Predicting Parkinson's Disease Risk through Protein and Peptide Level Analysis: An Evidence from EDA and Machine Learning based Approach

PROGRESS REPORT OF THE TERM PROJECT

Submitted by:

Kushagra Gupta-01114811621

BACHELOR OF TECHNOLOGY
IN

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Under the Guidance

of

Dr. Tripti Lamba

(Associate Prof., AIML)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Maharaja Agrasen Institute of Technology,
PSP area, Sector – 22, Rohini, New Delhi – 110085
(Affiliated to Guru Gobind Singh Indraprastha, New Delhi)

(JUNE 2023)

Introduction

Parkinson's disease (PD) is a debilitating brain disorder that affects movement, cognition, and mood. It currently has no known cause or cure, but symptoms can be managed through treatments. PD has a global impact, affecting approximately 7 million people worldwide, and this number is expected to reach 10 million by 2030. Recent studies have investigated the potential of protein and peptide levels as biomarkers for PD, influencing the Unified Parkinson's Disease Rating Scale (UPDRS) scores. By analyzing data from 1,019 patients, researchers utilized exploratory data analysis (EDA) and machine learning to identify changes in protein and peptide levels that could serve as valuable biomarkers for assessing the risk of PD. Early detection of PD is vital for effective management and intervention. Machine learning algorithms utilize protein and peptide levels obtained from routine health screenings to detect PD at an early stage by uncovering patterns and correlations that enable accurate predictive models for UPDRS scores. The results of the study demonstrated the association between several proteins and peptides and the risk of PD, significantly impacting UPDRS scores. Additionally, the levels of these proteins and peptides were found to be higher in individuals with PD compared to healthy individuals. In conclusion, while PD remains a complex disorder without a cure, the integration of machine learning and biomarker analysis shows promise for early detection. These techniques aid in the identification of PD biomarkers, leading to improved diagnosis and treatment strategies. Incorporating machine learning into routine screenings has the potential to transform PD management, facilitating timely interventions and personalized care.

Problem Statement

we are looking at the Unified Parkinson's Disease Rating Scale (UPDRS) that was revised by the Movement Disorder Society (MDS) in 2008. This new scale - the MDS-UPDRS (which we will refer to within this EDA as simply the UPDRS) - consists of 4 separate parts. Each part consists of a questionnaire that rates signs or symptoms of Parkinson's Disease (PD). According to Holden et al (2018), the individual parts consist of:

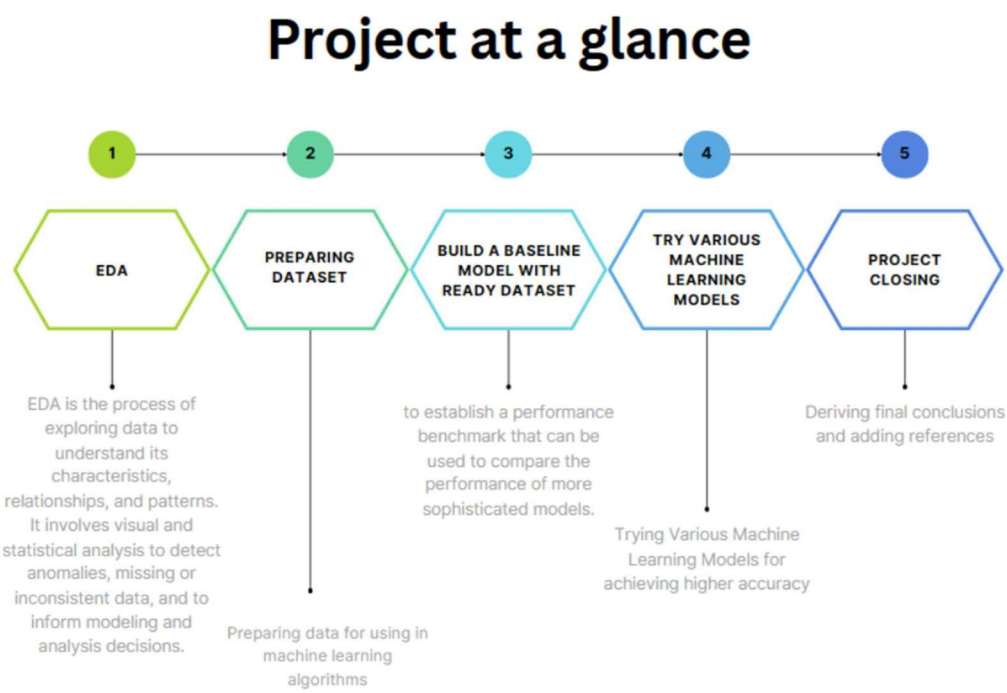
Part I - Non-Motor Aspects of Experiences of Daily Living

Part II - Motor Aspects of Experiences of Daily Living

Part III - Motor Examination

Part IV - Motor Complications

Questions within each part are scored on a 5 point scale ranging in values from 0 (normal) to 4 (most severe impairment). The maximum score that a patient may be assigned is 272 points. The challenge in this competition is to predict the UPDRS scores for parts 1 - 4 for each month that the patient had a visit and evaluation with a physician.



Objective

Our Objectives for the Project The objective of this research project is twofold:

- Firstly, it aims to explore the relationship between different features and data through exploratory data analysis (EDA). By analyzing various factors, including protein and peptide levels, the project aims to gain insights into their associations and understand the dynamics of Parkinson's disease. The goal is to uncover meaningful patterns and correlations that can enhance our understanding of the disease and its progression.
- (ii) Secondly, the project aims to identify the most suitable machine learning algorithm for predicting the SMAPE of UPDRS. By comparing and evaluating different algorithms, such as Support Vector Machines (SVM), Random Forest, K-Nearest Neighbours (KNN), Baseline CatBoost Model, Lasso Regression, Decision Tree Regression, TensorFlow Regression, PyTorch Regression, and others, the project seeks to determine which algorithm demonstrates the highest predictive performance and generalizability in estimating disease progression.

Technology Used

1. Python: a popular programming language used for developing the system.
2. Web Scrapping Tools: Given Dataset from Kaggle
3. Machine learning algorithms: used for training the Machine learning model
4. Data Visualization Libraries: Seaborn and Matplotlib

Methodology

The whole methodology can be easily represented through this workflow diagram:

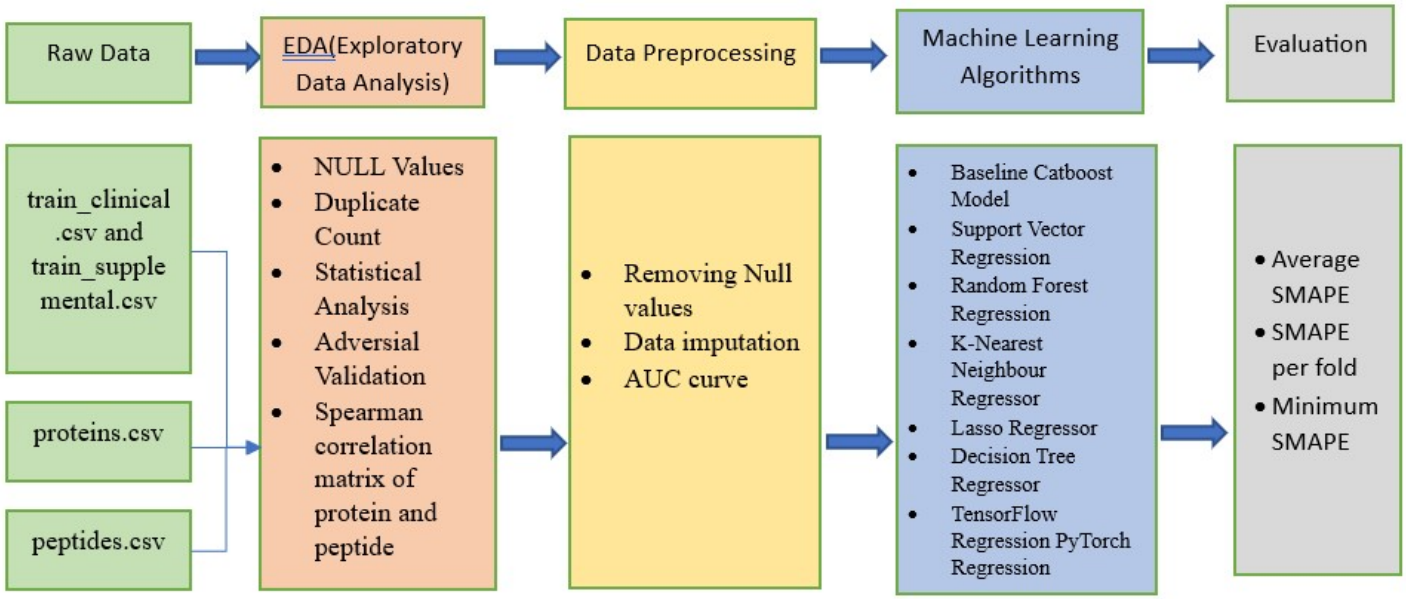


Fig. Workflow of Analysis and Model Evaluation

About the Data Set

Dataset	Unique Biomarkers	Number of Features
train_clinical.csv and train_supplemental.csv	1019 unique patients	**visit_id, patient_id, visit_month, updrs_1, updrs_2, updrs_3, updrs_4, upd23b_clinical_state_on_medication = 8 features**
proteins.csv	227 unique Protein values	**visit_id, patient_id, visit_month, UniProt, NPX =5 Features**
peptides.csv	968 unique Peptide values	**visit_id, patient_id, visit_month, UniProt, Peptide, PeptideAbundance =6 Features**

Files

train_peptides.csv Mass spectrometry data at the peptide level. Peptides are the component subunits of proteins.

- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The UniProt ID code for the associated protein. There are often several peptides per protein.
- `Peptide` - The sequence of amino acids included in the peptide. See this table for the relevant codes. Some rare annotations may not be included in the table. The test set may include peptides not found in the train set.
- `PeptideAbundance` - The frequency of the amino acid in the sample.

train_proteins.csv Protein expression frequencies aggregated from the peptide level data.

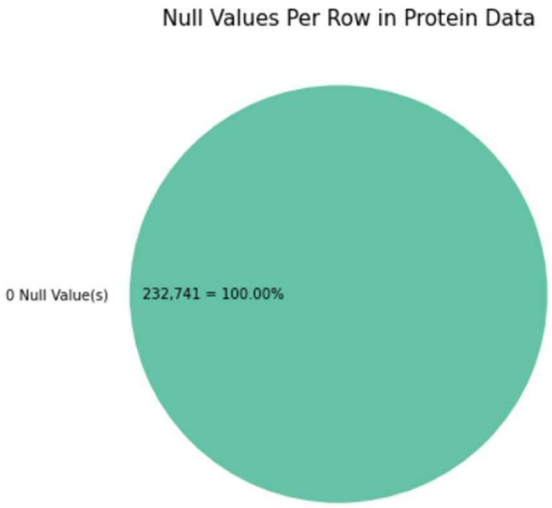
- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The UniProt ID code for the associated protein. There are often several peptides per protein. The test set may include proteins not found in the train set.
- `NPX` - Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.

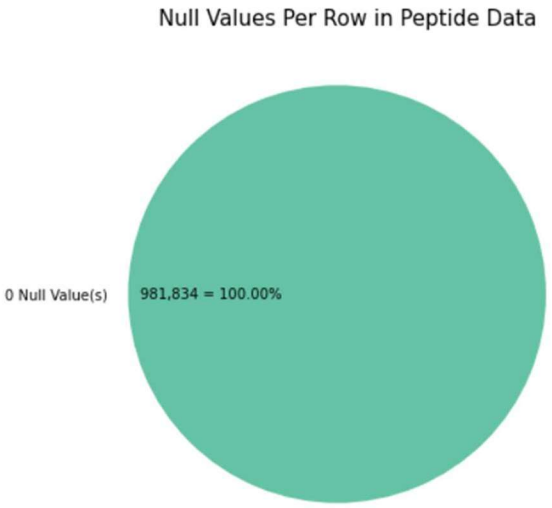
train_clinical_data.csv

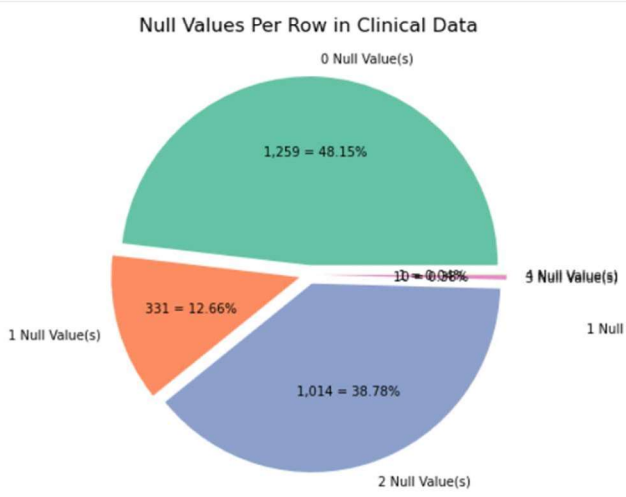
- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `updrs_[1-4]` - The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
- `upd23b_clinical_state_on_medication` - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function). These medications wear off fairly quickly (on the order of one day) so it's common for patients to take the motor function exam twice in a single month, both with and without medication.

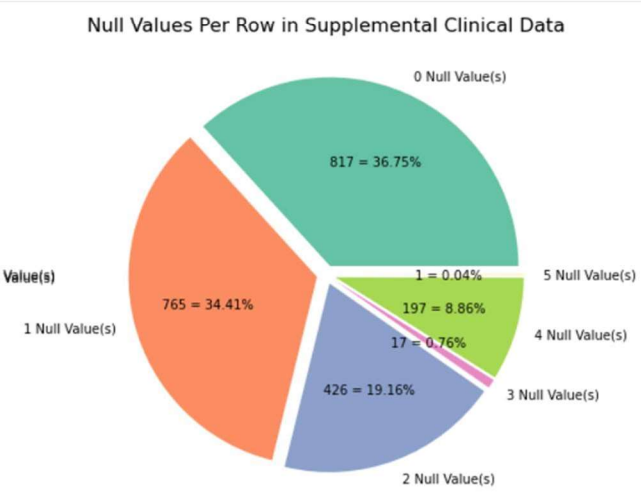
Observation and Analysis:

Exploring the issue of missing values in the dataset to see if there are systemic problems with data representation.









Null Values Per Row in Protein Data

Null Values Per Row in Peptide Data

Making Observations Based on Our Analysis

Rows with 1 Null Value

When a row contains only one null value, it is typically in the feature "upd23b_clinical_state_on_medication," which should have a value of either "On" or "Off." A null value in this field can be interpreted in two ways: the patient may have been off medication at the time of assessment, or the assessment may have failed to capture the patient's medication status. There are also 7 null values in the "updrs_3" column and 21 null values in the "updrs_4" column. According to Goetz et al (2008), "updrs_3" assesses motor function and "updrs_4" assesses motor complications, both with a minimum score of 0. A null value in either column suggests that the assessment was not performed, which is important because a score of 0 indicates that the patient was assessed and had normal responses.

Rows with 2 Null Values

When a row has two null values, they usually correspond to updrs_4 and upd23b_clinical_state_on_medication. As valid responses for the medication status field are limited to On or Off, a null value in this column is of interest as it is unclear whether the medication status was not recorded or if the patient was not taking any medication. The majority of other null values are found in the updrs_4 field, which assesses motor complications, while updrs_3 and updrs_2 have few null values. However, a null value in updrs_3 cannot be assumed to be a score of 0 as this indicates normal function. Similarly, null values in updrs_2 may indicate that the assessment was not performed.

Rows with 3 Null Values

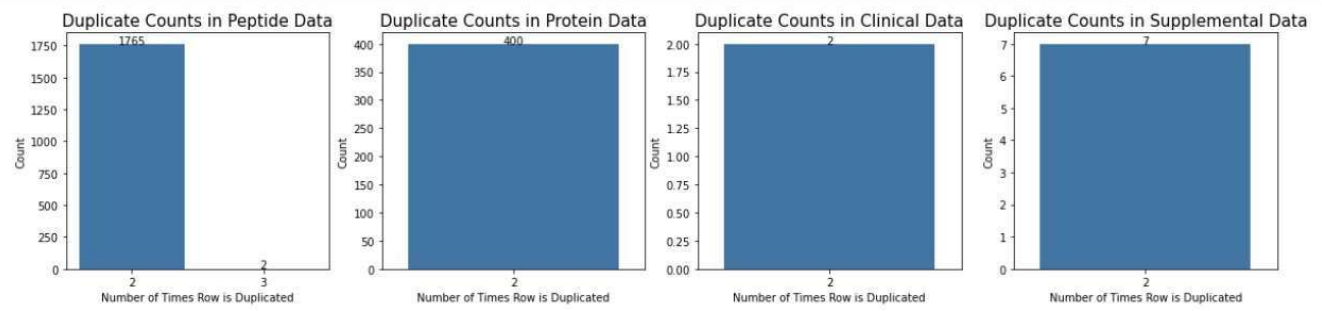
In 10 cases, rows have 3 null values, which correspond to missing information in updrs_3, updrs_4, and upd23b_clinical_state_on_medication. It should be noted that the missing values cannot be considered as 0 values.

Rows with 4 Null Values

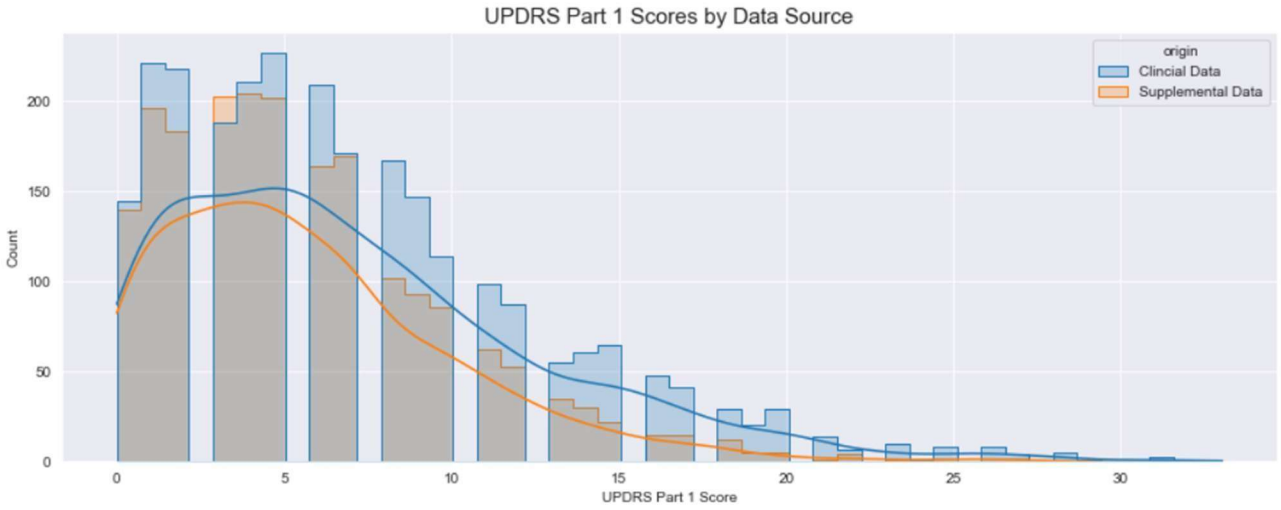
There is only one row in the dataset where four features contain null values. This suggests that only the UPDRS part 3 assessment was conducted during the visit, as the other features with null values correspond to motor complications and medication status. It is important to note that null values cannot be assumed to have a value of 0, as a score of 0 indicates normal function.

It is important to examine the supplementary information for any null values.

Handling Duplicate data

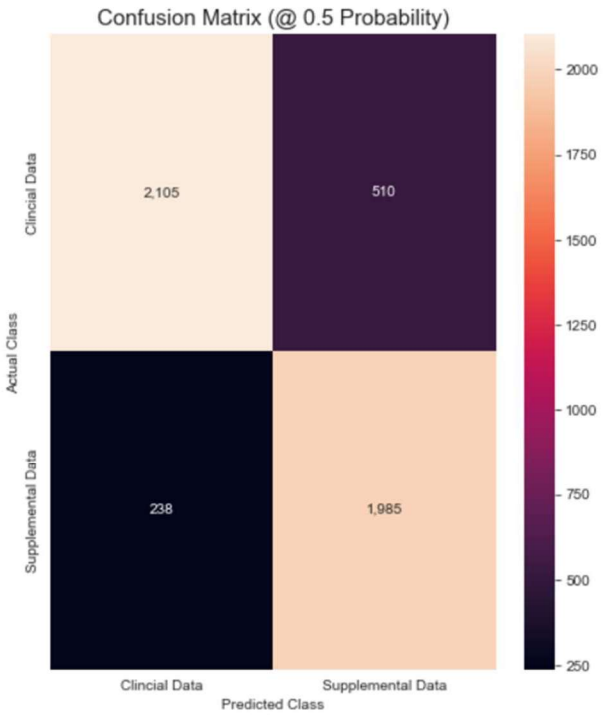


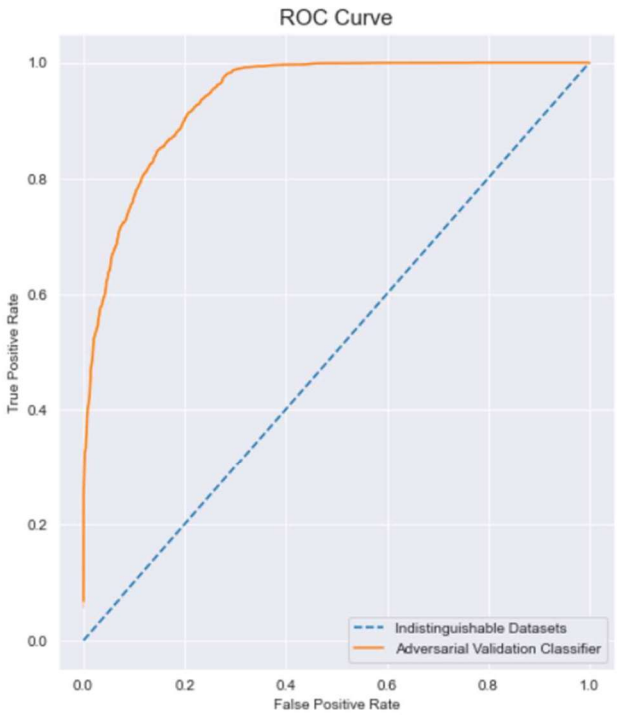
Analyzing UPRDS Score



Advertial Validation

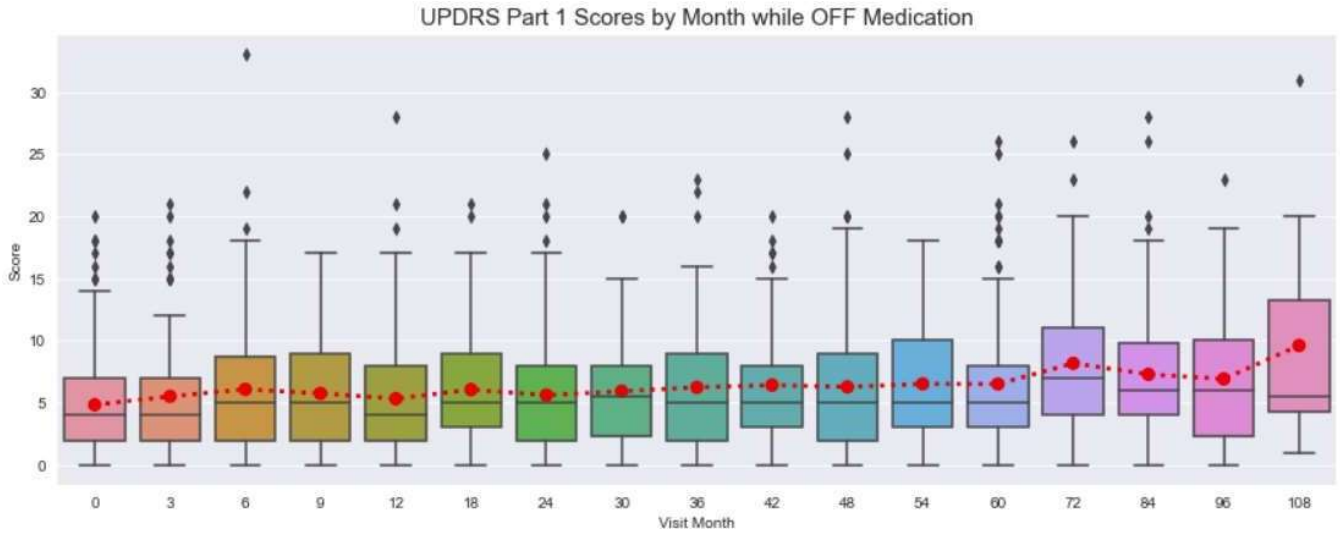
```
--> Overall results for out of fold predictions
: AUC ROC = 0.9389806206451068
```





We have trained a LightGBM classifier using clinical data and supplemental clinical data, and performs adversarial validation to check whether the two datasets are similar or not. It concatenates the two datasets, adds a label to identify the origin of each observation, and then performs 5-fold cross-validation. During each fold, it trains the model on the training set and evaluates it on the validation set. It uses early stopping and log evaluation callbacks to stop the training process when there is no improvement in the validation AUC ROC score for 50 consecutive iterations. It then computes the out-of-fold predictions and probabilities for each observation and stores them in arrays. Finally, it calculates the overall AUC ROC score, generates a confusion matrix, and plots the ROC curve.

Performing Statistical Analysis



	count	mean	std	min	25%	50%	75%	max
visit_month	2223.000000	12.910481	13.060532	0.000000	0.000000	6.000000	24.000000	36.000000
updrs_1	2010.000000	5.684080	4.366964	0.000000	2.000000	5.000000	8.000000	27.000000
updrs_2	2009.000000	6.507715	4.968132	0.000000	2.000000	5.000000	10.000000	34.000000
updrs_3	2218.000000	22.917944	12.342596	0.000000	14.000000	22.000000	31.000000	72.000000
updrs_4	1295.000000	0.840154	1.860247	0.000000	0.000000	0.000000	0.000000	12.000000

Fig.5. Statistical analysis of train_supplemental_data

Addressing the Limitations of "Defence1"

Upon receiving advice to expand the range of machine learning models in my research, I have made significant progress. Previously, I had only implemented one machine learning model. However, I am pleased to report that in my upcoming research paper, I have successfully incorporated eight different machine learning models. These models include the Baseline Catboost Model, Baseline CatBoost (UPDRS_4 as a constant), Support Vector Regression, Random Forest Regression, K-Nearest Neighbour Regression, Lasso Regression, Decision Tree Regression, TensorFlow Regression, and PyTorch Regression.

Among these models, the Random Forest regression model stands out with its exceptional performance, achieving an impressively low average SMAPE score of about 0.37. Additionally, to provide a more comprehensive understanding of the forecasting performance of the models, I have incorporated techniques such as 10-fold cross-validation or epochs to calculate the lowest SMAPE scores and SMAPE per fold.

The inclusion of these diverse machine learning models and the evaluation of their performance metrics enhance the robustness and reliability of my research findings. By expanding the scope of models and incorporating comprehensive evaluation techniques, I aim to provide valuable insights and contribute to the field of machine learning in the context of my research topic.

Models	Baseline Catboost Model	Baseline CatBoost Model [Setting UPDRS_4 constant]	Support Vector Regression	Random Forest Regression	K-Nearest Neighbour Regressor	Lasso Regressor	Decision Tree Regressor	TensorFlow Regression	PyTorch Regression
Average SMAPE	96.34	68.96	1.04	0.37	1.2454	0.74	0.62	0.60	0.93
Lowest SMAPE	93.59	66.96	1.02	0.32	1.08	0.71	0.57	-	0.91
SMAPE PER FOLD (k_fold=10 folds) /EPOCHS	97.64 95.88 97.15 95.36 97.51 94.36 95.63 97.02 93.59 99.23	67.44 68.25 70.93 70.58 67.30 66.96 68.63 68.83 69.39 71.26	1.06 1.05 1.03 1.07 1.08 1.02 1.05 1.02 1.07 1.04	0.32 0.37 0.37 0.34 0.35 0.40 0.40 0.38 0.36 0.36	1.20 1.21 1.30 1.35 1.17 1.24 1.32 1.25 1.28 1.08	0.71 0.74 0.72 0.75 0.71 0.76 0.82 0.75 0.74 0.72	0.57 0.62 0.66 0.58 0.63 0.65 0.65 0.63 0.61 0.57	-	0.92 0.95 0.92 0.94 0.91 0.93 0.94 0.93 0.95 0.94

Table: SMAPE scores of different Machine Learning algorithms used

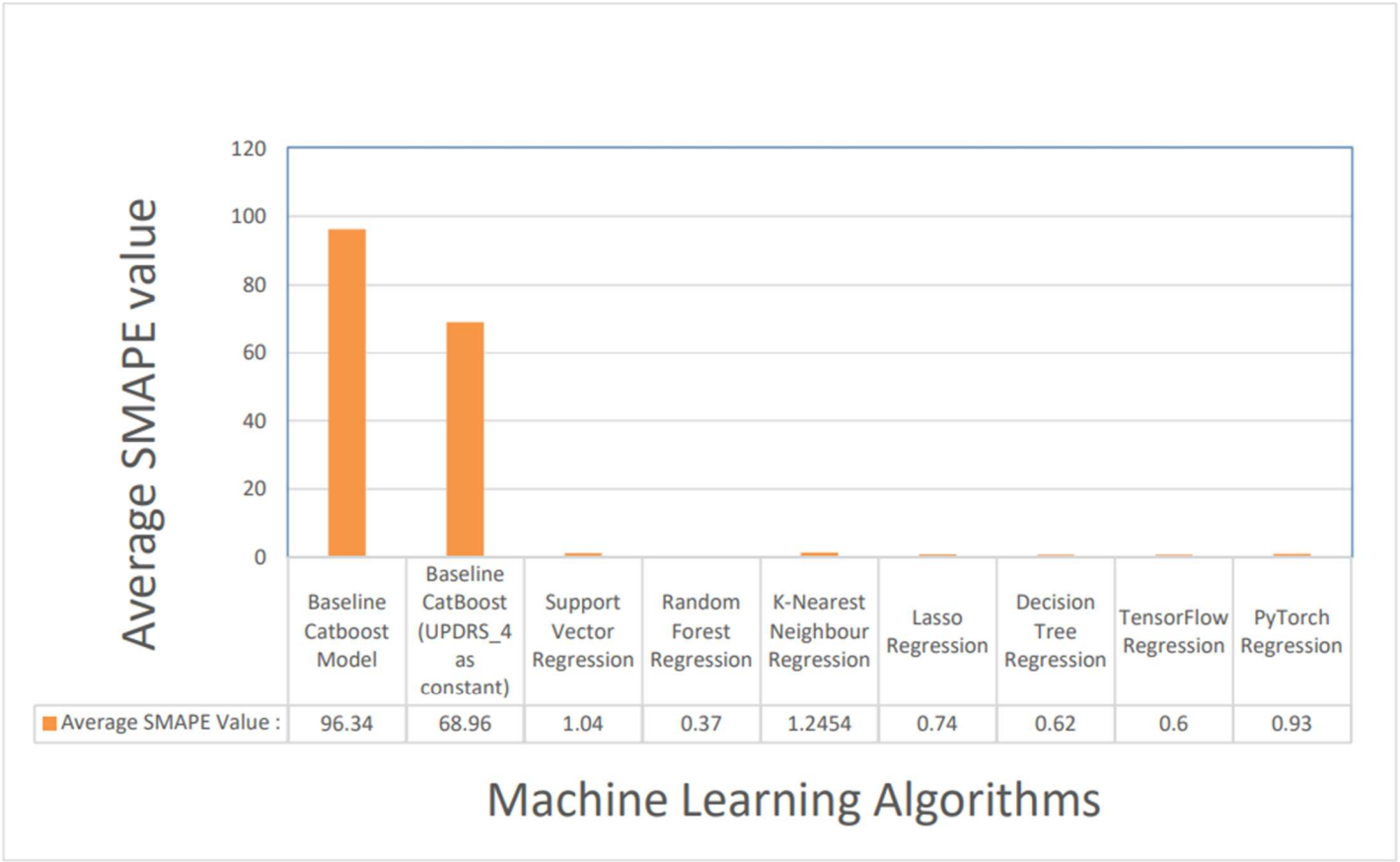
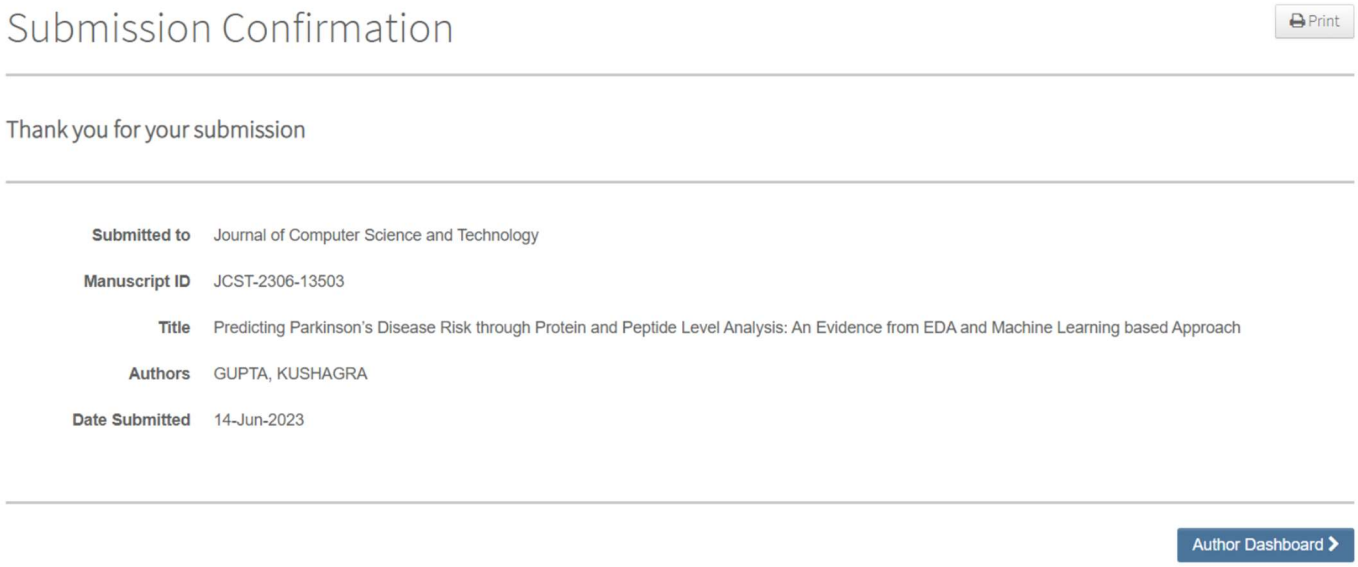


Fig.13. Graphical representation of Average SMAPE value of different Machine Learning Algorithms

Status of Research Paper

I am pleased to share that the current status of my research paper is that it has been officially submitted to JCST (Journal of Computer Science and Technology). JCST is a highly regarded journal in the field of computer science and technology, known for its reputation and credibility. It is a freely accessible journal indexed in Scopus, which adds to its prestige. JCST specifically welcomes papers in the domain of my research, making it an ideal platform to disseminate and contribute to the advancements in my field. I am hopeful that my submission will undergo a rigorous review process and, if accepted, will be published in JCST, thereby reaching a wide audience(international) and making a valuable contribution to the scientific community.



Submission Confirmation

Print

Thank you for your submission

Submitted to	Journal of Computer Science and Technology
Manuscript ID	JCST-2306-13503
Title	Predicting Parkinson's Disease Risk through Protein and Peptide Level Analysis: An Evidence from EDA and Machine Learning based Approach
Authors	GUPTA, KUSHAGRA
Date Submitted	14-Jun-2023

Author Dashboard

Conclusion

In conclusion, this research paper highlights the efficacy of utilizing protein and peptide level analysis, coupled with exploratory data analysis (EDA) and machine learning (ML) techniques, for predicting the risk of Parkinson's disease. The study showcases the superior performance of the Random Forest Regression model in accurately forecasting Parkinson's risk based on protein and peptide data. These findings underscore the significance of incorporating relevant variables and assessing model performance using metrics such as SMAPE. Ultimately, this research contributes to the advancement of Parkinson's research and holds promising implications for early diagnosis and personalized treatment approaches.