Hybrid Stacked Ensemble Regression Model for Predicting Parkinson's Progression on Protein Data

K. Shastry Aditya[1*], M. Mohan[2], and K. Deepthi[3]
[1,3]Department of Information Science and Engineering, Nitte Meenakshi Institute of Technology
Bengaluru, India 560064
[2]Department of Computer Science and Engineering, Amity University
Bengaluru, India 562110
Email: [1]adityashastry.k@nmit.ac.in, [2]mohanm@blr.amity.edu, [3]deepthi.k@nmit.ac.in

***Abstract*—Parkinson's Disease (PD) is a progressive neurological disorder marked by both motor and non-motor symptoms. Accurate prediction of disease progression is critical for effective patient management. The research presents a Hybrid Stacked Ensemble Regression (HSER) model for predicting PD progression using protein and peptide data measurements, leveraging the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) scores. The researchers integrate three datasets: clinical data, protein data, and peptide data into a comprehensive feature-engineered dataset. The dataset is split into training and testing sets in four configurations for predicting the four UPDRS scores, namely updrs_1, updrs_2, updrs_3, updrs_4. The hybrid approach combines stacking and blending techniques. The researchers select ridge regression, gradient boosting, and extra trees as base models. A meta-model is trained using the algorithms' out-of-fold estimates (ridge regression). The final predictions are obtained by averaging the predictions of the base models on the test data. The proposed HSER model exhibits enhanced performance compared to baseline models. These results underscore the promise of the hybrid model to enhance the prediction of PD progression, providing valuable insights for personalized treatment strategies. Future research can focus on refining model weights and exploring additional biomarkers to improve predictive accuracy.**

***Index Terms*—Parkinson's Disease, Hybrid Stacked Ensemble Regression, Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Scores, Protein and Peptide Data, Predictive Modeling**

I. Introduction

PARKINSON'S Disease (PD) is a progressive neurological disorder affecting millions worldwide, characterized by motor symptoms (such as tremors, rigidity, bradykinesia, and postural instability) and non-motor symptoms (including cognitive decline, mood disorders, sleep disturbances, and autonomic dysfunction) [1]. The complexity and variability of PD symptoms make it challenging to manage and treat effectively. Accurate prediction of disease progression is crucial for tailoring personalized treatment strategies, optimizing patient outcomes, and improving quality of life [2]. The Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is a comprehensive tool widely used to assess the severity and progression of PD. It encompasses four parts: non-motor experiences of daily living (updrs_1), motor experiences of daily living (updrs_2), motor examination (updrs_3), and motor complications (updrs_4) [3]. Despite its clinical utility, predicting PD progression using MDS-UPDRS scores remains challenging due to the complex interplay of various biological, environmental, and genetic factors [4].

Recent advancements in biomedical research have highlighted the promise of protein and peptide measurements as biomarkers for several illnesses, comprising of PD. Proteins and peptides can reflect underlying pathological processes and offer valuable insights into disease mechanisms [5]. However, current predictive models often fail to leverage the full spectrum of available biological data, particularly protein and peptide measurements, which hold significant potential for enhancing the accuracy of progression predictions in PD [6]. Previous exploration have surveyed various Machine Learning (ML) models for predicting PD progression, including Linear Regression, Random Forest (RF), and Gradient Boosting [7]. Whilst these simulations exhibit promise, they regularly face challenges such as overfitting, inability to handle high-

Received: Aug. 08, 2024; received in revised form: Nov. 12, 2024; accepted: Nov. 12, 2024; available online: April 14, 2025.
*Corresponding Author

dimensional data, and lack of integration of heterogeneous data sources [8]. Ensemble learning methods, that integrate multiple models to improve prediction accuracy and robustness, have shown potential in other domains but are underutilized in progression prediction of PD [9].

The research introduces a novel Hybrid Stacked Ensemble Regression (HSER) model that uniquely combines stacking and blending techniques to integrate clinical, protein, and peptide data for predicting PD progression. In contrast to earlier works which frequently rely on single data sources or simpler models, this approach leverages the strengths of multiple models and data types to enhance predictive accuracy. This effective approach not only improves prediction performance but also offers a comprehensive framework for incorporating diverse biological measurements in disease progression modeling.

The research aim to build a HSER model to predict the MDS-UPDRS scores (updrs_1, updrs_2, updrs_3, updrs_4) using integrated clinical, protein, and peptide data. By combining stacking and blending techniques, the research aims to enhance predictive accuracy and provide deeper insights into PD progression. The researchers hypothesize that the HSER model will outperform traditional single-model approaches in predicting MDS-UPDRS scores. The points steering the research are: Can the integration of clinical, protein, and peptide data improve the accuracy of PD progression predictions? How does the hybrid stacked ensemble model compare to baseline models in terms of predictive performance? The research has substantial consequences for the domain of PD management. By developing a more accurate predictive model, clinicians can better anticipate disease progression and tailor treatment plans accordingly. Additionally, the research addresses existing knowledge gaps by demonstrating the value of integrating heterogeneous biological data using advanced ensemble learning techniques.

The contributions of the research are as follows:

- Preparation of a comprehensive and cleaned medical dataset appropriate for training regression models to predict UPDRS scores based on protein and peptide data measurements.
- Development of a sophisticated ML-based framework for UPDRS score prediction using protein and peptide data measurements.
- Design of a novel HSER model for UPDRS score prediction using protein and peptide data measurements.
- Comparison of the designed HSER model with baseline models: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, RF, Gradient Boosting, AdaBoost, Bagging, and Extra Trees with respect to Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

A. Related Work

In this section, the researchers discuss the relevant research carried out in the domain of PD detection using ML/Deep Learning (DL), and ensemble methods. The researchers observe that less research is done on PD detection using protein and peptide data. The previous research addresses the challenge of accurately diagnosing PD, which shares symptoms with several other neurodegenerative diseases. Traditional diagnosis relies on patient history and symptoms, but these can overlap with conditions like Progressive Supranuclear Palsy (PSP), Multiple System Atrophy (MSA), essential tremor, and Parkinson's tremor. It utilizes neuroimaging biomarkers to assess dopamine levels in the brain, which is indicative of disease severity and progression, to improve diagnostic accuracy. It employs ML algorithms to classify patients based on this neuroimaging data. Specifically, it develops a stacked ML model that combines predictions from several algorithms, including K-Nearest Neighbor (KNN), RF, and Gaussian Naive Bayes (GANB). This approach achieves a 92.5% accuracy rate, surpassing traditional diagnostic methods [10].

Another previous research explores the potential of high throughput sequencing technologies, specifically RNA-Sequencing (RNA-Seq) data, to predict the progression of PD. Despite existing analytical representations which use medical data for this purpose, no models have previously been based on RNA-Seq data from PD patients. It aims to predict the progression of PD for a patient's subsequent medical visit by analyzing temporal patterns in RNA-Seq data. Using data from the Parkinson Progression Marker Initiative (PPMI) involving 423 PD patients over a span of 4 years and 34,682 predictor variables, the previous researchers develop a predictive model. This model employs a deep Recurrent Neural Network (RNN) enhanced with dense connections and batch normalization. The proposed model demonstrates the ability to predict PD progression with a RMSE of 6.0 and a significant rank-order correlation ($r = 0.83$, $p < 0.0001$) between predicted and actual disease status, indicating strong predictive performance [11].

Previous researchers have investigated main components which may forecast depression in patients with PD using a stacking ensemble approach. The goal is to provide foundational data for creating a nomogram prognostic index to identify high-risk groups for depression among these patients. They classify

depression into "with depression" and "without depression" categories using the Geriatric Depression Scale-30 (GDS-30). The team develops and tests nine ML models, including combinations of Artificial Neural Networks (ANN), RF, NB, and Decision Trees (DT) with Logistic Regression (LR). The models' analytical performance is assessed utilizing 10-fold cross-validation, and the RF combined with LR emerged as the best-performing model with an RMSE of 0.16, an Index of Agreement (IA) of 0.73, and an Explained Variance (EV) of 0.48. The analysis highlights ten significant predictors of depression in PD patients, including cognitive and motor assessments, daily living activity scales, and sleep-related disorders. It underscores the need for developing interpretable ML models that can be practically applied in the medical field to predict depression in PD patients [12].

Next, previous research has focused on predicting the severity of PD using protein and peptide biomarkers. PD, a debilitating neurological disorder affecting movement, cognition, and mood, impacts millions worldwide, with cases expected to rise significantly by 2030. It employs information from 1,019 patients to discover the relationship between biomarker levels and the UPDRS scores, which measure PD severity. The previous researchers employ Exploratory Data Analysis (EDA) and ML to identify biomarkers that can predict UPDRS scores, aiding in early PD detection and management. Their assessment demonstrates that numerous proteins and peptides are extensively correlated with PD risk and that these biomarkers are more prevalent in people with PD compared to healthy controls. Utilizing Symmetric Mean Absolute Percentage Error (SMAPE) to assess the predictive accuracy of various ML algorithms, the previous researchers find that the RF algorithm executes best, attaining a SMAPE score of 0.37. The findings suggest that biomarker analysis through ML holds promise for early PD detection, potentially leading to more effective intervention and management strategies, thereby offering promise for better results for PD patients [13].

Another previous research aims to enhance the early diagnosis and prediction of PD progression through an innovative ML approach. It develops a stack ensemble model that combines several ML algorithms, including DT, KNN, NB, and RF, to create a more accurate and robust predictive tool. It meticulously processes and combines diverse datasets, including clinical records, genetic information, and neuroimaging data, to extract valuable features for the model. Through extensive experiment and validation using a comprehensive dataset, the stack ensemble model exihibits effective analytical performance in comparison to individual algorithms. This model not only achieves higher accuracy but also improves interpretability by highlighting the key features contributing to the prediction of PD [14].

Next, previous research also leverages the extensive and heterogeneous dataset from the PPMI to enhance the prediction and diagnosis of PD using advanced ML techniques. By integrating and processing diverse data sources—including clinical records, genetic information, and neuroimaging data—it develops robust protocols for data handling and analysis. The ML methods, particularly Adaptive Boosting and Support Vector Machines, significantly outperform traditional model-based approaches, achieving high accuracy, sensitivity, and specificity. It underscores the importance of UPDRS scores in predicting PD and demonstrates that effective rebalancing of data cohorts can enhance predictive analytics [15].

II. Research Method

The researchers perform EDA to comprehend the description of the datasets. Subsequent tasks are performed in the EDA phase. First, the researchers plot clinical data. For example, in a random patient_id 1517, Fig. 1 plots the target label values (updrs_1, updrs_2, updrs_3, updrs_4) of the patient in each month (visit_month). Second, the researchers plot protein data. For example, a random patient_id 1517 in Fig. 2 shows the first 40 Protein entries (UniProt) of the patient and their Normalized Protein eXpression (NPX) value against patient's visit month (visit_month).

Then, Fig. 3 shows the proposed framework for prediction of MDS-UPDRS scores using protein and peptide measurements. Three datasets [16] are used for experimentation in the proposed framework. They are clinical data containing 8 columns and 2,615 records, proteins data containing 5 columns and 232,741 entries, and peptides data containing 6 columns and 981,834 entries.

There are three datasets used. First, the clinical data columns are visit_id, patient_id, visit_month, pd23b_clinical_state_on_medication, updrs_1, updrs_2, updrs_3, updrs_4. In the data, updrs_1, updrs_2, updrs_3, updrs_4 are the target variables that need to be predicted. Second, the columns in proteins dataset are visit_id, visit_month, patient_id, UniProt, and NPX. NPX is the protein concentration in shells. Third, the columns in peptides dataset are visit_id, visit_month, patient_id, UniProt, Peptide, and PeptideAbundance, showing the peptide concentration of each patient. These datasets are subjected to feature engineering. In the feature engineering phase, polynomial features up to the 2^{nd} degree are created out of the primary features. Polynomial features

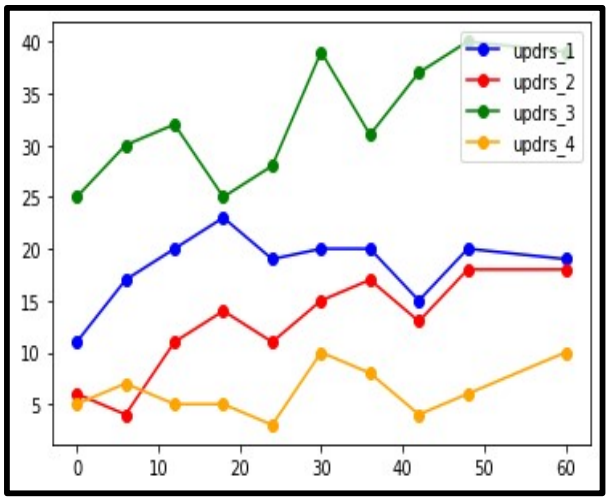

Fig. 1. Plot of Unified Parkinson’s Disease Rating Scale (UPDRS) scores for a random patient id: 1517.

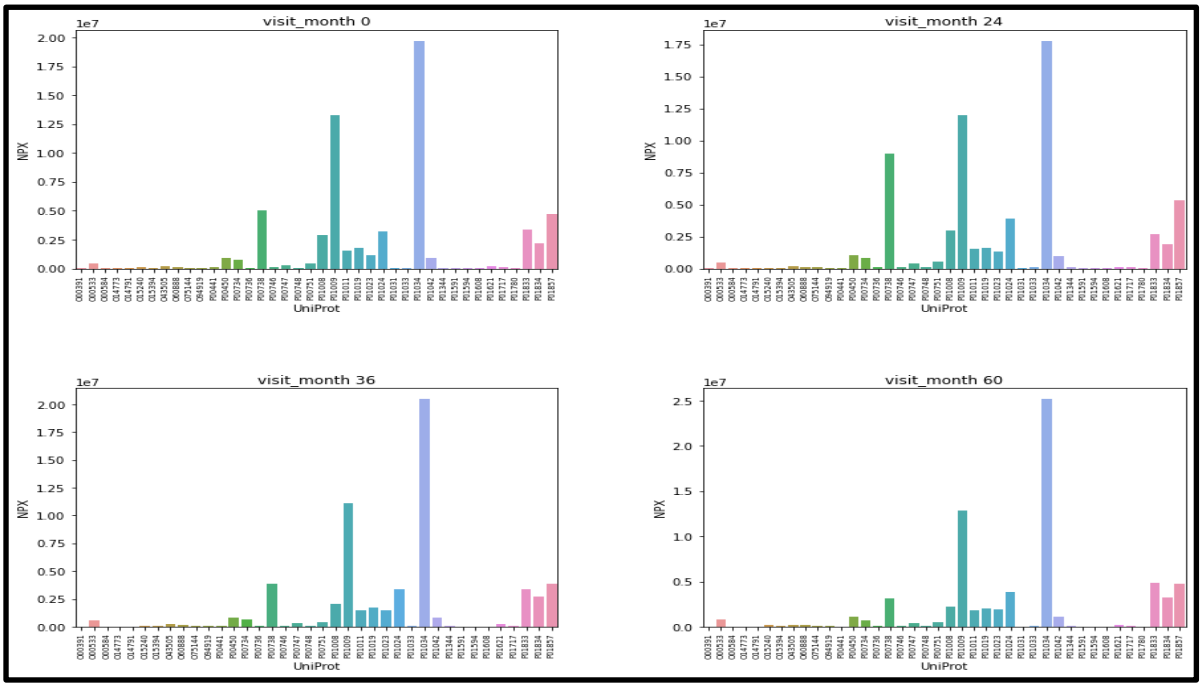

Fig. 2. Plot of first 40 entries of protein data for a random patient id: 1517.

are generated to let the model to capture non-linear relationships and interactions between features. It often delivers enhanced model performance by providing a richer set of features that can help in better approximating the essential patterns in the data. Subsequently, scaling is performed on these attributes. The proposed HSER model performs stacking by training multiple base models (Ridge Regression, Gradient Boosting Regressor, and Extra Trees Regressor) on clinical, protein, and peptide data, using their out-of-fold predictions to train a meta-model. In the blending phase, it averages the predictions from the base models for final robust and accurate UPDRS score predictions.

A. Dataset Preparation

The summary of the raw dataset used in the research is as follows:

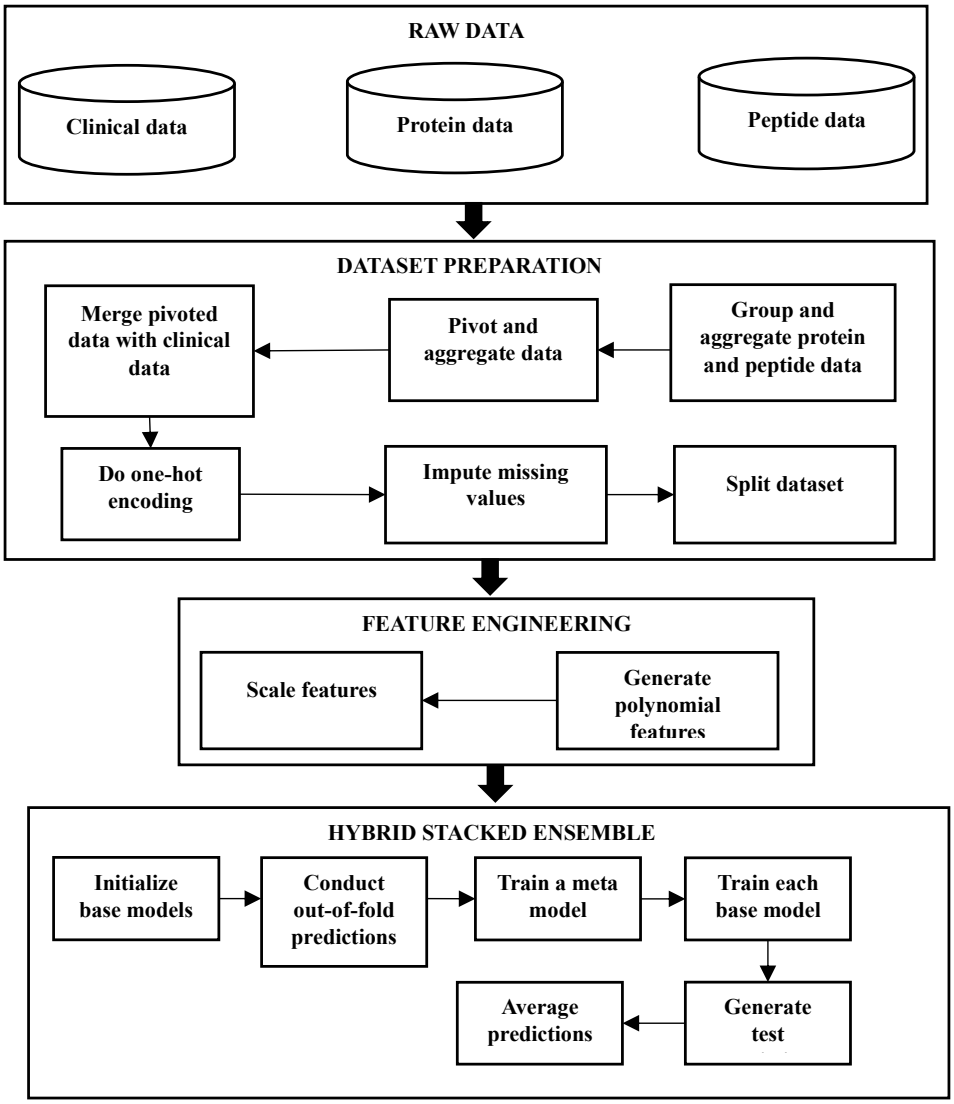

Fig. 3. Proposed framework for Unified Parkinson's Disease Rating Scale (UPDRS) score prediction using protein and peptide measurements.

- Train_clinical_data: The train_clinical data compose of eight attributes (visit_id, patient_id, visit_month, updrs_1, updrs_2, updrs_3, updrs_4, and upd23b_clinical_state_on_medication) and 2,615 records. UPDRS is a rating instrument used to measure the the severity and progression of PD in patients. When a patient visits the clinic, the clinic will record how the patient scored on 4 parts of UPDRS test. These records are available in train_clinical. The ratings for the the first four segments of UPDRS are available as updrs_1, updrs_2, updrs_3 and updrs_4 in this dataset. The goal is to train a model to predict these UPDRS ratings.
- Train_proteins_data: The train_proteins data have five attributes (visit_id, visit_month, patient_id, UniProt, and NPX) and 232,741 records. The clinic will also record the patient's NPX value for all the proteins relevant to PD during each visit. NPX is nothing but the value representing the protein concentration in shells.
- Train_peptides_data: Proteins are long molecules made up of multiple peptides. The clinic will record the peptide abundance of each peptide in proteins relevant to PD. It shows the peptide concentration, similar to NPX for proteins. The train_peptides data compose of six attributes (visit_id, visit_month, patient_id, UniProt, Peptide, and PeptideAbundance) and 981,834 records.

These three datasets (protein, peptide, and clinical

TABLE I
SAMPLE OF FEATURE ENGINEERED DATASET.

	Visit_Id	391	533	584	14498	14773	14791	15240
0	10053_0	9104.27	402321	NaN	NaN	7150.57	2497.84	83002.9
1	10053_12	10464.2	435586	NaN	NaN	NaN	NaN	197117.0
2	10053_18	13235.7	507386	7126.96	24525.7	NaN	2372.71	126506.0
3	10138_12	12600.2	494581	9165.06	27193.5	22506.10	6015.90	156313.0
4	10138_24	12003.2	522138	4498.51	17189.8	29112.40	2665.15	151169.0

datasets) are combined into a single comprehensive dataset for predicting UPDRS scores. The researchers have to predict the ratings for the the first four segments of UPDRS (updrs_1, updrs_2, updrs_3 and updrs_4) that are likely to be recorded by the clinic during a patient visit. Hence, these are the labels. Next, the researchers prepare the dataset for training models to forecast the four labels. To forecast the target labels (updrs_1, updrs_2, updrs_3, updrs_4) for a given visit, the researchers use the recorded protein and peptide data of the patient during that visit. The rows in train_proteins data are grouped by visit ids (visit id) and protein ids (UniProt). Then, the researchers replace the NPX values of each row in a group with the mean of the its values of all rows in that group.

Similarly, the researchers group the rows in train_peptides data by their visit ids (visit_id) and peptide ids (Peptide). Then, the researchers replace the PeptideAbundance values of each row in a group with the mean of the its values of all rows in that group. Subsequently, the researchers spread the rows of the grouped datasets into columns. For this, the researchers use the Pandas pivot function. The protein dataset is pivoted that unique values of visit_id become the indices, and the values of UniProt1 in the dataset become columns.

For each visit (row), the NPX values corresponding to the different UniProt1 values recorded that the visit is captured in the columns. Then, peptipe dataset is pivoted that unique values of visit_id become the indices, and the values of peptide in the dataset become columns. For each visit (row), the PeptideAbundance values corresponding to the different peptide values recorded for the visit captured in the columns. The researchers then merged the pivoted peptide dataset with the pivoted protein dataset on visit_id. The Yggdrasil DF handles the missing values in the numerical columns. The feature engineered dataset is composed of 1,196 columns and 1,113 entries. Its first five entries are shown in Table I.

For grouping the protein data, the researchers calculate the average NPX for each combination of visit_id and UniProt as shown in Eq. (1). It has v as visit_id, u as UniProt, and i as indexes of the individual measurements. For grouping the peptide data, the researchers calculate the average PeptideAbundance for each combination of visit_id and peptide as shown in Eq. (2), which p is peptide. For pivoting the protein data, the researchers reshape the protein data so that each UniProt is a separate column, and visit_id is the index using Eq. (3). The peptide data are pivoted by reshaping it so that each peptide is a separate column, and visit_id is the index. It is shown in Eq. (4).

TABLE II
PARTITIONED PROTEIN AND PEPTIDE DATASETS.

Datasets	Number of Training Records	Number of Test Records	Target Variable
Dataset-1	852	216	updrs_1
Dataset-2	838	230	updrs_2
Dataset-3	843	215	updrs_3
Dataset-4	464	105	updrs_4

$$\text{NPX}_{\text{mean}}(v,u) = \frac{1}{n}\sum_{i=1}^{n}\text{NPX}(v,u,i), \quad (1)$$

$$\text{PeptideAbundance}_{\text{mean}}(v,p) = \frac{1}{n}\sum_{i=1}^{n} \text{PeptideAbundance}(v,p,i), \quad (2)$$

$$df\_protein(v,u) = \text{NPX}_{\text{mean}}(v,u), \quad (3)$$

$$df\_peptide(v,p) = \text{PeptideAbundance}_{\text{mean}}(v,p). \quad (4)$$

The protein and peptide data are then combined based on visit_id. The combined protein-peptide data are merged with the clinical data based on 'visit_id'. The rows with missing UPDRS scores are dropped. The categorical values are then converted into binary form using one-hot encoding technique. Then, missing values are then filled by using the mean imputation method. Here, the missing values are replaced by the attribute mean as shown in Eq. (5) that $\bar{X}_{ij}$ is the new dataset value, X_{ij} is the i-th value of the j-th attribute, and μ_j is the average of feature j. Finally, the dataset is separated into training and testing datasets as shown in Table II.

Algorithm 1 Feature Engineering: `feature_engineering`(X_{train}, X_{valid})

Initialize PolynomialFeatures with degree = 2, include_bias = False
Generate Polynomial Features for Training Data:
for each sample $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ in X_{train} **do**
    Generate original features: $x_{i1}, x_{i2}, \ldots, x_{im}$
    Generate squares of features: $x_{i1}^2, x_{i2}^2, \ldots, x_{im}^2$
    Generate pairwise interactions: $x_{i1}x_{i2}, x_{i1}x_{i3}, \ldots, x_{i(m-1)}x_{im}$
end for
Resulting in X_{train_poly} with $\frac{m(m+1)}{2}$ features
Generate Polynomial Features for Validation Data:
for each sample $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ in X_{valid} **do**
    Generate original features: $x_{i1}, x_{i2}, \ldots, x_{im}$
    Generate squares of features: $x_{i1}^2, x_{i2}^2, \ldots, x_{im}^2$
    Generate pairwise interactions: $x_{i1}x_{i2}, x_{i1}x_{i3}, \ldots, x_{i(m-1)}x_{im}$
end for
Resulting in X_{valid_poly} with $\frac{m(m+1)}{2}$ features
Feature Scaling:
for each feature j in X_{train_poly} **do**
    Compute mean μ_j and standard deviation σ_j
    for each element X_{ij} **do**
        Transform: $\bar{X}_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$
    end for
end for
for each feature j in X_{valid_poly} **do**
    Compute mean μ_j and standard deviation σ_j
    for each element X_{ij} **do**
        Transform: $\bar{X}_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$
    end for
end for
Return X_{train_scaled}, X_{valid_scaled}

Algorithm 2 HSER Algorithm

Require: Training dataset (X_{train}, y_{train}), Test dataset X_{test}
Ensure: Predictions for the test dataset
Initialize base models:
base_models ← [RidgeRegression(),
                GradientBoostingRegressor(),
                ExtraTreesRegressor()]
Initialize variables:
num_models ← length(base_models)
K ← number of folds for cross-validation
predictions_out_of_fold ← array of shape (num_models, len(X_{train}))
Perform Stacking using K-fold cross-validation:
for each base_model in base_models **do**
    for each fold in K-fold cross-validation **do**
        Split X_{train} into training and validation sets
        Train base_model on training set
        Predict validation set with base_model
        Store predictions in predictions_out_of_fold
    end for
end for
Train Meta Model:
meta_model ← RidgeRegression()
Train meta_model using predictions_out_of_fold as input and y_{train} as target
Perform Blending (generate predictions for test set):
predictions_test ← array of shape (num_models, len(X_{test}))
for each base_model in base_models **do**
    Train base_model on entire X_{train} and y_{train}
    Predict X_{test} with base_model
    Store predictions in predictions_test
end for
Calculate final predictions for X_{test}:
final_predictions ← mean(predictions_test, axis = 0)
Output final_predictions as the predicted values for the test dataset

$$\bar{X}_{ij} = \begin{cases} X_{ij} & \text{if } X_{ij} \neq \text{NaN} \\ \mu_j & \text{if } X_{ij} = \text{NaN}. \end{cases} \tag{5}$$

B. Feature Engineering

In this phase, the prepared dataset is subjected to polynomial conversion and scaling. The step by step process is shown in Algorithm 1. It processes the combined protein and peptide dataset by first generating polynomial features (up to the second degree) and scaling these features to have zero mean and unit variance. This approach captures non-linear relationships and interactions between the features, potentially improving the model's performance. Initially, polynomial features up to the second degree are generated from the input training and validation datasets. It involves creating new features that include original features, their squares, and pairwise interactions, enabling the model to capture non-linear relationships.

Subsequently, the generated polynomial features are standardized using StandardScaler, ensuring each feature has zero mean and unit variance based on statistics computed from the training information. This scaling step prepares the information for ML algorithms that perform better with standardized inputs. Ultimately, the algorithm returns the scaled polynomial features for both the training and validation datasets, facilitating robust model training and evaluation.

III. Results and Discussion

A. Design of the Hybrid Stacked Ensemble Regression (HSER) Model

In the proposed work, the researchers develop a hybrid approach by combining stacking and blending for predicting the UPDRS scores (updrs_1, updrs_2, updrs_3, updrs_4). Algorithm 2 shows the proposed HSER model. The initial stage in Algorithm 2 involves the selection of models for stacking. In the research, the three best performing diverse models Viz. Ridge Regression, Gradient Boost, and Extra Trees are chosen for the stacking process. The second step involves the implementation of the stacking models. The researchers initialize the base models. Let M_1 denote the ridge regression model, M_2 represent the Gradient Boosting model, and M_3 signify the Extra-Trees model. Subsequently, the out of fold predictions are performed.

For each base model of M_i, the researchers split the training data into k folds. For each fold, the researchers train the base model on $k-1$ folds and predict the fold that is left out. The researchers then store these out-of-fold predictions. After 'k' iterations, the researchers obtain the out-of-fold predictions for each data point from each base model. Let $\hat{y}_{M_i}^{(k)}(x)$ denote the prediction of base model of M_i on the k^{th} fold. Now, the researchers use these out-of-fold predictions to train a meta-model (Ridge Regression).

The researchers construct a new training set for the meta-model using the out-of-fold predictions from all the base models. The researchers denote the out of fold predictions as $\hat{Y}_{M_i}^{(k)}$, where $\hat{Y}_{M_i}^{(k)}$ is a vector of predictions for base model M_i on fold 'k'.

Then, The researchers train the meta-model (Ridge Regression) to predict the actual target UPRDS scores y using these out-of-fold predictions. The ridge regression minimizes the following objective function in Eq. (6). It has β_1 , β_2, and β_3 as the coefficients for combining predictions and α as a regularization parameter.

$$\min_{\beta} \sum_{k=1}^{K} \left\| y - \sum_{i=1}^{3} \beta_i \hat{Y}_{M_i}^{(k)} \right\|_2^2 + \alpha \|\beta\|_2^2. \quad (6)$$

The third step, the blending phase, aims to combine predictions from multiple base models to produce final predictions for the UPDRS scores (updrs_1, updrs_2, updrs_3, and updrs_4) using the clinical, protein, and peptide datasets. In the blending phase, the researchers initialize an array to store predictions from each base model for the test dataset shown in Eq. (7). It consists of n as the quantity of samples in the test set and m as the number of base models. Each entry $\hat{y}_{ij}$ represents the prediction for the i^{th} test sample by the j^{th} as base model. Then, each base model M_j is trained on the entire training dataset (X_{train},y_{train}). For instance, if X_{train} includes features such as pd23b_clinical_state_on_medication, aggregated NPX, and PeptideAbundance, the model training can be represented as depicted in Eq. (8). After training, each base model predicts the test dataset X_{test} (Eq. (9)).These predictions are stored in the '$predictions\_test$' array.

$$\text{predictions\_test} = \begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1m} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \cdots & \hat{y}_{nm} \end{bmatrix}, \quad (7)$$

$$M_j \leftarrow \text{train}(X_{\text{train}}, y_{\text{train}}), \quad (8)$$

$$\hat{y}_{ij} = M_j(X_{\text{test}})_i. \quad (9)$$

To obtain the final predictions, the researchers calculate the mean of the predictions from all base models for every test sample (Eq. (10)). It has $\hat{y}_{\text{updrs}_{k,i}}$ as the final predicted value for the i^{th} test sample, k^{th} as UPDRS score, m as the number of base models, and $\hat{y}_{\text{updrs}_{k,i}}^{(j)}$ as the prediction for the i^{th} test sample by the j^{th} base model for the k^{th} UPDRS score. The final prediction vector for the whole test dataset for each UPDRS score is shown in Eq. (11). The final predictions for all UPDRS scores are the output as the predicted values for the test dataset. It can be represented in Eqs. (12) to (15). This blending phase ensures that the final predictions for the UPDRS scores are robust and accurate by combining the diverse strengths of multiple base models trained on comprehensive feature sets derived from clinical, protein, and peptide information.

$$\hat{y}_{\text{updrs}_{k,i}} = \frac{1}{m} \sum_{j=1}^{m} \hat{y}_{\text{updrs}_{k,i}}^{(j)}, \quad (10)$$

$$\hat{y}_{\text{updrs}_k} = \left[\hat{y}_{\text{updrs}_{k,1}}, \hat{y}_{\text{updrs}_{k,2}}, \ldots, \hat{y}_{\text{updrs}_{k,n}} \right], \quad (11)$$

$$\text{Final\_Predictions\_UPDRS1} = \frac{1}{m} \sum_{j=1}^{m} \text{Predictions\_Test}, \quad (12)$$

$$\text{Final\_Predictions\_UPDRS2} = \frac{1}{m} \sum_{j=1}^{m} \text{Predictions\_Test}, \quad (13)$$

$$\text{Final\_Predictions\_UPDRS3} = \frac{1}{m} \sum_{j=1}^{m} \text{Predictions\_Test}, \quad (14)$$

$$\text{Final\_Predictions\_UPDRS4} = \frac{1}{m} \sum_{j=1}^{m} \text{Predictions\_Test}. \quad (15)$$

B. Experiment Results

The experiments are conducted on Windows 10 operating system using Python as the programming language. The researchers discuss the prediction results of the proposed HSER model. Figures A1 to A4 as seen in Appendix demonstrate the model results for updrs_1, updrs_2, updrs_3, and updrs_4 prediction with respect to the performance metrics of RMSE and MAE respectively. Then, Table III shows the comparison of the models for UPDRS prediction.

Here are the inferences which can be done from Figs. A1 to A4 (in Appendix) and Table III. First, regarding the prediction results of updrs_1, the proposed HSER model exihibits RMSE improvements ranging from 0.13% to 13.46% and MAE improvements from 0.13% to 9.83%. Second, in updrs_2 prediction, RMSE improves from 0.22% to 8.46%, and MAE has improvements of 0.24% to 6.51% in the proposed HSER model. Third, in updrs_3 prediction, the proposed HSER model has RMSE improvements from 1.03% to 23.18% and MAE improvements from 1.15% to 17.57%. Last, for the updrs_4 prediction, the designed HSER model is performed with RMSE improvements from 0.03% to 4.48% and MAE improvements from 0.04% to 3.89%.

TABLE III
PERFORMANCE METRICS (ROOT MEAN SQUARED ERROR (RMSE) AND MEAN ABSOLUTE ERROR (MAE)) FOR VARIOUS MODELS ACROSS UNIFIED PARKINSON'S DISEASE RATING SCALE (UPDRS) METRICS.

Models	UPDRS_1		UPDRS_2		UPDRS_3		UPDRS_4	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Linear Regression	17.930	13.360	13.410	10.380	35.050	26.720	6.690	5.290
Bagging	5.380	4.280	5.800	4.690	14.800	11.940	2.470	2.090
AdaBoost	5.220	4.180	5.670	4.790	14.520	11.670	2.470	1.530
Lasso Regression	4.980	3.880	5.360	4.250	13.910	11.100	2.450	1.510
Random Forest	4.940	3.860	5.340	4.290	13.840	11.050	2.370	1.510
ElasticNet	4.875	3.863	5.310	4.230	13.380	10.990	2.310	1.490
Ridge Regression	4.690	3.650	5.230	4.170	12.990	10.440	2.260	1.450
Extra Trees	4.630	3.640	5.190	4.130	12.940	10.380	2.250	1.450
Gradient Boost	4.600	3.620	5.170	4.110	12.900	10.300	2.240	1.440
Proposed HSER	4.470	3.490	4.950	3.870	11.870	9.150	2.210	1.400

C. Results Discussion

The proposed HSER model for predicting PD progression outperforms several baseline models for several key reasons. First, it leverages multiple data sources. The HSER model integrates clinical, protein, and peptide data into a comprehensive feature-engineered dataset. This holistic approach allows the model to capture complex interactions and patterns that single-source models may miss, providing a richer and more nuanced dataset for training. Second, there are advanced ensemble techniques. The HSER model combines both stacking and blending techniques. Stacking involves training multiple base models (Ridge Regression, Gradient Boosting, and Extra Trees). It uses out-of-fold predictions to train a meta-model (Ridge Regression). This approach helps to capture diverse aspects of the data. Blending involves averaging the predictions of the base models on the test data, which reduce overfitting and improve generalization. Third, it has diverse base models. The use of three distinct base models (Ridge Regression, Gradient Boosting, and Extra Trees) provides complementary strengths. Ridge Regression offers a linear perspective with regularization to prevent overfitting. Gradient Boosting provides the capability to handle non-linear relationships and interactions within the data. Meanwhile, Extra Trees adds robustness and reduces variance through ensemble techniques. Fourth, there is meta-model refinement. Using a ridge regression as the meta-model allows the system to combine the strengths of the base models effectively. Ridge Regression's regularization properties manage potential overfitting from the base models' predictions.

A detailed explanation of how the proposed HSER model outperforms each of the baseline models individually is explained. First, Linear Regression assumes a simple linear relationship between the input features and the target variable. While it is easy to interpret, it struggles with complex, non-linear relationships [17]. The HSER model, with its combination of Ridge Regression, Gradient Boosting, and Extra Trees, captures both linear and non-linear patterns in the data, leading to more accurate predictions. Second, Ridge Regression introduces regularization to Linear Regression, which reduces overfitting. However, it still operates under the assumption of linear relationships [18]. The HSER model benefits from ridge regression's regularization while also leveraging the non-linear modeling capabilities of Gradient Boosting and Extra Trees, providing a more robust prediction. Third, Lasso Regression implements both attribute choice and regularization, which can improve the model's performance by excluding irrelevant features. However, like Ridge Regression, it assumes linear relationships [19]. The HSER model combines the strengths of Ridge Regression's regularization and the feature selection of Lasso with the non-linear capabilities of tree-based methods, leading to improved predictive power. Fourth, Elastic Net is a compromise between Ridge and Lasso Regressions, combining their regularization techniques. While it can handle some non-linearity, it still primarily assumes a linear relationship [20]. The HSER model outperforms Elastic Net by incorporating Gradient Boosting and Extra Trees, that are well suited for capturing complex and non-linear interactions in the data.

Fifth, RF is an ensemble of DT, which can capture non-linear relationships and interactions between features. However, it can suffer from excessive variance [21]. The HSER model mitigates this by combining predictions from Ridge Regression and Gradient Boosting, which reduces variance and improves generalization. Sixth, Gradient Boosting is effective at handling non-linear relationships and can model complex interactions. However, it can be prone to overfitting, especially with noisy data [22]. The HSER model benefits from the regularization properties of Ridge Regression and the robustness of Extra Trees,

which together help to control overfitting and enhance predictive accuracy. Seventh, AdaBoost is another boosting method that focuses on improving the performance of weak learners. While powerful, it can be sensitive to noisy data and outliers [23]. The HSER model combines the robustness of Ridge Regression, Gradient Boosting, and Extra Trees, which helps to manage noise and outliers more effectively, leading to better overall performance. Eighth, Bagging (Bootstrap Aggregating) reduces variance by averaging the predictions of multiple DT. However, it can still be limited by the individual performance of its base learners [24]. The HSER model enhances this by using a diverse set of base models (Ridge Regression, Gradient Boosting, and Extra Trees) and employing a meta-model for further refinement, which provides a more accurate and stable prediction. Last, Extra Trees are similar to random forests but with more randomness introduced in the splitting process, which can lead to lower variance but also potential underfitting [25]. The HSER model balances this by combining the strengths of Extra Trees with the more deterministic gradient boosting and the regularization provided by Ridge Regression, resulting in superior performance.

The HSER model outperforms each of these baseline models by effectively combining their strengths and mitigating their weaknesses through a hybrid ensemble approach. By integrating Ridge Regression (for regularization), Gradient Boosting (for capturing nonlinear relationships), and Extra Trees (for robustness) and refining these predictions with a meta-model, the HSER model achieves better predictive accuracy for PD progression. This comprehensive approach allows it to handle complex interactions, reduce overfitting, and improve generalization compared to the individual baseline models. Overall, the proposed HSER model provides a more integrated, accurate, and clinically relevant approach to predicting PD progression, leveraging diverse datasets and advanced ML techniques to deliver noteworthy enhancements over prevailing approaches.

D. Research Benefit

In comparison to the existing research, the researchers observe several research benefits as follows:

- Integration of Diverse Data Sources: the research combines clinical, protein, and peptide data into a single comprehensive feature-engineered dataset. This multi-dimensional approach offers a more holistic view of PD progression compared to previous studies [10, 11] which focus on neuroimaging and RNA-Seq data respectively.
- HSER Model: by designing a HSER model that combines Ridge Regression, Gradient Boosting, and Extra Trees, the researchers enhance predictive accuracy. This method outperforms the individual ML models used in previous studies [12, 14, 15], offering superior performance and robustness.
- Predictive Performance: the model's superior performance, as indicated by consistently high accuracy in predicting UPDRS scores, surpasses the results of previous studies [14, 15], which use RF and Adaptive Boosting, respectively. The hybrid approach ensures better generalization and reliability of predictions.
- Feature Engineering and Data Handling: the extensive feature engineering and meticulous data processing in the approach address the complexities and heterogeneity of the datasets more effectively than the protocols outlined in previous research [15].
- Clinical Relevance: by focusing on MDS-UPDRS scores, the research aligns thoroughly with clinical standards for evaluating PD progression. This relevance to clinical practice is more direct compared to the broader biomarker analysis [13] and the neuroimaging focus [10].
- Interpretability and Practical Application: the model not only achieves superior predictive accuracy but also offers enhanced interpretability by identifying significant features contributing to PD progression. This aspect of the research is particularly effective for designing tailored therapy, an area less emphasized in previous studies [11, 12].
- Versatility and Reproducibility: the comprehensive protocol developed for data characterization, manipulation, and analysis is adaptable and may be adapted for other neurodegenerative diseases, similar to the scope suggested in previous studies [14, 15]. Nonetheless, the approach provides more detailed steps for reproducibility and practical application.

IV. Conclusion

The researchers propose a HSER model for predicting the progression of PD using protein and peptide data measurements. By leveraging the MDS-UPDRS scores, the researchers integrate clinical, protein, and peptide data into a comprehensive feature-engineered dataset. The hybrid approach, which combines Stacking and Blending techniques with Ridge Regression, Gradient Boosting, and Extra Trees as base models, demonstrates superior performance compared to baseline models, comprising Linear Regression, Ridge

Regression, Lasso Regression, Elastic Net, RF, Gradient Boosting, AdaBoost, Bagging, and Extra Trees. The HSER model's capability to precisely estimate the four UPDRS scores (updrs_1, updrs_2, updrs_3, and updrs_4) underscores its potential to enhance PD progression prediction, providing valuable insights for personalized treatment strategies.

The findings have significant implications for both clinical practice and future research. By demonstrating the effectiveness of the HSER model, the researchers pave the way for its application in clinical settings, where accurate prediction of PD progression can inform personalized treatment strategies and improve patient outcomes. Moreover, the integration of diverse data types illustrates the importance of a multimodal approach in understanding complex diseases like PD, encouraging further exploration of such methodologies in other neurodegenerative disorders.

However, there are several research limitations. First, the reliance on a specific dataset may limit the generalizability of the HSER model across diverse populations and settings. Second, the complexity of the model, while advantageous in terms of performance, may pose challenges in interpretability, potentially hindering its clinical adoption. Additionally, the research primarily focuses on protein and peptide data, which, although valuable, may not encompass the full spectrum of biomarkers relevant to PD progression. Future research will focus on refining the model weights to further increase the performance of the HSER model. Additionally, the researchers aim to explore the inclusion of more biomarkers, such as genetic and imaging data, to provide a more comprehensive understanding of PD progression. By encompassing these surplus information resources, the researchers hope to improve the analytical precision of the model. Further, the researchers will investigate the application of more sophisticated ML methods, such as DL and Transfer Learning, to capture even more complex patterns in the data. Finally, longitudinal studies will also be conducted to validate the model's predictive capabilities over time, ensuring its robustness and applicability in clinical settings.

Author Contribution

Conceived and designed the analysis, K. S. A.; Collected the data, K. S. A., M. M., and K. D.; Contributed data or analysis tools, K. S. A. and M. M.; Performed the analysis, K. S. A. and K. D.; Wrote the paper, K. S. A. and M. M.; and Reviewed the paper, M. M. and K. D.

Data Availability

Due to the nature of the research [ethical/legal/commercial], supporting data are not available.

References

[1] K. Kurihara, R. Nakagawa, M. Ishido, Y. Yoshinaga, J. Watanabe, Y. Hayashi, T. Mishima, S. Fujioka, and Y. Tsuboi, "Impact of motor and nonmotor symptoms in Parkinson disease for the quality of life: The Japanese Quality-of-Life Survey of Parkinson Disease (JAQPAD) study," *Journal of the Neurological Sciences*, vol. 419, pp. 1–6, 2020.

[2] J. G. Goldman, D. Volpe, T. D. Ellis, M. A. Hirsch, J. Johnson, J. Wood, A. Aragon, R. Biundo, A. Di Rocco, G. S. Kasman *et al.*, "Delivering multidisciplinary rehabilitation care in Parkinson's disease: An international consensus statement," *Journal of Parkinson's disease*, vol. 14, no. 1, pp. 135–166, 2024.

[3] C. G. Goetz, "Unified Parkinson's Disease Rating Scale (UPDRS) and the Movement-Disorder Society Sponsored-unified Parkinson's Disease Rating Scale (MDS-UPDRS)," *Encyclopedia of Movement Disorders*, pp. 307–309, 2010.

[4] R. Z. U. Rehman, L. Rochester, A. J. Yarnall, and S. Del Din, "Predicting the progression of Parkinson's disease MDS-UPDRS-III motor severity score from gait data using deep learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, Nov. 1–5, 2021, pp. 249–252.

[5] A. Ahmad, M. Imran, and H. Ahsan, "Biomarkers as biomedical bioindicators: Approaches and techniques for the detection, analysis, and validation of novel biomarkers of diseases," *Pharmaceutics*, vol. 15, no. 6, pp. 1–36, 2023.

[6] W. F. Zeng, X. X. Zhou, S. Willems, C. Ammar, M. Wahle, I. Bludau, E. Voytik, M. T. Strauss, and M. Mann, "AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics," *Nature Communications*, vol. 13, no. 1, pp. 1–14, 2022.

[7] M. Junaid, S. Ali, F. Eid, S. El-Sappagh, and T. Abuhmed, "Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 234, 2023.

[8] M. Martínez-García and E. Hernández-Lemus, "Data integration challenges for machine learning

in precision medicine," *Frontiers in Medicine*, vol. 8, pp. 1–21, 2022.

[9] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," *Healthcare*, vol. 11, no. 12, pp. 1–21, 2023.

[10] J. Hathaliya, H. Modi, R. Gupta, S. Tanwar, F. Alqahtani, M. Elghatwary, B.-C. Neagu, and M. S. Raboaca, "Stacked model-based classification of Parkinson's disease patients using imaging biomarker data," *Biosensors*, vol. 12, no. 8, pp. 1–17, 2022.

[11] S. Ahmed, M. Komeili, and J. Park, "Predictive modelling of Parkinson's disease progression based on RNA-Sequence with densely connected deep recurrent neural networks," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[12] H. Byeon, "Development of a stacking-based ensemble machine learning for detection of depression in Parkinson's disease: Preliminary research," *Biology and Life Sciences Forum*, vol. 9, no. 1, pp. 1–7, 2021.

[13] K. Gupta, T. Lamba, and N. Garg, "Predicting Parkinson's disease risk through protein and peptide level analysis: An evidence from EDA and machine learning based approach," in *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. Greater Noida, India: IEEE, Nov. 3–4, 2023, pp. 662–668.

[14] S. Bharathidason and C. Sujdha, "Prediction of Parkinson's disease using a stack ensemble modelling," *Tuijin Jishu/Journal of Propulsion Technology*, vol. 45, no. 2, pp. 509–517, 2024.

[15] I. D. Dinov, B. Heavner, M. Tang, G. Glusman, K. Chard, M. Darcy, R. Madduri, J. Pa, C. Spino, C. Kesselman *et al.*, "Predictive big data analytics: A study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations," *PLOS ONE*, vol. 11, no. 8, pp. 1–28, 2016.

[16] FNIH, "AMP Parkinson's disease." [Online]. Available: https://fnih.org/our-programs/accelerating-medicines-partnership-amp/amp-parkinsons-disease/

[17] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, 2021.

[18] S. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman, D. Awwad, and Z. Abdalla, "Optimizing linear regression models with Lasso and Ridge regression: A study on uae financial behavior during COVID-19," *Migration Letters*, vol. 20, no. 6, pp. 139–153, 2023.

[19] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, "A critical review of Lasso and its derivatives for variable selection under dependence among covariates," *International Statistical Review*, vol. 90, no. 1, pp. 118–145, 2022.

[20] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, no. 2, pp. 201–230, 2009.

[21] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[22] Y. Park and J. C. Ho, "Tackling overfitting in boosting for noisy healthcare data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 7, pp. 2995–3006, 2019.

[23] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, "Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences," *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1–25, 2020.

[24] P. Mahajan, S. Uddin, F. Hajati, M. A. Moni, and E. Gide, "A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets," *Health and Technology*, vol. 14, no. 3, pp. 597–613, 2024.

[25] M. Arya, H. Sastry G, A. Motwani, S. Kumar, and A. Zaguia, "A novel Extra Tree Ensemble Optimized DL Framework (ETEODL) for early detection of diabetes," *Frontiers in Public Health*, vol. 9, pp. 1–13, 2022.

APPENDIX

The Appendix can be seen in the next page.

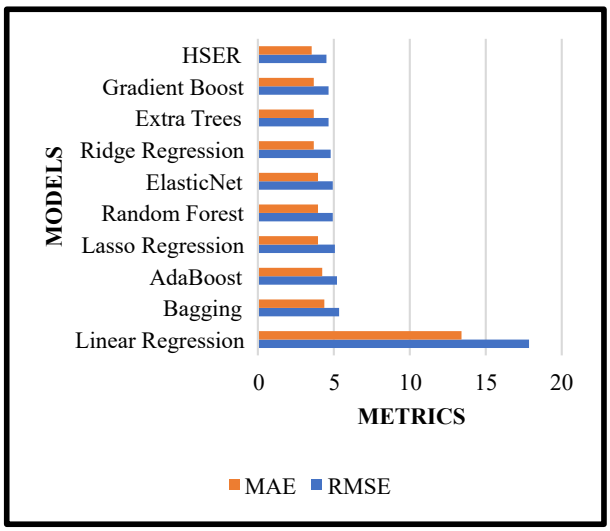

Fig. A1. Unified Parkinson’s Disease Rating Scale (UPDRS)-1 prediction results.

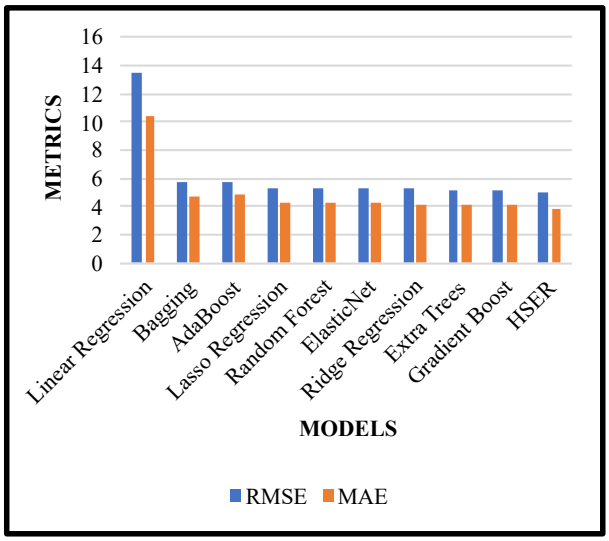

Fig. A2. Unified Parkinson’s Disease Rating Scale (UPDRS)-2 prediction results.

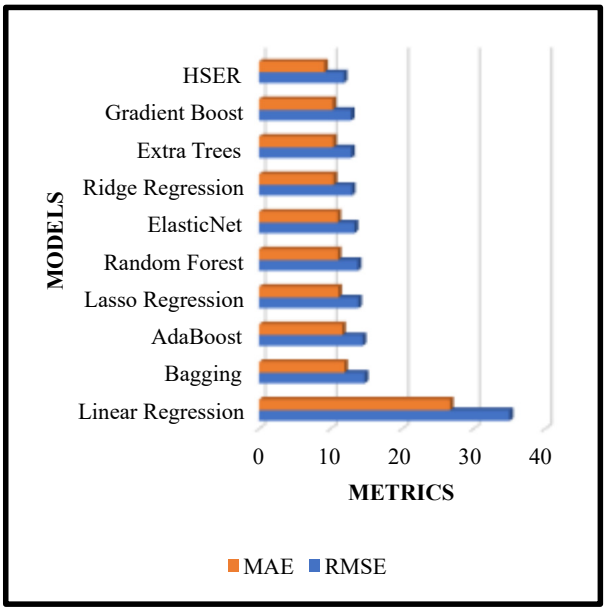

Fig. A3. Unified Parkinson’s Disease Rating Scale (UPDRS)-3 prediction results.

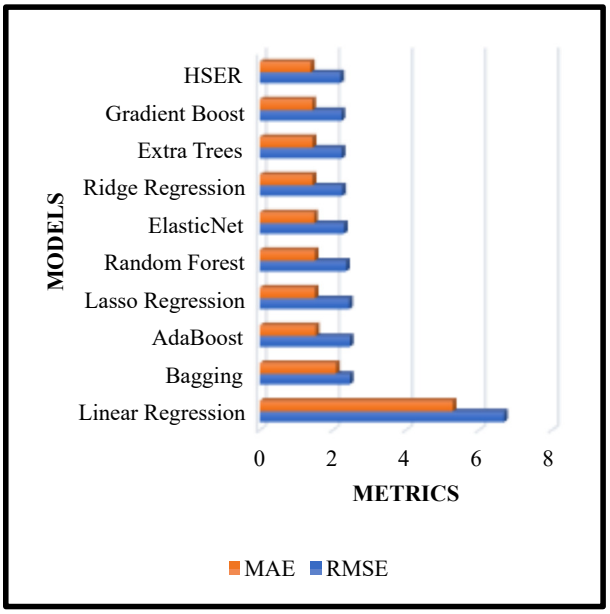

Fig. A4. Unified Parkinson’s Disease Rating Scale (UPDRS)-4 prediction results.