Joshua Wilson

CDS 492 Final Paper

Date: 5/11/2023

Improving Parkinson's Disease Prediction Models with the Novel "Lift" Variable

Abstract

This project investigates the addition of the novel variable "lift" to machine learning models for predicting Parkinson's disease using the spiral drawing test. The lift variable represents the frequency of pen lifts during the test. The project aims to understand how this new variable can improve the accuracy of certain models and compare its effectiveness with other variables. The results so far indicate that the lift variable significantly improves the prediction of Parkinson's disease, with the best model achieving a cross-validation accuracy of 0.875. The ultimate goal of the project is to create a working application to help predict patients' chances of developing Parkinson’s based on drawings.

Introduction

Due to the early detection and diagnosis of Parkinson's disease being critical to disease management and quality of life of patients, there has been increasing interest in utilizing handwriting and drawing tasks as a means of diagnosing the disease. This is due to the fine motor skill impairments that are symptomatic of the disease(NIH). A promising area of investigation is in spiral sketching, specifically focusing on the speed and pen-pressure during the task.

Zham et al. (2017) demonstrated that speed and pen-pressure of spiral sketching are negatively correlated with the severity of Parkinson's disease. Furthermore, they found that the frequency of pen lifts

during the task may influence both speed and pen-pressure. A higher frequency of pen lifts was associated with slower speed and reduced pen-pressure, while fewer pen lifts were associated with faster speed and more consistent pen-pressure.

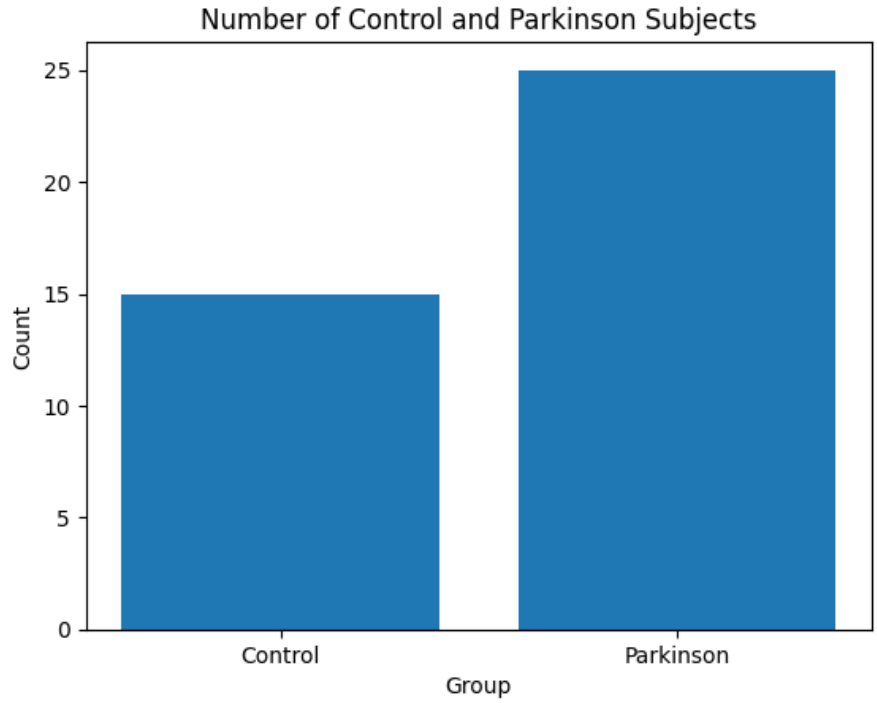
Testing spiral drawings was further explored by Kamble et al. (2021) who developed a machine learning model to classify spiral drawing images from both healthy individuals and Parkinson's patients. The model was able to achieve an accuracy of 91.6% and an area under the curve (AUC) of 98.1% using logistic regression. Other machine learning models such as the Random Forest Classifier, Support Vector Classifier (SVC), and KNeighborsClassifier were also utilized.

Building upon these works, the current project sought to investigate the addition of the novel variable "lift" to these machine learning models. "Lift" represents the frequency of pen lifts during the spiral sketching task. This variable was found to have a statistically significant relationship with Parkinson's disease, as evidenced by a Mann-Whitney U test p-value of 0.00017. The variable was added to the machine learning models, which led to notable improvements in model performance. With the inclusion of the "lift" variable, the best model achieved a cross-validation accuracy of 87.5% and an AUC of 91.6%, comparable to current literature for the Random Forest model. While not beating the 91.6% accuracy given by Kamble et al. it shows potential as a new feature that could further help researchers in prediction.

Methods

The project uses two datasets. The first dataset, from UCI Machine Learning, contains spiral drawings, divided into control and Parkinson's groups(1,2). The second dataset, from Zham et al. (2017), includes 15 control images and 25 Parkinson's images and focuses on time and pressure during the spiral test. The first data set contains CSV files that have information on x,y,z,pressure,timestamp,gripangle , and Testid; all of these values are held as integers.

Figure 3: Group Size comparison



Exploratory Data Analysis

Refer to Figure 1 for feature statistics of the control group, Figure 2 for feature statistics of the Parkinson's group, and Figure 3 for group size comparison. Figure 4 shows the X and Y axis comparison between the control and Parkinson's groups, and Figure 5 presents histogram plots for the combined data frame.

Figure 1: Feature Statistics of Control Group

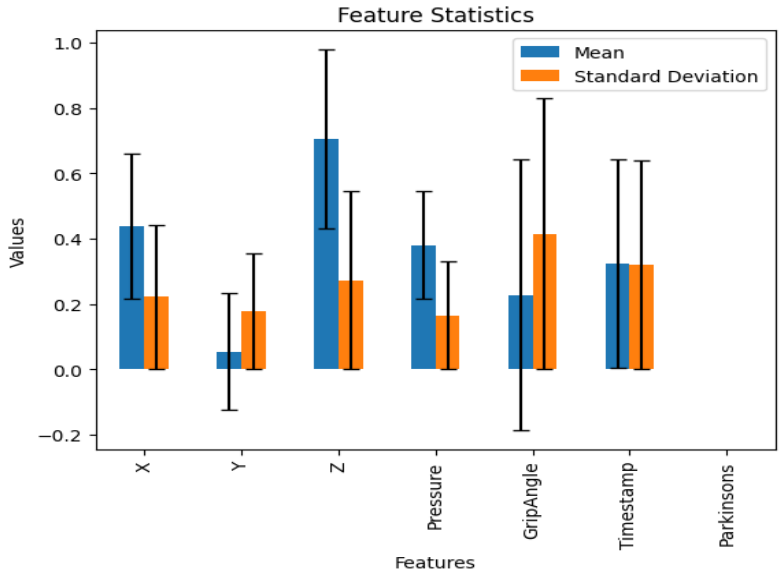
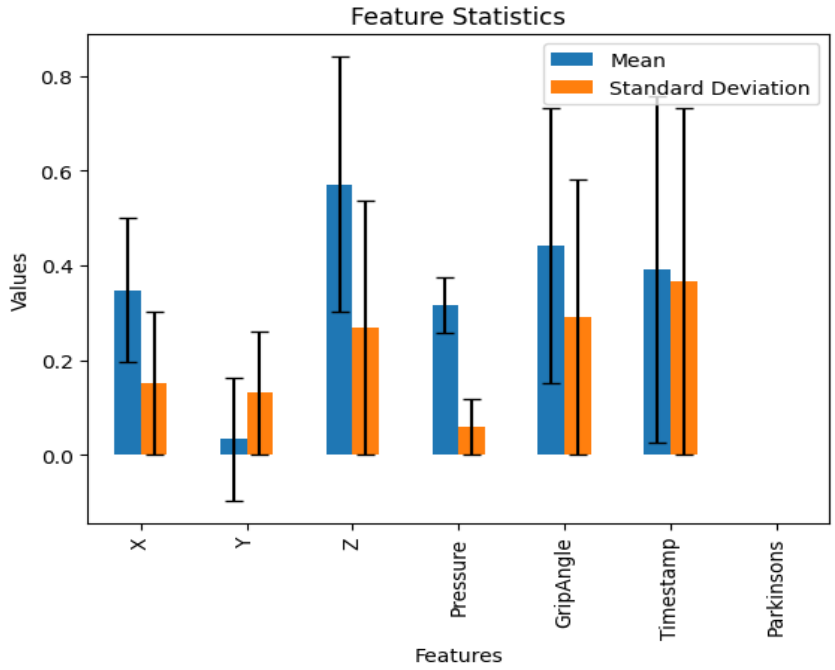


Figure 2: Feature Statistics Parkinson's groups



Upon examination of the dataset, consisting of 25 Parkinson's and 15 control CSV files, it was clear that there were some differences between the two groups. Notably, Parkinson's patients exhibited less pressure on average when compared with the control group. This initial observation prompted a deeper investigation into the data using pairwise plots and other graphical representations to identify correlations between the variables.X,Y,Z,Pressure,GripAngle,Timestamp, and TestID. For the control group the mean values for each were X: 206.202 , Y:221.156 , Z:37.565 , Pressure:721.933. In the Parkinsons group the values were X: 221.703 , Y: 224.328, Z:35.116, Pressure: 635.89. From a simple look at the data there is already a clear pattern showing that Parkinson's patients tend to have less pressure applied on average this can be seen in figure 1 and 2 while a difference in the X and Y can be seen in figure 4.

.Figure 4: X and Y Axis Comparison of Control Vs Parkinsons

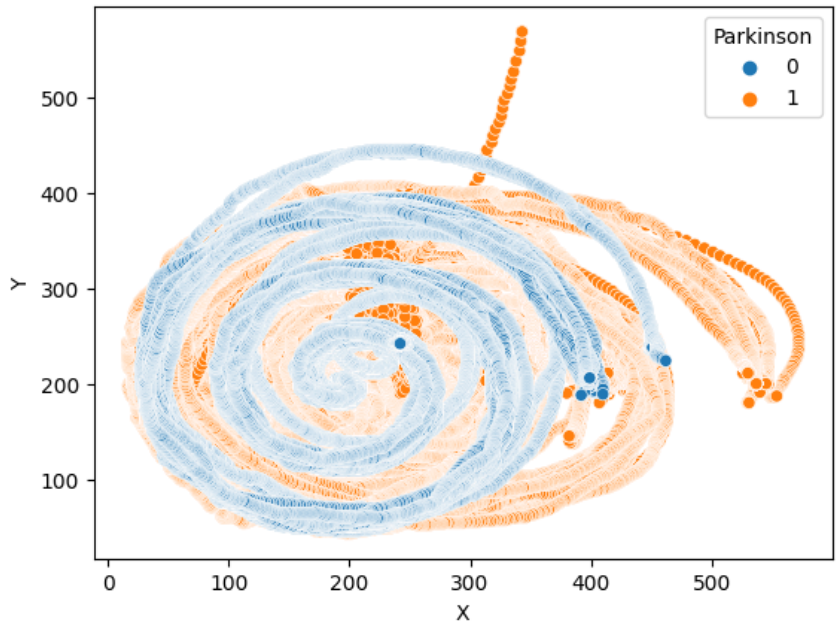
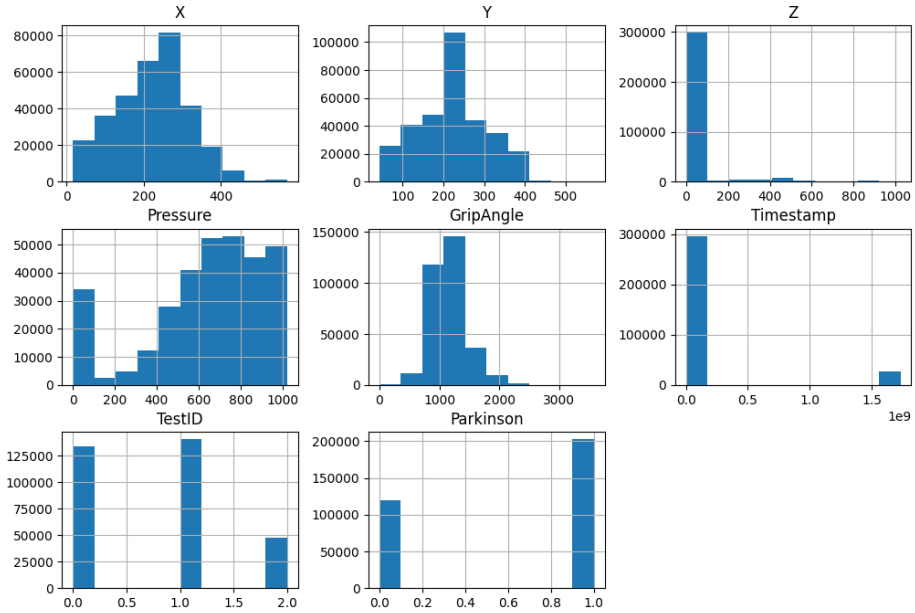


Figure 5: Histogram Plots for Combined DataFrame

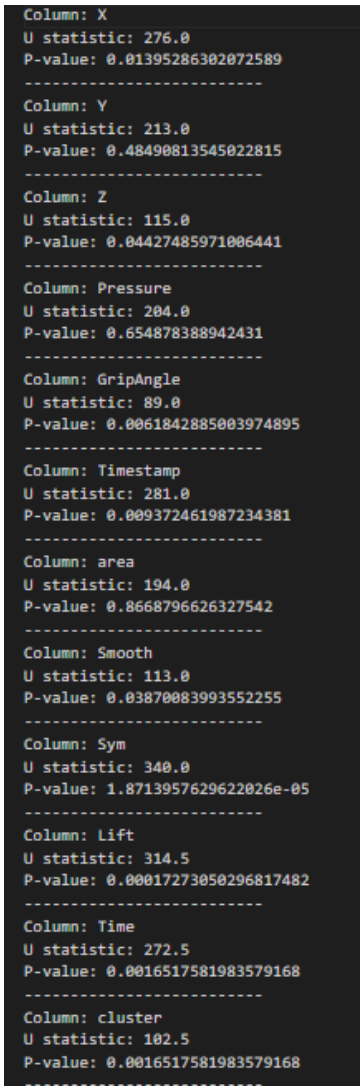


In this paper, the Mann-Whitney U test was employed to assess the significance of the relationship between the novel variable "lift" and Parkinson's disease. By comparing the distributions of

lift values in the control group and the Parkinson's group, the test generated a p-value of 0.00017. This low p-value indicated a statistically significant difference between the two groups, suggesting that the lift variable was indeed related to Parkinson's disease and could potentially improve the accuracy of the machine learning models.Compared to other variables such as Y with a P-value of .484 lift is one of the best variables in this grouping.

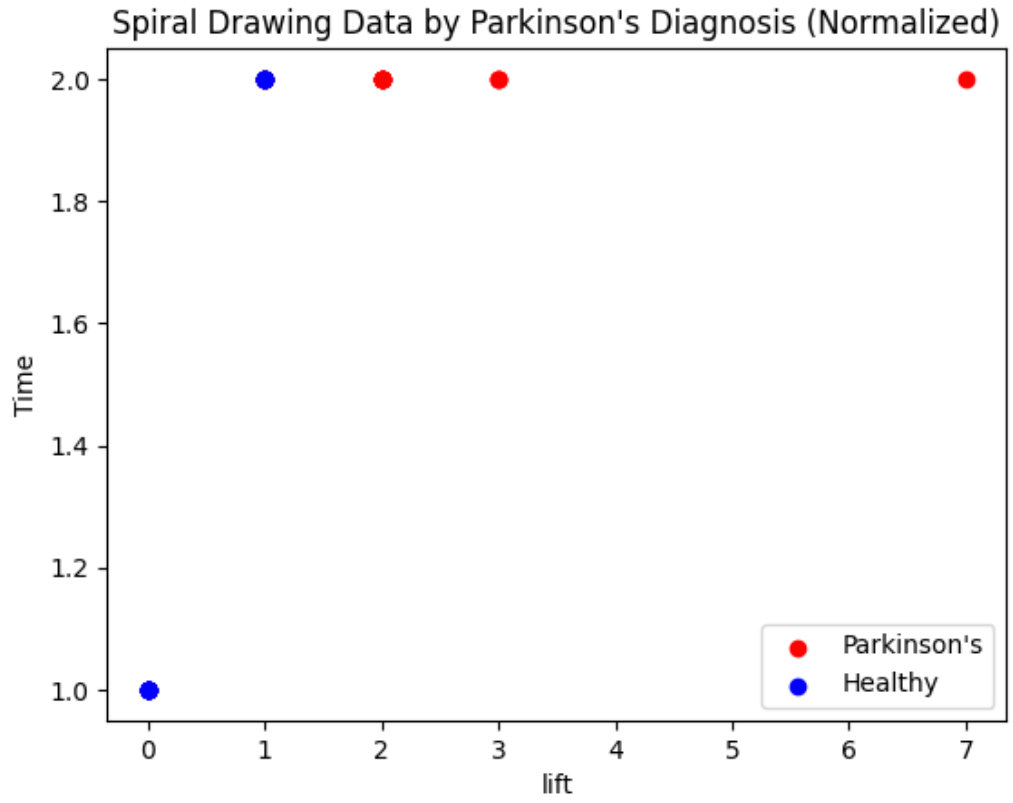
Figure 7 shows the Mann-Whitney U test results for the "lift" variable.

Figure 7: Mann-Whitney U test



In an effort to further enhance the predictive model's performance an additional preprocessing and feature engineering step was introduced: clustering. In the context of this project, clustering was used to create a new feature based on the "Time" and "Lift" variables. By grouping instances that exhibited similar times and lifts during the spiral drawing task, I added an extra layer of separation in the data, which proved to be beneficial for the model's predictive performance

Figure 9 presents the K-means clustering results for scaled "lift" versus "time."

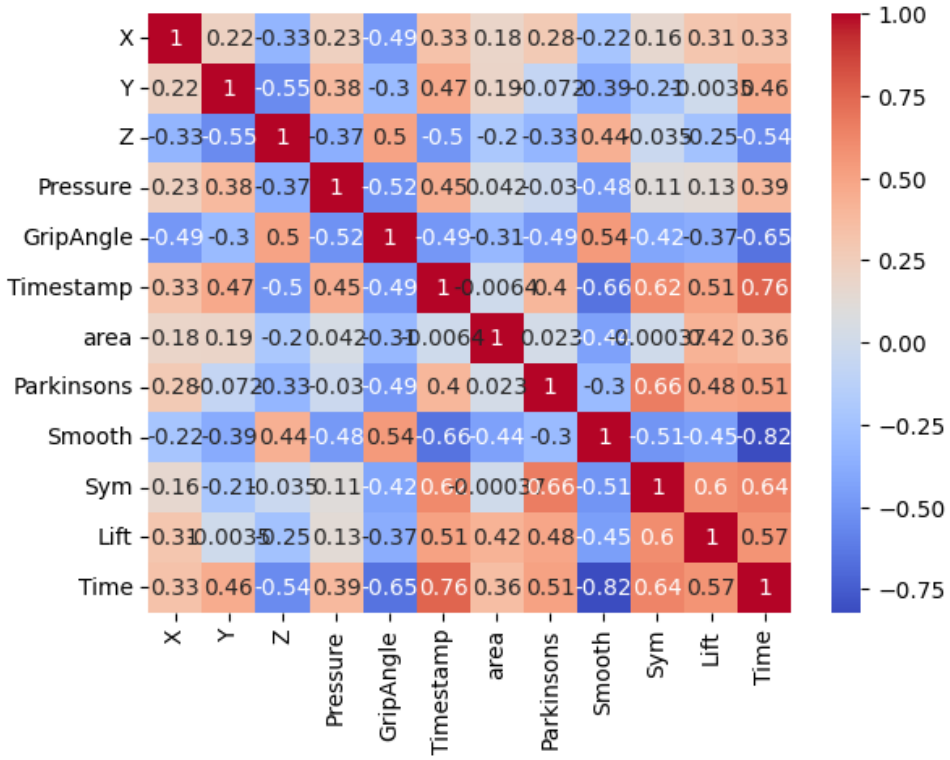


Feature Engineering and Selection

In addition to the existing variables, several new variables were engineered to enrich the dataset and potentially improve prediction accuracy. These included "lift," "symmetry," "area," "length," and "smoothness.""Lift," for instance, was determined by counting unique instances where the pressure value was zero, with consecutive zeros counted as a single lift. The average lift values for the control and

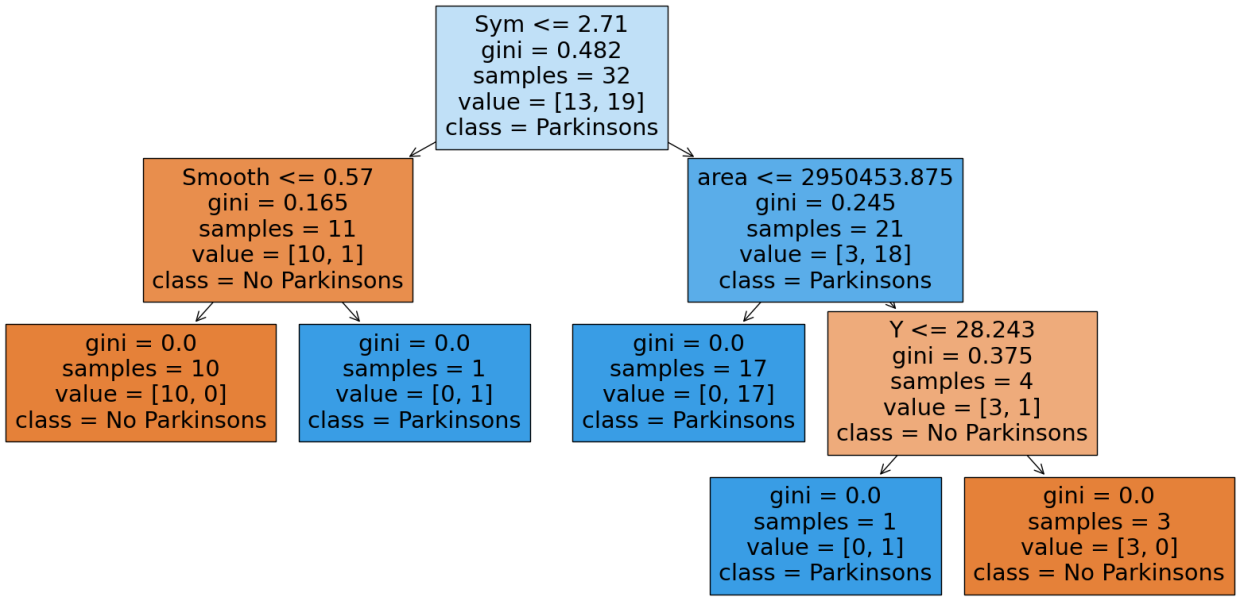
Parkinson's groups were 0.4666 and 1.72, respectively, marking a significant difference. "Area" was calculated using the x and y values of points to determine the area under the spiral, using integration from the Scipy package."Symmetry" and "smoothness" were more complex to calculate but provided additional valuable insights into the data. A higher symmetry score indicated a greater disparity between the maximum and minimum average distances of data points from the center, suggesting lower symmetry in the handwriting sample, while smoothness took into account the standard deviation of the sample to give a regularity score.The process of feature selection involved creating a correlation matrix to test the relationships between variables. The highest correlation coefficients were found with pressure, grip angle, time,and lift variables, suggesting their potential significance in the prediction mode with lift having a correlation value of .48.

Figure 6: Correlation Matrix



In addition to correlation analysis, a decision tree model was built to better interpret the variables and their influence on the prediction. This model identified symmetry,area,Y, and smooth as crucial variables, reaching a prediction accuracy of 75% for Parkinson's.

Figure 8: Decision Tree



To further refine the selection, a random forest model was employed. The model's feature importance output revealed X ,Smooth, Symmetry, Lift and Time as the most important features. Furthermore, an ANOVA test using F-values with K-best select in Python confirmed X ,Smooth, Symmetry, Lift and Time as the most significant variables.Based on these feature selection procedures, the variables chosen for the construction of the prediction models were X ,Smooth, Symmetry, Lift and Time.

Machine Learning Models and Results:

Multiple machine learning algorithms were implemented, including logistic regression, random forest, support vector machine (SVM), k-nearest neighbors (KNN), neural network, and gradient boosting machine (GBM). Each model was tuned using a grid search for hyperparameter optimization.

Model selection was conducted in a two-step process. The first step involved training and validating each model on the training dataset using 5-fold cross-validation. The performance of each model was evaluated based on four different metrics: accuracy, precision, F1 score, and AUC. These metrics provide a comprehensive evaluation of the model's performance. For all of the models a grid search was implemented to help optimize the hyperparameters in each of the models.

The results from the machine learning models showed that lift, a novel variable, helped significantly on the prediction of Parkinson's disease. The models, with lift included, achieved a high degree of accuracy and AUC, with the best models being Random Forest and XGBoost (Figure 10 and 11).

Figure 10:Model Comparison for Training Set

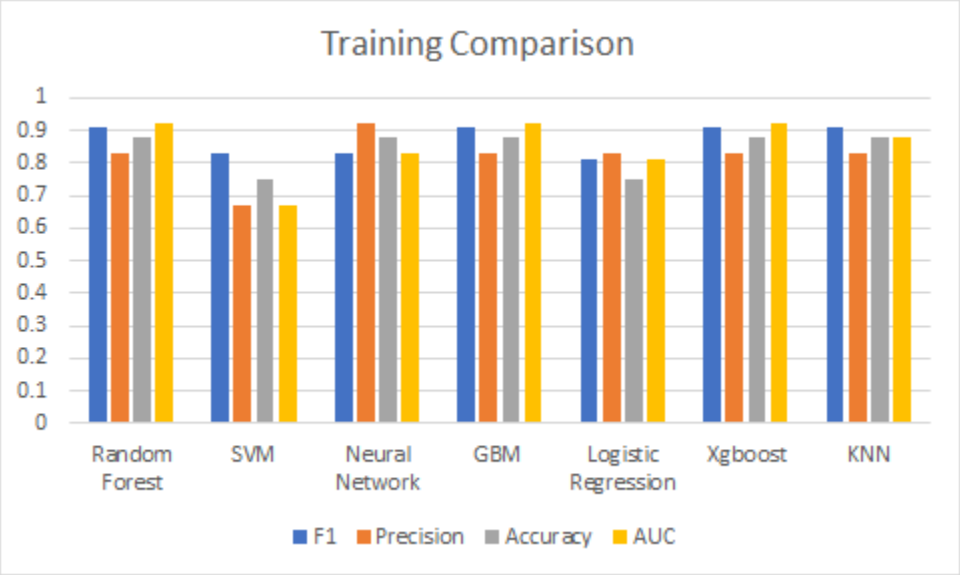
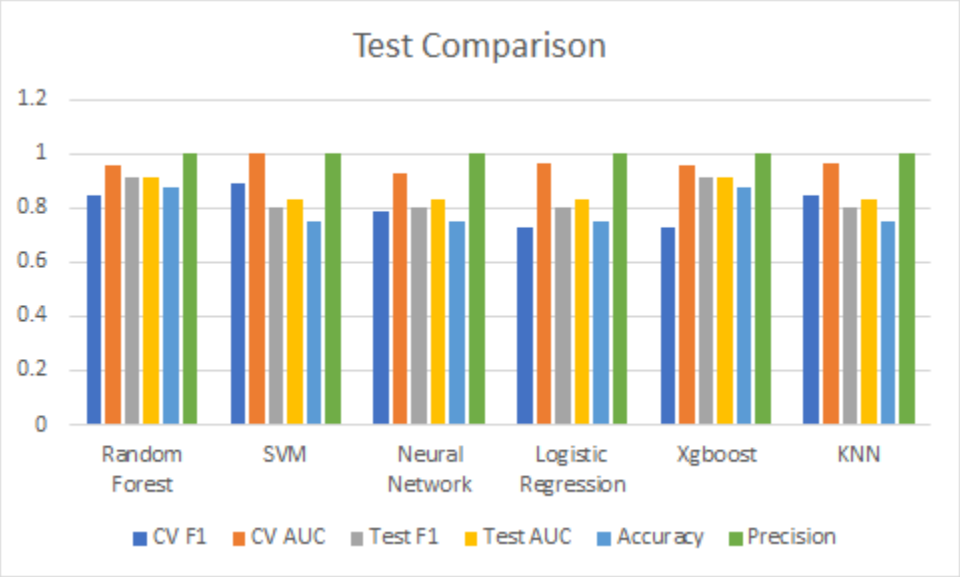


Figure 11: Model Comparison For Test Set



The Random Forest model, with lift included, achieved an accuracy of 87.5% and an AUC of 91.6%. This performance was comparable to the current literature for Random Forest. Similarly, the XGBoost model achieved an accuracy of 87.5% and an AUC of 91.6%. These results indicate that the addition of lift enhances the performance of machine learning models and provides additional insight for early detection of Parkinson's disease.

Figure 12 and Figure 13 display the ROC and AUC for the XGBoost and Random Forest models, respectively. Figure 14 and Figure 15 present the training and test results for all models.

Figure 12: Xgboost AUC ROC

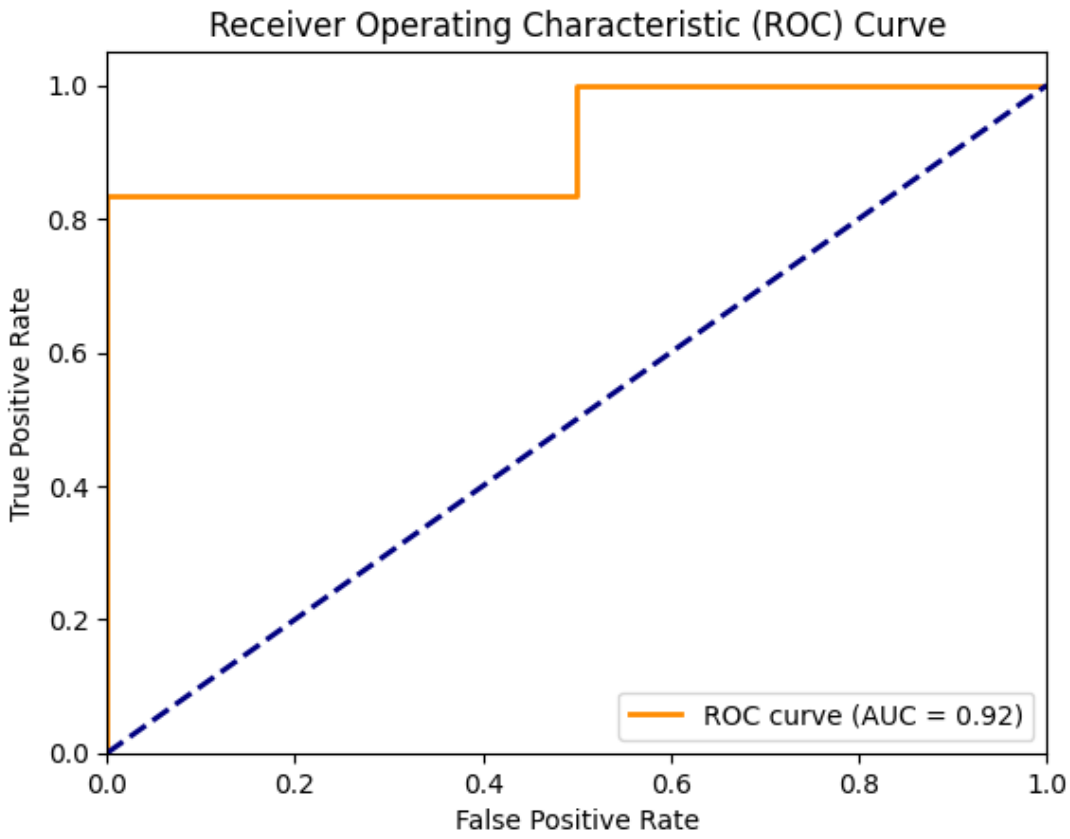


Figure 13: Random Forest AUC ROC

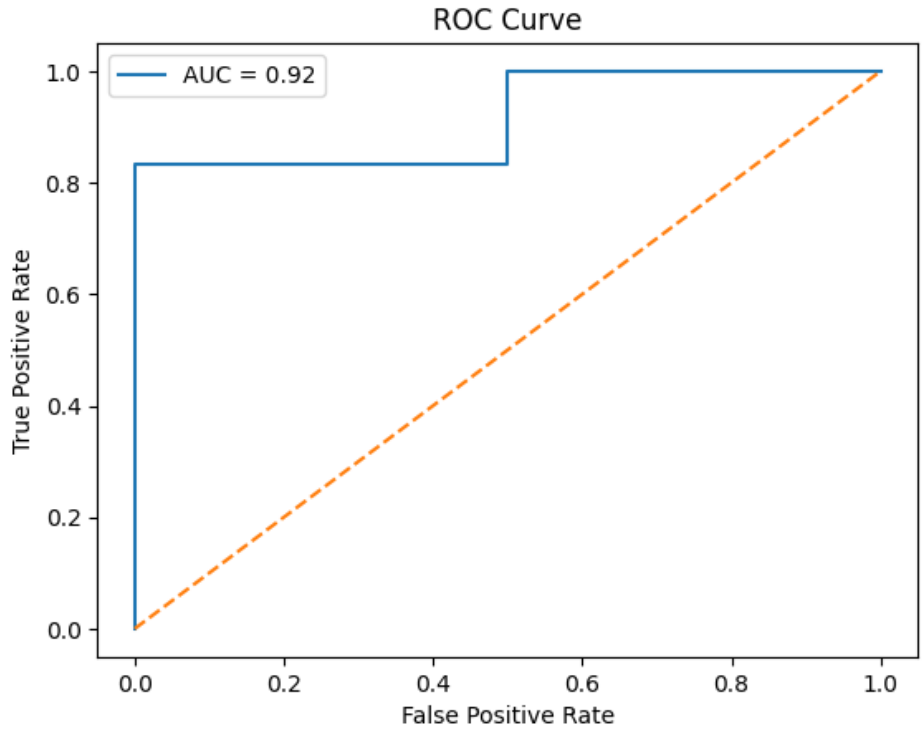


Figure 15: Model Test Set Results

Test						
Model	CV F1	CV AUC	Test F1	Test AUC	Accuracy	Precision
Random Forest	0.848	0.96	0.909	0.916	0.875	1
SVM	0.888	1	0.8	0.83	0.75	1
Neural Network	0.79	0.93	0.8	0.83	0.75	1
Logistic Regression	0.727	0.966	0.8	0.83	0.75	1
Xgboost	0.727	0.96	0.909	0.916	0.875	1
KNN	0.845	0.966	0.8	0.83	0.75	1

Figure 14: Model Training set Results

Train				
Model	F1	Precision	Accuracy	AUC

Random Forest	0.91	0.83	0.88	0.92
SVM	0.83	0.67	0.75	0.67
Neural Network	0.83	0.92	0.88	0.83
GBM	0.91	0.83	0.88	0.92
Logistic Regression	0.81	0.83	0.75	0.81
Xgboost	0.91	0.83	0.88	0.92
KNN	0.91	0.83	0.88	0.88

Further Work

Building upon the successful integration of the lift variable in the predictive models, this project had the ambitious goal of creating a working application to predict the likelihood of Parkinson's disease based on the user drawing a spiral on a screen. The application was designed to capture the drawing data, process it, and then run it through the optimized machine learning model to provide a prediction.

However, during the development and testing stages of the application, several challenges emerged. The main issue encountered was a discrepancy between how the application obtained data and how the original dataset was collected. These differences led to the model occasionally providing inaccurate predictions, creating a barrier to the application's deployment.

Future work on this project will aim to address these issues. This could involve refining the data capture process within the application to more closely align with the data in the original study. Another potential solution is to incorporate more robust data preprocessing and feature engineering techniques that can handle a wider variety of data inputs.

Conclusion

The addition of the lift variable provides a new dimension for Parkinson's disease prediction. The results show an improvement in the performance of machine learning models, especially Random Forest and XGBoost. The performance of these models suggests that the inclusion of the lift variable enhances model performance and provides a new tool for researchers and clinicians in the early detection and diagnosis of Parkinson's disease. The enhanced performance of the machine learning models also underlines the benefits of feature engineering in the process of disease prediction.

The results of this project open new possibilities for further research. For instance, the lift variable could be used in combination with other motor activity data to predict different stages of Parkinson's disease. Moreover, the developed models could be integrated into digital health platforms and applications to enable real-time Parkinson's disease detection using handwriting and drawing tasks.

It is essential to note that while the project yielded promising results, there are limitations to consider. The relatively small sample size might have influenced the accuracy of the models. Future work could use larger datasets to validate these findings.

In conclusion, the project demonstrates the potential of the lift variable in improving the accuracy of Parkinson's disease prediction models. It also highlights the importance of feature engineering and selection in machine learning model development.

References

Zham P, Kumar DK, Dabnichki P, Poosapadi Arjunan S and Raghav S (2017) Distinguishing Different Stages of Parkinson’s Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral. Front. Neurol. 8:435. doi: 10.3389/fneur.2017.00435

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

Isenkul, M.E.; Sakar, B.E.; Kursun, O. . 'Improved spiral test using digitized graphics tablet for

monitoring Parkinson's disease.' The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014), pp. 171-175, 2014.

.Erdogdu Sakar, B., Isenkul, M., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O., 'Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings', IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013.

Thomas M, Lenka A, Kumar Pal P. Handwriting Analysis in Parkinson's Disease: Current Status and Future Directions. Mov Disord Clin Pract. 2017 Nov 1;4(6):806-818. doi: 10.1002/mdc3.12552. PMID: 30363367; PMCID: PMC6174397.

U.S. Department of Health and Human Services. (n.d.). *Parkinson's disease: Causes, symptoms, and treatments*. National Institute on Aging. Retrieved April 14, 2023, from https://www.nia.nih.gov/health/parkinsons-disease