TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Diana Kuntsmann 233322IAIB
Sofia Loginova 233324IAIB

# Explainable Artificial Intelligence Analysis of Drawing Tests for Assessment of Cognitive State

Bachelor's Thesis

Supervisor: Sven Nõmm
PhD

Client: Aaro Toomela
PhD

Tallinn 2026

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Diana Kuntsmann 233322IAIB
Sofia Loginova 233324IAIB

# Selgitatava tehisintellekti analüüs joonistustestidest kognitiivse seisundi hindamiseks

Bakalaureusetöö

Juhendaja: Sven Nõmm
PhD

Klient: Aaro Toomela
PhD

Tallinn 2026

## Author's declaration of originality

We hereby certify that we are the sole authors of this thesis and that this thesis has not been presented for examination or submitted for defense anywhere else. All used materials, references to the literature, and work of others have been cited.

Authors: Diana Kuntsmann and Sofia Loginova

01.06.2026

# Abstract

This thesis investigates whether fine motor patterns affected by Parkinson's disease share measurable kinematic similarities with motor patterns of individuals who lack formal education and thus writing practice. Both groups exhibit movement slowness and irregularity, arising from fundamentally different mechanisms: progressive neurodegeneration in one case, and underdeveloped fine motor skills in the other.

Two digitized drawing datasets are used: DraWritePD, containing data from 17 PD (Parkinson's disease patients) and 12 HC (healthy controls) collected in Estonia, and DraWriteEduUniPampa, containing data from 51 ILL (illiterate) and 33 LIT (literate) individuals from Brazil. Both datasets contain recordings of Luria's alternating series tracing task. Since the datasets were collected with different devices, the data was harmonized. 97 features were extracted from each recording to perform feature selection and train binary and multiclass classifiers. Eight types of machine learning models were trained using nested cross-validation, and SHAP (SHapley Additive exPlanations) was applied to explain predictions. The best HC vs PD model achieved F1-score of 0.885, with velocity-based features as the primary discriminators. The ILL vs LIT task proved more challenging, reaching F1-score of 0.707. The cross-domain experiment supported the hypothesis: PD patients were classified mostly as illiterate by the ILL vs LIT model, and illiterate individuals were misclassified as PD more often than literate individuals. Multiclass experiments confirmed that PD remains consistently separable, while the illiterate group shows partial overlap with both healthy and pathological groups. In conclusion, results support the hypothesis that the kinematic overlap between pathological disease and lack of education exists and is detectable with machine learning.

The client for the Bachelor's thesis is Aaro Toomela, a professor of cultural and neuropsychology at Tallinn University, who proposed the central hypothesis examined in this thesis.

The thesis is in English and contains 37 pages of text, 8 chapters, 15 figures, 29 tables.

# Lühikokkuvõte

## Selgitatava tehisintellekti analüüs joonistustestidest kognitiivse seisundi hindamiseks

Käesolev töö uurib hüpoteesi, et Parkinsoni tõve poolt mõjutatud peenmotoorsed mustrid jagavad mõõdetavaid kinemaatilisi sarnasusi nende inimeste liigutusmallidega, kellel puudub formaalne haridus ja seeläbi ka kirjutamispraktika. Mõlemas grupis esineb liigutuste aeglustumine ja ebaühtlus, kuigi erinevatel põhjustel: ühel juhul on tegemist neurodegeneratiivse haigusega, teisel juhul arenguperioodil omandamata jäänud peenmotoorikaga.

Töös kasutatakse kahte digitaliseeritud joonistusandmestikku: *DraWritePD*, mis sisaldab 17 PD (*Parkinson's disease*, Parkinsoni tõbi) patsientide ja 12 HC (*healthy controls*, tervete kontrollisikute) andmeid Eestist, ning *DraWriteEduUniPampa*, mis koosneb 51 ILL (*illiterat*e, kirjaoskamatute) ja 33 LIT (*literate*, kirjaoskajate) isikute andmetest Brasiiliast. Mõlemad andmestikud sisaldavad Luria vahelduvate seeriate testide salvestused ($\Pi\Lambda$ tüüpi jälgimisülesanne), mis hindab peenmotoorika sooritust ilma planeerimiskomponentita. Kuna andmestikud koguti erinevate seadmetega, andmeid harmoniseeriti. Masinõppemudelite treenimisel ja tunnuste valikul rakendati *nested cross-validation* ja SHAP (*SHapley Additive exPlanations*) rakendati ennustuste selgitamiseks.

Tulemused kinnitavad, et binaarsed klassifikaatorid eristavad Parkinsoni tõvega patsiente kontrollrühmast suure täpsusega (parima mudeli F1-skoor 0,885), kusjuures tähtsad tunnused on kiiruspõhised tunnused. Kirjaoskamatute ja kirjaoskajate eristamine osutus keerukamaks (F1-skoor 0,707), tuginedes peamiselt pliiatsi surve dünaamikale. Valdkonnaülene eksperiment (*cross-domain experiment*), kus ühel andmestikul treenitud mudelid rakendati teise andmestiku peale, kinnitas hüpoteesi: märkimisväärne osa hariduseta indiviididest klassifitseeriti mudelite poolt Parkinsoni tõvega patsientideks ja vastupidi. SHAP selgitas, et mõlema grupi puhul on ennustuste peamiseks ajendiks just madal joonistamiskiirus ja ebaühtlane surve. Mitmeklassilised katsed kinnitasid, et PD jääb järjepidevalt eraldatavaks, samas kui kirjaoskamatute rühm kattub osaliselt nii tervete kui

ka patoloogiliste rühmadega.

Kokkuvõttes saadud tulemused toetavad hüpoteesi, et kinemaatiline kattuvus patoloogilise haiguse ja hariduse puudumise vahel on reaalne ning masinõppega tuvastatav.

Selle uurimise tellija on Aaro Toomela, Tallinna Ülikooli kultuuri- ja neuropsühholoogia professor, kes esitas käesolevas töös uuritava keskse hüpoteesi.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 37 leheküljel, 8 peatükki, 15 joonist, 29 tabelit.

# List of abbreviations and terms

| | |
|---|---|
| AB | AdaBoost |
| AST | Luria's Alternating Series Test |
| CSV | Comma-Separated Values |
| DT | Decision Tree |
| FDR | False Discovery Rate |
| HC | Healthy control |
| ILL | Illiterate |
| JSON | JavaScript Object Notation |
| KNN | K-Nearest Neighbours |
| LIME | Local Interpretable Model-Agnostic Explanations |
| LIT | Literate |
| LR | Logistic Regression |
| ML | Machine Learning |
| NB | Naive Bayes |
| PD | Parkinson's disease |
| pltrace | Luria's alternating ΠΛ series drawing task |
| RF | Random Forest |
| RQ | Research Question |
| SHAP | SHapley Additive exPlanations |
| SVM | Support Vector Machine |
| UMAP | Uniform Manifold Approximation and Projection |
| XAI | Explainable Artificial Intelligence |
| XGB | XGBoost |

# Table of contents

# List of figures

# List of tables

# 1 Introduction

PD (Parkinson's disease) is one of the most common neurodegenerative disorders. Progressive loss of fine motor skills leads to the development of core motor symptoms, which are rest tremor, bradykinesia (slowness of movement), rigidity, and micrographia (decreased handwriting size) [1]. When diagnosed in the early stages, PD can be controlled with medication and physical therapy, helping to preserve the patient's quality of life [2]. However, the diagnosis of PD remains a complex and subjective process: it depends on the visual evaluation and the expertise of the clinician performing the analysis; the disease does not have specific biomarkers and its manifestations can vary from one patient to another. Furthermore, early symptoms of PD often overlap with those of other conditions that produce similar motor abnormalities [1].

Today, digitization of clinical data has introduced new possibilities for assessing motor and cognitive states using drawing tests. Digital tablets equipped with styluses offer a promising alternative to classical paper-and-pencil assessments, as they enable the acquisition of dynamics of the drawing process. That is, capturing not only the pen's trajectory but also its dynamic parameters such as velocity, applied pressure, pen orientation, etc., providing crucial kinematic information that is invisible to the naked eye [3].

As a result, a significant number of studies appeared that have focused on distinguishing patients with PD from HC (healthy controls) using features extracted from digitized drawing and handwriting tasks [3], [4], [5], [6], [7], [8]. However, one of the main challenges of automated detection of PD is that irregular fine motor patterns are not unique to the disease. Similar irregularities can arise for a range of other reasons: physiological aging, essential tremor, drug-induced parkinsonism, fatigue, or lack of formal education and writing practice [1], [9], [10].

Although previous studies have shown that digital drawing tasks can distinguish PD from HC with reasonable accuracy [3], [8], comparisons between PD motor patterns and

non-pathological sources of motor irregularity have received little attention. However, for these research results to be applicable in clinical practice, it is necessary to move toward differential analysis, the ability to distinguish not just PD from HC, but PD from the full spectrum of conditions and circumstances that produce similar motor patterns.

## 1.1 Problem Statement

Pen trajectories produced by PD patients and by individuals without formal education appear visually similar [11]. Both groups tend to draw slowly, with irregularities and less smooth directional changes. Despite arising from entirely distinct mechanisms, progressive neurodegeneration in one case, and underdeveloped fine motor skills in the other, the kinematic features appear to overlap. This observation finds a theoretical base in the concept of motor reserve [9]. The idea is that the more a person practises fine motor activities such as writing, the more efficiently the brain learns to execute movement, creating reserve capacity that can partially compensate for neurological damage. Studies have shown that among Parkinson's disease patients with the same degree of brain damage, those with lower levels of education are more likely to have more severe movement disorders [9], suggesting that education contributes to this reserve. Therefore, pen movements of people who lack formal education may, at a kinematic level, resemble the motor patterns observed in PD, even in the complete absence of neurological disease. The hypothesis of this thesis is thus that fine motor patterns affected by progressive Parkinson's disease share measurable kinematic similarities with those underdeveloped due to the absence of formal education.

Based on this hypothesis, the following RQ (research questions) are formulated:

- RQ1: Which kinematic features are most discriminative in distinguishing PD from HC and ILL from LIT?
- RQ2: To what extent do the motor patterns of PD patients overlap with those of illiterate individuals at the kinematic level?
- RQ3: Which kinematic features drive the model decisions in each classification task, and do the same features contribute to predictions across different domains?
- RQ4: How separable are the motor patterns of HC, PD, ILL, and LIT groups from one another, and which groups show the greatest kinematic overlap?

To investigate these research questions using ML (machine learning) methods, the following objectives are defined:

1. To extract a set of tremor-related features from the digitized AST (Luria's alternating series test) drawing task for both groups.
2. To train ML models to distinguish PD patients from HC and ILL (illiterate) from LIT (literate) individuals, establishing classification performance for each domain and identifying the kinematic features most relevant to each task.
3. To conduct cross-domain experiment: ML models trained on one domain are applied to samples from the other. If the hypothesis holds, illiterate individuals would likely be systematically misclassified as PD, and PD individuals as illiterate.
4. To train multiclass ML models (three- and four-class) on all combinations of the four groups using feature sets from the binary tasks, to assess group separability and analyse misclassification patterns across groups.
5. To use SHAP (SHapley Additive exPlanations) to provide interpretable explanations of model decisions, showing how selected features contribute to predictions.

The complete workflow is visually summarized in Figure 1.

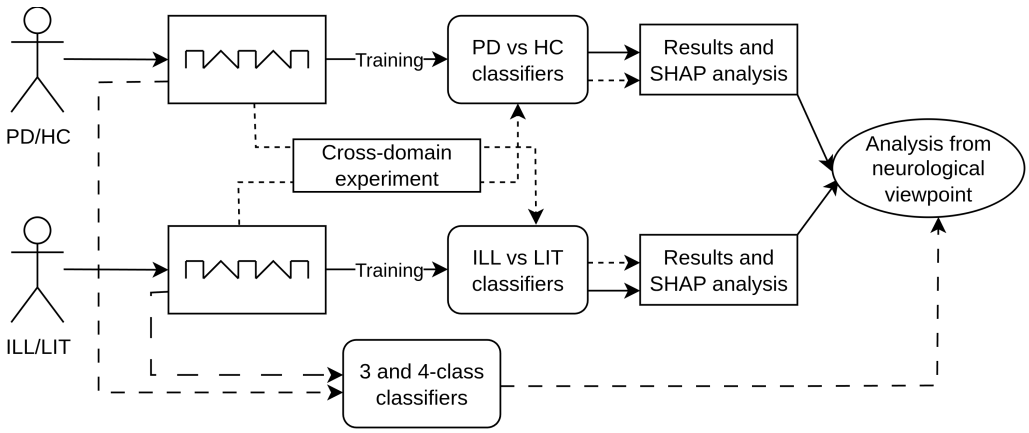



Figure 1. Workflow of experiments.

The thesis is structured as follows. Chapter 2 provides background information on Parkinson's disease, fine motor tests, and Luria's alternating series test. Chapter 3 describes the datasets and the preprocessing pipeline. Chapter 4 presents the methodology. Chapter 5 reports the experimental results. Chapter 6 discusses the findings and limitations. Chapter 7 discusses directions for future research. Finally, Chapter 8 summarizes the thesis.

# 2 Background

This chapter provides background information needed to understand the motivation behind this thesis. Section 2.1 gives an overview of Parkinson's disease, its symptoms, and the challenges of its diagnosis. Section 2.2 describes Luria's alternating series test and its role in motor assessment.

## 2.1 Parkinson's Disease

Parkinson's disease is a neurodegenerative disorder in which neurons in the central nervous system die or are damaged, causing severe disabilities [12]. It is one of the fastest-growing neurodegenerative disorders worldwide: the number of affected individuals increased from 2.5 million in 1990 to 11.77 million in 2021, with projections indicating a further increase to 17.27 million by 2035 [13]. This growth is largely driven by the aging of global population, because age is the greatest risk factor for PD. Prevalence and incidence increase nearly exponentially after the age of 60 [2], [12], [14].

The causes of PD remain largely unknown. Literature on this topic cites possible causes such as a combination of genetic and environmental factors, particularly in individuals with a hereditary predisposition. Identified risk factors include exposure to pesticides, herbicides, and industrial chemicals [13].

Since no definitive biomarker test exists for this disease, diagnosis heavily depends on the expertise of the clinician and remains a subjective process [1], [12]. The main condition for diagnosis is the presence of at least two of the following motor symptoms: rest tremor, bradykinesia, rigidity, or postural imbalance. Also other causes of parkinsonism must be excluded before confirming a diagnosis of PD [13]. In addition to motor symptoms, PD is also associated with non-motor symptoms such as cognitive impairment, depression, sleep disturbances, and loss of smell [1], [12]. All of these symptoms affect the patient's quality of life. While drug treatment can help manage symptoms in the early stages, its effectiveness typically decreases as the disease progresses, so early diagnosis is crucial [2].

An additional challenge in diagnosis is that tremor associated with Parkinson's disease is not unique and can be mistaken for other types of tremor. Essential tremor, drug-induced parkinsonism, and physiological tremor can all produce movement patterns very similar to those seen in patients with Parkinson's disease, making differential diagnosis particularly challenging [1]. Furthermore, a patient's high motor reserve can mask early signs of the disease, delaying detection [9]. However, studies have shown that automated analysis of writing and drawing tasks can help distinguish PD from other tremor-causing conditions with promising results [10].

This highlights the need for more objective and standardised assessment tools to support the diagnostic process.

## 2.2 Luria's Alternating Series Test

In 1966 Alexander Luria published a simple clinical battery, where he described twelve methods to investigate graphomotor functions [15]. One of these methods is AST, in which the subject is asked to draw a pattern of alternating squares and triangles. It is universal and can be easily used worldwide, because this test is a purely graphomotor task and does not involve linguistic or script-specific knowledge, which makes it popular among researchers [16], [17], [18]. For this thesis, the use of such a test is especially important because the study involves participants from two different countries: Brazil and Estonia.

AST is targeted to assess the state of planning and execution functions of fine motor motions [8]. Planning refers to the ability to organise and sequence a movement before performing it, while execution refers to the ability to accurately carry out the movement itself.

Initially, three types of tests were introduced: copying, tracing and continuing a periodic pattern. The test requiring to copy the series is targeted on assessment of the state of planning function. The tracing test assesses the execution function. For the test requiring to continue the line only part of the series should be shown and both planning and execution functions are tested [8]. Furthermore, three types of series were suggested: $\Pi$-type, $\Pi\Lambda$-type and Sinusoidal line, shown in Figure 2.

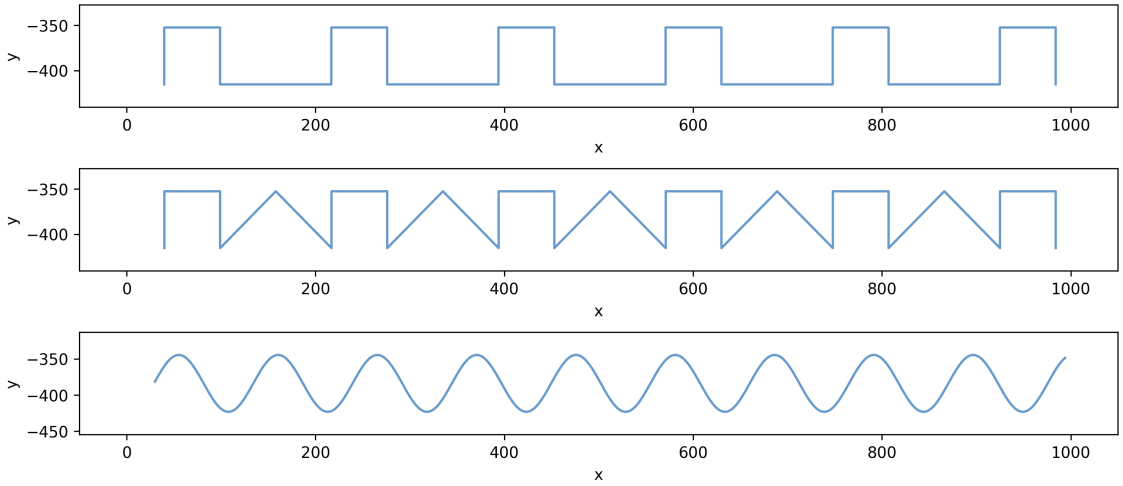



Figure 2. Examples of different series types.

In this thesis, only the pltrace ($\Pi\Lambda$) type of test is considered, as it isolates the motor execution component by removing the planning component, allowing for a more direct comparison of motor functions across all studied groups.

# 3 Data

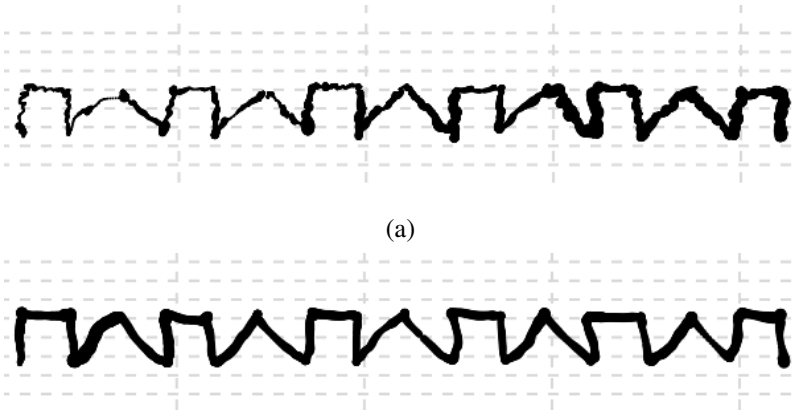
This chapter describes the data and preprocessing pipeline used in this thesis. Section 3.1 introduces datasets, including their acquisition setup and group compositions. Section 3.2 describes the preprocessing steps applied to each dataset. Section 3.3 describes the technical differences between the two acquisition devices. Section 3.4 presents the harmonisation pipeline used to resolve these differences. Section 3.5 summarizes the configurations of the dataset used in the experiments.

## 3.1 Datasets

In this thesis, two digitized drawing datasets are considered: DraWritePD [8] and DraWriteEduUniPampa [19]. Both contain recordings of 12 different drawing and handwriting tasks, acquired on digital tablets using a stylus. In this thesis only Luria's alternating series drawing task pltrace is considered. Raw data from both datasets are in JSON (JavaScript Object Notation) format, where each file contains metadata (session id, test type, anonymous identification number) and an array of recorded points. Each point is described by six parameters: x-coordinate (`x`), y-coordinate (`y`), timestamp (`t`), pressure (`p`), altitude (`a`), and azimuth (`l`). The participants ranged in age from 40 to 80 years, with a balanced gender distribution across all groups.

### 3.1.1 DraWritePD

The DraWritePD dataset was collected in Tartu, Estonia using an Apple iPad Pro (2016) tablet and an Apple Pencil stylus. Data were recorded using a custom iOS application developed at Tallinn University of Technology [20]. The dataset contains 12 HC and 17 patients diagnosed with PD. Data collection was approved by the Research Ethics Committee of the University of Tartu (No. 327/T-9). Examples of pltrace drawings from the PD and HC groups are shown in Figure 3.



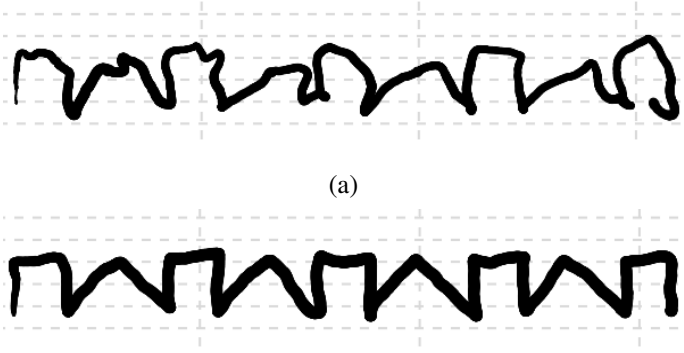
(a)

(b)

Figure 3. Pltrace drawn by a PD (a) and a HC (b).

### 3.1.2 DraWriteEduUniPampa

The DraWriteEduUniPampa dataset was collected in the Pampa region of Brazil using a Microsoft Surface (2018) tablet and a Surface Pen stylus. Data were recorded using an application developed also at Tallinn University of Technology [21]. The dataset contains 33 LIT and 51 ILL individuals, none of whom have a neurodegenerative diagnosis. LIT individuals have completed higher education, while ILL individuals have received no formal education at all. Data collection was approved by Universidade da regiao de Joinville (UNIVILLE) nr. 4.787.687 on June 17, 2021. Examples of pltrace drawings from the ILL and LIT groups are shown in Figure 4.



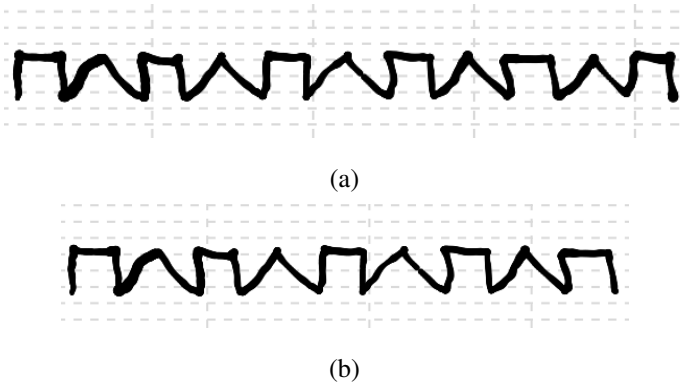
(a)

(b)

Figure 4. Pltrace drawn by an ILL (a) and a LIT (b).

## 3.2 Data Preprocessing

In the DraWritePD dataset, the data collection protocol slightly differed from DraWriteEduUniPampa: the tracing template included one additional $\Pi\Lambda$ segment. To ensure consistency between the datasets, this final segment was manually removed from each file in the DraWritePD dataset. This was done using a custom interactive visualization tool that allows selecting and deleting specific points. Figure 5 illustrates an example of an HC pltrace drawing before and after the removal of the extra $\Pi\Lambda$ segment.



(a)

(b)

Figure 5. Example of a HC pltrace drawing before (a) and after (b) removal of the additional $\Pi\Lambda$ segment.

## 3.3 Device Differences

The two datasets were collected on separate devices. To perform a cross-dataset analysis, the following technical differences had to be addressed:

- The iPad records data in millimeters, while the Microsoft Surface uses screen pixels.
- The iPad outputs abstract force values ranging from $0.0$ to $6.0$, whereas the Surface normalizes pressure between $0.0$ and $1.0$.
- The iPad relies on the `NSDate` format (seconds since Jan 1, 2001), while the Surface uses the Unix format (seconds since Jan 1, 1970).
- The iPad operates at a rate of 240 points per second, but due to burst delivery of timestamps its effective rate is approximately 60 points per second. The Surface records at approximately 200 points per second.

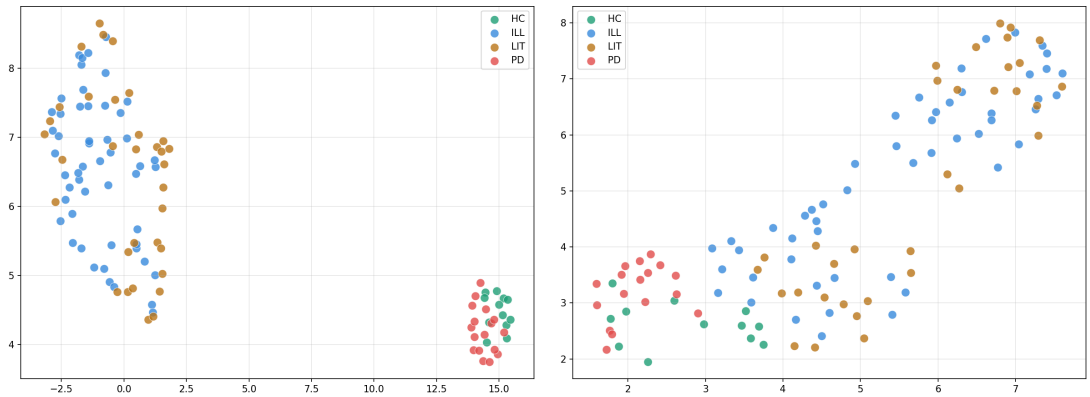
## 3.4 Data Harmonisation

To resolve the hardware and software differences described above, a data harmonisation pipeline was developed. This pipeline processes the raw files and transforms them into a unified format using the following steps:

1. All timestamps are converted into milliseconds, starting from zero for the first point of each drawing.
2. Pressure values are rescaled to a range of $[0.0, 1.0]$. To reduce sensitivity to random hardware spikes, the 5th and 95th percentiles of pressure are computed across the entire dataset for each device.
3. All recordings are resampled to 120 points per second, chosen as a compromise between the sampling rates of the iPad (60 Hz) and the Surface (200 Hz).
4. To account for different screen sizes and coordinate units, every drawing is moved to the center of a standard abstract canvas. The coordinates are then scaled so that all drawings have the same relative size regardless of the device.

The harmonised data are saved as CSV (Comma-Separated Values) files.

The effect of harmonisation on the feature space is illustrated in Figure 6 using UMAP (Uniform Manifold Approximation and Projection) [22] dimensionality reduction technique. Before harmonization, UMAP projects the data into two clearly separated clusters corresponding to the two devices, with groups intermixed within each cluster. After harmonisation, the device-driven separation is reduced.




(a) (b)

Figure 6. UMAP projections of all features before (a) and after (b) harmonisation.

## 3.5 Dataset Summary

The experiments in this thesis use three configurations of the data. For the binary classification task HC vs PD, all available subjects were used. For the binary classification task ILL vs LIT, a balanced subset was constructed by randomly sampling 33 subjects from the ILL group to match the number of LIT subjects. For three- and four-class classification tasks, a further balanced subset was constructed by randomly sampling 15 subjects from both the ILL and LIT groups, while retaining all HC and PD subjects. The resulting configurations are summarised in Table 1.

Table 1. Dataset configurations used in experiments.

| **Group** | **HC vs PD** | **ILL vs LIT** | **3/4-class** |
|---|---|---|---|
| HC | 12 | — | 12 |
| PD | 17 | — | 17 |
| LIT | — | 33 | 15 |
| ILL | — | 33 | 15 |

# 4 Methodology

This chapter describes the methodology used in this thesis. Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 present the theoretical background of the chosen methods and describe practical implementation of proposed methods. The proposed pipelines were implemented as a Python 3 code with the use of libraries such as: `pandas`, `scikit-learn` and `NumPy`.

## 4.1 Fisher's Score

The Fisher's score measures the ratio of the average interclass separation to the average intraclass separation. The larger the Fisher score, the greater the discriminatory power of the feature [23]. In this thesis Fisher's score was used as a filter-based feature selection method for model training to find features with the most discriminative power and filter out irrelevant features within nested cross-validation. It was chosen because of its popularity among research papers related to this topic [3] [8].

The formula for calculating Fisher's score is the following [23]:

$$F = \frac{\sum_{j=1}^{k} p_j(\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2} \tag{4.1}$$

where $k$ - number of classes, $p_j$ - the fraction of samples belonging to class $j$, $\mu_j$ - the mean of the feature value in class $j$, $\sigma_j^2$ is the variance of the feature value in class $j$, $\mu$ - the global mean of the feature being evaluated.

## 4.2 Pearson's Correlation Coefficient

Since Fisher's score does not account for interactions between features, a correlation filter was additionally applied to ensure that models are trained on diverse and non-redundant information. Therefore, Pearson's correlation coefficient was chosen as a measure of linear dependence between two features and was used as additional filter feature selection method.

Pearson's correlation between two variables X and Y is calculated as [23]:

$$\rho = \frac{E[X \cdot Y] - E[X] \cdot E[Y]}{\sigma(X) \cdot \sigma(Y)} \tag{4.2}$$

where $E[X]$ and $E[Y]$ are the expected values of variables $X$ and $Y$, $E[X \cdot Y]$ is the expected value of their product, and $\sigma(X)$ and $\sigma(Y)$ denote the standard deviations of $X$ and $Y$.

## 4.3 Classification Models

To evaluate and compare classification performance across the discrimination tasks, the following ML algorithms were trained: AdaBoost (AB), Decision Tree (DT), K-Nearest Neighbours (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB) [23], [24], [25]. These models were chosen to represent a diverse range of classification approaches: linear, tree-based, ensemble, distance-based, and probabilistic, in order to compare their performances on the same data and identify the most suitable approach for the given tasks.

## 4.4 Feature Extraction and Selection

The same feature set as in [3] was used in this thesis, as that study was focused on extracting features describing tremor and motor impairments. However, angular features such as azimuth and altitude were excluded, as they are rarely used in similar studies and the results obtained using these features would not have had a direct point of comparison. In total, 97 features were extracted from the following recorded signals: pen position (x- and y-coordinates), timestamp, and pen pressure. The extracted features can be grouped into two categories:

- Kinematic features: velocity and its derivatives - acceleration, jerk, snap, crackle, and pop
- Pressure-based features: pressure and its derivatives - yank, tug, snatch and shake

For each feature, five statistical measures were computed: mean, median, standard deviation, minimum, and maximum. Additionally, motion mass was computed for each feature, which was introduced in [26] to describe the amount and smoothness of motion and is defined as the sum of absolute values across all observation points.

The full list of features is provided in Appendix 2.

The feature selection process for binary classification task took place in the inner loop of nested cross-validation. All features were ranked by their Fisher score in descending order, and any feature with a correlation coefficient higher than 0.7 with an already selected feature was excluded, ensuring that among correlated features, the one with the least discriminatory power was removed. Then, the $k$ best remaining features were selected based on their Fisher score. Each model was trained four times, once for each value of $k$, where $k$ is the number of selected features, $k \in \{2, 3, 4, 5\}$. This range was chosen to keep the number of features small relative to the dataset size and to reduce the risk of overfitting.

## 4.5 Nested Cross-Validation

In [3], the authors compare nested cross-validation and non-nested cross-validation for a feature selection task and prove that non-nested approach leads to data leakage and overoptimistic training results. The nested cross-validation described in [25] was used in this thesis. In the outer loop, the dataset is split into three folds, where in each iteration two folds are used for training and one is held out for testing. Within each outer training fold, an inner cross-validation loop is used to perform feature selection and hyperparameter tuning using grid search. The final model is then evaluated on the held-out outer test fold. Nested cross-validation was used to obtain accurate training results and since the datasets considered in this work are small, this approach allows the use of the entire dataset for training. All metrics are reported as mean ± standard deviation across the outer cross-validation folds. The overall experimental workflow, starting from initial data loading and feature extraction to the feature selection and model training, is summarized in Figure 7.

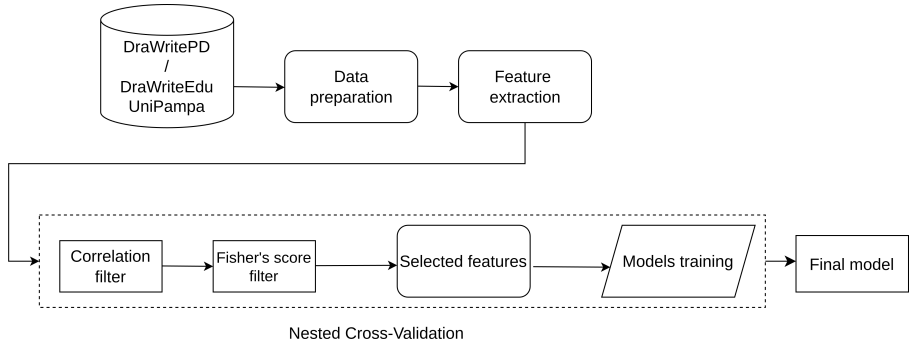



Figure 7. Experimental pipeline for feature selection and model training.

## 4.6 Explainable Artificial Intelligence

In the context of this study, it is extremely important to understand on what basis the model made its decisions: which features had the most significant impact on the classification. For this thesis, SHAP [27] was selected to interpret model predictions. It estimates the contribution of each feature to the model output, using the concept of Shapley values from cooperative game theory, where features are treated as players and the contributions are fairly distributed [28]. SHAP is a model-agnostic method, i.e. it can be used for any classifier, and allows for both local explanations of individual predictions and global explanations of the overall model behaviour.

LIME (Local Interpretable Model-Agnostic Explanations) [29] is another widely used approach for interpreting model predictions, which works by approximating the behaviour of a complex model locally around a specific prediction using a simpler, interpretable model. However, unlike SHAP, LIME only supports explanations of individual predictions, which is why it was not selected for this study.

The SHAP library [30] provides a variety of visualisations based on SHAP values. In this thesis, three types are used: beeswarm plot, cohort bar plot and decision plot.

- The beeswarm plot displays the distribution of SHAP values for each feature across all samples. Each point represents one sample: its position on the x-axis shows the contribution of the feature to the prediction (positive values push toward class 1, negative toward class 0), and its colour reflects the feature value (red for high, blue for low). Features are ordered by their mean absolute SHAP value, with the most influential feature at the top.
- The cohort bar plot shows the mean absolute SHAP value per feature for each group separately. Each bar represents the average magnitude of a feature's contribution for one group, allowing a direct comparison of feature importance between classes.
- The decision plot shows how the model reaches its final prediction for each individual sample. Each sample is represented by a separate line, starting at the expected model output (baseline) and built by sequentially adding each feature's SHAP value, shifting the line left or right depending on the sign and magnitude of the contribution. The final position of the line on the x-axis corresponds to the model's output for that

sample. Features are ordered by their mean absolute SHAP value, with the most influential feature at the top.

The beeswarm plot was chosen because it shows the influence of each feature on predictions in detail. The cohort bar plot enables a direct comparison of feature importance between the two groups. The decision plot was chosen to observe how the model handles individual samples in general, and whether different groups' predictions follow distinct patterns.

## 4.7 Cross-domain Experiment and Statistical Analysis

To investigate whether fine motor patterns affected by progressive PD share measurable kinematic similarities with those underdeveloped due to the absence of formal education, a cross-domain experiment was conducted. In particular, models trained to distinguish one pair of groups were applied to samples from the other domain: HC vs PD models classified samples from DraWriteEduUniPampa dataset, while ILL vs LIT models classified samples from DraWritePD dataset. To statistically control the cross-domain classification patterns, one-sided Fisher's Exact Test was applied. Fisher's Exact Test is a non-parametric test used to assess whether the predicted class label is independent of the true group and it computes an exact $p$-value rather than an approximation, making it appropriate when sample sizes are small [31].

The test operates on a $2 \times 2$ contingency table, where rows represent the two true groups and columns represent the two possible predicted classes. The effect size is quantified using the odds ratio $\theta$, defined as:

$$\hat{\theta} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} \tag{4.3}$$

where $n_{11}$ and $n_{12}$ are the counts of the first group predicted as class 1 and class 0 respectively, and $n_{21}$, $n_{22}$ are the corresponding counts for the second group. The null hypothesis corresponds to $\theta = 1$, which means that both groups are equally likely to be classified as class 1. The alternative hypothesis is $\theta > 1$, meaning that the first group is classified as class 1 more often than the second [32].

In this experiment, for the HC vs PD model applied to the ILL and LIT groups, the first group is ILL and class 1 is PD, so the alternative hypothesis states that ILL subjects are classified as PD more often than LIT subjects, i.e. $P(\hat{y} = \text{PD} \mid y = \text{ILL}) > P(\hat{y} = \text{PD} \mid y = \text{LIT})$.

For the ILL vs LIT model applied to the HC and PD groups, the first group is PD and class 1 is ILL, so the alternative hypothesis states that PD patients are classified as ILL more often than HC subjects, i.e. $P(\hat{y} = \text{ILL} \mid y = \text{PD}) > P(\hat{y} = \text{ILL} \mid y = \text{HC})$, where $\hat{y}$ denotes the predicted class label and $y$ the true group label.

The test requires the following assumptions to be satisfied [31]:

- *Both variables are categorical and binary.* In this study, the two variables are the true group label (ILL or LIT, HC or PD) and the predicted class (PD or ILL). Both are binary.
- *Independence of observations.* Since the two datasets were collected independently and no subject appears in both groups, observations are independent.
- *Mutually exclusive groups.* A subject cannot belong to more than one group simultaneously. ILL and LIT are separate individuals, as are HC and PD.

Since the test was applied separately for each of the 3 best-performing models from binary classification task, a correction for multiple comparisons was applied to control the risk of false positives arising from conducting several simultaneous tests. The Benjamini-Hochberg procedure [33] was chosen to control the FDR (False Discovery Rate). FDR is an expected proportion of falsely rejected hypotheses among all rejected hypotheses [33] at level $\alpha = 0.05$. The procedure works as follows: the $m$ obtained $p$-values are sorted in ascending order and assigned ranks $i = 1, \ldots, m$. The largest rank $k$ for which $p_{(i)} \leq \frac{i}{m} \cdot \alpha$ is identified, and all hypotheses with rank $\leq k$ are rejected. This approach was preferred over the Bonferroni correction [34], which divides $\alpha$ equally across all tests, as it is less conservative while still maintaining control over the proportion of false discoveries [33].

## 4.8 Multiclass Classification

In addition to binary classification, three- and four-class experiments were conducted. The same classifiers and nested cross-validation approach for training and hyperparameter tuning were used as in the binary experiments. For SVM, the one-vs-rest strategy was applied, in which a separate binary classifier is trained for each class against all remaining classes [35].

For these experiments feature sets were constructed from the four most discriminative

features identified in the binary classification experiments: one containing the top features from the HC vs PD task, and one containing the top features from the ILL vs LIT task. Multiclass classifiers were trained using both feature sets to assess which set better separates the groups. Each model was trained three times, once for each value of $k$, where $k$ is the number of selected features, $k \in \{2, 3, 4\}$.

Performance of multiclass classifiers is evaluated using overall accuracy, as well as macro F1-score, macro precision, and macro recall. For the macro metrics, the value is first computed independently for each class $k$, treating it as the positive class against all others, and then the unweighted average is calculated across all $K$ classes [36]. The macro F1-score is calculated as the unweighted average of the per-class F1-scores:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{4.4}$$

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{4.5}$$

$$\text{Macro Precision} = \frac{1}{K}\sum_{k=1}^{K} \text{Precision}_k \tag{4.6}$$

$$\text{Macro Recall} = \frac{1}{K}\sum_{k=1}^{K} \text{Recall}_k \tag{4.7}$$

$$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{4.8}$$

$$\text{Macro F1} = \frac{1}{K}\sum_{k=1}^{K} F1_k \tag{4.9}$$

# 5 Results

This chapter shows experimental results. Sections 5.1 and 5.2 report classification results. Section 5.3 shows results of cross-domain experiment and Sections 5.1.1, 5.2.1, 5.3.2, describe the analysis of the results carried out using SHAP. Section 5.3.1 provides statistical analysis of results of cross-domain experiment.

Throughout the SHAP analysis, positive SHAP values correspond to predictions toward PD/ILL (class 1), and negative values toward HC/LIT (class 0). Section 5.4 reports the results of three-class and four-class classification experiments.

## 5.1 Healthy Control vs Parkinson's Disease Classification

The best-performing model for the task of discriminating PD (class 1) from HC (class 0) is LR. It achieved an accuracy of 0.863, a recall of 0.878 and a specificity of 0.856. The number of features $k = 2$ was selected during nested cross-validation and these features are: `velocity_x_mean` with Fisher's score 1.044 and `velocity_y_max` with Fisher's score 0.497. Appendix 3 contains selected features and Fisher's scores for the remaining models.

Table 2. Top 5 models for HC vs PD classification.

| Model | k | F1 | ACC | REC | SPEC | PREC |
|---|---|---|---|---|---|---|
| LR | 2 | **0.885 ± 0.021** | **0.863 ± 0.045** | 0.878 ± 0.088 | **0.856 ± 0.075** | **0.917 ± 0.118** |
| AB | 3 | 0.868 ± 0.097 | 0.833 ± 0.125 | **0.889 ± 0.079** | 0.819 ± 0.138 | 0.849 ± 0.117 |
| LR | 4 | 0.860 ± 0.023 | 0.830 ± 0.042 | 0.878 ± 0.088 | 0.814 ± 0.063 | 0.861 ± 0.104 |
| RF | 4 | 0.856 ± 0.062 | 0.830 ± 0.092 | 0.822 ± 0.016 | 0.828 ± 0.114 | 0.905 ± 0.135 |
| AB | 4 | 0.831 ± 0.049 | 0.796 ± 0.077 | 0.822 ± 0.016 | 0.786 ± 0.095 | 0.849 ± 0.117 |

### 5.1.1 Explanation of Model Predictions

RF with $k = 4$ features was chosen for the SHAP analysis because it uses more features than the best-performing model, allowing for a more detailed examination of the contribution of all features. Figure 9 shows the beeswarm plot, and Figure 8 shows the cohort bar chart

for the HC vs PD classification task. The x-axis on these plots represents SHAP values expressed in probability units.

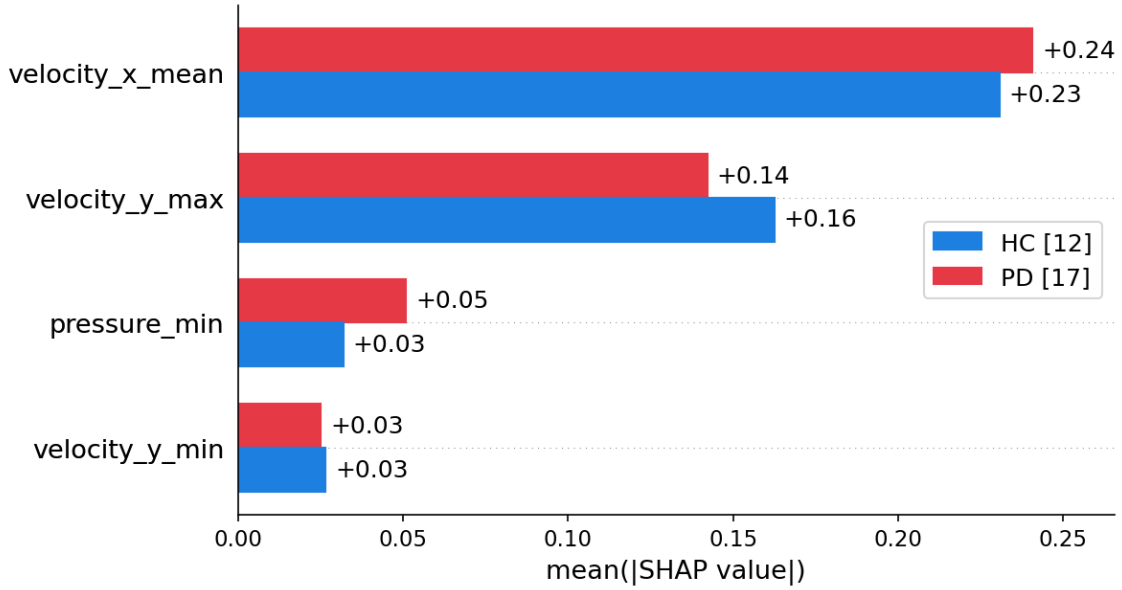



Figure 8. Cohort bar plot — HC vs PD (RF $k = 4$).

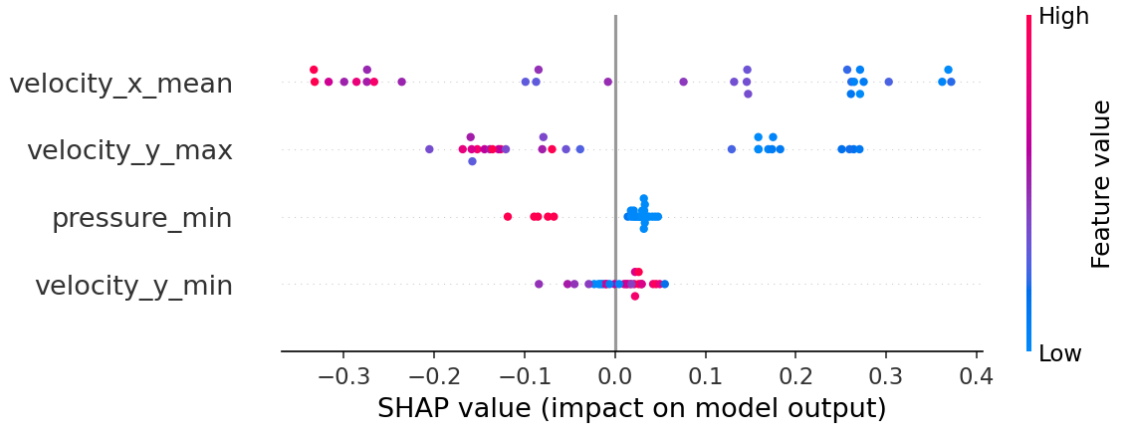



Figure 9. Beeswarm plot — HC vs PD (RF $k = 4$).

As shown in the bar plot, `velocity_x_mean` makes the largest contribution to the model's predictions, followed by `velocity_y_max`. The `pressure_min` and `velocity_y_min` features show significantly smaller contributions. The beeswarm plot shows that lower values of `velocity_x_mean`, `velocity_y_max`, and `pressure_min` push the prediction toward PD, while higher values push it toward HC. For `velocity_y_min`, however, points are clustered close to the baseline, confirming its limited influence on the model's output.

## 5.2 Illiterate vs Literate Classification

The best-performing model for the task of discriminating ILL (class 1) from LIT (class 0) is XGB. It achieved an accuracy of 0.697, a recall of 0.727 and a specificity of 0.697. The number of features $k = 4$ was selected during nested cross-validation and these features are: `yank_mean` with Fisher's score 0.223, `pressure_max` with Fisher's score 0.151, `velocity_mean` with Fisher's score 0.116 and `velocity_x_mean` with Fisher's score 0.088.

Table 3. Top 5 models for LIT vs ILL classification.

| Model | k | F1 | ACC | REC | SPEC | PREC |
|---|---|---|---|---|---|---|
| XGB | 4 | **0.707 ± 0.091** | 0.697 ± 0.107 | **0.727 ± 0.129** | 0.697 ± 0.107 | 0.709 ± 0.137 |
| AB | 4 | 0.701 ± 0.094 | **0.727 ± 0.064** | 0.667 ± 0.155 | **0.727 ± 0.064** | **0.773 ± 0.075** |
| RF | 2 | 0.695 ± 0.063 | 0.697 ± 0.057 | 0.697 ± 0.086 | 0.697 ± 0.057 | 0.696 ± 0.047 |
| LR | 2 | 0.689 ± 0.063 | 0.682 ± 0.037 | 0.727 ± 0.149 | 0.682 ± 0.037 | 0.673 ± 0.042 |
| LR | 4 | 0.680 ± 0.095 | 0.697 ± 0.043 | 0.697 ± 0.227 | 0.697 ± 0.043 | 0.713 ± 0.053 |

Appendix 4 contains selected features and Fisher's scores for the remaining models.

The classifiers of ILL vs LIT groups showed lower performance compared to HC vs PD, with the best F1 score of 0.707 and notably lower Fisher's scores across all selected features.

### 5.2.1 Explanation of Model Predictions

XGB with $k = 4$ features was selected for SHAP analysis as it is the best-performing model for the ILL vs LIT classification task. Figure 11 presents the beeswarm plot and Figure 10 presents the cohort bar plot. The x-axis on these plots represents SHAP values expressed in log-odds units.

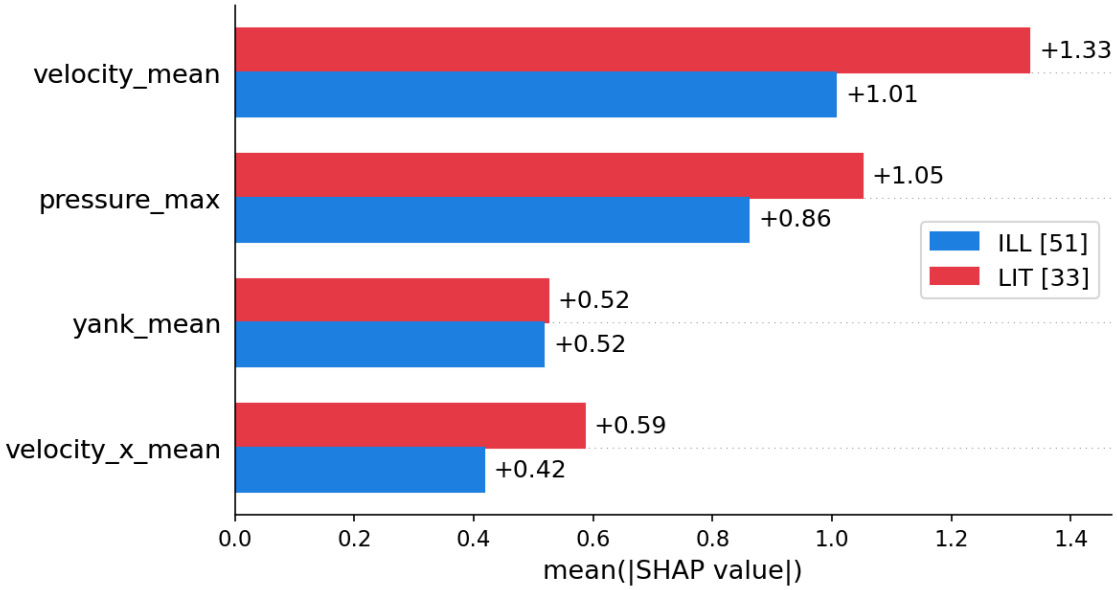



Figure 10. Cohort bar plot — ILL vs LIT (XGB $k = 4$).

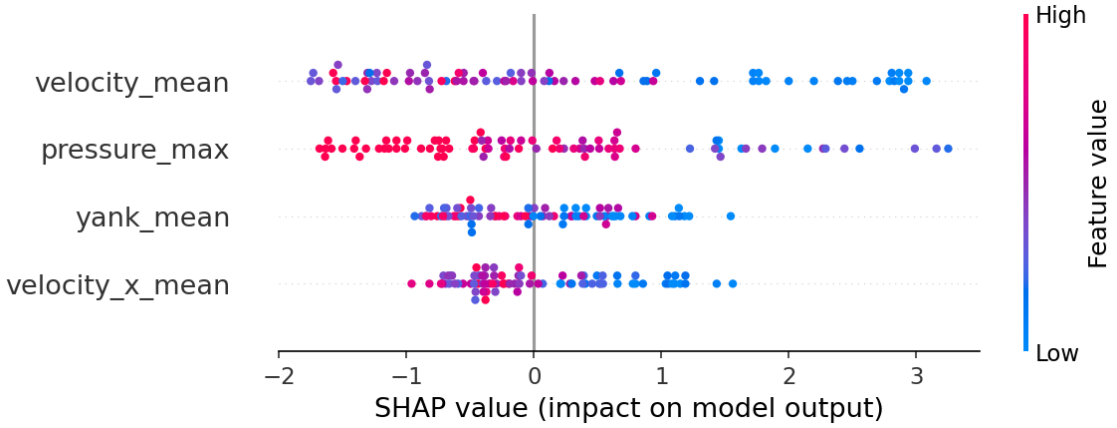



Figure 11. Beeswarm plot — ILL vs LIT (XGB $k = 4$).

As shown in the cohort bar plot, the contribution of the selected features differs slightly between the two groups. All features have a larger contribution for LIT group, compared to ILL.

The beeswarm plot shows that lower values of all four features push the prediction toward ILL, while higher values push it toward LIT.

## 5.3 Cross-domain Experiment

The results of cross-domain experiment described in Section 4.7 are shown using top 3 best-performing models in Tables 4, 5, 6, 7.

Table 4. Top 3 HC vs PD models applied to ILL.

| Model | k | Pred HC | Pred PD | % HC |
|---|---|---|---|---|
| LR | 2 | 36 | 15 | 70.6 |
| AB | 3 | 19 | 32 | 37.3 |
| LR | 4 | 37 | 14 | 72.5 |

Table 5. Top 3 HC vs PD models applied to LIT.

| Model | k | Pred HC | Pred PD | % HC |
|---|---|---|---|---|
| LR | 2 | 28 | 5 | 84.8 |
| AB | 3 | 22 | 11 | 66.7 |
| LR | 4 | 29 | 4 | 87.9 |

Table 6. Top 3 ILL vs LIT models applied to HC.

| Model | k | Pred LIT | Pred ILL | % LIT |
|---|---|---|---|---|
| XGB | 4 | 7 | 5 | 58.3 |
| AB | 4 | 8 | 4 | 66.7 |
| RF | 2 | 6 | 6 | 50.0 |

Table 7. Top 3 ILL vs LIT models applied to PD.

| Model | k | Pred LIT | Pred ILL | % LIT |
|---|---|---|---|---|
| XGB | 4 | 2 | 15 | 11.8 |
| AB | 4 | 2 | 15 | 11.8 |
| RF | 2 | 5 | 12 | 29.4 |

Given that the DraWriteEduUniPampa participants have no neurodegenerative diagnosis, HC classification would be the expected outcome for both ILL and LIT groups. However, the results presented in Tables 4 and 5 show that both ILL and LIT groups are predominantly classified as HC, with ILL being classified as HC less often (72.5–37.3%), compared to LIT (84.8–87.9%), which suggests that in some cases the model classifies ILL and LIT as PD, and more often such misclassifications occur in the case of ILL. The reverse experiment results presented in Tables 6 and 7 show that PD patients are predominantly classified as ILL (only 11.8–29.4% classified as LIT), while HC tend to be classified as LIT (58.3-66.7%).

### 5.3.1 Statistical Analysis

To assess whether the observed differences in cross-domain classification rates were statistically significant, a one-sided Fisher's Exact Test was applied for each of the top 3 models in both experiments. The Benjamini-Hochberg procedure was used to correct for multiple comparisons across the three models ($m = 3$, $\alpha = 0.05$). The hypotheses are defined as follows:

$H_0$: $\theta = 1$: the two groups are equally likely to be classified as the same predicted class.

$H_1$: $\theta > 1$: one group is more likely to be classified as the predicted class than the other.

Specifically, for the HC vs PD model applied to the ILL and LIT groups, $H_1$ states that ILL subjects are classified as PD more often than LIT subjects. For the ILL vs LIT model applied to the HC and PD groups, $H_1$ states that PD patients are classified as ILL more often than HC subjects. $\theta$ denotes the odds ratio. Results are presented in Tables 8 and 9.

Table 8. Fisher's Exact Test results — HC vs PD models applied to ILL/LIT group (Benjamini-Hochberg correction, $m = 3$, $\alpha = 0.05$).

| **Model** | **k** | **F1** | **% pred. PD (ILL)** | **% pred. PD (LIT)** | ***p*** | $\boldsymbol{p}_{\textbf{adj}}$ |
|---|---|---|---|---|---|---|
| LR | 2 | 0.885 | 29.4 | 15.2 | 0.107 | 0.107 |
| AB | 3 | 0.868 | 62.7 | 33.3 | **0.008** | **0.023** |
| LR | 4 | 0.860 | 27.5 | 12.1 | 0.078 | 0.107 |

Table 9. Fisher's Exact Test results — ILL vs LIT models applied to HC/PD group (Benjamini-Hochberg correction, $m = 3$, $\alpha = 0.05$).

| **Model** | **k** | **F1** | **% pred. ILL (PD)** | **% pred. ILL (HC)** | ***p*** | $\boldsymbol{p}_{\textbf{adj}}$ |
|---|---|---|---|---|---|---|
| XGB | 4 | 0.707 | 88.2 | 41.7 | **0.012** | **0.017** |
| AB | 4 | 0.701 | 88.2 | 33.3 | **0.004** | **0.011** |
| RF | 2 | 0.695 | 70.6 | 50.0 | 0.23 | 0.23 |

For the HC vs PD model applied to the ILL and LIT groups, $H_0$ is rejected for AB ($k = 3$, $p_{\text{adj}} = 0.023$), with ILL subjects classified as PD more frequently (62.7%) than LIT subjects (33.3%). The remaining two models showed the same trend but failed to reject $H_0$ after correction.

For the ILL vs LIT model applied to the HC and PD groups, $H_0$ is rejected for two out of three models: XGB ($p_{\text{adj}} = 0.017$) and AB ($p_{\text{adj}} = 0.011$). In both cases, PD patients were classified as ILL substantially more often (88.2%) than HC subjects (41.7% and 33.3% respectively). RF showed the same direction of effect but failed to reject $H_0$ ($p_{\text{adj}} = 0.23$).

Overall, the statistical analysis confirms that the observed cross-domain misclassification patterns are not random, providing support for the hypothesis that fine motor patterns of PD and ILL share measurable kinematic similarities.

### 5.3.2 Explanation of Model Predictions

For the SHAP analysis of the cross-domain experiment, the same models were used as in the previous sections: RF $k = 4$ for the HC vs PD classification and XGB $k = 4$ for the ILL vs LIT classification. Figures 12 and 13 present the beeswarm and decision plots for the RF $k = 4$ model applied to the ILL/LIT group, while Figures 14 and 15 present the corresponding plots for the XGB $k = 4$ model applied to the HC/PD group.

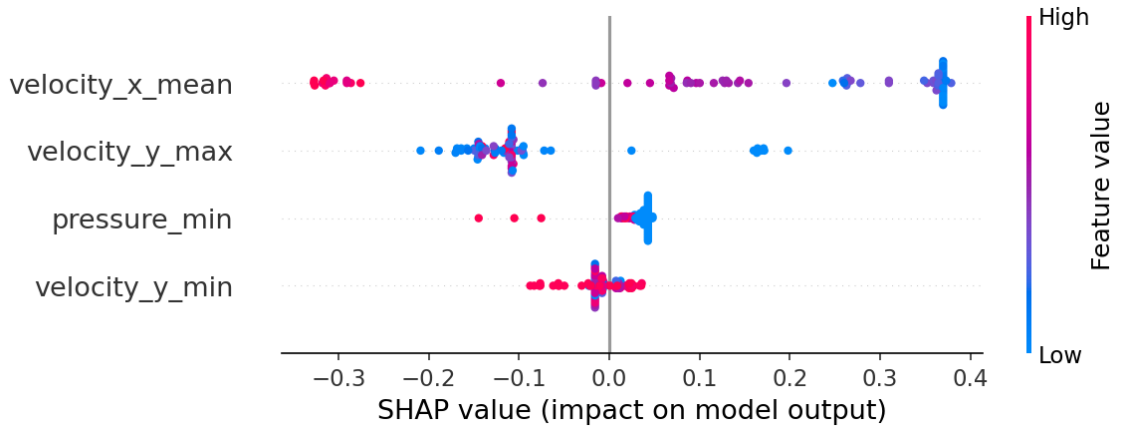



Figure 12. Beeswarm plot — HC vs PD model applied to ILL/LIT group (RF $k = 4$).

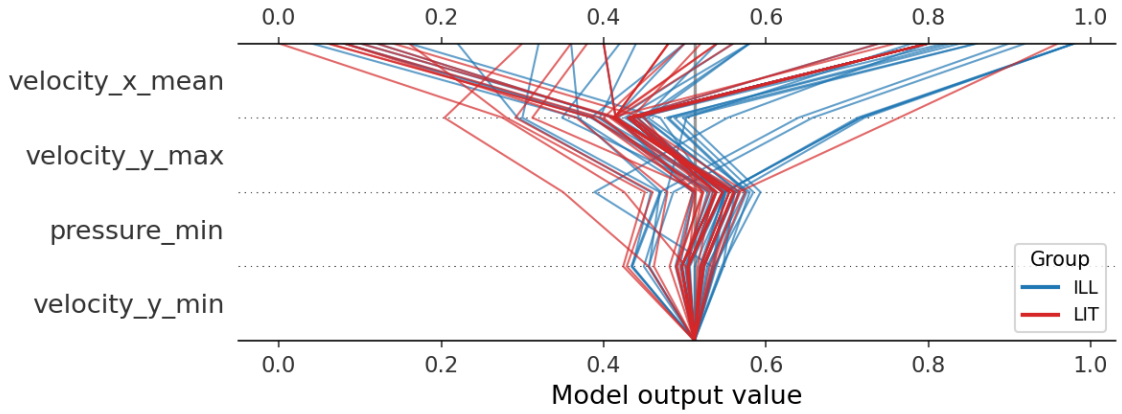



Figure 13. Decision plot — HC vs PD model applied to ILL/LIT group (RF $k = 4$).

As shown in the beeswarm plot (Figure 12), lower values of `velocity_x_mean` and `pressure_min` push the predictions towards PD, while higher values push them towards HC. SHAP values for `velocity_y_min` are clustered around baseline, confirming its limited influence on the model's output — in line with the pattern observed in Section 5.1.1. However, the behaviour of `velocity_y_max` differs: in Figure 9, lower values of `velocity_y_max` pushed predictions towards PD, whereas in this beeswarm plot both high

and low feature values are associated with negative SHAP values, suggesting that the entire range of this feature in the ILL/LIT group falls within the region the model associates with HC.

The decision plot (Figure 13) shows that ILL and LIT lines are intermixed, although most lines fall in the 0.0–0.6 range, consistent with the cross-domain classification results, where both ILL and LIT were mostly classified as HC.

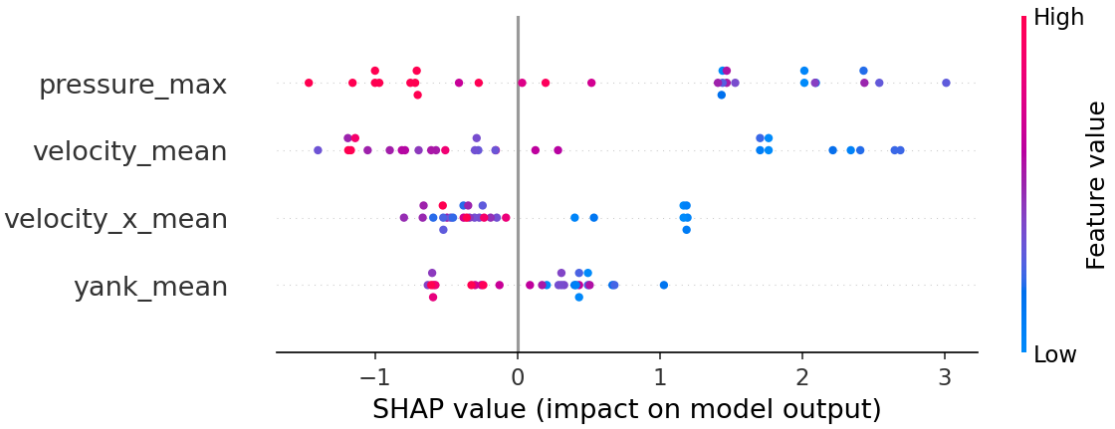



Figure 14. Beeswarm plot — ILL vs LIT model applied to HC/PD group (XGB $k = 4$).

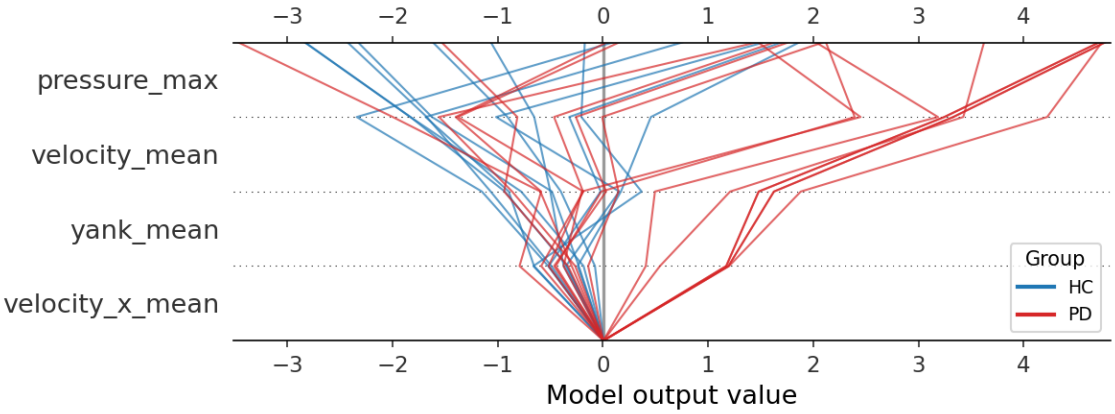



Figure 15. Decision plot — ILL vs LIT model applied to HC/PD group (XGB $k = 4$).

The beeswarm plot (Figure 14) shows that all four features behave as in Figure 11: lower values of these features push the predictions towards ILL, but the order of features has changed, now in first place there is `pressure_max` instead of `velocity_mean`.

The decision plot (Figure 15) shows that predictions begin to separate already at the `velocity_x_mean` level. Finally, HC participants are mostly classified as LIT and PD participants as ILL, confirming cross-domain experiment results.

## 5.4 Multiclass Classification

Four three-class tasks are considered: HC vs PD vs ILL, HC vs PD vs LIT, LIT vs ILL vs PD, and LIT vs ILL vs HC, as well as one four-class task: HC vs PD vs ILL vs LIT. Each task was performed twice: using the HC/PD feature set (`velocity_x_mean`, `velocity_y_max`, `pressure_min`, `velocity_y_min`) and the ILL/LIT feature set (`yank_mean`, `pressure_max`, `velocity_mean`, `velocity_x_mean`), and results for both are reported. For each task, the top three models are ranked by macro F1 score, overall accuracy, macro recall, and macro precision, along with the confusion matrix of the best-performing model.

### 5.4.1 Healthy Control vs Parkinson's Disease vs Illiterate classification

For the three-class task distinguishing HC, PD, and ILL, the best-performing model using HC/PD features is SVM ($k = 4$) with a macro F1 of 0.708, accuracy of 0.748, macro recall of 0.733 and macro precision of 0.839. Using ILL/LIT features, NB ($k = 4$) achieves a slightly higher macro F1 of 0.740, accuracy of 0.751, macro recall of 0.756 and macro precision of 0.834. The top 3 models for both feature sets are shown in Tables 10 and 11.

Table 10. Top 3 models — HC vs PD vs ILL (HC/PD features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| SVM | 4 | **0.708 ± 0.083** | **0.748 ± 0.074** | **0.733 ± 0.049** | **0.839 ± 0.052** |
| LR | 4 | 0.657 ± 0.008 | 0.705 ± 0.028 | 0.683 ± 0.041 | 0.824 ± 0.019 |
| DT | 4 | 0.654 ± 0.080 | 0.682 ± 0.059 | 0.681 ± 0.077 | 0.721 ± 0.100 |

Table 11. Top 3 models — HC vs PD vs ILL (ILL/LIT features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| NB | 4 | **0.740 ± 0.019** | **0.751 ± 0.025** | **0.756 ± 0.042** | **0.834 ± 0.040** |
| KNN | 4 | 0.698 ± 0.020 | 0.705 ± 0.028 | 0.709 ± 0.027 | 0.752 ± 0.057 |
| XGB | 4 | 0.677 ± 0.090 | 0.706 ± 0.078 | 0.711 ± 0.067 | 0.809 ± 0.044 |

Table 12. Confusion matrix — HC vs PD vs ILL (SVM $k = 4$, HC/PD features).

| | **Pred HC** | **Pred PD** | **Pred ILL** |
|---|---|---|---|
| **HC** | 8 | 3 | 1 |
| **PD** | 1 | 16 | 0 |
| **ILL** | 3 | 3 | 9 |

Table 13. Confusion matrix — HC vs PD vs ILL (NB $k = 4$, ILL/LIT features).

| | **Pred HC** | **Pred PD** | **Pred ILL** |
|---|---|---|---|
| **HC** | 10 | 2 | 0 |
| **PD** | 1 | 14 | 2 |
| **ILL** | 4 | 2 | 9 |

As shown in Tables 12 and 13, PD is the best-classified group across both feature sets, with 16/17 (16 out of 17) samples correctly classified using HC/PD features and 14/17 using ILL/LIT features, where 2 PD samples are misclassified as ILL. HC and ILL show more confusion: with HC/PD features, 8/12 HC samples are correctly classified, 3 ILL samples are misclassified as HC and 3 as PD, while with ILL/LIT features, HC classifications improve to 10/12, but 4 ILL samples are misclassified as HC. The consistent misclassification of ILL toward both HC and PD across both feature sets suggests that motor patterns of ILL lie between those of HC and PD.

### 5.4.2 Healthy Control vs Parkinson's Disease vs Literate classification

For the HC vs PD vs LIT task, the best-performing model using HC/PD features is SVM ($k = 2$) with a macro F1 of 0.590, accuracy of 0.614, macro recall of 0.596 and macro precision of 0.604. Using ILL/LIT features, LR ($k = 2$) achieves a comparable macro F1 of 0.559, accuracy of 0.592, macro recall of 0.563 and macro precision of 0.568. The top 3 models for both feature sets are shown in Tables 14 and 15.

Table 14. Top 3 models — HC vs PD vs LIT (HC/PD features).

| **Model** | **k** | **F1** | **ACC** | **REC** | **PREC** |
|---|---|---|---|---|---|
| SVM | 2 | **0.590 ± 0.034** | **0.614 ± 0.020** | **0.596 ± 0.039** | 0.604 ± 0.023 |
| XGB | 4 | 0.563 ± 0.091 | 0.594 ± 0.085 | 0.583 ± 0.096 | **0.629 ± 0.078** |
| XGB | 3 | 0.548 ± 0.178 | 0.618 ± 0.131 | 0.587 ± 0.153 | 0.593 ± 0.209 |

Table 15. Top 3 models — HC vs PD vs LIT (ILL/LIT features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| LR | 2 | **0.559 ± 0.057** | 0.592 ± 0.045 | 0.563 ± 0.057 | 0.568 ± 0.054 |
| NB | 4 | 0.557 ± 0.083 | **0.633 ± 0.098** | **0.598 ± 0.079** | **0.593 ± 0.184** |
| XGB | 3 | 0.507 ± 0.168 | 0.567 ± 0.119 | 0.548 ± 0.136 | 0.525 ± 0.164 |

Table 16. Confusion matrix — HC vs PD vs LIT (SVM $k = 2$, HC/PD features).

| | Pred HC | Pred PD | Pred LIT |
|---|---|---|---|
| **HC** | 6 | 2 | 4 |
| **PD** | 1 | 14 | 2 |
| **LIT** | 4 | 4 | 7 |

Table 17. Confusion matrix — HC vs PD vs LIT (LR $k = 2$, ILL/LIT features).

| | Pred HC | Pred PD | Pred LIT |
|---|---|---|---|
| **HC** | 4 | 1 | 7 |
| **PD** | 1 | 14 | 2 |
| **LIT** | 4 | 3 | 8 |

As shown in Tables 16 and 17, PD remains the best-classified group across both feature sets, with 14/17 correct classifications. LIT classifications prove more difficult: with HC/PD features, 4 LIT samples are misclassified as HC and 4 as PD, while ILL/LIT features improve LIT classifications to 8/15, with HC dropping to 4/12, mostly misclassified as LIT. This is expected since HC and LIT are both neurologically healthy and literate, unlike PD group, whose motor impairment produces distinctly different movement patterns.

### 5.4.3 Literate vs Illiterate vs Parkinson's Disease Classification

For the LIT vs ILL vs PD task, the best-performing model using HC/PD features is SVM ($k = 3$) with a macro F1 of 0.686, accuracy of 0.700, macro recall of 0.689 and macro precision of 0.742. Using ILL/LIT features, SVM ($k = 4$) achieves a higher macro F1 of 0.725, accuracy of 0.725, macro recall of 0.730 and macro precision of 0.778. The top 3 models for both feature sets are shown in Tables 18 and 19.

Table 18. Top 3 models — LIT vs ILL vs PD (HC/PD features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| SVM | 3 | **0.686 ± 0.058** | **0.700 ± 0.071** | **0.689 ± 0.063** | **0.742 ± 0.064** |
| LR | 3 | 0.668 ± 0.052 | 0.679 ± 0.061 | 0.670 ± 0.055 | 0.722 ± 0.041 |
| LR | 4 | 0.666 ± 0.052 | 0.679 ± 0.061 | 0.670 ± 0.055 | 0.695 ± 0.029 |

Table 19. Top 3 models — LIT vs ILL vs PD (ILL/LIT features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| SVM | 4 | **0.725 ± 0.049** | **0.725 ± 0.053** | **0.730 ± 0.050** | **0.778 ± 0.043** |
| NB | 4 | 0.671 ± 0.062 | 0.682 ± 0.044 | 0.678 ± 0.051 | 0.689 ± 0.083 |
| SVM | 3 | 0.648 ± 0.083 | 0.661 ± 0.072 | 0.667 ± 0.079 | 0.694 ± 0.118 |

Table 20. Confusion matrix — LIT vs ILL vs PD (SVM $k = 3$, HC/PD features).

| | Pred LIT | Pred ILL | Pred PD |
|---|---|---|---|
| **LIT** | 9 | 1 | 5 |
| **ILL** | 4 | 8 | 3 |
| **PD** | 1 | 0 | 16 |

Table 21. Confusion matrix — LIT vs ILL vs PD (SVM $k = 4$, ILL/LIT features).

| | Pred LIT | Pred ILL | Pred PD |
|---|---|---|---|
| **LIT** | 12 | 3 | 0 |
| **ILL** | 3 | 12 | 0 |
| **PD** | 4 | 3 | 10 |

As shown in Tables 20 and 21, with HC/PD features PD is classified most reliably (16/17 correct), while there are 9/15 correct classifications for LIT and 8/15 for ILL, with 4 ILL samples misclassified as LIT and 3 as PD. With ILL/LIT features, both LIT and ILL classifications improve substantially to 12/15 each. However, correct classifications of PD decrease, with 4 samples misclassified as LIT and 3 as ILL, suggesting that pressure- and yank-based features are less effective at capturing the motor symptoms of PD when all three groups are present simultaneously.

### 5.4.4 Literate vs Illiterate vs Healthy Control classification

For the LIT vs ILL vs HC task, the best-performing model using HC/PD features is NB ($k = 2$) with a macro F1 of 0.535, accuracy of 0.571, macro recall of 0.589 and macro precision of 0.604. Using ILL/LIT features, NB ($k = 4$) achieves a slightly higher macro F1 of 0.593, accuracy of 0.595, macro recall of 0.589 and macro precision of 0.661. The top 3 models for both feature sets are shown in Tables 22 and 23.

Table 22. Top 3 models — LIT vs ILL vs HC (HC/PD features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| NB | 2 | **0.535 ± 0.038** | **0.571 ± 0.000** | **0.589 ± 0.016** | **0.604 ± 0.108** |
| KNN | 2 | 0.529 ± 0.089 | 0.524 ± 0.089 | 0.517 ± 0.085 | 0.592 ± 0.121 |
| LR | 3 | 0.511 ± 0.195 | 0.571 ± 0.154 | 0.572 ± 0.159 | 0.478 ± 0.217 |

Table 23. Top 3 models — LIT vs ILL vs HC (ILL/LIT features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| NB | 4 | **0.593 ± 0.170** | **0.595 ± 0.178** | **0.589 ± 0.166** | **0.661 ± 0.187** |
| NB | 2 | 0.570 ± 0.105 | 0.571 ± 0.117 | 0.567 ± 0.109 | 0.594 ± 0.096 |
| DT | 4 | 0.554 ± 0.149 | 0.571 ± 0.154 | 0.578 ± 0.142 | 0.595 ± 0.176 |

Table 24. Confusion matrix — LIT vs ILL vs HC (NB $k = 2$, HC/PD features).

| | Pred LIT | Pred ILL | Pred HC |
|---|---|---|---|
| **LIT** | 4 | 2 | 9 |
| **ILL** | 2 | 10 | 3 |
| **HC** | 2 | 0 | 10 |

Table 25. Confusion matrix — LIT vs ILL vs HC (NB $k = 4$, ILL/LIT features).

| | Pred LIT | Pred ILL | Pred HC |
|---|---|---|---|
| **LIT** | 9 | 3 | 3 |
| **ILL** | 3 | 10 | 2 |
| **HC** | 6 | 0 | 6 |

As shown in Tables 24 and 25, this combination proves the most challenging among the three-class tasks, with the lowest F1 scores across both feature sets. With HC/PD features, classification of LIT shows the weakest performance (4/15 correct), with the majority of LIT samples misclassified as HC. ILL/LIT features improve classification of LIT to 9/15 and classification of ILL to 10/15, while classification of HC improves to 10/12 with HC/PD features and remains moderate at 6/12 with ILL/LIT features. Unlike other three-class tasks, this one consists entirely of healthy groups, which explains the overall lower performance compared to tasks involving PD.

### 5.4.5 Four-class Classification

The four-class task is the most challenging configuration, combining all groups simultaneously. The best-performing model using HC/PD features is NB ($k = 2$) with a macro F1 of 0.546, accuracy of 0.625, macro recall of 0.617 and macro precision of 0.615. Using ILL/LIT features, NB ($k = 4$) achieves a lower macro F1 of 0.549, accuracy of 0.593, macro recall of 0.571 and macro precision of 0.562. The top 3 models for both feature sets are shown in Tables 26 and 27.

Table 26. Top 3 models — HC vs PD vs ILL vs LIT (HC/PD features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| NB | 2 | **0.546 ± 0.116** | **0.625 ± 0.107** | **0.617 ± 0.085** | **0.615 ± 0.140** |
| SVM | 3 | 0.506 ± 0.021 | 0.542 ± 0.011 | 0.531 ± 0.014 | 0.539 ± 0.032 |
| SVM | 4 | 0.502 ± 0.135 | 0.524 ± 0.125 | 0.518 ± 0.135 | 0.518 ± 0.122 |

Table 27. Top 3 models — HC vs PD vs ILL vs LIT (ILL/LIT features).

| Model | k | F1 | ACC | REC | PREC |
|---|---|---|---|---|---|
| NB | 4 | **0.549 ± 0.010** | **0.593 ± 0.042** | **0.571 ± 0.021** | 0.562 ± 0.044 |
| LR | 3 | 0.538 ± 0.036 | 0.541 ± 0.052 | 0.532 ± 0.041 | **0.577 ± 0.030** |
| KNN | 3 | 0.515 ± 0.049 | 0.541 ± 0.052 | 0.526 ± 0.044 | 0.557 ± 0.080 |

Table 28. Confusion matrix — HC vs PD vs ILL vs LIT (NB $k = 2$, HC/PD features).

| | Pred HC | Pred PD | Pred ILL | Pred LIT |
|---|---|---|---|---|
| **HC** | 8 | 2 | 0 | 2 |
| **PD** | 1 | 16 | 0 | 0 |
| **ILL** | 5 | 3 | 7 | 0 |
| **LIT** | 6 | 2 | 1 | 6 |

Table 29. Confusion matrix — HC vs PD vs ILL vs LIT (NB $k = 4$, ILL/LIT features).

| | Pred HC | Pred PD | Pred ILL | Pred LIT |
|---|---|---|---|---|
| **HC** | 3 | 2 | 0 | 7 |
| **PD** | 0 | 13 | 2 | 2 |
| **ILL** | 3 | 2 | 8 | 2 |
| **LIT** | 1 | 1 | 2 | 11 |

As shown in Tables 28 and 29, with HC/PD features PD is again classified most reliably (16/17 correct), while ILL and LIT show notable confusion with HC, with 5 ILL and 6 LIT samples misclassified as HC. With ILL/LIT features, LIT classification improves substantially to 11/15 and ILL remains stable at 8/15, while HC drops to 3/12, with 7 samples misclassified as LIT. PD classification decreases slightly to 13/17, with 2 samples misclassified as ILL and 2 as LIT.

# 6 Discussion

This chapter discusses the findings presented in Chapter 5 and provides answers to the research questions of this thesis.

The binary classification results show a notable performance gap between the two tasks. Classifiers distinguished PD patients from healthy controls with high accuracy, with the best model achieving F1 = 0.885, whereas the best ILL vs LIT model reached F1 = 0.707. This performance gap can be explained by the fact that PD patients have consistent and progressive motor symptoms, primarily slowness and irregularity of movement, that can be captured using tremor-related features. Values of these features contrast with HC movements, making the two groups reliably separable. In contrast, the motor differences between illiterate and literate individuals arise from the absence of fine motor practice rather than neurological damage, and are therefore more gradual and variable across individuals. This explains both the lower classification accuracy and the weaker Fisher scores observed in the ILL vs LIT classification task.

A notable observation is that during the independent training of the models, the nested cross-validation process selected similar features for two different classification tasks. The top five features ranked by Fisher's score for the HC vs PD task were: `velocity_x_mean` [F=1.0438], `velocity_y_max` [F=0.4969], `pressure_min` [F=0.4000], `velocity_y_min` [F=0.1841], `accel_mass` [F=0.1782] and for the ILL vs LIT task: `yank_mean` [F=0.2225], `pressure_max` [F=0.1508], `velocity_mean` [F=0.1161], `velocity_x_mean` [F=0.0884], `accel_x_median` [F=0.0527]. The feature `velocity_x_mean` appears among the top features in both tasks, suggesting that horizontal drawing speed is a discriminative kinematic marker across both domains (RQ1). Moreover, SHAP analysis showed that the velocity-based features were the most influential ones for both tasks and showed a consistent directional effect on predictions: lower values of velocity pushed the models output towards PD and ILL classes (RQ2, RQ3). Physiologically, this aligns with expectations. Slowness of movement is one of the primary motor symptoms of Parkinson's disease. Likewise,

for illiterate individuals, executing a complex geometric drawing task requires a cognitive and physical effort, resulting in a slower drawing speed. Regarding the pressure features, which played an important role in the ILL vs LIT classification, their high discriminative power can be explained by the lack of handwriting practice. Illiterate individuals have less experience with holding a pen or stylus, which leads to inconsistent and irregular application of pressure during the drawing process. Overall, these matching results provide evidence for the hypothesis of this thesis: fine motor patterns affected by progressive Parkinson's disease indeed share measurable kinematic similarities with those underdeveloped due to the absence of formal education.

The cross-domain experiment provides additional evidence for the hypothesis of this thesis. As expected, models trained on the HC vs PD dataset classified the majority of the DraWriteEduUniPampa participants as HC, since they do not have neurological diseases. However, a pattern was observed: illiterate individuals were misclassified as PD much more often (up to nearly 30%) than literate individuals. The reverse experiment, when models trained to detect illiteracy were applied to PD patients, classified 88.2% of them as ILL (Table 7). This asymmetry can be explained with the fact that the Parkinson's disease affects hand movements much more than the absence of writing practice, so PD patterns fall more deeply into the ILL region, while ILL patterns partially overlap with PD. These misclassifications show that the algorithms found kinematic reasons to group these different conditions together (RQ2). To control the significance of these observations, Fisher's Exact Test with Benjamini-Hochberg FDR correction was applied to the best-performing models in both experiments. In the first experiment, the HC vs PD model applied to the ILL and LIT group, AB showed a statistically significant misclassification: it classified illiterate subjects as PD more often than literate subjects ($p_{adj} = 0.023$). In the second experiment, the ILL vs LIT model applied to the HC and PD group, XGB ($p_{adj} = 0.017$) and AB ($p_{adj} = 0.011$) significantly misclassified PD subjects as illiterate.

SHAP analysis (Section 5.3.2) further explained these patterns. When the ILL vs LIT model was applied to PD patients, low values of `velocity_mean` and `pressure_max` pushed the predictions toward the ILL class. This confirms that the slowness and uneven pen pressure typical for PD look very similar to the drawing style of individuals without handwriting experience (RQ3).

A similar pattern was seen in the reverse experiment. When the HC vs PD model was applied to the illiterate group, low values of `velocity_x_mean` and `pressure_min` pushed the predictions toward the PD class. This shows that the slow drawing speed and unstable pen grip of illiterate individuals closely resemble the motor symptoms of Parkinson's disease. Overall, even though Parkinson's disease affects hand movements more severely, the basic kinematic similarities between the two groups are clearly visible.

The multiclass classification experiments further support these findings. Across all multiclass classifications, PD was consistently the best-classified group, showing the distinctiveness of its motor impairment. ILL was frequently misclassified as healthy groups and PD, suggesting that its motor patterns lie between healthy individuals and PD. The choice of k = 4 as the upper bound for multiclass experiments was informed by the binary classification results, where k = 5 did not appear among the top-performing models, suggesting that this number of features is sufficient for these small datasets (RQ4).

However, several limitations should be acknowledged. First, the small sample size in the HC vs PD dataset (29 samples), limits the generalisability of the results. While nested cross-validation mitigates overfitting and makes full use of the available data, larger datasets will be required to confirm these results in wider populations. Second, for the HC vs PD task, class balancing was not applied, as the HC group represented the smaller class (12 samples) and further reducing the PD group would have resulted in a small training set and loss of important training data. Third, the two datasets were collected in different countries on different devices, introducing potential confounding variables in addition to the technical differences that were addressed during the harmonization process. Cultural background, differences in drawing habits, and demographic factors may have influenced the results and cannot be fully controlled for in the current experimental setup. Finally, this study considered only the pltrace type of the Luria's alternating series test. While this choice was motivated by the need to isolate the motor execution component across all groups, extending the analysis to other test types could provide a more comprehensive understanding of the kinematic similarities between the studied groups.

# 7 Future Work

The most immediate direction for future work is the collection of a new version of DraWritePD dataset using the same device as was used for collecting DraWriteEduUni-Pampa dataset. This would eliminate the need for cross-device harmonisation and allow the analysis pipeline developed in this thesis to be applied directly to fully comparable data. Future work could also extend the study population to include patients with multiple sclerosis and children who have not yet developed handwriting skills. Unlike illiterate adults, whose fine motor control is underdeveloped due to lack of writing practice, pre-literate children represent a case where fine motor skills are still developing. Including these groups would allow the pipeline to be tested on a broader range of conditions that affect or reflect fine motor control, moving toward differential analysis of motor disorder.

Additionally, including Parkinson's disease patients within Brazil would enable a within-country comparison between PD and illiterate individuals, fully eliminating cultural background and device as confounding variables. This within-country design would provide the cleanest possible test of whether the motor signatures of Parkinson's disease and illiteracy are distinguishable by machine learning algorithms.

# 8 Conclusion

This thesis investigated whether fine motor patterns affected by Parkinson's disease share measurable kinematic similarities with those underdeveloped due to the absence of formal education, using ML and XAI for analysis of digitised Luria's alternating series drawing tests.

The results of this study answered the research questions and supported the central hypothesis. In particular, the binary classifications demonstrated that velocity and pressure dynamics are the primary discriminative features for both conditions (RQ1). The overlap between the most discriminative features for PD and illiteracy, along with the directional effect of low velocity observed in SHAP analysis, provides evidence which supports the hypothesis of this thesis (RQ2, RQ3). The cross-domain experiment showed that the kinematic overlap between PD and illiteracy is systematic and statistically significant (RQ2). Models trained to classify PD from HC misclassified illiterate individuals as PD more often than literate individuals, while models trained to detect illiteracy classified PD patients mostly as illiterate. The multiclass experiments showed that PD remained the most separable group, and the illiterate was frequently misclassified as both HC and PD, suggesting that this group's motor patterns lie between the healthy and PD groups (RQ4).

Overall, the results of the present thesis show that the kinematic overlap between Parkinson's disease and the absence of formal education is detectable by ML, and that education level should be considered in diagnostics of Parkinson's disease.

# References


[1] J. Jankovic. „Parkinson's disease: clinical features and diagnosis“. In: *J. Neurol., Neurosurg. Psychiatry* 79.4 (2008), pp. 368–376. DOI: 10.1136/jnnp.2007.131045.

[2] L. V. Kalia and A. E. Lang. „Parkinson's disease“. In: *Lancet* 386.9996 (2015), pp. 896–912. DOI: 10.1016/S0140-6736(14)61393-3.

[3] E. Valla et al. „Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics“. In: *Biomed. Signal Process. Control* 75 (2022), p. 103551. DOI: 10.1016/j.bspc.2022.103551.

[4] P. Drotar et al. „Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease“. In: *Artif. Intell. Med.* 67 (2016), pp. 39–46. DOI: 10.1016/j.artmed.2016.01.004.

[5] P. Drotar et al. „Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease“. In: *Comput. Methods Programs Biomed.* 117.3 (2014), pp. 405–411. DOI: 10.1016/j.cmpb.2014.08.007.

[6] E. J. Smits et al. „Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinson's Disease“. In: *PLoS ONE* 9 (2014), pp. 1–8. DOI: 10.1371/journal.pone.0097614.

[7] M. Thomas, A. Lenka, and P. K. Pal. „Handwriting Analysis in Parkinson's Disease: Current Status and Future Directions“. In: *Mov. Disord. Clin. Pract.* 4.6 (2017), pp. 806–818. DOI: 10.1002/mdc3.12552.

[8] S. Nomm et al. „Detailed Analysis of the Luria's Alternating Series Tests for Parkinson's Disease Diagnostics“. In: *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*. 2018, pp. 1347–1352. DOI: 10.1109/ICMLA.2018.00219.

[9] S. J. Chung et al. „Emerging Concepts of Motor Reserve in Parkinson's Disease“. In: *J. Mov. Disord.* 13.3 (2020), pp. 171–184. DOI: 10.14802/jmd.20029.

[10] A. Tolonen et al. „Distinguishing Parkinson's disease from other syndromes causing tremor using automatic analysis of writing and drawing tasks“. In: *Proc. IEEE 15th Int. Conf. Bioinf. Bioeng. (BIBE)*. 2015, pp. 1–4. DOI: 10.1109/BIBE.2015.7367690.

[11] C. Zhang et al. „Hand copy performance of young children and the illiterate, semi-illiterate, and literate adults“. In: *Curr. Psychol.* 43.9 (2024), pp. 8018–8028. DOI: 10.1007/s12144-023-05009-x.

[12] G. DeMaagd and A. Philip. „Parkinson's disease and its management: Part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis“. In: *P & T* 40 (2015), pp. 504–532.

[13] Y. Luo et al. „Global, regional, national epidemiology and trends of Parkinson's disease from 1990 to 2021: findings from the Global Burden of Disease Study 2021". In: *Front. Aging Neurosci.* 16 (2025), p. 1498756. DOI: 10.3389/fnagi.2024.1498756.

[14] L. M. L. de Lau and M. M. B. Breteler. „Epidemiology of Parkinson's disease". In: *Lancet Neurol.* 5.6 (2006), pp. 525–535. DOI: 10.1016/S1474-4422(06)70471-9.

[15] A. R. Luria. *Higher Cortical Functions in Man*. New York, NY, USA: Springer, 1995. DOI: 10.1007/978-1-4684-7741-2.

[16] P. Stepien et al. „Computer Aided Feature Extraction in the Paper Version of Luria's Alternating Series Test in Progressive Supranuclear Palsy". In: *Information Technology in Biomedicine*. Ed. by E. Pietka et al. Cham, Switzerland: Springer, 2019, pp. 561–570. DOI: 10.1007/978-3-319-91211-0_49.

[17] S. Nomm et al. „Determining Necessary Length of the Alternating Series Test for Parkinson's Disease Modelling". In: *Proc. Int. Conf. Cyberworlds*. Oct. 2019, pp. 261–266. DOI: 10.1109/CW.2019.00050.

[18] E. Valla et al. „Deep Learning Based Segmentation of Luria's Alternating Series Test to Support Diagnostics of Parkinson's Disease". In: *Proc. Int. Conf. Mach. Learn. Appl. (ICMLA)*. 2023, pp. 1066–1071. DOI: 10.1109/ICMLA58977.2023.00158.

[19] S. Nomm et al. „Drawing Strategies Analysis for the Embedded Figure Tests". In: *Recent Challenges in Intelligent Information and Database Systems*. Singapore: Springer Nature Singapore, 2025, pp. 3–14. DOI: 10.1007/978-981-96-5884-8_1.

[20] K. Bardos. „Analysis of interpretable anomalies and kinematic parameters in Luria's alternating series tests for Parkinson's disease modeling". MA thesis. Tallinn, Estonia: Sch. Inf. Technol., Tallinn Univ. Technol., 2018.

[21] S. Zarembo. „Adaptation Affect on Fine Motor Motions". B.S. thesis. Tallinn, Estonia: Sch. Inf. Technol., Tallinn Univ. Technol., 2019.

[22] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426. [Online]. Available: https://arxiv.org/abs/1802.03426. 2018.

[23] C. C. Aggarwal. *Data Mining: The Textbook*. Cham, Switzerland: Springer, 2015.

[24] T. Chen and C. Guestrin. „XGBoost: A Scalable Tree Boosting System". In: *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. 2016. DOI: 10.1145/2939672.2939785.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York, NY, USA: Springer, 2009.

[26] S. Nomm and A. Toomela. „An alternative approach to measure quantity and smoothness of the human limb motions". In: *Estonian J. Eng.* 19.4 (2013), pp. 298–308. DOI: 10.3176/eng.2013.4.05.

[27] S. M. Lundberg and S.-I. Lee. *A unified approach to interpreting model predictions*. arXiv:1705.07874. 2017.

[28] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/. Leanpub, 2019.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv:1602.04938. 2016.

[30] SHAP Documentation. *SHAP API Examples*. Accessed: May 10, 2026. [Online]. Available: https://shap.readthedocs.io/en/latest/api_examples.html. 2026.

[31] N. Faizi and Y. Alvi. „Categorical Variables". In: *Biostatistics Manual for Health Research*. Academic Press, 2023, pp. 127–148. DOI: 10.1016/B978-0-443-18550-2.00001-3.

[32] A. Agresti. *Categorical Data Analysis*. 2nd. Hoboken, NJ, USA: Wiley-Interscience, 2002.

[33] Y. Benjamini and Y. Hochberg. „Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *J. Roy. Stat. Soc. B* 57.1 (1995), pp. 289–300. DOI: 10.2307/2346101.

[34] R. A. Armstrong. „When to Use the Bonferroni Correction". In: *Ophthalmic Physiol. Opt.* 34.5 (2014), pp. 502–508. DOI: 10.1111/opo.12131.

[35] C. M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[36] M. Grandini, E. Bagli, and G. Visani. *Metrics for Multi-Class Classification: an Overview*. arXiv:2008.05756. 2020.

# Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

We Diana Kuntsmann and Sofia Loginova

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for our thesis "Explainable Artificial Intelligence Analysis of Drawing Tests for Assessment of Cognitive State", supervised by Sven Nõmm and Aaro Toomela
    1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
    1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright
2. We are aware that the authors also retain the rights specified in clause 1 of the nonexclusive licence.
3. We confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

01.06.2026

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive licence shall not be valid for the period.

# Appendix 2 – Extracted Features

This appendix lists all 97 features extracted from the raw drawing signals, grouped into kinematic features (velocity and its derivatives) and pressure-based features (pen pressure and its derivatives).

Table 30. Extracted features and their descriptions.

| **Feature** | **Description** |
|---|---|
| `duration` | Time interval between the first and the last recorded point. |
| `velocity_mean` | Average velocity. |
| `velocity_median` | Median velocity. |
| `velocity_std` | Standard deviation of velocity. |
| `velocity_min` | Minimal velocity. |
| `velocity_max` | Maximal velocity. |
| `velocity_mass` | Velocity mass: $V_N = \sum_{k=1}^{N} \lvert v_k \rvert$. |
| `velocity_x_mean` | Average horizontal velocity ($\Delta x / \Delta t$). |
| `velocity_x_median` | Median horizontal velocity. |
| `velocity_x_std` | Standard deviation of horizontal velocity. |
| `velocity_x_min` | Minimal horizontal velocity. |
| `velocity_x_max` | Maximal horizontal velocity. |
| `velocity_x_mass` | Horizontal velocity mass: $\sum_{k=1}^{N} \lvert v_{x,k} \rvert$. |
| `velocity_y_mean` | Average vertical velocity ($\Delta y / \Delta t$). |
| `velocity_y_median` | Median vertical velocity. |
| `velocity_y_std` | Standard deviation of vertical velocity. |
| `velocity_y_min` | Minimal vertical velocity. |
| `velocity_y_max` | Maximal vertical velocity. |
| `velocity_y_mass` | Vertical velocity mass: $\sum_{k=1}^{N} \lvert v_{y,k} \rvert$. |
| `accel_mean` | Average acceleration. Rate of change in velocity with respect to time; second time derivative of displacement. |
| `accel_median` | Median acceleration. |
| `accel_std` | Standard deviation of acceleration. |
| `accel_min` | Minimal acceleration. |

| Feature | Description |
|---|---|
| accel_max | Maximal acceleration. |
| accel_mass | Acceleration mass: $\sum_{k=1}^{N} \lvert a_k \rvert$. |
| accel_x_mean | Average horizontal acceleration. |
| accel_x_median | Median horizontal acceleration. |
| accel_x_std | Standard deviation of horizontal acceleration. |
| accel_x_min | Minimal horizontal acceleration. |
| accel_x_max | Maximal horizontal acceleration. |
| accel_x_mass | Horizontal acceleration mass. |
| accel_y_mean | Average vertical acceleration. |
| accel_y_median | Median vertical acceleration. |
| accel_y_std | Standard deviation of vertical acceleration. |
| accel_y_min | Minimal vertical acceleration. |
| accel_y_max | Maximal vertical acceleration. |
| accel_y_mass | Vertical acceleration mass. |
| jerk_mean | Average jerk. Rate of change in acceleration; third time derivative of displacement. |
| jerk_median | Median jerk. |
| jerk_std | Standard deviation of jerk. |
| jerk_min | Minimal jerk. |
| jerk_max | Maximal jerk. |
| jerk_mass | Jerk mass: $\sum_{k=1}^{N} \lvert j_k \rvert$. |
| snap_mean | Average snap. Rate of change in jerk; fourth time derivative of displacement. |
| snap_median | Median snap. |
| snap_std | Standard deviation of snap. |
| snap_min | Minimal snap. |
| snap_max | Maximal snap. |
| snap_mass | Snap mass: $\sum_{k=1}^{N} \lvert sn_k \rvert$. |
| crackle_mean | Average crackle. Rate of change in snap; fifth time derivative of displacement. |
| crackle_median | Median crackle. |
| crackle_std | Standard deviation of crackle. |
| crackle_min | Minimal crackle. |
| crackle_max | Maximal crackle. |
| crackle_mass | Crackle mass: $\sum_{k=1}^{N} \lvert cr_k \rvert$. |

| Feature | Description |
|---|---|
| pop_mean | Average pop. Rate of change in crackle; sixth time derivative of displacement. |
| pop_median | Median pop. |
| pop_std | Standard deviation of pop. |
| pop_min | Minimal pop. |
| pop_max | Maximal pop. |
| pop_mass | Pop mass: $\sum_{k=1}^{N} \|po_k\|$. |
| pressure_mean | Average force applied by the pen on the tablet surface. |
| pressure_median | Median pressure. |
| pressure_std | Standard deviation of pressure. |
| pressure_min | Minimal pressure. |
| pressure_max | Maximal pressure. |
| pressure_mass | Pressure mass: $P_N = \sum_{k=1}^{N} \|p_k\|$. |
| pressure_diff_mean | Average change in pressure between points $[p_i, p_{i+1}]$. |
| pressure_diff_median | Median pressure difference. |
| pressure_diff_std | Standard deviation of pressure difference. |
| pressure_diff_min | Minimal pressure difference. |
| pressure_diff_max | Maximal pressure difference. |
| pressure_diff_mass | Pressure difference mass. |
| yank_mean | Average yank. Rate of change in pressure; first time derivative of force applied on the surface. |
| yank_median | Median yank. |
| yank_std | Standard deviation of yank. |
| yank_min | Minimal yank. |
| yank_max | Maximal yank. |
| yank_mass | Yank mass: $\sum_{k=1}^{N} \|y_k\|$. |
| tug_mean | Average tug. Rate of change in yank; second time derivative of force applied on the surface. |
| tug_median | Median tug. |
| tug_std | Standard deviation of tug. |
| tug_min | Minimal tug. |
| tug_max | Maximal tug. |
| tug_mass | Tug mass: $\sum_{k=1}^{N} \|tg_k\|$. |
| snatch_mean | Average snatch. Rate of change in tug; third time derivative of force applied on the surface. |

| Feature | Description |
|---|---|
| snatch_median | Median snatch. |
| snatch_std | Standard deviation of snatch. |
| snatch_min | Minimal snatch. |
| snatch_max | Maximal snatch. |
| snatch_mass | Snatch mass: $\sum_{k=1}^{N} \lvert sn_k \rvert$. |
| shake_mean | Average shake. Rate of change in snatch; fourth time derivative of force applied on the surface. |
| shake_median | Median shake. |
| shake_std | Standard deviation of shake. |
| shake_min | Minimal shake. |
| shake_max | Maximal shake. |
| shake_mass | Shake mass: $\sum_{k=1}^{N} \lvert sh_k \rvert$. |

# Appendix 3 – Selected Features for Parkinson's Disease vs Healthy Control Models

This appendix lists the features selected during nested cross-validation for each of the top 5 HC vs PD models and their Fisher scores.

Table 31. Selected features with Fisher scores for top 5 HC vs PD models.

| Model | k | Feature | F |
|---|---|---|---|
| LR | 2 | velocity_x_mean | 1.044 |
| | | velocity_y_max | 0.497 |
| AB | 3 | velocity_x_mean | 1.044 |
| | | velocity_y_max | 0.497 |
| | | pressure_min | 0.400 |
| LR | 4 | velocity_x_mean | 1.044 |
| | | velocity_y_max | 0.497 |
| | | pressure_min | 0.400 |
| | | velocity_y_min | 0.184 |
| RF | 4 | velocity_x_mean | 1.044 |
| | | velocity_y_max | 0.497 |
| | | pressure_min | 0.400 |
| | | velocity_y_min | 0.184 |
| AB | 4 | velocity_x_mean | 1.044 |
| | | velocity_y_max | 0.497 |
| | | pressure_min | 0.400 |
| | | velocity_y_min | 0.184 |

# Appendix 4 – Selected Features for Illiterate vs Literate Models

This appendix lists the features selected during nested cross-validation for each of the top 5 LIT vs ILL models and their Fisher scores.

Table 32. Selected features with Fisher scores for top 5 LIT vs ILL models.

| Model | k | Feature | F |
|---|---|---|---|
| XGB | 4 | yank_mean | 0.223 |
| | | pressure_max | 0.151 |
| | | velocity_mean | 0.116 |
| | | velocity_x_mean | 0.088 |
| AB | 4 | yank_mean | 0.223 |
| | | pressure_max | 0.151 |
| | | velocity_mean | 0.116 |
| | | velocity_x_mean | 0.088 |
| RF | 2 | yank_mean | 0.223 |
| | | pressure_max | 0.151 |
| LR | 2 | yank_mean | 0.223 |
| | | pressure_max | 0.151 |
| LR | 4 | yank_mean | 0.223 |
| | | pressure_max | 0.151 |
| | | velocity_mean | 0.116 |
| | | velocity_x_mean | 0.088 |