Parkinson's Disease Progression Prediction

Summary

The project aims to address the challenge of predicting Parkinson's disease (PD) progression using protein abundance data. The problem to be solved is identifying potential indicators of disease progression and severity that could contribute to future research and therapeutic approaches for PD. This project builds upon related work in the field of PD research, particularly in the area of protein and peptide analysis. The dataset consists of mass spectrometry readings of cerebrospinal fluid (CSF) samples collected from patients over multiple years, including protein and peptide data, clinical data with UPDRS scores, and supplemental clinical data without CSF samples. The project goals include exploring the data, evaluating various machine-learning models, and identifying potential relationships between proteins, peptides, and UPDRS scores. A range of machine learning models was employed, such as linear regression, ridge regression, Bayesian ridge, ARD regression, SVR, decision tree regressor, random forest regressor, k-neighbors regressor, and SGD regressor. These models were compared based on their Mean Squared Error (MSE) scores and Symmetric Mean Absolute Percentage Error (SMAPE) metric. In summary, the project seeks to predict PD progression using protein abundance data, aiming to identify potential indicators of disease progression and severity. A variety of machine learning models were explored, and their performance was evaluated based on MSE scores and SMAPE metrics. The insights gained from this project could contribute to future PD research and treatment approaches.

Methods

The detailed Methods section, outlining the variousapproaches,machine learning models, and evaluation techniques used in this project, is available separately.

Results

Initially, we performed an adversarial validation to assess the differences between the clinical data and supplemental data. Adversarial validation aims to determine if a classifier can distinguish between the two datasets. We utilized the Receiver Operating Characteristic Area Under the Curve (ROC AUC) score to quantify these differences. If the datasets are highly similar, the classifier would struggle to differentiate between them, resulting in a ROC AUC score of 0.5. Conversely, if the datasets are considerably dissimilar, the ROC AUC score would approach 1, indicating that the classifier can easily distinguish between the two datasets.

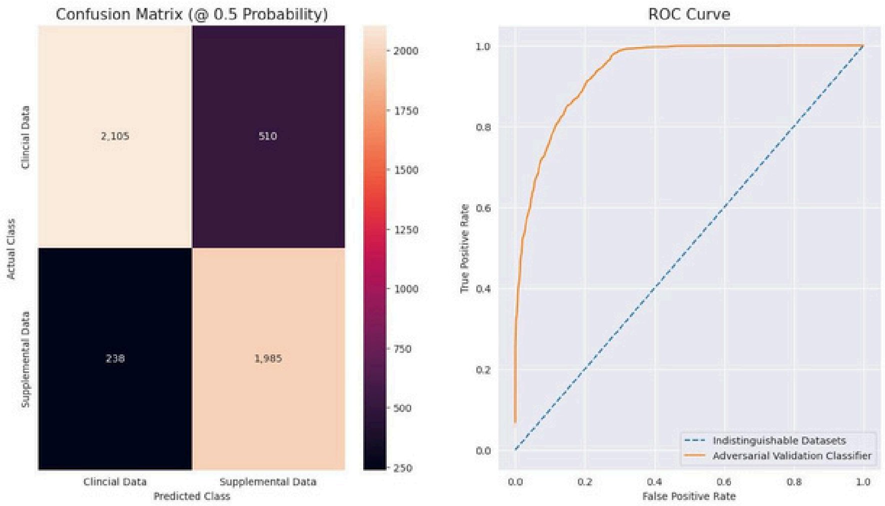


Figure 1: Confusion Matrix and ROC Curve for Adversarial Validation Classifier

Based on the AUC ROC score of 0.939 that we obtained from the image, it is clear that the classifier can effectively distinguish between the two datasets. This high score emphasizes that the datasets are significantly dissimilar. As a result, we were cautious when integrating these datasets, given their distinct characteristics.

Moving forward, we aimed to create a series of models to explore the potential benefits of leveraging our knowledge of features and their properties in Parkinson's disease analysis. We started with a simple Baseline CatBoost Model that did not include any protein or peptide data, relying only on the visit month information for future predictions. This model's performance was evaluated using the Symmetric Mean Absolute Percentage Error (SMAPE) score. The initial SMAPE score of 95.7 was suboptimal, prompting us to consider improvements. One issue we encountered was the model's inability to learn enough about UPDRS 4 due to missing and null values. By setting these values to zero, we observed an improved SMAPE score, and we continued to use this zeroed-out UPDRS score for subsequent models. We then considered incorporating additional clinical data, which slightly improved the model's performance. This could be particularly useful when hidden test data lacks protein information. We also explored the potential impact of the medication state on UPDRS scores and found that including this information improved the model's performance further. Finally, we tried adding raw protein and peptide data to our model. However, this did not result in any significant performance improvement. We then compared the results of all models constructed to determine the best approach. This iterative process allowed us to identify areas of improvement and gain insights into the factors that contribute to better predictions for Parkinson's disease progression.

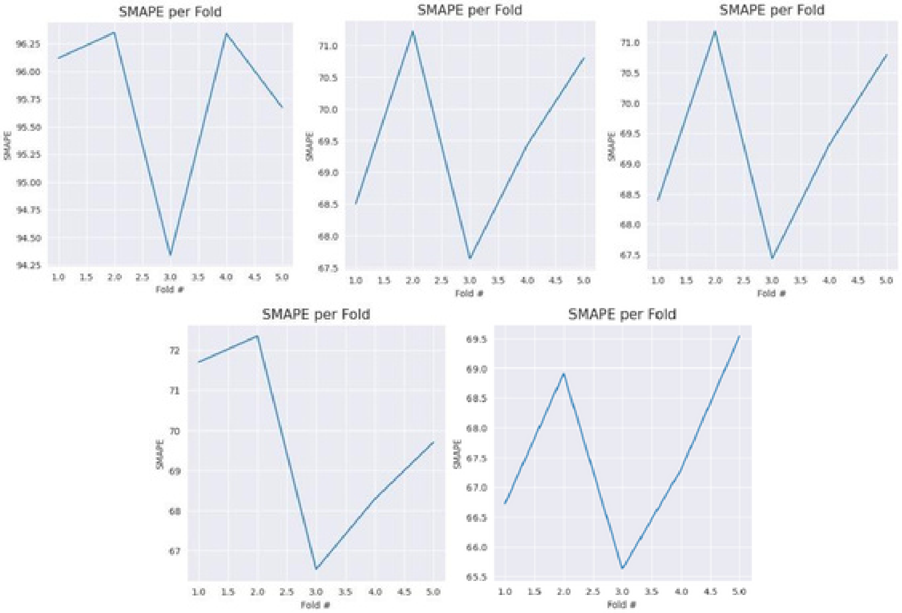


Figure 2: A visual representation of SMAPE scores for each iterative model, arranged in a grid starting from the top-left corner and progressing to the right, then continuing on to the next row

By comparing the performance of each model, we can identify the best-performing one. The red dashed line represents the baseline performance, which serves as the benchmark we aim to surpass.

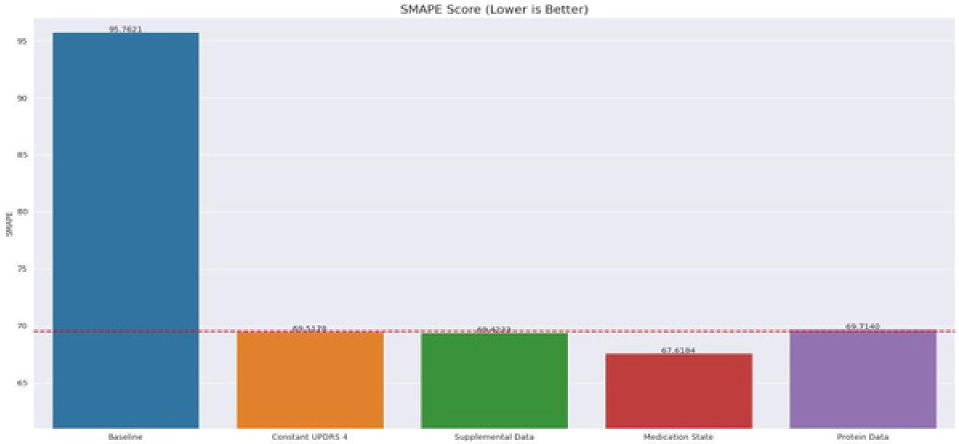


Figure 3: SMAPE score comparison

Comparing the performance of the models demonstrates that each one exceeds the baseline SMAPE score of 95.7621. The Medication State model yields the best results, with the lowest SMAPE score of 67.6184. Following this are the Supplemental Data model (69.4233), the Constant UPDRS 4 model (69.5178), and the Protein Data model (69.7140). Following the initial analysis, we employed additional methods, such as linear regression, ridge regression, Bayesian ridge, ARD regression, SVR, decision tree regressor, random forest regressor, k-neighbors regressor, and SGD regressor. We then measured their mean squared error (MSE) scores for updrs_1, updrs_2, updrs_3, updrs_4, and the average of all UPDRS MSE scores to further evaluate the performance of each model.

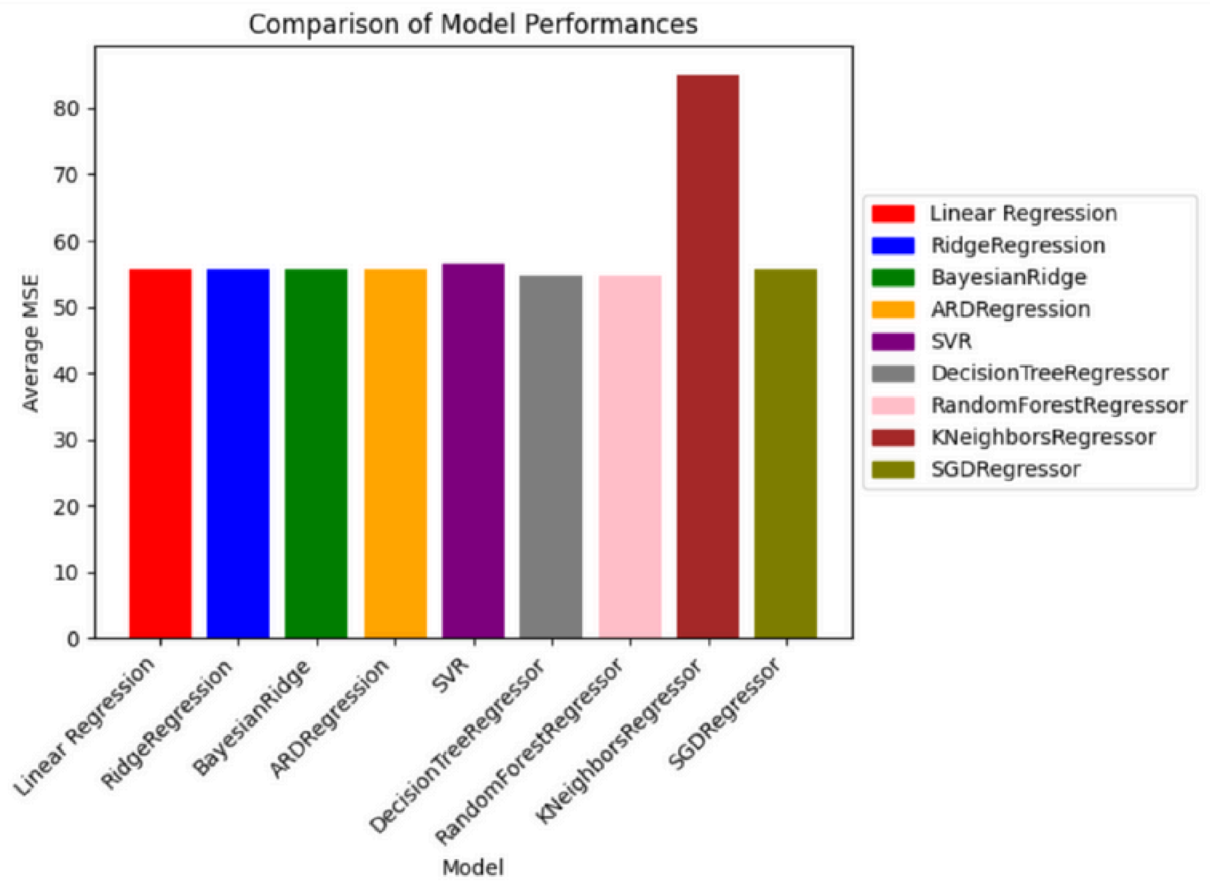
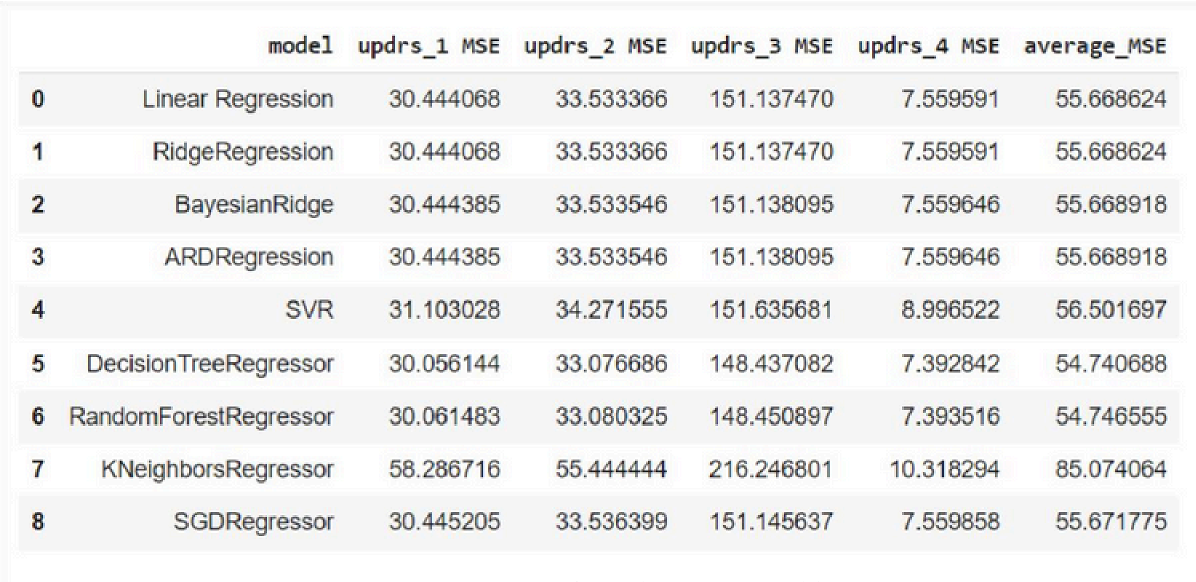


Figure 4: Performances of Machine Learning Models

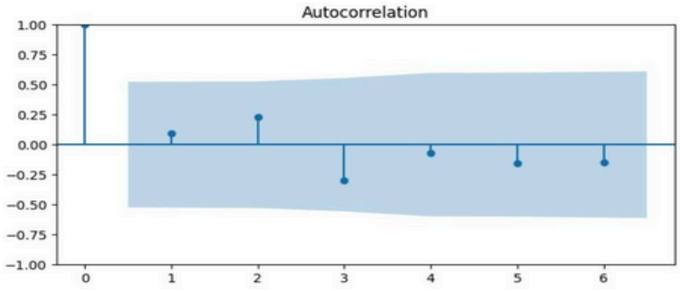


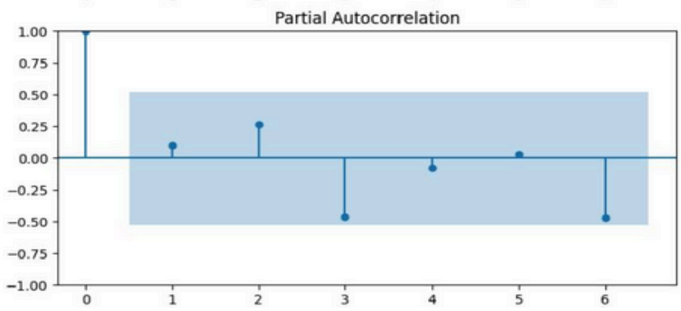
	model	updrs_1 MSE	updrs_2 MSE	updrs_3 MSE	updrs_4 MSE	average_MSE
0	Linear Regression	30.444068	33.533366	151.137470	7.559591	55.668624
1	RidgeRegression	30.444068	33.533366	151.137470	7.559591	55.668624
2	BayesianRidge	30.444385	33.533546	151.138095	7.559646	55.668918
3	ARDRegression	30.444385	33.533546	151.138095	7.559646	55.668918
4	SVR	31.103028	34.271555	151.635681	8.996522	56.501697
5	DecisionTreeRegressor	30.056144	33.076686	148.437082	7.392842	54.740688
6	RandomForestRegressor	30.061483	33.080325	148.450897	7.393516	54.746555
7	KNeighborsRegressor	58.286716	55.444444	216.246801	10.318294	85.074064
8	SGDRegressor	30.445205	33.536399	151.145637	7.559858	55.671775

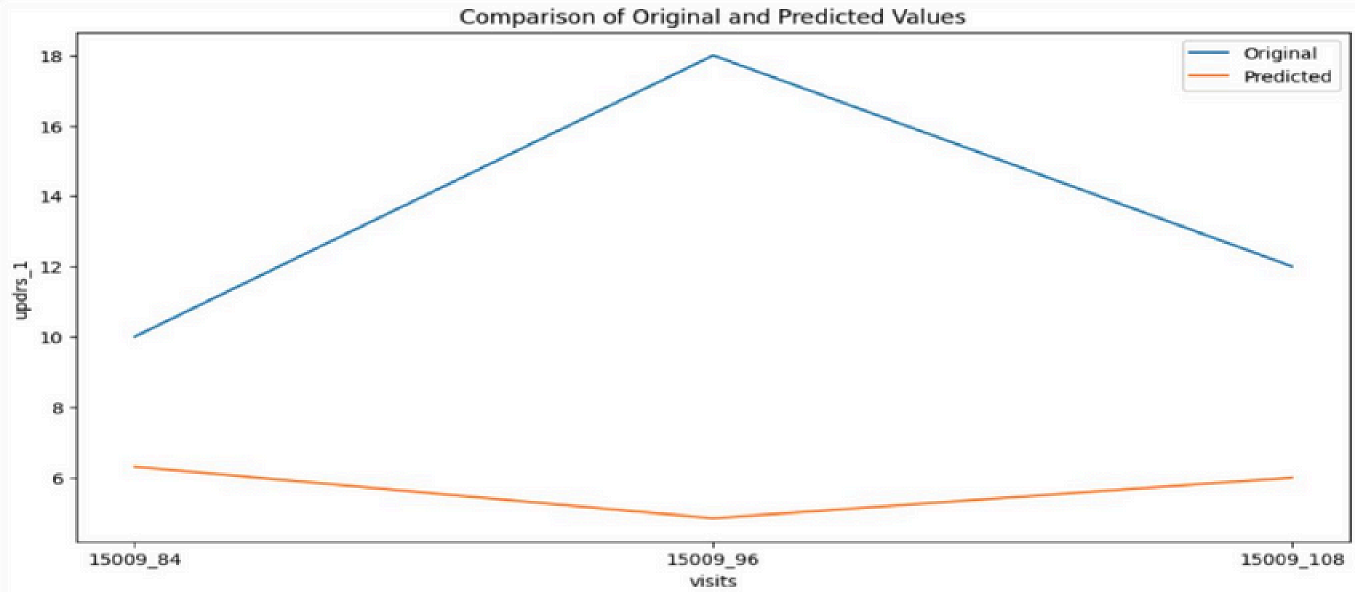
Figure 5: MSE scores for Machine Learning Models

The bar chart illustrates that the DecisionTreeRegressor model has the lowest average MSE value compared to all other models. Consequently, it demonstrates the best performance among the evaluated models. In summary, the

DecisionTreeRegressor model outperforms the other models, as evidenced by its lowest average MSE value when tested. We also tried the ARIMA model to predict future values by preparing a univariate model trained on the “updrs_1” of each patient. Since this univariate model requires the train and test data to be in the correct temporal order, we had to train the model on only one patient at a time. For each patient, there are very less records, so it is very tough for the model to understand and generalize the trend in the data. Still, based on the p,d, and q values found from the ACF and PACF plots, some predictions were made. Below are the RMSE and MAE values of the ARIMA model created for patient_id: 15009 predicting the values of “updrs_1”:
Mean absolute error: 7.6198173556472, Root
mean squared error: 8.619919963270238.







Discussion

The results of this project provide valuable insights into the progression of Parkinson's disease and its relationship with protein and peptide data. The impact of these findings is significant for researchers, clinicians, and patients alike, as it contributes to a deeper understanding of the disease and aids in more informed decision-making.

Researchers can use the results to further investigate the connections between protein and peptide data, medication state, and disease progression, ultimately contributing to the development of new diagnostic tools and therapies. Clinicians can benefit from the project by utilizing the generated models to better assess patient conditions, track the disease's progression, and personalize treatment plans. Patients, in turn, can receive more targeted and effective treatments, improving their quality of life.

The results can be used to make better-informed decisions by identifying the most relevant factors affecting Parkinson's disease progression, guiding the allocation of resources for research, and helping healthcare professionals monitor and manage the disease more effectively.

In future work, several aspects of the project could be improved. First, addressing the issue of missing and null values in the dataset could lead to more accurate models. Second, incorporating additional features, such as genetic data or patient lifestyle factors, might improve the predictive power of the models. Lastly, exploring more advanced machine learning techniques, such as deep learning or ensemble methods, could further enhance the models' performance.

Overall, the project has the potential to positively impact the lives of those affected by Parkinson's disease by offering valuable insights and contributing to better decision-making in diagnosis, treatment, and disease management.

References

1. https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction
2. https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-ScaleMDS-UPDRS.htm
3. https://www.kaggle.com/competitions/tabular-playground-series-jan-2022/discussion/298201
4. https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mds.22340
5. https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mdc3.12553 6. https://www.mcponline.org/article/S1535-9476(20)33221-7/fulltext