	5	26	98.95	5	3	96.90	5	21	98.26
6	25	95.04	6	34	93.98	6	2	98.52	6	20	96.34	6	36	98.18
7	17	94.35	7	32	93.28	7	18	98.25	7	17	95.89	7	30	98.09
8	20	93.23	8	26	93.23	8	30	98.00	8	8	95.30	8	19	98.08
9	30	93.19	9	31	92.41	9	33	97.99	9	36	95.19	8	27	98.08
10	33	93.04	10	9	92.20	10	19	97.95	10	30	95.10	10	26	97.21

Table 11. Top-10 ranked tasks for GNB, KNN, LGBM, MLP, and XGB.

and were therefore ranked first. These tasks are shown in Fig. 10b. For Tasks 18 and 28, the ensemble model achieved higher accuracy than any individual model, suggesting that probability-based stacking contributed to improved classification performance for these tasks. In contrast, for Task 12, the ensemble model matched the best individual-model performance. Among the top-ranked ensemble tasks, Task 18 is particularly notable because it appeared among the top 5 tasks in 7 out of 10 individual models, making it the most consistently highly ranked task across the individual models. Task 28 also appeared among the top 5 tasks in 4 out of 10 individual models, suggesting that it may contain handwriting-related information useful for PD classification across both individual and ensemble-based analyses. Although Task 33 appeared among the top 5 tasks in 5 out of 10 individual models, it was ranked lower in the ensemble model than Tasks 12, 18, and 28. Task 34 also ranked highly in the ensemble model, but its ensemble accuracy did not exceed the best individual-model accuracy. These results indicate that the ensemble model identified a partially different set of highly ranked

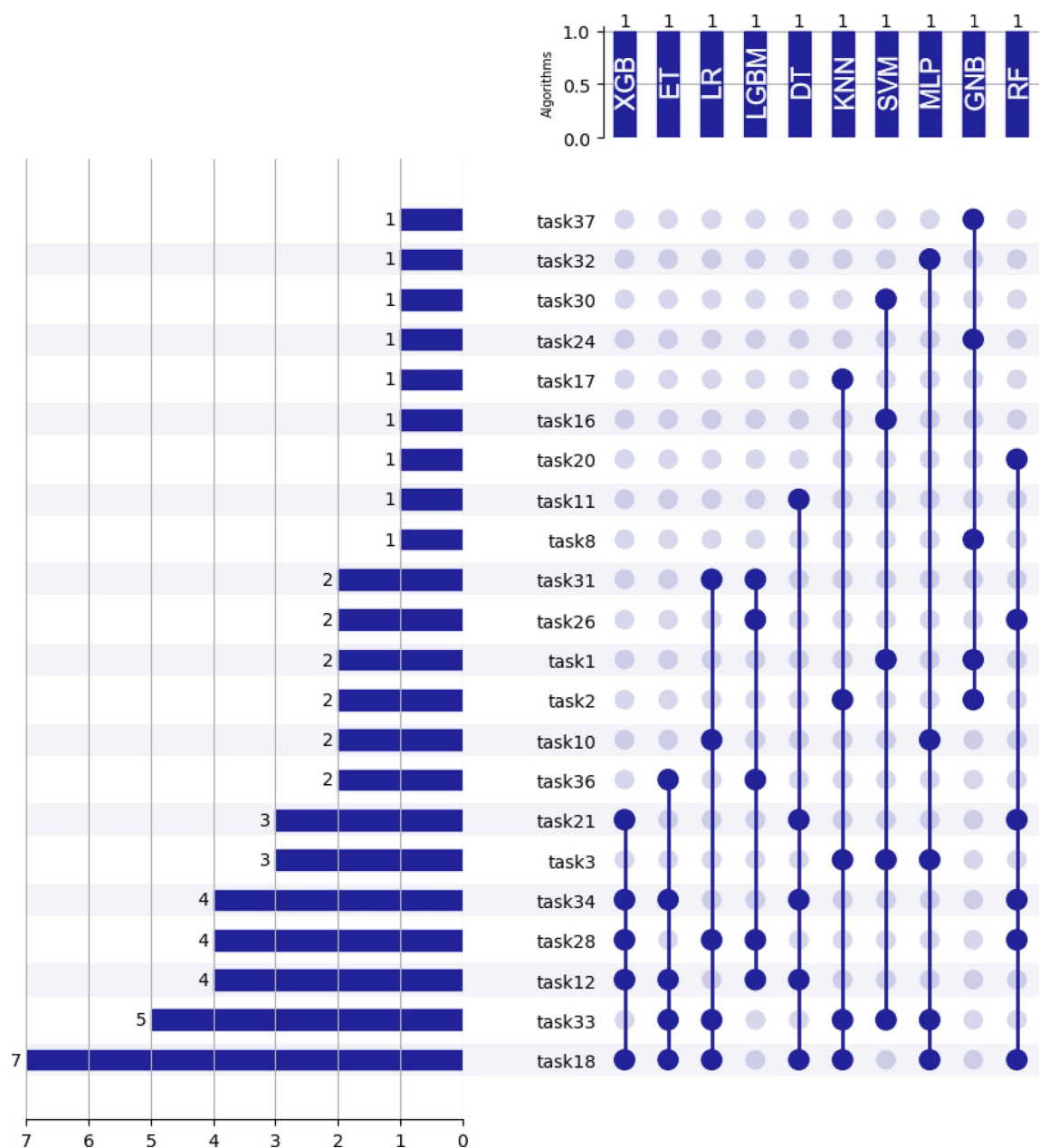


Fig. 9. The number of times it appeared in the top 5 tasks.

tasks compared with the individual-model analysis. Overall, these results suggest that the ensemble model may capture complementary information from the outputs of the individual models. Although the top-ranked tasks differed partly between the individual-model analysis and the ensemble-model analysis, Task 18 was consistently informative across individual models and also achieved the highest accuracy in the ensemble model. Therefore, Task 18 may be particularly relevant for PD classification based on handwriting analysis, while Tasks 12 and 28 may also represent highly informative tasks in the ensemble-based analysis.

From a medical perspective, the high ranking of Tasks 18, 28, 33, and 34 remains clinically meaningful because these tasks reflect motor symptoms commonly observed in Parkinson's disease, such as tremor, bradykinesia, rigidity, reduced movement smoothness, impaired fine motor control, and micrographia. Task 18, which involves spiral drawing, requires continuous visuomotor control and is sensitive to tremor-related trajectory irregularities. Task 28, which involves tracing a complex Japanese character, requires precise visuomotor coordination, stroke sequencing, and stable pen-pressure control; therefore, it can reveal impairments in fine motor control, movement planning, and handwriting stability in PD patients. Task 33, involving wave drawing, evaluates rhythmic and continuous hand movement, which can be disrupted by tremor and bradykinesia. Task 34, which includes Archimedean spiral drawing and repetitive cursive "l" writing, is particularly relevant because repetitive and continuous strokes can reveal reduced movement amplitude, handwriting instability, and micrographia. Therefore, the strong performance of these tasks is not only statistically important but

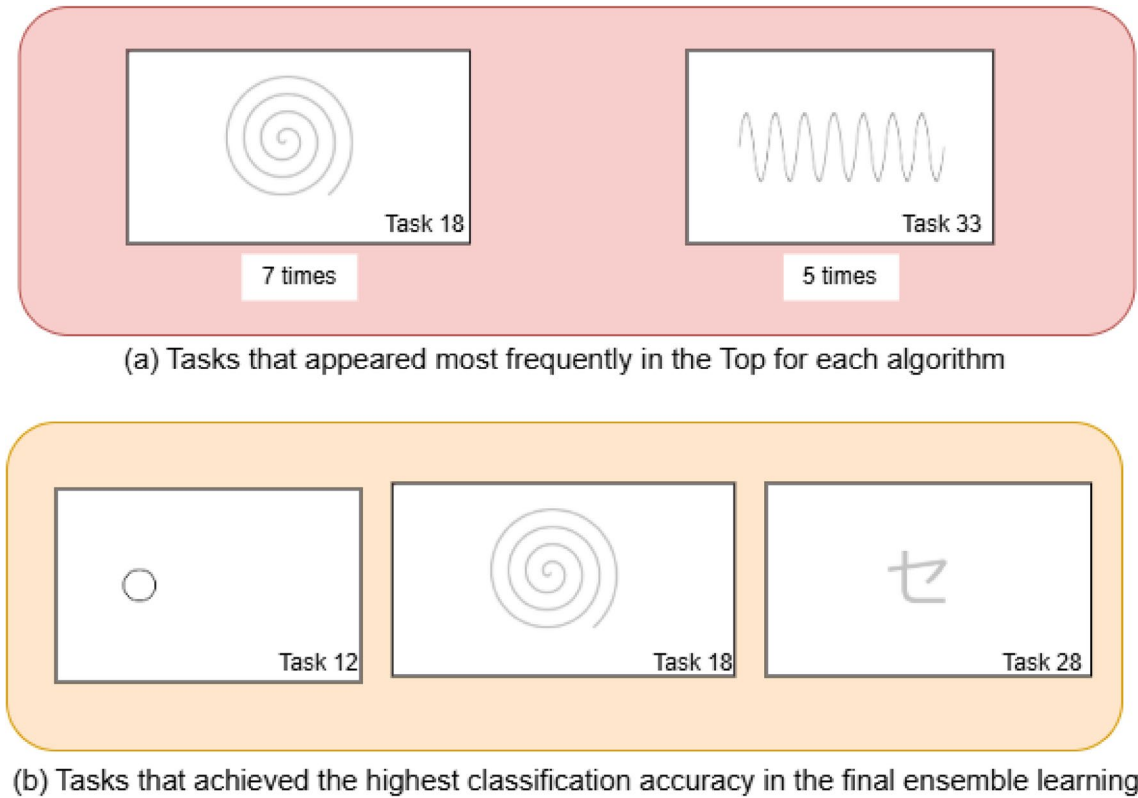


Fig. 10. Effective top tasks: (a) Each algorithm and (b) the ensemble algorithm.

Final accuracy ranking			(continued)			(continued)		
Rank	Task	Accuracy (%)	Rank	Task	Accuracy (%)	Rank	Task	Accuracy (%)
1	12	100.00	13	10	98.46	25	32	96.05
1	18	100.00	14	11	98.43	26	8	95.93
1	28	100.00	15	7	98.33	27	17	95.89
4	2	99.29	16	26	98.08	28	5	95.43
5	13	99.17	17	30	98.05	29	15	95.39
6	19	99.13	18	1	97.66	30	3	95.36
6	35	99.13	19	29	97.10	31	20	95.16
8	34	99.05	20	16	96.89	32	4	94.47
8	36	99.05	21	9	96.89	33	14	94.46
10	31	99.00	22	6	96.76	34	24	94.29
11	27	98.95	23	21	96.34	35	25	94.17
11	33	98.95	24	37	96.20	36	22	93.36
						37	23	92.00

Table 12. Ranking of final result (accuracy score in percent format).

also medically interpretable, as they capture handwriting abnormalities associated with PD-related motor dysfunction.

Effectiveness of feature selection

As shown in Table 13, the mean classification accuracy was compared across different feature selection strategies. Overall, the highest mean accuracy was obtained when all features were used. In contrast, applying feature selection based on the U test, the correlation coefficient filter, or their combination slightly decreased the mean classification accuracy. These results indicate that feature selection did not improve the mean classification accuracy in the revised analysis. Therefore, removing features based only on statistical significance and inter-feature correlation did not necessarily lead to improved classification performance. However, feature selection still substantially reduced the number of input features. Although the degree of reduction varied depending on

Methodology of feature selection	Average accuracy (%)
All features	97.10
Common features across all tasks: U-test ($p < 0.01$)	89.09
Common features across all tasks: U-test ($p < 0.05$)	91.09
Common features across all tasks: U-test ($p < 0.10$)	91.37
Significant features for each task: U-test ($p < 0.01$)	96.13
Significant features for each task: U-test ($p < 0.05$)	96.66
Significant features for each task: U-test ($p < 0.10$)	96.66
Selected features for each task: Correlation Filter Only	96.75
Significant features for each task: U-test ($p < 0.01$) and Correlation Filter	96.11
Significant features for each task: U-test ($p < 0.05$) and Correlation Filter	96.56
Significant features for each task: U-test ($p < 0.10$) and Correlation Filter	96.68

Table 13. The feature selection methodologies and corresponding average of 37 task classification accuracy.

the task, the number of features was reduced to approximately one-half to one-third of the original feature set. Considering this substantial reduction in feature dimensionality, the decrease in mean classification accuracy was relatively small. Thus, these results suggest a trade-off between dimensionality reduction and classification performance. When common features with p -values less than 0.10 across all tasks were used, the average accuracy decreased about 5% compared with the task-specific feature selection approach. This suggests that using only common features across all tasks may be inadequate for effective classification. In other words, the HandPD37 dataset may contain task-specific feature characteristics, and the features contributing to classification may differ depending on the task.

Implications for medical practice

The proposed HandPD37-based system may have potential as a simple, non-invasive, and objective decision-support tool for PD assessment. In clinical practice, selected handwriting tasks could be performed on a tablet during routine hospital visits, rehabilitation assessment, or follow-up monitoring. By analyzing handwriting features such as writing speed, pen pressure, movement irregularity, tremor-related changes, and writing smoothness, the system may provide additional quantitative information related to fine motor impairment. However, this system is not intended to replace clinical diagnosis by medical specialists. The results should be interpreted together with neurological examination, patient history, medication status, disease progression, and other standard clinical assessments. Further validation using larger multi-center clinical datasets and direct comparison with established clinical diagnostic standards will be necessary before real clinical application

Conclusion

This study proposed an optimized methodology for PD detection that integrates newly developed dynamic kinematic-based features with advanced machine learning (ML) techniques to capture the movement dynamics during handwriting tasks. The primary aim of this work was to improve the early diagnosis of PD by leveraging a comprehensive handwriting HandPD37 dataset that contains 37 carefully designed tasks. These tasks were designed to assess both fine motor skills and linguistic complexity, providing a rich and diverse set of data that captures the subtle motor impairments characteristic of PD. We introduced a novel PD handwriting dataset, comprising handwriting samples from 132 subjects, including 78 PD patients and 54 healthy controls. The data was recorded using a Wacom Cintiq Pro 16 LCD tablet, providing time-series data with attributes such as pen pressure, X–Y coordinates, azimuth, and altitude. This precise data collection enabled the extraction of 171 features, including 21 kinematic features and 8 statistical features for each kinematic measure, offering deep insights into the dynamic aspects of handwriting and movement impairments in PD patients. For classification, we employed a stacking ensemble of 10 ML algorithms, utilizing LR as a meta-classifier to integrate the predictions from the individual models. This ensemble approach allowed for the combination of strengths from multiple models, leading to improved classification accuracy and robustness in distinguishing PD patients from healthy controls. The results demonstrated that kinematic and pressure-based handwriting features, when combined with ensemble learning, provide valuable insights for the early, non-invasive detection of PD. The tasks were ranked based on the performance of individual models and the ensemble approach, highlighting the most effective handwriting features for PD diagnosis. We found that Tasks 18, 28, 33, and 34 were the top effective tasks. This ranking helps identify the handwriting tasks that most significantly contribute to distinguishing PD patients from healthy controls. In conclusion, the proposed approach not only enhances the understanding of handwriting disturbances in PD patients but also offers a powerful, data-driven framework for early diagnosis. By combining dynamic kinematic features with ensemble ML models, we provide a robust solution that can contribute significantly to the detection and management of Parkinson’s Disease. Moving forward, we plan to expand the dataset with additional handwriting tasks, particularly those that can capture even finer tremor-related features. Furthermore, we aim to explore the clinical applicability of the proposed methodology in future studies, improving its accessibility and usability for healthcare professionals. Another area for future work is the integration of other sensor modalities, such as accelerometers and gyroscopes, to complement the handwriting data and provide a more comprehensive view of PD symptoms. In addition, we plan to incorporate image-

based handwriting analysis using the visual patterns of written traces, such as stroke shape, size variation, spatial deformation, tremor-induced irregularity, and overall writing legibility. Combining image-based features with dynamic tablet signals may further improve the robustness and clinical applicability of PD detection.

Data availability

Correspondence and requests for materials should be addressed to J.S. (<https://u-aizu.ac.jp/labs/is-pp/> pplab/shin.html; email: Jpshin@u-aizu.ac.jp). The dataset used and/or analyzed during the current study is available from the corresponding author upon reasonable request.

Received: 6 March 2026; Accepted: 15 July 2026

Published online: 01 October 2026

References

- De Lau, L. M. & Breteler, M. M. Epidemiology of Parkinson's disease. *Lancet Neurol.* **5**, 525–535 (2006).
- Shin, J., Miah, A. S. M., Kabir, M. H., Rahim, M. A. & Al Shiam, A. A methodological and structural review of hand gesture recognition across diverse data modalities. *IEEE Access.* (2024).
- Hassan, N., Miah, A. S. M., Okuyama, Y. & Shin, J. Neurological disorder recognition via comprehensive feature fusion by integrating deep learning and texture analysis. *IEEE Open J. Comput. Soc.* (2025).
- Matsumoto, M., Miah, A. S. M., Asai, N. & Shin, J. Machine learning-based differential diagnosis of Parkinson's disease using kinematic feature extraction and selection. *IEEE Access* **13**, 54090–54104. <https://doi.org/10.1109/ACCESS.2025.3553528> (2025).
- Miah, A. S. M., Hassan, N., Hossain, M. M. A., Okuyama, Y. & Shin, J. Multi class parkinsons disease detection based on finger tapping using attention-enhanced CNN BILSTM. *arXiv preprint arXiv:2510.10121* (2025).
- Von Campenhausen, S. et al. Prevalence and incidence of Parkinson's disease in Europe. *Eur. Neuropsychopharmacol.* **15**, 473–490 (2005).
- Contreras-Vidal, J. & Stelmach, G. E. Effects of Parkinsonism on motor control. *Life Sci.* **58**, 165–176 (1995).
- Hughes, A. J., Daniel, S. E., Ben-Shlomo, Y. & Lees, A. J. The accuracy of diagnosis of Parkinsonian syndromes in a specialist movement disorder service. *Brain* **125**, 861–870. <https://doi.org/10.1093/brain/awf080> (2002).
- Ammenwerth, E., Nykänen, P., Rigby, M. & de Keizer, N. Clinical decision support systems: need for evidence, need for evaluation (2013).
- Dreiseitl, S. & Binder, M. Do physicians value decision support? a look at the effect of decision support systems on physician opinion. *Artif. Intell. Med.* **33**, 25–30 (2005).
- Mattison, H. A., Stewart, T. & Zhang, J. Applying bioinformatics to proteomics: Is machine learning the answer to biomarker discovery for PD and MSA?. *Mov. Disord.* **27**, 1595–1597 (2012).
- Saunders-Pullman, R. et al. Validity of spiral analysis in early Parkinson's disease. *Mov. Disord. J. Mov. Disord. Soc.* **23**, 531–537 (2008).
- Broderick, M. P., Van Gemmert, A. W., Shill, H. A. & Stelmach, G. E. Hypometria and bradykinesia during drawing movements in individuals with Parkinson's disease. *Exp. Brain Res.* **197**, 223–233 (2009).
- Bidet-Ildei, C., Pollak, P., Kandel, S., Fraix, V. & Orliaguet, J.-P. Handwriting in patients with Parkinson disease: effect of l-dopa and stimulation of the sub-thalamic nucleus on motor anticipation. *Hum. Mov. Sci.* **30**, 783–791 (2011).
- Contreras-Vidal, J. L., Poluha, P., Teulings, H.-L. & Stelmach, G. E. Neural dynamics of short and medium-term motor control effects of levodopa therapy in Parkinson's disease. *Artif. Intell. Med.* **13**, 57–79 (1998).
- Drotár, P. A new modality for quantitative evaluation of Parkinson's disease: In-air movement. In *13th IEEE International Conference on Bioinformatics and Bioengineering* 1–4 (IEEE, 2013).
- Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I. & Schlesinger, I. Handwriting as an objective tool for Parkinson's disease diagnosis. *J. Neurol.* **260**, 2357–2361 (2013).
- Ünlü, A., Brause, R. & Krakow, K. Handwriting analysis for diagnosis and prognosis of Parkinson's disease. In *International Symposium on Biological and Medical Data Analysis*, 441–450 (Springer, 2006).
- Hughes, A. J., Daniel, S. E., Ben-Shlomo, Y. & Lees, A. J. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain* **125**, 861–870 (2002).
- Kraus, P. H. & Hoffmann, A. Spiralometry: computerized assessment of tremor amplitude on the basis of spiral drawing. *Mov. Disord.* **25**, 2164–2170 (2010).
- Stanley, K. et al. Digitized spiral analysis is a promising early motor marker for Parkinson disease. *Parkinson. Relat. Disord.* **16**, 233 (2010).
- Teulings, H.-L., Contreras-Vidal, J. L., Stelmach, G. E. & Adler, C. H. Parkinsonism reduces coordination of fingers, wrist, and arm in fine motor control. *Exp. Neurol.* **146**, 159–170 (1997).
- Sesa-Nogueras, E., Faundez-Zanuy, M. & Mekyska, J. An information analysis of in-air and on-surface trajectories in online handwriting. *Cogn. Comput.* **4**, 195–205 (2012).
- Pereira, C. R., Weber, S. A., Hook, C., Rosa, G. H. & Papa, J. P. Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics. In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–346 (IEEE, 2016).
- Shrivastava, A., Chakkaravarthy, M. & Asif Shah, M. A novel approach using learning algorithm for Parkinson's disease detection with handwritten sketches. *Cybern. Syst.* **55**, 2388–2404 (2024).
- Drotár, P. et al. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artif. Intell. Med.* **67**, 39–46 (2016).
- Shin, J., Miah, A. S. M., Hirooka, K., Hasan, M. A. M. & Maniruzzaman, M. Parkinson disease detection based on in-air dynamics feature extraction and selection using machine learning. *Sci. Rep.* **15**, 28027 (2025).
- Miah, A. S. M., Islam, M. R. & Molla, M. K. I. Motor imagery classification using subband tangent space mapping. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1–5 (IEEE, 2017).
- Miah, A. S. M., Rahim, M. A. & Shin, J. Motor-imagery classification using Riemannian geometry with median absolute deviation. *Electronics* **9**, 1584 (2020).
- Miah, A. S. M. et al. Motor-imagery bci task classification using riemannian geometry and averaging with mean absolute deviation. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1–7 (IEEE, 2019).

Author contributions

J.S., A.S.M.M., K.H., A.H., and M.A.M.H. contributed to conceptualization and project development. J.S., A.S.M.M., A.H., K.H., and N.H. prepared the original draft. J.S., M.A.M.H., A.H., N.H., M.A., R.T., and S.K. contributed to review and editing. J.S. and S.K. supervised the study. J.S. acquired funding. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

Declarations

Ethics approval and consent to participate

This study involved handwriting data collected from human participants. All procedures were reviewed and approved by the Ethics Review Committee of The University of Aizu, Japan (Approval No. 92, June 2021). All experiments were conducted in accordance with the ethical standards of the institutional research committee and with the Declaration of Helsinki. Informed consent was obtained from all participants involved in the study.

Competing Interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.S. or S.K.

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) 2026