Developing a Diagnostic Tool for Parkinson's Disease Using a Machine Learning Approach

by

Isabelle Southon

300597453

Submitted to

Victoria University of Wellington

Supervisor: Alejandro Frery
Group Members: Amelia Bentley & Viyanka Moodley

October 22, 2024

Executive Summary

Parkinson's disease is a degenerative neurological disorder that leads to involuntary movements, such as tremors, and difficulties with balance and coordination (Cleveland Clinic, 2022; Kamble et al., 2021). Common symptoms include shaking involuntarily, slow movement, stiff muscles, and difficulty speaking (Parkinson's Foundation, 2024). Given these symptoms, completing a handwritten spiral drawing would be a challenging task for an individual with Parkinson's disease. Drawing a spiral is an activity that requires coordination and, according to past literature, can serve as an effective assessment of motor function (Kamble et al., 2021).

This analysis examines a dataset containing features of handwritten spirals drawn by healthy individuals and patients with Parkinson's disease. The purpose of this analysis is to find the best features of handwritten spirals to develop a diagnostic tool that can help determine whether a participant has Parkinson's disease or not. The goal of this report is to provide a list of variables of handwritten spirals that can best the determine prevalence of Parkinson's disease. New features were derived (created) from the original features for this analysis.

When testing the initial model on new unseen data, results illustrated 85.71% accuracy in separating patients with Parkinson's disease from healthy individuals. After reducing the number of features in the dataset, results still displayed an accuracy of 85.71%. This suggests that the removed features were redundant or irrelevant and did not significantly contribute to the model's ability to distinguish Parkinson's patients from healthy individuals.

This analysis finds that the most influential variables that determine the prevalence of Parkinson's disease are the average pressure, the average azimuth, the count of pen lifts, the drawing duration, and the average acceleration.

Contents

1 Background

Parkinson's disease affects over 10 million people globally, creating a demand for improved diagnostic tools (AlMahadin et al., 2020). Current assessments are limited, relying on a combination of neurological and physical exams (Mayo Clinic, 2023; Parkinson's Foundation, 2024). Research shows there is no definitive method for diagnosis, highlighting the need for more robust tools (AlMahadin et al., 2020).

Research from Kamble et al. (2021) highlights that digitized drawings of spirals can support differential diagnosis of Parkinson's disease through the separation of Parkinson's patients from healthy individuals (Kamble et al., 2021). A differential diagnosis is not the final diagnosis a person receives but is a list of potential conditions that share mutual symptoms they possess. Kamble et al. (2021) demonstrated that digitized spiral drawings can support the differential diagnosis of Parkinson's by distinguishing patients from healthy individuals (Kamble et al., 2021). Wang et al. (2008) introduced a computerized technique to analyse spirals, improving accuracy in detecting motor disorders by reducing low-frequency noise and using automated centre detection. The authors found that incorporating automated centre detection into the analysis of spirals significantly improved the accuracy in assessing the prevalence of Parkinson's disease (Wang et al., 2008).

The focus of this analysis shifts from phase 1's research question, "Using a regression model, what are the best features of handwritten spirals that classify between control and Parkinson's patients?" to the evolved scientific question, "What are the top derived features from the original spirals dataset that most effectively distinguish healthy patients from those with Parkinson's disease, identified through recursive feature elimination (RFE) in a support vector machine (SVM)?" The data is better suited for a classification model, reflecting the evolution of this analysis from phase 1.

2 Data Description

The dataset included multiple SVC files, each representing time series data from patients' handwritten spirals. Each patient's folder contained at least 1 SVC file with several observations recorded over time. There was no structure to the missing data, thus variables with missing data were removed. The original dataset contains 9 features: 6 numerical features (`X`, `Y`, `timestamp`, `azimuth`, `altitude`, `pressure`) and 3 categorical features (`pen_state`, `patient_id`, `patient_group` as the target variable). Note that `timestamp` has been input as a numerical variable, but it is actually temporal. The original dataset had 121299 observations among 33 patients.

Feature derivation was applied to prepare the data for machine learning algorithms. New features (`avg_pressure`, `avg_azimuth`, and `avg_altitude`) were calculated as the average values for each patient. Additional derived features included `pen_lift_count` (counting pen lift instances from the `pen_state`) and `drawing_duration` (calculated as the time taken to complete the spiral based on the `timestamp`).

Derived features such as `avg_velocity`, `avg_acceleration`, and `avg_jerk`, were cre-

ated using changes in `X` and `Y` coordinates over time. Velocity measured the pen's speed, calculated from changes in position over time. Acceleration reflected changes in velocity, while jerk measured changes in acceleration, representing the smoothness of the pen's movement. All these features were computed using standard formulas for velocity, acceleration, and jerk, seen in Table 1.

Table 1: Summary of derived features, formulas, and explanations.

Feature	**Derived From**	**Formula**	**Explanation**
Velocity	X, Y, timestamp	$v_x = \frac{\Delta X}{\Delta t}$, $v_y = \frac{\Delta Y}{\Delta t}$	Where ΔX and ΔY are the changes in position for the 'X' and 'Y' coordinates and Δt is the time difference.
Total	v_x, v_y	$v = \sqrt{v_x^2 + v_y^2}$	The combined speed in both X and Y directions.
Acceleration	$v_x, v_y, timestamp$	$a_x = \frac{\Delta v_x}{\Delta t}$, $a_y = \frac{\Delta v_y}{\Delta t}$	Where Δv_x and Δv_y are the changes in velocity along the 'X' and 'Y' coordinates and Δt is the time difference.
Total	a_x, a_y	$a = \sqrt{a_x^2 + a_y^2}$	The combined change in speed across both directions.
Jerk	$a_x, a_y, timestamp$	$j_x = \frac{\Delta a_x}{\Delta t}$, $j_y = \frac{\Delta a_y}{\Delta t}$	Where Δa_x and Δa_y are the changes in acceleration along the 'X' and 'Y' coordinates and Δt is the time difference.
Total	j_x, j_y	$j = \sqrt{j_x^2 + j_y^2}$	The combined change in acceleration across both directions.

3 Ethics, Privacy, and Security

3.1 Ethical Considerations

Findings must be accurate and relevant to the scientific question. Although consent was given for using the dataset for learning purposes, it must be confirmed that data was collected with informed consent, and participants were reminded of their right to withdraw without negative consequences. The use of the spirals dataset for learning purposes has been authorised by a supervisor.

To minimize harm, it's critical to avoid biases that could lead to unfair treatment of patient groups and ensure transparency in data usage and interpretation, promoting fairness and trust. As the dataset can indicate Parkinson's disease status, privacy must be safeguarded to protect individual from stereotypes or stigma.

This project should be reviewed by an ethics committee to ensure compliance with ethical standards, and any discovered errors should be corrected before public release. The ethical considerations for phase 2 are consistent with phase 1.

3.2 Privacy Concerns

This project must comply with the NZ Privacy Act (2020). While the spirals dataset is anonymised and lacks direct identifiers, there is still a risk of re-identification if it is combined with other datasets. Features like the `timestamp` could potentially identify individuals by linking them to a specific location and time, exposing their Parkinson's

disease status, which would violate their right to privacy.

Given the sensitive nature of the dataset, unauthorised disclosure could lead to stigma or discrimination. To mitigate this, results should be aggregated into tables or graphs before being shared to protect patient privacy. If the dataset is used beyond its intended educational purpose, it could violate the NZ Privacy Act 2020, particularly Principle 10. Furthermore, as 10% of Parkinson's cases are hereditary, patients' results may impact family members, so sensitive data must be protected (Parkinsons UK, 2020).

3.3 Security Steps

To ensure that personal information remains secure, access is restricted to authorized individuals only. Two-factor authentication is used to protect access to project data and results, such as requiring a password and PIN for GitHub access. GitHub uses HTTPS encryption for secure file transfers, and laptops require fingerprint ID for access.

Additional measures, like seeking advice from Cert NZ for cybersecurity threats, could be implemented. Firewalls and intrusion detection systems can further prevent unauthorized access by monitoring network traffic and potential threats (Cisco, 2024). Keeping a record of who accesses the data, real-time monitoring, and educating team members on security practices would strengthen data protection.

When data is no longer needed, secure deletion methods should be used to prevent recovery, thus protecting the privacy of participants and ensuring the confidentiality of research results. Staying vigilant and encouraging security awareness is important in protecting the data, as it will aid in mitigating privacy and ethical concerns.

4 Exploratory Data Analysis

The target `patient_group` was added to the dataset, and it distinguishes between healthy individuals and patients with Parkinson's disease. The dataset is imbalanced with 23 individuals having Parkinson's and 10 individuals being healthy. This data is time series, therefore derived features were created to answer the research question. Table 2 describes the variables included in the original dataset.

Table 2: Original variables, their type, and definition.

Variable	Type	Definition
`X`	Integer	The horizontal coordinate of the spiral.
`Y`	Integer	The vertical coordinate of the spiral.
`timestamp`	Numerical	The time at which the point of the spiral was drawn.
`pen_state`	Integer	The on/off state of the pen.
`azimuth`	Integer	The horizontal angle of the pen drawing the spiral.
`altitude`	Integer	The vertical angle of the pen relative to the paper or drawing surface.

`pressure`	Integer	The amount of force applied by the pen on the drawing surface.
`patient_id`	Character	The unique identifier for each patient.
`patient_group`	Character	Whether the patient has Parkinson's disease or not.

Figure 1 shows strong positive correlations between `pressure` and `pen_state` (0.8), and `altitude` and `pen_state` (0.35). Moderate negative correlations are observed between the `X` and `Y` coordinates (-0.44) and between `timestamp` and `azimuth` (-0.21).

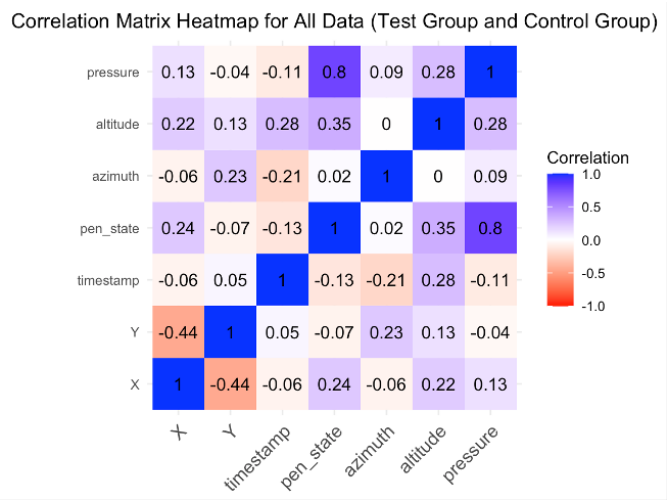

Figure 1: Heat-map displaying correlations for original data.

Figure 2 shows the `X` and `Y` coordinates for a Parkinson's patient's spiral. The irregularities in the spiral path highlight motor issues indicating the patients tremors and loss of fine motor control.

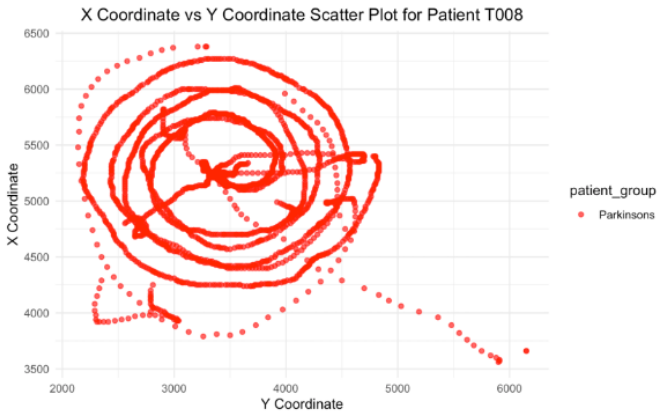

Figure 2: Plot of `X` and `Y` coordinates of a spiral drawn by a Parkinson's patient.

Figure 3 presents a scatterplot of a healthy individual's spiral, showing a smooth, consistent pattern indicative of controlled motor function. In contrast to Figure 2, it highlights the differences between healthy individuals and Parkinson's patients, reinforcing the importance of identifying optimal features for developing a diagnostic tool.

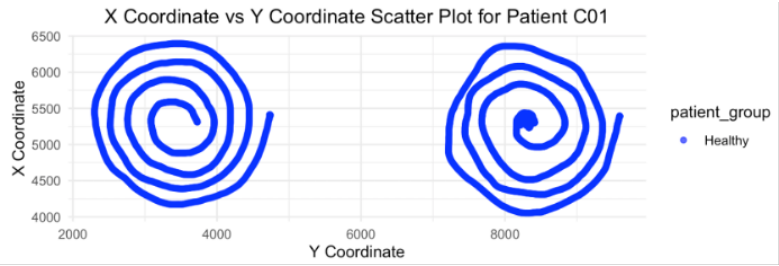

Figure 3: Plot of `X` and `Y` coordinates of a spiral drawn by a healthy patient.

To prepare the dataset for analysis techniques, new features were derived from the original features. The equations that calculated these features are provided in Section 2. The final variables used in the analysis, their variable types, and definitions are provided in Table 3 below.

Table 3: Derived variables, their type, and definition.

Variable	**Type**	**Definition**
`patient_id`	Character	The unique identifier for each patient.
`patient_group`	Character	Whether the patient has Parkinson's disease or not.
`avg_pressure`	Numerical	The average amount of force applied by the pen on the drawing surface.
`avg_azimuth`	Numerical	The average horizontal angle of the pen.
`avg_altitude`	Numerical	The average vertical angle of the pen.
`pen_life_count`	Integer	The total number of times the pen was lifted off the drawing surface.
`drawing_duration`	Numerical	The total time taken to complete the spiral drawing.
`avg_velocity`	Numerical	The average speed of the pen's movement across the drawing surface.
`avg_acceleration`	Numerical	The average rate of change of velocity.
`avg_jerk`	Numerical	The average rate of change of acceleration.

Figure 4 shows strong positive correlations between `avg_jerk` and `avg_acceleration` (0.99), and `avg_jerk` and `avg_velocity` (0.83). Negative correlations are observed between `drawing_duration` and `avg_altitude` (-0.64), and `avg_pressure` and `drawing_duration` (-0.45).

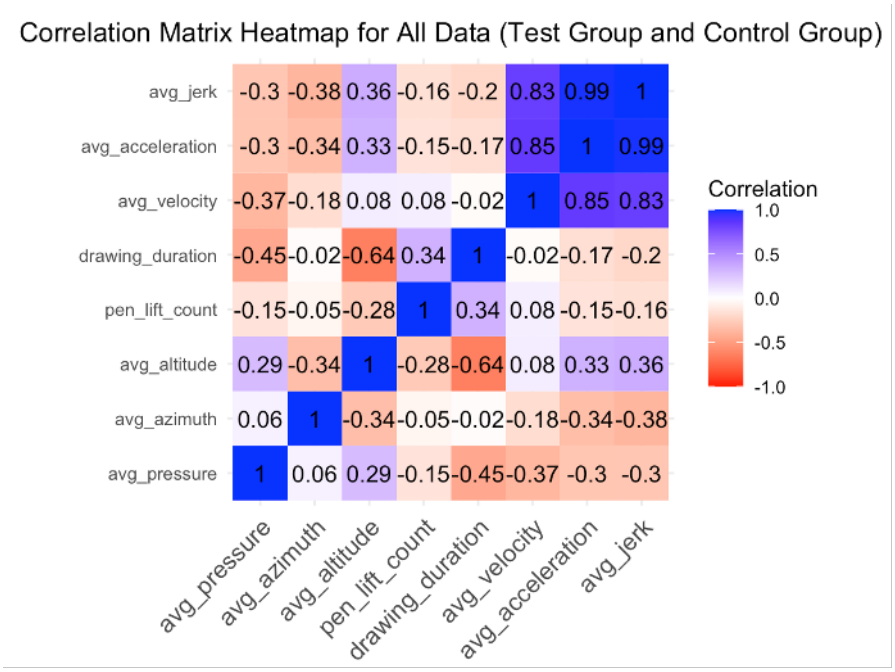

Figure 4: Heatmap displaying correlations for derived features.

5 Detailed Analysis Results

The main analysis technique used was a support vector machine (SVM), a machine learning algorithm designed for binary classification tasks by finding the optimal hyperplane to separate different classes (Kamble et al., 2021). The dataset, initially constructed in RStudio, was exported to Python for analysis.

SVMs are designed to effectively handle binary classification tasks, reflecting how this analysis technique is a suitable model for the spirals dataset. SVMs can also handle the separation of linear and non-linear data through different kernel functions (Tabsharani, 2023). SVMs also work well with small to medium-sized datasets. The dataset containing the derived features has 33 observations, further reflecting the suitability of an SVM model for this data.

The SVM model was used to separate healthy individuals from Parkinson's patients, while RFE ranked and selected the most important features for classification. This approach was crucial for identifying optimal derived features for detecting Parkinson's disease. RFE operated by removing the least important features based on their contribution to classification (Awad & Fraihat, 2023). This analysis technique was necessary to answer the research question because its goal was to identify the most important derived features for detecting Parkinson's disease by separating Parkinson's patients from healthy individuals.

5.1 Preprocessing

The `patient_id` feature was removed, and `patient_group` was mapped to numerical labels (0 = Healthy, 1 = Parkinson's). The data was split into 80% training and 20% testing sets, and numeric features were standardized.

5.2 Baseline Model

The baseline SVM model included all 8 features such as the average pressure, average azimuth, average altitude, the count of pen lifts, the drawing duration, the average velocity, the average acceleration, and the average jerk.

5.3 Final Model

The final model used RFE to reduce the number of features. The top 5 features were the average pressure, the average azimuth, the count of pen lifts, the drawing duration, and the average acceleration. This answers the research question by identifying the most effective features for distinguishing Parkinson's patients from healthy individuals.

5.4 Model Comparison

Table 4 compares the performance of the baseline model (all features) and the final model (after feature removal).

Table 4: Comparison of baseline and final SVM model performance metrics.

Metric	**Baseline Model**	**Final Model**
Train Accuracy	92.31%	96.15%
Test Accuracy	85.71%	85.71%
Test Precision	83.33%	83.33%
Test Recall	100.0%	100.0%
Test F1 Score	90.91%	90.91%

The SVMs generalization to unseen data was not compromised after removing redundant features using RFE. Training accuracy increased from 92.31% to 96.15%, but test accuracy remained the same, indicating that RFE effectively removed less relevant features. This simplified the model, allowing it to focus on the most informative variables, thereby answering the research question. Figure 5 shows the inverted feature importance, with the average altitude, velocity, and jerk being less important than the features retained in the final model.

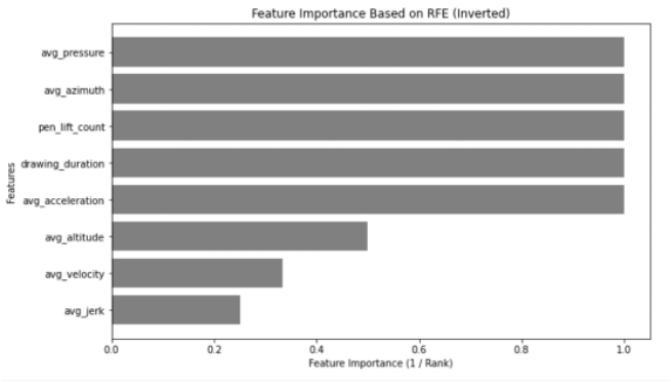

Figure 5: Feature importance using RFE with a SVM model.

The top-ranked features contribute most to the SVM model's performance, while lower-ranked ones may add noise. Although estimates of uncertainty could be obtained, they are not essential for answering the research question, as the SVM provides definitive outputs. Cross-validation could more reliably assess stability across data subsets, adding transparency and reliability to the results by indicating how stable performance metrics are.

A useful measure of uncertainty is cross-validation for RFE, as it shows the variability in feature importance and the consistency of the top features. This would provide insight into the reliability of the selected features and help ensure that the most important features are consistently identified. Although cross-validation for RFE was not part of this analysis, it represents valuable future work to enhance the consistency and reliability of the findings (Awad & Fraihat, 2023).

5.5 Biases

Potential biases in the dataset and results must be acknowledged, particularly in medical contexts like diagnosing Parkinson's. Selection bias may arise if the dataset used for training the model is not representative of the general population, limiting the model's ability to generalize. This could occur for this dataset as participants may be from a particular geographic region and, thus do not generalise well to other populations. Sampling bias could occur if certain groups (e.g., gender) are over- or underrepresented. This means that the model may learn patterns that are not relevant to the general population. For example, if males are underrepresented, the model might not detect Parkinson's as accurately for males.

The SVM model is sensitive to feature scaling, so if features are not scaled correctly, this could cause algorithmic bias, where some features dominate the model, even if they are less important to detecting the prevalence of Parkinson's disease (Sotelo, 2017). Classes are imbalanced for this dataset, as there are more participants with Parkinson's disease than healthy individuals. This means that the SVM may be biased toward detecting the majority class, thus the model might fail to correctly identify the minority class (e.g., healthy individuals). The confusion matrix (Figure 6) shows that a healthy

individual was incorrectly classified, reflecting the model's tendency to misclassify the minority class.

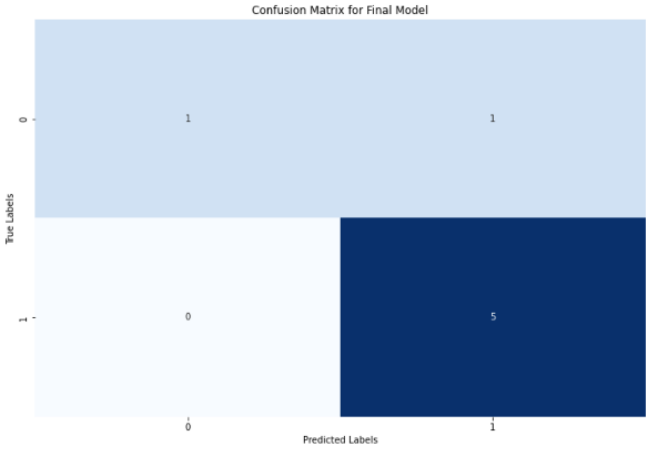

Figure 6: Confusion matrix for the final model.

6 Conclusion

RFE with SVM concluded that the most important features for distinguishing between healthy individuals and Parkinson's patients are the average pressure, the average azimuth, the count of pen lifts, the drawing duration, and the average acceleration. The final SVM model with reduced features achieved a high accuracy and precision for the test set, illustrating that the selected features are effective for the classification task.

6.1 Recommendations and Limitations

This analysis could be used to develop diagnostic tools for Parkinson's disease, and features could be continuously monitored over time which may provide insights into disease progression. Limitations of this analysis are that the dataset was small with only 33 observations and there was a class imbalance for the target class. Small datasets can lead to overfitting and limits the generalizability of the model to different populations, and the class imbalance could result in biased predictions toward the majority class. There are also limited features as features used were solely derived from spiral drawings, which may not capture all relevant aspects of Parkinson's motor control issues.

6.2 Future Work

Future work could involve increasing the size of the dataset which would improve the model's generalizability and reduce the risk of overfitting. Other machine learning algorithms such as Random Forest could be explored to see if they outperform the SVM. Future work could also involve using RFE with cross-validation to confirm if the optimal features selected remain consistent across different data splits.

References

AlMahadin, G., Lotfi, A., Zysk, E., Siena, F., Carthy, M., & Breedon, P. (2020). Parkinson's disease: current assessment methods and wearable devices for evaluation of movement disorder motor symptoms—a patient and healthcare professional perspective. *BMC Neurology*, *20*(1), 1–10. doi: 10.1186/s12883-020-01996-7

Awad, M., & Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67. Retrieved from https://doi.org/10.3390/jsan12050067 doi: 10.3390/jsan12050067

Cisco. (2024). *What is a firewall?* Retrieved from https://www.cisco.com/c/en_ca/products/security/firewalls/what-is-a-firewall.html (Accessed: 2024-09-06)

Cleveland Clinic. (2022). *Parkinson's disease: Causes, symptoms, stages, treatment, support.* Retrieved from https://my.clevelandclinic.org/health/diseases/8525-parkinsons-disease-an-overview (Accessed: 2024-09-06)

Kamble, M., Shrivastava, P., & Jain, M. (2021). Digitized spiral drawing classification for parkinson's disease diagnosis. *Measurement: Sensors*, *16*, 100047. doi: 10.1016/j.measen.2021.100047

Mayo Clinic. (2023). *Parkinson's disease—diagnosis and treatment.* Retrieved from https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062 (Accessed: 2024-05-26)

Parkinson's Foundation. (2024). *Getting diagnosed.* Retrieved from https://www.parkinson.org/understanding-parkinsons/getting-diagnosed (Accessed: 2024-09-02)

Parkinsons UK. (2020). *How is parkinson's diagnosed?* Retrieved from https://www.parkinsons.org.uk/information-and-support/how-parkinsons-diagnosed (Accessed: 2024-09-06)

Sotelo, D. (2017). *Effect of feature standardization on linear support vector machines. Towards Data Science*. Retrieved from https://towardsdatascience.com/effect-of-feature-standardization-on-linear-support-vector-machines-13213765b812 (Accessed: 2024-10-21)

Tabsharani, F. (2023). *What is support vector machine (svm)?* Retrieved from https://www.techtarget.com/whatis/definition/support-vector-machine-SVM (Accessed: 2023-08-10)

Wang, H., Yu, Q., Kurtis, M., Floyd, A., Smith, W., & Pullman, S. (2008). Spiral analysis-improved clinical utility with center detection. *Journal of Neuroscience Methods*, *171*(2), 264–270. doi: 10.1016/j.jneumeth.2008.03.009