TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Maria Rizo 194152 IAIB

XAI based analysis of drawing tests for the diagnosis of Parkinson's disease

Bachelor's Thesis

Supervisor: Sven Nõmm
PhD

Co-supervisor: Rajesh Kalakoti
MSc

Tallinn 2025

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Maria Rizo 194152 IAIB

XAI-PÕHINE JOONISTAMISTESTIDE ANALÜÜS PARKINSONI TÕVE DIAGNOOSIMISEL

Bakalaureusetöö

Juhendaja: Sven Nõmm
PhD

Kaasjuhendaja: Rajesh Kalakoti
MSc

Tallinn 2025

Author’s Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Maria Rizo

04.06.2025

Abstract

This thesis explores the application of Explainable Artificial Intelligence (XAI) in analyzing digital drawing tests to support the diagnosis of Parkinson's disease (PD). The study focuses on kinematic and pressure-based features extracted from drawing data collected via tablets. A machine learning workflow was implemented, featuring feature extraction, Fisher Score - based selection, and classifier training (Logistic Regression, SVM, Random Forest), with Random Forest achieving the highest accuracy (81.45 percent).

To enhance interpretability XAI methods (LIME and SHAP) were integrated, providing local explanations for model predictions. Their fidelity was quantitatively evaluated using faithfulness and monotonicity metrics. Results indicated that SHAP explanations were more consistent (faithfulness up to 0.91) than LIME, particularly for accurate classifications. Mechanical drawing tasks (e.g., spirals) outperformed cognitive tasks (e.g., digits) in both accuracy (avg. 0.63 vs. 0.52) and explanation quality.

Key contributions include:

- A reproducible, interpretable pipeline for PD diagnosis using drawing tests.
- Quantitative validation of XAI methods.
- Task-specific insights, highlighting the superiority of motion mass features in mechanical tasks.

Limitations include small sample sizes and variability in LIME's explanations. The work lays a foundation for future research into transparent AI-assisted diagnostics.

The thesis is written in english and is 33 pages long, including 6 chapters, 13 figures and 2 tables.

Annotatsioon

XAI-põhine joonistamistestide analüüs Parkinsoni tõve diagnoosimisel

Käesolev bakalaureusetöö uurib seletatava tehisintellekti (XAI) rakendamist Parkinsoni tõve diagnoosimisel digitaalsete joonistustestide põhjal. Uuringus analüüsitakse joonistamise käigus kogutud kineetilisi ja survenäitajaid (nt kiirusmass, tõukemass, rappumismass), mis on eristusvõimelised Parkinsoni tõve korral. Töös kasutatud masinõppemudelid (Logistiline regressioon, SVM, Random Forest) treeniti Fisher Score'iga valitud tunnustel, millest parima tulemuse (81,45 protsendi täpsus) andis Random Forest.

Selgitamiseks rakendati XAI meetodeid (LIME ja SHAP), mille usaldusväärsust hinnati faithfulness ja monotonicity meetrikute abil. SHAP-i selgitused olid stabiilsemad (kuni 0,91 faithfulness) ning mehaanilised testid (nt spiraal) andsid paremaid tulemusi kui kognitiivsed (nt numbrid).

Töö peamised panused:

- Selgitustega varustatud masinõppeprotsess Parkinsoni tõve diagnoosimiseks.
- XAI meetodite kvantitatiivne hindamine.
- Tõestus, et liigutustunnused on efektiivsemad mehaanilistes ülesannetes.

Piiranguteks on väike andmemaht ja LIME-i ebastabiilsus. Töö annab aluse edasisteks uuringuteks arvutiabistatud diagnoosimisel.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 33 leheküljel, 6 peatükki, 13 joonist, 2 tabelit.

List of Abbreviations and Terms

XAI	Explainable Artificial Intelligence
PD	Parkinson's Disease
KT	Known Typical
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations

Table of Contents

1 Introduction . . . 9

2 Problem Statement . . . 10

2.1 Main goal and objectives . . . 10

3 Background . . . 11

3.1 Hardware and software . . . 11

3.2 Motion Mass Parameters . . . 12

3.3 Explainable AI (XAI) . . . 12

4 Methodology . . . 13

4.1 Feature Extraction . . . 13

4.1.1 Feature Extraction pipeline . . . 14

4.2 Feature Selection . . . 15

4.3 Machine Learning Workflow . . . 16

4.3.1 Selected Features . . . 16

4.3.2 Classifier Selection . . . 16

4.3.3 Cross-Validation . . . 17

4.4 Model Explainability with LIME and SHAP . . . 17

4.4.1 LIME . . . 18

4.4.2 SHAP . . . 18

4.5 Automatization for other tests . . . 19

4.6 LIME and SHAP evaluation . . . 21

4.6.1 Faithfulness Metric . . . 22

4.6.2 Monotonicity Metric . . . 22

4.6.3 Output . . . 23

5 Results . . . 25

5.1 Performance-Explanation Relationship . . . 26

5.2 Test-Type Patterns . . . 28

5.3 Conclusion . . . 28

5.4 Limitations and Considerations . . . 28

5.5 Contribution and Significance . . . 29

6 Summary . . . 30

References **31**

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis **33**

List of Figures

1 *Workflow.* 10

2 *Part of json file.* 11

3 *Machine Learning Workflow.* 16
4 *LIME Explanation.* 18
5 *SHAP Explanation.* 19
6 *All Model Performances.* 20
7 *Lines test case picture* 21
8 *LIME Explanation.* 21
9 *SHAP Explanation.* 21

10 *Faithfulness values.* 25
11 *Monotonicity values.* 26
12 *Faithfulness values and models accuracy.* 27
13 *Monotonicity values and models accuracy.* 27

List of Tables

1 *Performance metrics of classifiers* 17

2 *Quantitative Evaluation of XAI Methods Across Models* 23

1. Introduction

Parkinson's disease is one of the most common neurodegenerative disorders worldwide, primarily affecting the motor system. Although there is currently no known cure, early diagnosis and appropriate treatment can significantly improve patients' quality of life. The disease manifests through a variety of motor symptoms, including tremors, rigidity, slowness of movement and impaired coordination. Because of these characteristics, tasks involving fine motor control—such as handwriting and drawing—can serve as useful indicators for early detection of the disease.

Fine-motor assessments have long been used in neurology to identify impairments, and technological advancements have brought these tests into the digital domain. Instead of relying on traditional paper-and-pencil tests, tablets and digital pens can now capture not only the geometry of drawn shapes but also additional parameters such as time, pressure, velocity and drawing angle. These parameters are not visible to the human eye but can provide deeper insights into the motor behavior of individuals. As a result, tablets and artificial intelligence have emerged as valuable tools for the objective analysis of Parkinson's symptoms.

However despite the potential of machine learning models, a critical challenge is their interpretability. Clinicians and researchers must understand the reasoning behind each prediction in order to trust AI-assisted diagnostic support. Explainable Artificial Intelligence (XAI) provides tools to expose internal decision logic and highlight which input features influence predictions the most.

The aim of this thesis is to evaluate whether interpretable machine learning methods can be effectively used to support the diagnosis of Parkinson's disease through drawing tests. By combining motion analysis with explainability tools, the study seeks to contribute toward the development of more transparent and reliable computer-aided diagnostic systems.

2. Problem Statement

High accuracy of machine learning models does not inherently translate into trust, especially in sensitive domains like healthcare, where understanding the reasoning behind predictions is essential. XAI methods can make model decisions more transparent, enabling clinicians to evaluate and potentially rely on AI-assisted diagnosis.

2.1 Main goal and objectives

The main goal of the present thesis is to implement a statistical machine learning workflow for the analysis of drawing and writing tests enriched by the post hoc explanation and explanation evaluation steps. To achieve the main goal, the following subproblems must be solved:

- Reproduce previously performed research aimed at feature engineering and selection, classifier training, and validation. This step serves two purposes. Validate previously performed research and provide the basis for explanations and their evaluation.
- Integrate LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) into the workflow.
- Calculate faithfulness and monotonicity metrics for LIME and SHAP.
- Apply the workflow to the available tests and evaluate the results.

For the validation of the first step, a two-stage procedure would be used. The first is a comparison of the selected features with the previously published results[1]. For the evaluation of the classifier's accuracy metrics, precision, recall and the F1 score will be used. However, values above 0.7 should be considered acceptable. LIME and SHAP explanations could not be compared to the previously published results, but their monotonicity and faithfulness values could provide information on the goodness of the explanations. To the best of knowledge of the author, there are no published results to compare the numerical values of such outputs, but a comparison between the tests should reveal the limitations of the proposed approach.

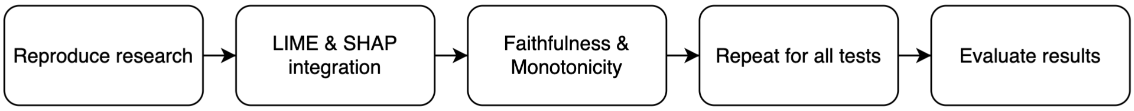


Figure 1. *Workflow.*

3. Background

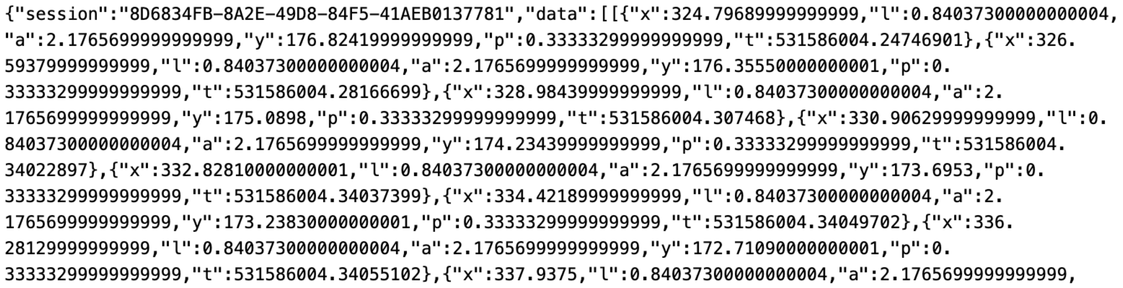
The present thesis is part of a larger research that studies human motor functions whose objectives are to support the diagnosis of neurodegenerative diseases, detect early cognitive impairments, and recognize signs of fatigue. The datasets used in this research were provided by supervisor. The findings of this work might be later used further in research.

The data acquisition process was carried out under strict privacy law guidelines. As such any and all files containing sensitive data are excluded from this work (all the json files used for model training and testing).

3.1 Hardware and software

Data used in present thesis was acquired during previous research[1]. For data gathering during writing tests, Apple iPad pro (2016) with a 9.7-inch screen and Apple pen (stylus) were used. To collect movements of the stylus tip, software and interface suitable for the task was developed. The coordinates of the apple pen tip and the pressure applied to the screen were saved to the matrix. The rows of the matrix correspond to the observation points acquired up to 200 times per second, and the columns contain information that describes each point. For each test collected data were saved for future processing in JavaScript Object Notation (JSON) files.

An example of data in a JSON file is shown below in (Figure 1). For each point was recorded information about the X coordinate (x) and the Y coordinate (y), pressure applied to the screen (p), stylus orientation altitude (l) and azimuth (a), time stamp (t).[2]



```
{"session":"8D6834FB-8A2E-49D8-84F5-41AEB0137781","data":[[{"x":324.79689999999999,"l":0.84037300000000004,
"a":2.1765699999999999,"y":176.82419999999999,"p":0.33333299999999999,"t":531586004.24746901},{"x":326.
59379999999999,"l":0.84037300000000004,"a":2.1765699999999999,"y":176.35550000000001,"p":0.
33333299999999999,"t":531586004.28166699},{"x":328.98439999999999,"l":0.84037300000000004,"a":2.
1765699999999999,"y":175.0898,"p":0.33333299999999999,"t":531586004.307468},{"x":330.90629999999999,"l":0.
84037300000000004,"a":2.1765699999999999,"y":174.23439999999999,"p":0.33333299999999999,"t":531586004.
34022897},{"x":332.82810000000001,"l":0.84037300000000004,"a":2.1765699999999999,"y":173.6953,"p":0.
33333299999999999,"t":531586004.34037399},{"x":334.42189999999999,"l":0.84037300000000004,"a":2.
1765699999999999,"y":173.23830000000001,"p":0.33333299999999999,"t":531586004.34049702},{"x":336.
28129999999999,"l":0.84037300000000004,"a":2.1765699999999999,"y":172.71090000000001,"p":0.
33333299999999999,"t":531586004.34055102},{"x":337.9375,"l":0.84037300000000004,"a":2.1765699999999999,
```

Figure 2. *Part of json file.*

3.2 Motion Mass Parameters

Recent studies[1] have demonstrated that kinematic and pressure-based features extracted from digital drawing tests can objectively quantify motor impairments in Parkinson's disease.

"Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics" (2022)[1] proposed "motion mass" parameters (velocity mass, jerk mass, etc.) derived from Archimedean spirals, achieving 84.3 percent accuracy in Parkinson's disease detection. These integral-like features capture cumulative deviations in motor control, which are less sensitive to noise than point-wise metrics. This metrics will be used to train models.

3.3 Explainable AI (XAI)

Transparency in machine learning is critical for medical applications. Explainable AI refers to a set of processes and methods that aim to provide a clear and human-understandable explanation for the decisions generated by AI and machine learning models.[3]

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.[4]

- Quantifies each feature's contribution to predictions using game theory.
- Advantages: Global consistency (features retain importance across samples).

LIME (Local Interpretable Model-agnostic Explanations):
Instead of providing a global understanding of the model on the entire dataset, LIME focuses on explaining the model's prediction for individual instances.[5][3]

- Approximates complex models with interpretable local linear models.
- Advantages: Intuitive for case-by-case analysis (e.g., "High jerk mass $\rightarrow$ Parkinson's disease" for a specific patient).

4. Methodology

In this thesis, digital drawings from Parkinson's patients and healthy individuals are analyzed to compute a set of kinematic and pressure-based features. In particular, a family of "motion mass" parameters — such as velocity mass, jerk mass, and shake mass — are extracted from raw drawing data. Feature relevance is assessed using the Fisher Score[6], and selected features are used to train multiple machine learning classifiers. The models are evaluated using nested cross-validation[7] and multiple performance metrics, such as accuracy, precision, recall and F1 score. Finally, LIME[5] and SHAP[8] are applied to interpret and compare the predictions of the trained models and their interpretations are assessed through faithfulness and monotonicity metrics.

The process is separated into two phases, learning and developing the process using only spiral drawing tests data, adjusting the process to generate models and explanations for any particular test.

4.1 Feature Extraction

To ensure complete and correct parsing, a recursive data extraction function was developed to extract all relevant values ("x", "y" pen coordinates, "p" pressure and "t" timestamp) for one file.

The collected data was processed to compute a series of kinematic and pressure-related features. These features capture the dynamic properties of the drawing process and serve as inputs to classification models.

- Velocity Mass: total magnitude of velocity throughout the drawing.
- Acceleration Mass: total magnitude of acceleration.
- Jerk Mass: total magnitude of jerk (change in acceleration).
- Yank Mass: change in pressure over time.
- Tug, Snatch, Shake Mass: higher-order pressure derivatives capturing variations in drawing force.

Features were computed using formulas from "Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics"[1], with small constants added where needed to prevent division by zero.

4.1.1 Feature Extraction pipeline

Process begins with raw digital drawing data consisting of:

- x: x-coordinate of pen position
- y: y-coordinate of pen position
- p: pressure value
- t: timestamp

First, we convert timestamps into time differences:

$$\Delta \mathbf{t} = [t_1 - t_0, t_2 - t_1, ..., t_n - t_{n-1}] \tag{4.1}$$

We compute successive derivatives of pressure with respect to time:

$$\begin{aligned}
&\text{Yank: } yank_i = \frac{p_{i+1} - p_i}{\Delta t_i} \\
&\text{Tug: } tug_i = \frac{yank_{i+1} - yank_i}{\Delta t_i} \\
&\text{Snatch: } snatch_i = \frac{tug_{i+1} - tug_i}{\Delta t_i} \\
&\text{Shake: } shake_i = \frac{snatch_{i+1} - snatch_i}{\Delta t_i}
\end{aligned} \tag{4.2}$$

From the positional data, we compute displacement and its derivatives:

$$\begin{aligned}
&\Delta x_i = x_{i+1} - x_i, \quad \Delta y_i = y_{i+1} - y_i \\
&\text{Displacement: } d_i = \sqrt{\Delta x_i^2 + \Delta y_i^2} \\
&\text{Velocity: } v_i = \frac{d_i}{\Delta t_i} \\
&\text{Acceleration: } a_i = \frac{v_{i+1} - v_i}{\Delta t_i} \\
&\text{Jerk: } j_i = \frac{a_{i+1} - a_i}{\Delta t_i}
\end{aligned} \tag{4.3}$$

The "mass" parameters represent the cumulative magnitude of each dynamic quantity:

$$\begin{aligned}
&\text{Velocity Mass: } \sum |v_i| \\
&\text{Acceleration Mass: } \sum |a_i| \\
&\text{Jerk Mass: } \sum |j_i| \\
&\text{Yank Mass: } \sum |yank_i| \\
&\text{Tug Mass: } \sum |tug_i| \\
&\text{Snatch Mass: } \sum |snatch_i| \\
&\text{Shake Mass: } \sum |shake_i|
\end{aligned} \tag{4.4}$$

This feature extraction pipeline captures both kinematic (movement-related) and pressure dynamics that are particularly relevant for characterizing Parkinson tremors and other motor symptoms.

The implementation of feature extraction is available at github repository, in particular folder Part 1 contains feature extraction process development, for a refined process refer to Part 5/modelCreator.py.[9]

4.2 Feature Selection

After feature extraction, not all features were equally informative for classification. To rank their discriminative power, the Fisher Score was applied to each feature across the two groups.

The Fisher Score, as defined by Aggarwal (2014)[6], measures the ratio of interclass separation to intraclass variance: The Fisher Score $F(X)$ for a feature X is computed as:

$$F(X) = \frac{\sum_{i=1}^{k} p_i(\mu_i - \mu)^2}{\sum_{i=1}^{k} p_i \sigma_i^2}$$

- k = number of classes
- p_i = proportion (or probability) of samples in class i
- μ_i = mean of feature X for class i
- μ = global mean of feature X
- σ_i^2 = variance of feature X within class i

In this study, 3 the top-ranked features were selected for use in classification.

4.3 Machine Learning Workflow

Development of machine learning pipeline used to classify Parkinson's patients based on their drawing tests. The process includes classifier selection, evaluation using cross-validation[7], and performance assessment with standard classification metrics. The goal is to compare different models trained on the most informative features and identify which ones offer the best predictive performance.

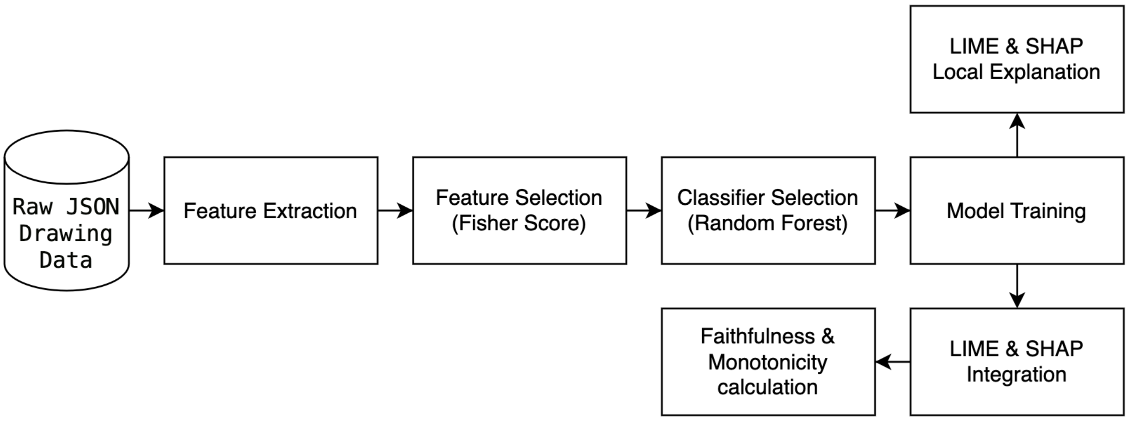

LIME & SHAP Local Explanation
Raw JSON Drawing Data
Feature Extraction
Feature Selection (Fisher Score)
Classifier Selection (Random Forest)
Model Training
Faithfulness & Monotonicity calculation
LIME & SHAP Integration

Figure 3. *Machine Learning Workflow.*

4.3.1 Selected Features

Based on Fisher Score ranking described previously, the three most discriminative features (Shake Mass, Jerk Mass, and Snatch Mass for spiral test) were selected. These features were extracted from each sample and used as input for training classifiers.

4.3.2 Classifier Selection

Three well-established classification algorithms were chosen for evaluation:

- Logistic Regression: A linear classifier used as a baseline due to its simplicity and interpretability.
- Support Vector Machine (SVM): Effective in high-dimensional spaces and robust to small datasets.
- Random Forest: An ensemble method that combines multiple decision trees for improved generalization.

Each classifier was tested using the same feature set and evaluation framework to ensure a fair comparison.

4.3.3 Cross-Validation

To obtain reliable estimates of model performance and reduce the risk of overfitting, nested cross-validation was used.[7]

Each model was evaluated using the following performance metrics:

- Accuracy: Overall percentage of correct predictions.
- Precision: Proportion of predicted positives that were actually positive.
- Recall (Sensitivity): Proportion of actual positives that were correctly identified.
- F1 Score: Harmonic mean of precision and recall, useful in imbalanced datasets.

These metrics were computed using Scikit-learn's evaluation utilities, and scores were averaged across folds in cross-validation. The results are as follows:

Table 1. *Performance metrics of classifiers*

	Accuracy	**Precision**	**Recall**	**F1-Score**
SVM:	0.7236	0.7400	0.5000	0.5467
Random Forest:	0.8145	0.8333	0.7000	0.7267
Logistic Regression:	0.7236	0.7400	0.5000	0.5467

Random Forest was chosen to proceed with as the best performing classifier. Hyperparameters for the model were tuned using GridSearchCV.

The implementation of feature and classifier selection is available at github repository, in particular folder Part 2, model hyperparameters tunning and creation is available in Part 3, for a refined process refer to Part 5/modelCreator.py.[9]

4.4 Model Explainability with LIME and SHAP

XAI provides tools to interpret model behavior. In this thesis, two XAI methods were applied: LIME[5] and SHAP[8]. Both techniques offer insights into which features influenced each prediction and to what extent, enabling human validation and understanding.

4.4.1 LIME

LIME explains individual predictions by approximating the classifier with a simpler, interpretable model (e.g. linear regression) around the vicinity of a specific instance. It perturbs the input data slightly and observes how the model's prediction changes, allowing it to assign importance scores to each feature.[5]

In this project, LIME was used to explain predictions made on selected samples. Visual outputs showed the contribution of each feature (e.g., Shake Mass, Jerk Mass) to the prediction probability. Below is explanation from LIME made on spiral test file from Parkinson's patient:

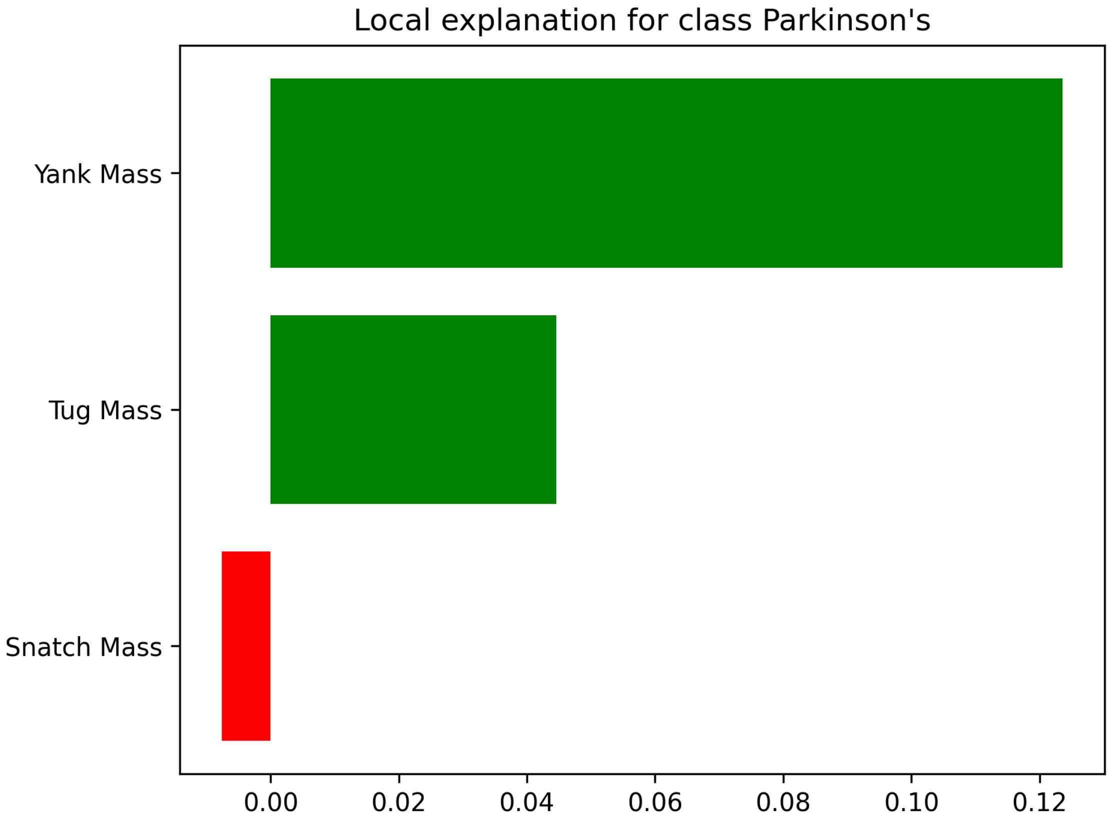


Figure 4. *LIME Explanation.*

4.4.2 SHAP

SHAP is based on cooperative game theory and assigns an additive importance value (Shapley value) to each feature based on how much it contributed to moving the model output from the baseline. Unlike LIME, SHAP ensures consistency and local accuracy, making it theoretically grounded and often more stable across different runs.[8]

SHAP values were computed for all samples and individual plots were created to show how specific values of a feature increased or decreased the probability of a sample being classified as Parkinson's.[10]

Below is explanation from SHAP made on spiral test file from Parkinson's patient:

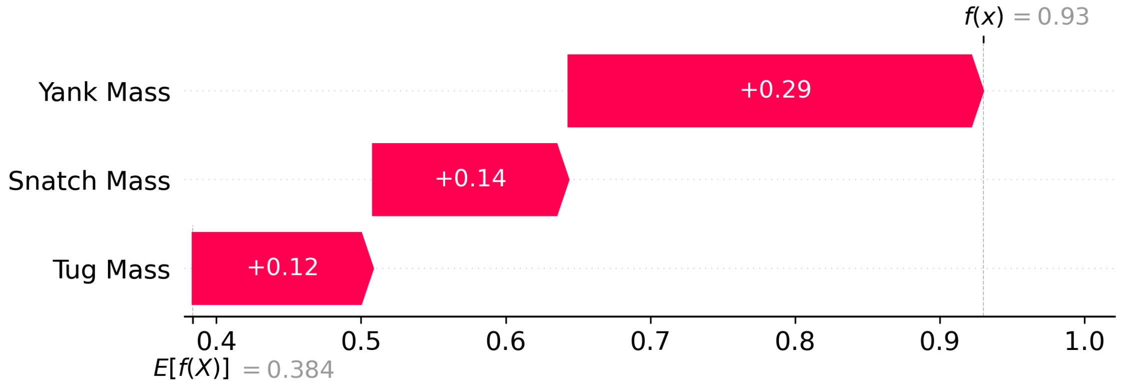


Figure 5. *SHAP Explanation.*

4.5 Automatization for other tests

The process was refined and consolidated into a single program that accepts paths to a particular drawing test dataset and creates a model based on it as well as returns said model performance. For example model for lines test is created:

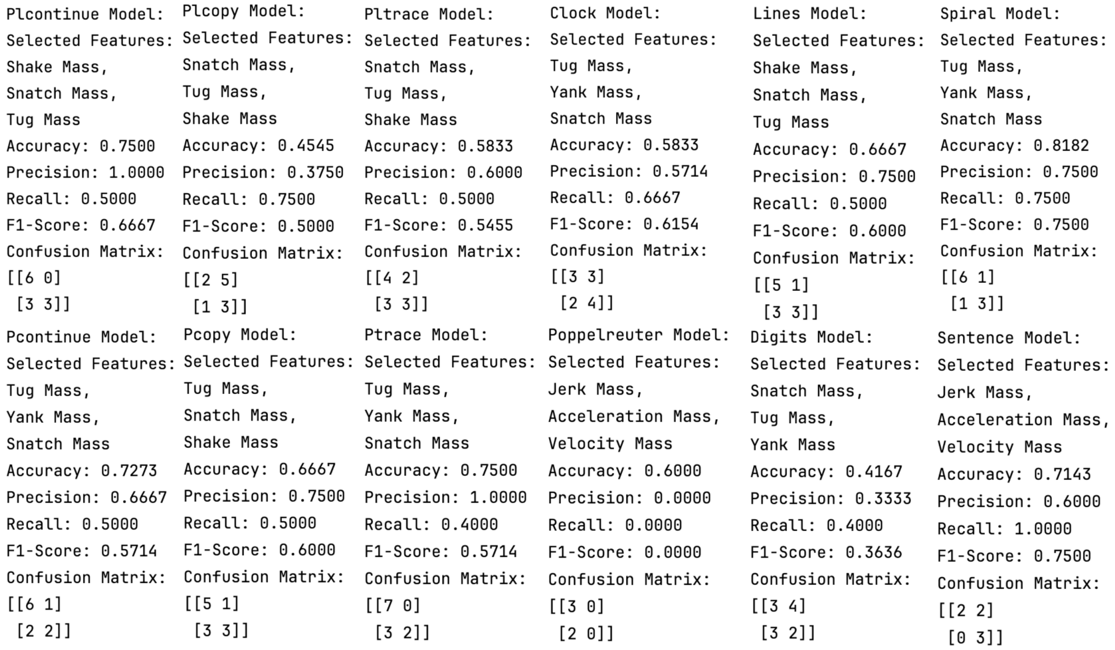
Created lines test model performance:

```
Lines Model:

Selected Features:
Shake Mass,
Snatch Mass,
Tug Mass

Accuracy: 0.6667
Precision: 0.7500
Recall: 0.5000
F1-Score: 0.6000
Confusion Matrix:
[[5 1]
 [3 3]]
```

And on figure 4 are all the created models performances. From which we can conclude that same mass metrics what were developed for spiral drawing test in particular can actually be utilized to determine Parkinson's cases in other tests done on same device with higher accuracy of predictions if said test is more of a mechanical nature.



```
Plcontinue Model:
Selected Features:
Shake Mass,
Snatch Mass,
Tug Mass
Accuracy: 0.7500
Precision: 1.0000
Recall: 0.5000
F1-Score: 0.6667
Confusion Matrix:
[[6 0]
 [3 3]]
```

```
Plcopy Model:
Selected Features:
Snatch Mass,
Tug Mass,
Shake Mass
Accuracy: 0.4545
Precision: 0.3750
Recall: 0.7500
F1-Score: 0.5000
Confusion Matrix:
[[2 5]
 [1 3]]
```

```
Pltrace Model:
Selected Features:
Snatch Mass,
Tug Mass,
Shake Mass
Accuracy: 0.5833
Precision: 0.6000
Recall: 0.5000
F1-Score: 0.5455
Confusion Matrix:
[[4 2]
 [3 3]]
```

```
Clock Model:
Selected Features:
Tug Mass,
Yank Mass,
Snatch Mass
Accuracy: 0.5833
Precision: 0.5714
Recall: 0.6667
F1-Score: 0.6154
Confusion Matrix:
[[3 3]
 [2 4]]
```

```
Lines Model:
Selected Features:
Shake Mass,
Snatch Mass,
Tug Mass
Accuracy: 0.6667
Precision: 0.7500
Recall: 0.5000
F1-Score: 0.6000
Confusion Matrix:
[[5 1]
 [3 3]]
```

```
Spiral Model:
Selected Features:
Tug Mass,
Yank Mass,
Snatch Mass
Accuracy: 0.8182
Precision: 0.7500
Recall: 0.7500
F1-Score: 0.7500
Confusion Matrix:
[[6 1]
 [1 3]]
```

```
Pcontinue Model:
Selected Features:
Tug Mass,
Yank Mass,
Snatch Mass
Accuracy: 0.7273
Precision: 0.6667
Recall: 0.5000
F1-Score: 0.5714
Confusion Matrix:
[[6 1]
 [2 2]]
```

```
Pcopy Model:
Selected Features:
Tug Mass,
Snatch Mass,
Shake Mass
Accuracy: 0.6667
Precision: 0.7500
Recall: 0.5000
F1-Score: 0.6000
Confusion Matrix:
[[5 1]
 [3 3]]
```

```
Ptrace Model:
Selected Features:
Tug Mass,
Yank Mass,
Snatch Mass
Accuracy: 0.7500
Precision: 1.0000
Recall: 0.4000
F1-Score: 0.5714
Confusion Matrix:
[[7 0]
 [3 2]]
```

```
Poppelreuter Model:
Selected Features:
Jerk Mass,
Acceleration Mass,
Velocity Mass
Accuracy: 0.6000
Precision: 0.0000
Recall: 0.0000
F1-Score: 0.0000
Confusion Matrix:
[[3 0]
 [2 0]]
```

```
Digits Model:
Selected Features:
Snatch Mass,
Tug Mass,
Yank Mass
Accuracy: 0.4167
Precision: 0.3333
Recall: 0.4000
F1-Score: 0.3636
Confusion Matrix:
[[3 4]
 [3 2]]
```

```
Sentence Model:
Selected Features:
Jerk Mass,
Acceleration Mass,
Velocity Mass
Accuracy: 0.7143
Precision: 0.6000
Recall: 1.0000
F1-Score: 0.7500
Confusion Matrix:
[[2 2]
 [0 3]]
```

Figure 6. *All Model Performances.*

After model creation one case file from dataset can be selected to receive explanation on, that includes drawing of a test, results of model prediction, accurate prediction, explanation from LIME, explanation from SHAP and their faithfulness and monotonicity metrics[11] for each feature that was selected for this dataset. More on that in next section. One case from lines test was explained as example, below are as follows, test drawing, LIME explanation, SHAP explanation:

Created explanations performances:

```
Model predicted: Parkinson's
Case was: Parkinson's
=== Parkinson's Case Lime ===
Shake Mass - Faithfulness: 0.9948, Monotonicity: [False]
Snatch Mass - Faithfulness: -0.4096, Monotonicity: [False]
Tug Mass - Faithfulness: -0.5852, Monotonicity: [False]
=== Parkinson's Case Shap ===
Shake Mass - Faithfulness: 0.9948, Monotonicity: [False]
Snatch Mass - Faithfulness: -0.4096, Monotonicity: [False]
Tug Mass - Faithfulness: -0.5852, Monotonicity: [False]
```

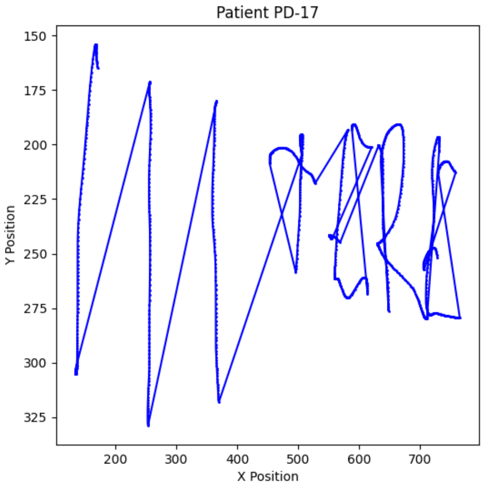


Figure 7. *Lines test case picture*

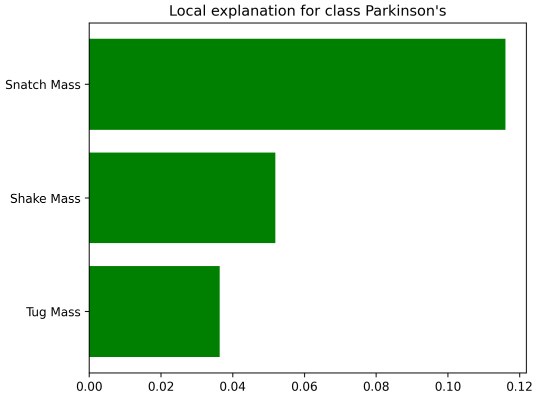


Figure 8. *LIME Explanation.*

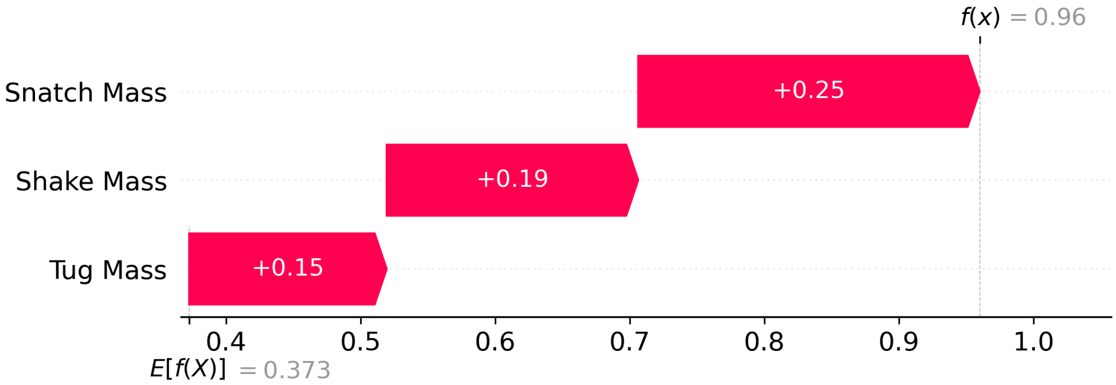


Figure 9. *SHAP Explanation.*

LIME and SHAP integration programms are available at github repository, in particular folder Part 4, for a refined process refer to Part 5/explanationPic.py.[9]

4.6 LIME and SHAP evaluation

Various quantitative metrics have been introduced in the literature to assess the results of explainability methods. Two primary metrics such as faithfulness and monotonicity[12] were employed as suitable criteria for local explanations of LIME and SHAP in this work. Their scores were computed as averages and separated between accurate classification and miss-classification cases.

4.6.1 Faithfulness Metric

Faithfulness measures how well the feature importance scores from explainers (LIME/SHAP) g reflect the actual importance of the features in the black-box model M for input x. It is computed using Pearson's correlation coefficient[13] between the sum of attributions and the corresponding difference in output values.[11]

$$\mu_F(M, g; x) = \rho_{B \in \binom{[d]}{|B|}} \left(\sum_{i \in B} g(M, x)_i,\ M(x) - M(x_B) \right)$$

- M: Black-box model.
- g: Explanation function (e.g., LIME/SHAP).
- x: Input instance.
- B: Subset of features set to baseline values.
- ρ: Pearson's correlation coefficient.
- x_B: Input x with features in B set to baseline.

A value close to 1 means high faithfulness — the explanation is a good reflection of the model's actual behavior.
A value near 0 means the explanation is no better than random.
A negative value, means the explanation is actively misleading: explainer said some features are important, but in reality, increasing them made the model less likely to predict what it predicted, or vice versa.

4.6.2 Monotonicity Metric

Evaluates whether incremental changes in input features lead to consistent changes in explanations.[11]

Given two input points $x, x' \in \mathbb{R}^d$ such that $x_i \leq x'_i$ for all $i \in \{1, \ldots, d\}$, the explanation g is said to be monotonic if, for any subset $S \subseteq \{1, \ldots, d\}$:

$$\sum_{i \in S} g(M, x)_i \leq \sum_{i \in S} g(M, x')_i$$

$$M(x) - M\left(x[x_S = \bar{x}_S]\right) \leq M(x') - M\left(x'[x'_S = \bar{x}_S]\right)$$

High monotonicity means explanations consistently reflect feature importance as inputs vary. Monotonicity was evaluated per test file as a binary outcome (True/False), for aggregate reporting, results were averaged into a single score (e.g., 2 out of 3 cases = 0.67), representing the proportion of monotonic behavior observed.

4.6.3 Output

In following table both faithfulness and monotonicity results averages and standard deviations separated in accurate and miss-classification cases for each test dataset

Table 2. *Quantitative Evaluation of XAI Methods Across Models*

Model	Explainer	Accuracy Class	Faithfulness	Monotonicity
Clock	LIME	Accurate	0.39 ± 0.61	0.57 ± 0.49
		Misclassified	-0.05 ± 0.83	0.40 ± 0.49
	SHAP	Accurate	0.82 ± 0.19	0.71 ± 0.45
		Misclassified	-0.59 ± 0.55	0.40 ± 0.49
Digits	LIME	Accurate	0.41 ± 0.72	0.80 ± 0.40
		Misclassified	-0.48 ± 0.64	0.43 ± 0.49
	SHAP	Accurate	0.32 ± 0.73	1.00 ± 0.00
		Misclassified	-0.49 ± 0.69	0.57 ± 0.49
Lines	LIME	Accurate	0.40 ± 0.61	0.62 ± 0.48
		Misclassified	0.26 ± 0.18	1.00 ± 0.00
	SHAP	Accurate	0.28 ± 0.76	0.75 ± 0.43
		Misclassified	-0.12 ± 0.86	0.75 ± 0.43
Pcontinue	LIME	Accurate	0.39 ± 0.74	1.00 ± 0.00
		Misclassified	0.45 ± 0.55	0.67 ± 0.47
	SHAP	Accurate	0.74 ± 0.18	0.88 ± 0.33
		Misclassified	-0.48 ± 0.15	0.67 ± 0.47
Pcopy	LIME	Accurate	0.45 ± 0.67	0.88 ± 0.33
		Misclassified	0.01 ± 0.84	0.50 ± 0.50
	SHAP	Accurate	0.62 ± 0.35	0.88 ± 0.33
		Misclassified	-0.86 ± 0.13	0.50 ± 0.50
Ptrace	LIME	Accurate	0.59 ± 0.45	1.00 ± 0.00
		Misclassified	0.60 ± 0.49	0.67 ± 0.47
	SHAP	Accurate	0.41 ± 0.56	0.89 ± 0.31
		Misclassified	-0.74 ± 0.12	0.00 ± 0.00
Plcontinue	LIME	Accurate	-0.17 ± 0.78	0.67 ± 0.47
		Misclassified	-1.00 ± 0.00	0.67 ± 0.47
	SHAP	Accurate	0.23 ± 0.57	0.67 ± 0.47
		Misclassified	-0.41 ± 0.00	0.67 ± 0.47
Plcopy	LIME	Accurate	0.58 ± 0.60	0.40 ± 0.49
		Misclassified	0.57 ± 0.29	0.17 ± 0.37
	SHAP	Accurate	0.56 ± 0.73	0.40 ± 0.49
		Misclassified	-0.63 ± 0.47	0.17 ± 0.37

Model	**Explainer**	**Accuracy Class**	**Faithfulness**	**Monotonicity**
Pltrace	LIME	Accurate	0.53 ± 0.64	0.86 ± 0.35
		Misclassified	-0.21 ± 0.93	0.80 ± 0.40
	SHAP	Accurate	0.76 ± 0.26	1.00 ± 0.00
		Misclassified	-0.58 ± 0.70	0.80 ± 0.40
Poppelreuter	LIME	Accurate	-0.30 ± 0.82	0.33 ± 0.47
		Misclassified	-0.01 ± 0.73	0.50 ± 0.50
	SHAP	Accurate	0.65 ± 0.10	0.33 ± 0.47
		Misclassified	-0.68 ± 0.24	0.50 ± 0.50
Sentence	LIME	Accurate	-0.67 ± 0.62	0.80 ± 0.40
		Misclassified	0.39 ± 0.39	1.00 ± 0.00
	SHAP	Accurate	0.91 ± 0.05	0.60 ± 0.49
		Misclassified	-0.22 ± 0.22	1.00 ± 0.00
Spiral	LIME	Accurate	0.07 ± 0.58	0.78 ± 0.42
		Misclassified	0.92 ± 0.02	1.00 ± 0.00
	SHAP	Accurate	0.72 ± 0.42	0.56 ± 0.50
		Misclassified	-0.95 ± 0.02	0.50 ± 0.50

LIME and SHAP evaluation programm used in this project is available at github repository, in particular folder Part 5/explanationAverage.py.[9]

5. Results

Previously performed research[1] aimed at feature engineering and selection was successfully reproduced. Classifiers were trained and validated. LIME[5] and SHAP[8] were integrated and faithfulness and monotonicity metrics[12] were calculated. The process was applied to all the available tests.

Explanations Evaluation revealed following:

SHAP generally produced more faithful explanations compared to LIME. Across almost all models, SHAP explanations for accurately classified samples have consistently high faithfulness (often >0.6, sometimes as high as 0.91). In contrast LIME shows much more fluctuation, including negative faithfulness scores in some cases (e.g., Sentence model: -0.67). For misclassified samples faithfulness drops sharply, especially for SHAP. Many SHAP values for misclassified inputs are strongly negative (e.g., Pltrace: -0.58, Spiral: -0.95, Pcopy: -0.86). LIME is also affected but often remains closer to zero or slightly positive. Strong negative faithfulness values indicate misleading explanations, which may reflect the model's incorrect predictions in misclassified cases.

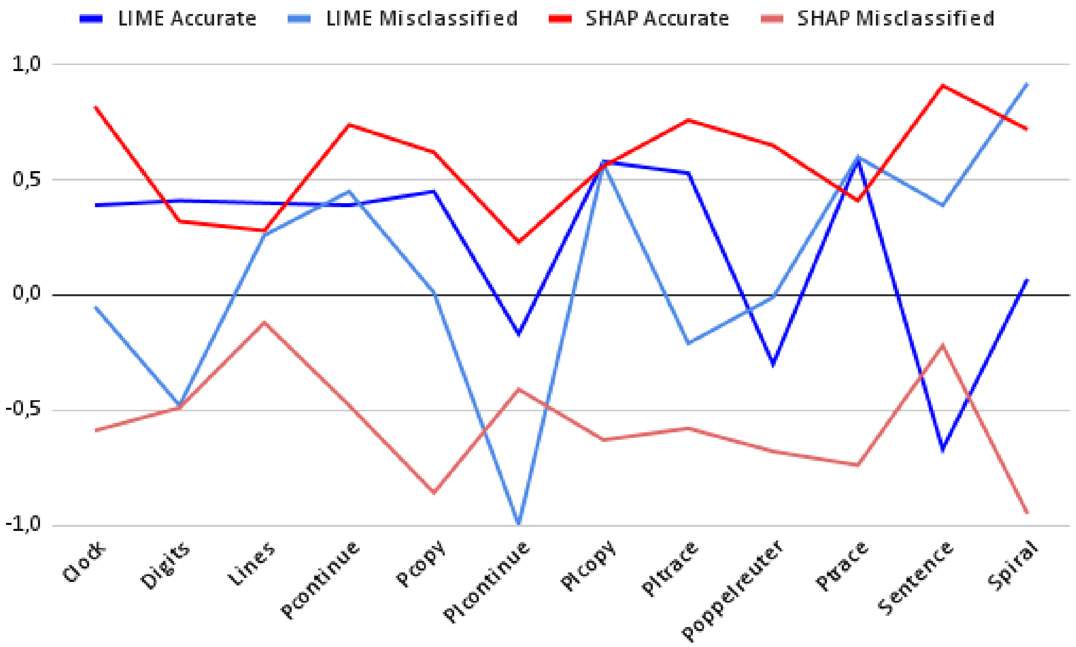


Figure 10. *Faithfulness values.*

Monotonicity was generally high for both methods (overall avg: 0.68) and with less variance than faithfulness scores. Perfect 1.00 scores occur frequently.

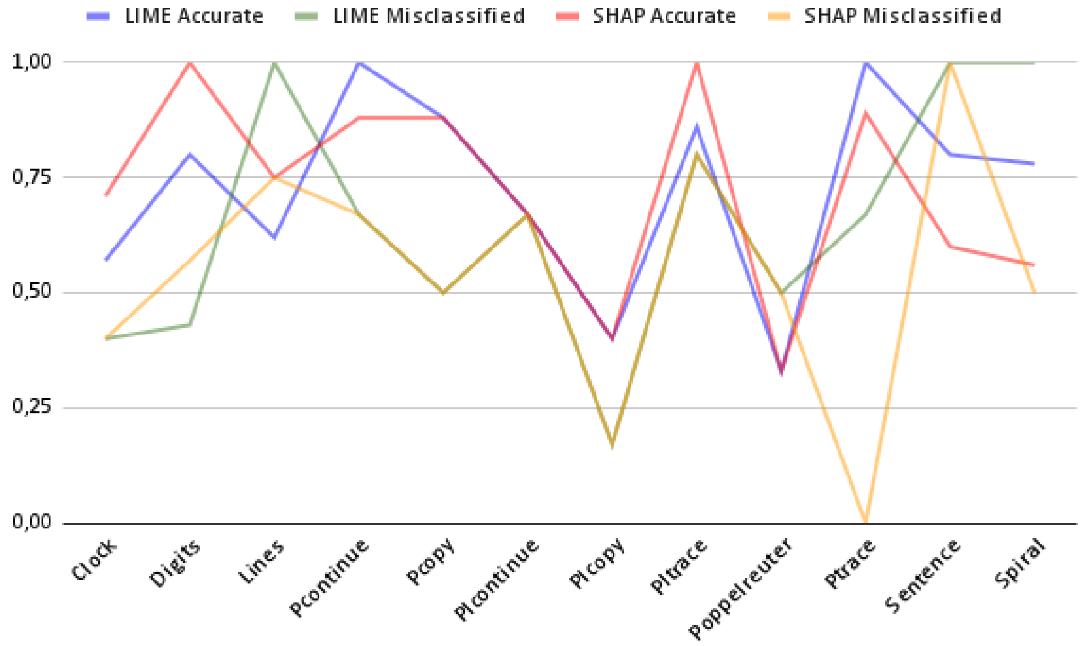


Figure 11. *Monotonicity values.*

From this can be concluded that SHAP is more aligned with the model's actual behavior on correct predictions, which supports its use for trustworthy explanation generation in accurate cases.

5.1 Performance-Explanation Relationship

Refer to Figure 12 and Figure 13 for visuals.
Higher accuracy models tend to have:

- Better faithfulness (Spiral: 0.82 acc → 0.72 faithfulness)
- Better monotonicity (Sentence: 0.71 acc → 1.00 monotonicity)

Lower accuracy models show:

- Bigger explanation variance in faithfulness (Digits: 0.42 acc → 0.32 ± 0.73 faithfulness)
- More frequent negative faithfulness
- Much bigger variation across the models in monotonicity while maintaining much smaller variation for one model

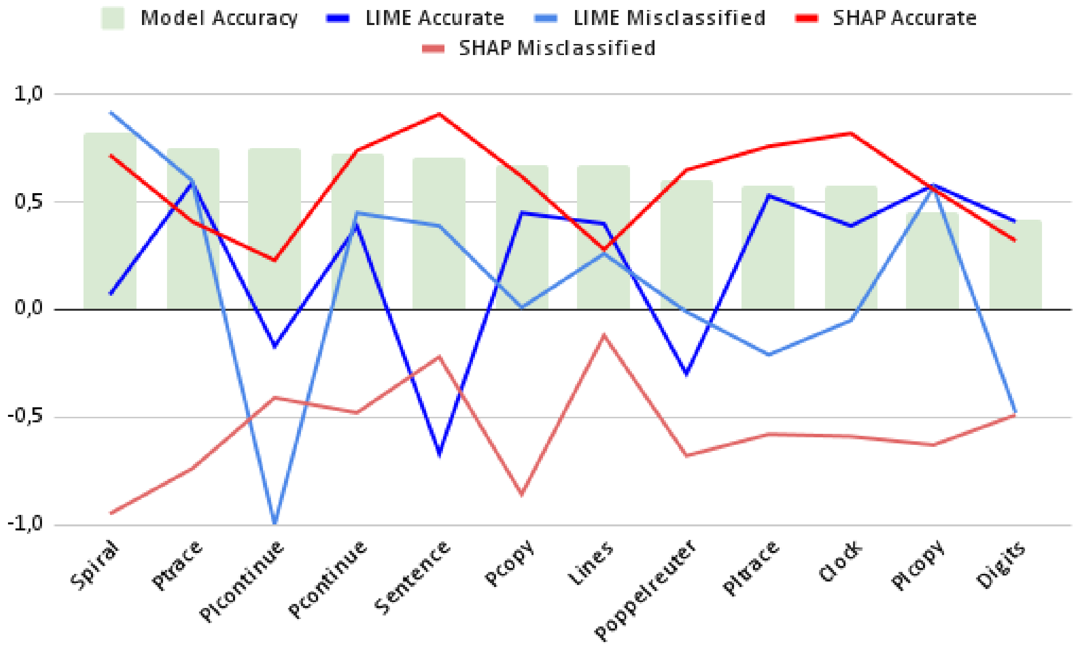


Figure 12. *Faithfulness values and models accuracy.*

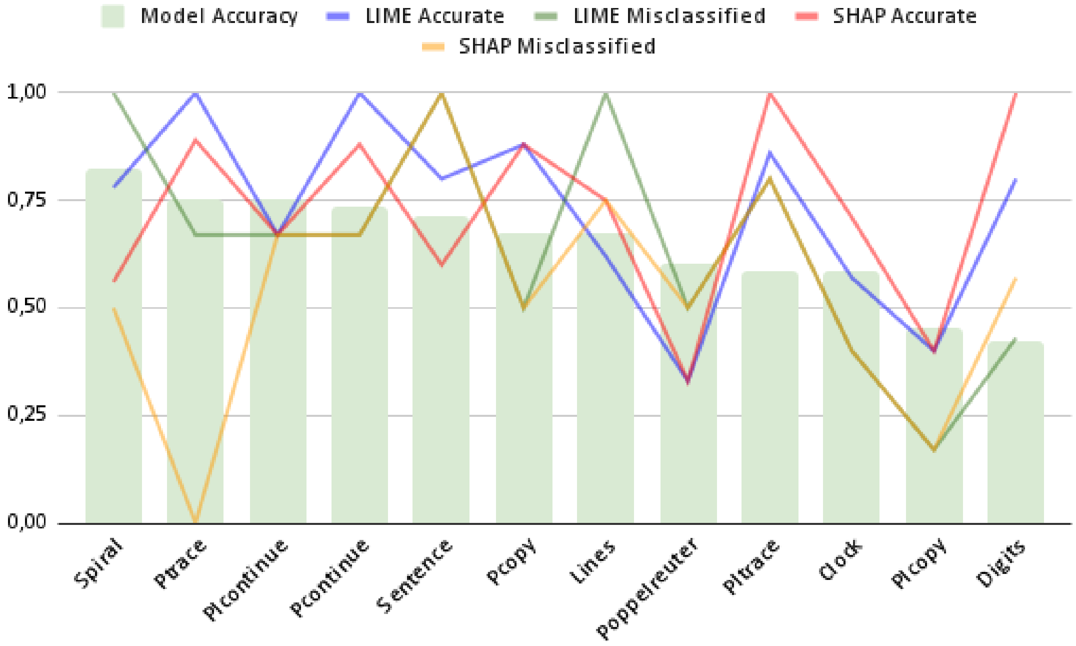


Figure 13. *Monotonicity values and models accuracy.*

Overall can be concluded that explanations done for higher accuracy models perform better.

5.2 Test-Type Patterns

Spiral tests show best overall metrics:

- Highest accuracy (0.818)
- Good faithfulness (0.72 SHAP)
- Stable monotonicity (0.78 LIME)

Which is expected as the process was developed with spiral tests and later applied on others.

Constrained/Mechanical tests (Ptrace, Pltrace, Pcopy, Plcopy, Pcontinue, Plcontinue, Lines, Spiral) outperform Open/Cognitive tests (Clock, Sentence, Digits, Poppelreuter):

- Avg accuracy: 0.73 (mechanical) vs 0.58 (cognitive)
- Better explanation quality for mechanical tests

5.3 Conclusion

The integration of LIME[5] and SHAP[8] was meant to provide insights into model decision-making. Both LIME and SHAP were evaluated with above average results, SHAP performing better in most cases. From that we can conclude that both LIME and SHAP can be used as explainers for Parkinson's drawing tests models and be trusted to provide accurate explanations. Faithfulness and monotonicity metrics[12] could be used to validate results of those explanations.

5.4 Limitations and Considerations

Several things should be noted:

- Relatively small sample size (<30 KT, <20 PD) may affect models performances.
- LIME explanations showed some instability between runs due to random perturbations.
- Performance varied across different drawing tasks, suggesting task-specific adaptation may be needed.
- The reliance on only top 3 motion mass parameters may miss patterns detectable with more parameters.

5.5 Contribution and Significance

This work makes several important contributions:

Methodological: Developed a complete, interpretable pipeline for Parkinson's diagnosis using drawing tests.

Technical: Successfully integrated statistical feature selection with modern XAI tools.

Practical: Provided quantitative metrics for evaluating explanation quality in clinical applications.

6. Summary

This thesis focuses on the application of interpretable machine learning to drawing tests as a tool for supporting the diagnosis of Parkinson's disease. The work is situated as an intermediate step within a larger research effort, where the goal is to support the diagnosis of neurodegenerative diseases, detect early cognitive impairments, and recognize signs of fatigue. The goal of this thesis was to explore motion-based features derived from digital drawing data and make classification results more transparent through explainability techniques.

The raw data used in the project consisted of drawings collected via tablets from both healthy individuals and Parkinson's patients. A significant portion of the work involved developing a robust data extraction pipeline capable of handling inconsistent JSON structures, nested artifacts, and other irregularities in the files. The accuracy of data parsing was confirmed through visualization tools developed specifically for internal verification.

Feature extraction was based on a scientific paper that defined a set of motion mass parameters[1]. These features were computed and tested for relevance using Fisher Score[6]. From these, the most informative three features were selected for classification.

Classifiers were evaluated, with Random Forest ultimately selected as the primary model due to better performance. The models were fine-tuned via hyperparameter optimization, and explainability was integrated using LIME[5] and SHAP[8] to provide insight into the model's decisions.

LIME and SHAP explanations were evaluated and found trustworthy in most cases, SHAP more so than LIME. As such these explainers can be trusted to provide meaningful insight on model predictions.

All the code used for this work is accessible at github repository.[9]

Although not a complete end-to-end diagnostic tool, this thesis contributes a validated, interpretable classification workflow that can serve as a building block for further research. The models may be tested or retrained on newly provided drawing datasets, and future improvements may include the integration of additional classifiers, more top selected features, and expansion to related tasks handled by other researchers and PhD students.

References

[1] Elli Valla et al. *Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics*. [Accessed: 18-05-2025]. URL: https://www.sciencedirect.com/science/article/pii/S1746809422000738.

[2] Gorbatšov Vassili. *Motion Freezing Analysis in Drawing and Writing Tests*. [Accessed: 18-05-2025]. URL: https://digikogu.taltech.ee/et/Item/6948a26d-cb52-47dc-a3f4-22d1fe903004.

[3] Zoumana Keita. *Explainable AI - Understanding and Trusting Machine Learning Models*. [Accessed: 18-05-2025]. URL: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models.

[4] *SHAP Documentation*. [Accessed: 18-05-2025]. URL: https://shap.readthedocs.io/en/latest/.

[5] Marco Tulio Correia Ribeiro. *LIME: Local Interpretable Model-agnostic Explanations*. [Accessed: 18-05-2025]. URL: https://github.com/marcotcr/lime.

[6] Charu C. Aggarwal. "10.2.1.3 Fisher Score". In: *Data Mining The Textbook*. 2014.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction(7.10 Cross-Validation)*. [Accessed: 18-05-2025]. URL: https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf.

[8] Scott Lundberg. *SHAP (SHapley Additive exPlanations)*. [Accessed: 18-05-2025]. URL: https://github.com/slundberg/shap.

[9] Maria Rizo. *XAI Parkinson*. [Accessed: 29-05-2025]. URL: https://github.com/MariaRizo/XAI_Parkinson/tree/main.

[10] *Waterfall plot*. [Accessed: 18-05-2025]. URL: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html.

[11] *aix360 Documentation*. [Accessed: 18-05-2025]. URL: https://aix360.readthedocs.io/en/latest/metrics.html.

[12] *AI Explainability 360 (AIX360) Toolkit*. [Accessed: 18-05-2025]. URL: https://github.com/Trusted-AI/AIX360/blob/master/aix360/metrics/local_metrics.py.

[13] *Pearson Correlation Coefficient*. [Accessed: 18-05-2025]. URL: https://www.geeksforgeeks.org/pearson-correlation-coefficient/.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Maria Rizo

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “XAI based analysis of drawing tests for the diagnosis of Parkinson’s disease”, supervised by Sven Nõmm and Rajesh Kalakoti
   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. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I 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.

04.06.2025

[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 license shall not be valid for the period.