Parkinson's Disease Severity Assessment on the Unified Parkinson's Disease Rating Scale Using Text Embeddings Generated from Speech

Christina Koumpogianni and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
email: {ckoumpo, costas}@csd.auth.gr

***Abstract*—Parkinson's disease (PD) is a movement disorder of the central nervous system that affects patients' mobility over time. PD can be assessed using the Unified Parkinson's Disease Rating Scale (UPDRS), a 50-item, 4-part assessment used by clinicians and researchers to evaluate the severity and progression of Parkinson's disease. It measures motor, non-motor, cognitive, and activities-of-daily-living functions, with higher scores indicating greater impairment; it is commonly used to guide treatment decisions. The paper aims to predict the UPDRS using a complete pipeline that integrates a hybrid audio-to-text approach, in which speech signals are processed with deep learning models for transcription and embedding generation, followed by a Support Vector Regression model for prediction. The experimental results show that this hybrid model is a sufficiently accurate tool for UPDRS prediction.**

Keywords-Parkinson's disease; Natural language processing; Transformer-based embeddings; Self-supervised learning; Regression.

I. Introduction

Parkinson's Disease (PD) is a neurological disorder caused by a combination of genetic and environmental factors. Although no definitive cure currently exists, available treatments focus on symptom management and are primarily effective in the early stages of the disease. A significant proportion of individuals with PD experience dysarthria, with reported prevalence ranging from 70% to 90% [1].

In this paper, patients with PD who exhibit dysarthria are assigned their corresponding Unified Parkinson's Disease Rating Scale (UPDRS) scores, and the objective is to predict these scores as accurately as possible using deep learning–based audio-to-text embeddings combined with a machine-learning regression model. The main contributions are the design of an end-to-end pipeline that converts pathological speech into semantic representations using a pre-trained speech recognition and transcription model, followed by deep learning-based sentence-embedding extraction. These embeddings are then used as input to the regression model to predict UPDRS scores. This approach moves beyond traditional handcrafted acoustic features by capturing higher-level linguistic and contextual information, while enabling a more robust evaluation of the relationship between speech impairments and clinical severity in PD. Recent studies have addressed the same regression task primarily using Support Vector Regression (SVR), audio-to-text embeddings, and handcrafted feature vectors [2].

A. Related Work

Speech analysis has been a topic of interest to many scientists and clinicians. The paper aims to automate PD screening using text embeddings generated from speech. A large percentage of patients with PD experience dysarthria. Prior studies have explored both acoustic and linguistic representations of speech for PD detection and severity estimation. Several studies have addressed this task by extracting acoustic features from voice signals. Speech features are studied that are related to how sounds are produced and articulated in both sustained vowels and continuous speech [2]. These features can reliably reflect speech-motor impairments associated with PD. In a similar line of research, non-negative matrix factorization [3] is used to extract time–frequency features from voice signals, thereby capturing patterns associated with vocal degradation in PD. These manually designed acoustic features are typically used with traditional machine learning models, such as Support Vector Machines (SVMs) or regression methods, to estimate the presence or severity of the disease.

However, most of these studies focus on PD detection rather than fine-grained severity prediction. The prediction of the modified Frenchay dysarthria level assessment (m-FDA) was addressed in [4], [5]. The Hilbert Cepstral Coefficients were used as input features in three regression models: Radial Basis Function (RBF) SVR, Random Forest Regressor [6], and Multilayer Perceptron Regressor [7]. Performance was evaluated by computing the Spearman correlation between the predicted and ground-truth m-FDA scores. This paper focuses on transcribing audio into text using a transformer [8]-based pretrained model, the Whisper [9], which is a large-scale pretrained model trained on multilingual and multitask data, demonstrating strong robustness across diverse speech conditions. Subsequently, textual representations are encoded with a transformer-based sentence-embedding model (e.g., mixedbread-ai/mxbai-embed-large-v1 [10]), which maps transcripts to high-dimensional semantic vectors. The final prediction in UPDRS is performed using SVR.

The remainder of the paper is organized as follows. Section II describes the dataset, data preparation, audio-to-text transcription, text embedding, regression, and UPDRS prediction. Section III presents the experimental results. Section IV concludes the paper and outlines directions for future research.

II. Methodology

A synopsis of the proposed pipeline is illustrated in Figure 1. The PC-GITA speech dataset [11] was loaded, and its metadata were extracted for each .wav recording. The clinical metadata from the provided Excel spreadsheet was merged with

the audio index using speaker identifiers, retaining only PD patients with complete information. Audio preprocessing included mono conversion, resampling to 16 kHz for compatibility with the Whisper model's input, amplitude normalization, and segmentation into overlapping 30-second chunks with a 5-second overlap. Speech transcriptions were generated in Spanish using the Whisper model [9]. Sentence-level semantic embeddings were subsequently extracted from the transcriptions using the `mixedbread-ai/mxbai-embed-large-v1` transformer model [10]. For each speaker, utterance embeddings were aggregated using the median representation and ℓ_2-normalized to obtain a robust speaker-level embedding vector. The PD severity prediction was performed using SVR within a pipeline consisting of standardization, Principal Component Analysis (PCA), and hyperparameter optimization via grid search. Model evaluation was performed using Leave-One-Out Cross-Validation (LOOCV), and performance was assessed using the Root Mean Squared Error (RMSE).

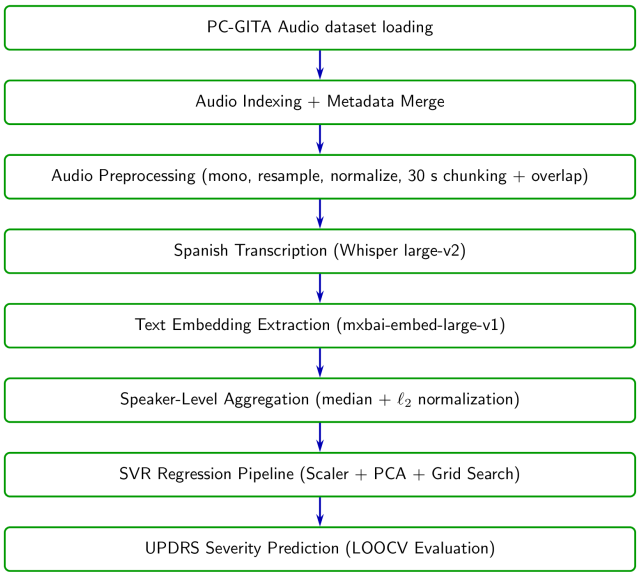

Figure 1. Proposed pipeline.

A. Dataset

Experiments have been conducted using the PC-GITA dataset [11], a Spanish-language balanced speech dataset used in PD research. It includes 50 speakers with PD and 50 healthy control (HC) speakers, comprising a total of 6405 recordings collected under controlled recording conditions. Specifically, the dataset contains 3214 raw recordings from speakers with PD and 3200 raw recordings from HC speakers. Each participant performed a standardized set of speech tasks designed to elicit a wide range of vocal behavior. Each participant's recording includes the following characteristics: unique ID, UPDRS_total, UPDRS_speech, Hoehn and Yahr scale, gender, age, and Time since diagnosis. The Hoehn and Yahr scale is a clinical staging scale used to describe the severity of PD. In the PC GITA dataset, this score is included as a clinical label for each patient, indicating disease stage. It describes the five stages of PD's symptoms, i.e., Stage 1: Symptoms on one side of the body only; Stage 2: Symptoms on both sides with no balance problems; Stage 3: Balance impairment but physically independent; Stage 4: Severe disability with the ability to walk or stand with assistance; and Stage 5: Wheelchair-bound or bedridden unless assisted [12]. The Time since diagnosis refers to the number of years since each participant was diagnosed with PD. UPDRS_total refers to the overall UPDRS score that clinicians have assigned based on a full clinical assessment of disease severity. UPDRS_speech is a sub-score within the UPDRS scale that focuses only on speech impairment. This sub-score typically ranges on a much smaller scale (e.g., 0–4), reflecting the severity of speech impairment due to PD. In the paper, only the UPDRS_total is used, which ranges from 6 (mild motor impairment) to 93 (severe disease progression). The PC-GITA dataset comprises a variety of speech tasks, including sustained vowel phonation, diadochokinetic (DDK) tasks, spontaneous monologue, and text reading, which together capture different aspects of dysarthria. Importantly, PD speakers are annotated with UPDRS scores, allowing for conducting a regression task. That is, to make a reasonable prediction of the UPDRS score using an automated method for assessing disease severity in patients. Since multiple recordings are available for each speaker, their corresponding embeddings are aggregated into a single representative feature vector by taking the median across all of these embeddings. However, the dataset remains relatively small and balanced, which will limit generalization.

B. Data Preprocessing

The PC-GITA dataset is first indexed by recursively scanning the directory structure and collecting all available 6405 audio files in Waveform Audio File Format (`.wav`) along with their corresponding task labels, speaker identifiers, and file paths, to create a structured `CSV` catalog of all `.wav` files. Speaker-level pathological condition labels (PD or HC) are inferred from folder and file naming conventions and later validated using the provided metadata. Clinical information, including total UPDRS score, Hoehn and Yahr scale, age, gender, and time since diagnosis, is extracted from the original PC-GITA metadata and merged with the audio file index using speaker identifiers. After merging with speaker-level metadata and filtering for complete information (UPDRS, age, gender, Hoehn-Yahr scale, and time since diagnosis), only 50 PD speakers retained all characteristics without any missing data, whereas all HC speakers were excluded. The `.wav` files associated with the 50 PD speakers total 1099. Samples with missing or inconsistent clinical information were excluded to maintain data quality. This accounts for the reduction in size in the PD and HC subsets. As shown in Table I, the dataset is indexed by task, pathological condition, and speaker. Each entry in the final dataset corresponds to a single speech recording produced by a specific speaker during a given task. For example, in the DDK analysis task, recordings consist of rapid syllable repetitions (e.g., "ka-ka-ka"), which are

TABLE I. EXAMPLE OF DATA STRUCTURE AFTER METADATE MERGING.

Task	Label	Speaker_ID	Relative Path	Full Path	UPDRS_total	UPDRS_Speech	Hoehn-Yahr scale	Gender	Age	Time since diagnosis
DDK analysis	PD	AVPEPUDEA0053	DDK analysis/ka-ka-ka/...	/content/drive/...	33.0	2.0	2.0	M	47.0	2.0
DDK analysis	PD	AVPEPUDEA0052	DDK analysis/ka-ka-ka/...	/content/drive...	23.0	1.0	2.0	F	70.0	12.0
DDK analysis	PD	AVPEPUDEA0042	DDK analysis/ka-ka-ka/...	/content/drive...	54.0	1.0	3.0	F	65.0	8.0
DDK analysis	PD	AVPEPUDEA0023	DDK analysis/ka-ka-ka/...	/content/drive...	14.0	0.0	1.0	M	68.0	1.0
DDK analysis	PD	AVPEPUDEA0047	DDK analysis/ka-ka-ka/...	/content/drive...	40.0	1.0	2.0	F	64.0	3.0

known to reflect motor control and articulatory precision. This structure allows multiple recordings from the same speaker to be linked to a single clinical profile, facilitating speaker-level aggregation. The UPDRS scores in the dataset range from 6 (mild motor impairment) to 93 (severe disease progression). This broad distribution of scores supports regression analysis rather than coarse disease classification.

C. Audio-to-Text Transcription

First, `Whisper-large-v2` [9], a pre-trained transcription model, is loaded. `Whisper` is a pre-trained model for Automatic Speech Recognition (ASR) and speech translation/ transcription. Trained on 680k hours of labeled data, Whisper models demonstrate a strong ability to generalize to many datasets without the need for fine-tuning. The updated large-v2 version was chosen for its improved recognition performance and greater model capacity. The audio `.wav` recordings are loaded using the `torchaudio` [13] library and, when necessary, converted to a single channel by averaging across channels. To be compatible with the transcription model, which is designed to operate on 16 kHz audio, all signals are resampled to 16 kHz. The waveform amplitudes are then normalized to reduce variability due to recording conditions. The preprocessed audio waveform is first converted into a one-dimensional array by removing the channel dimension, yielding a continuous sequence of samples representing the entire recording. To facilitate transcription by the Whisper model, the waveform is then segmented into fixed-length chunks. Each chunk corresponds to a specified duration of 30 seconds, which is converted to the equivalent number of samples based on the 16 kHz sampling rate (30×16000 samples). However, after this process, it must be ensured that no recordings are lost or truncated due to 30-second chunking, which is why an overlapping strategy was used. Consecutive chunks are designed to partially overlap for 5 seconds, ensuring that no speech information is lost. Specifically, the starting index of each new chunk is advanced by the chunk length minus the overlap, producing overlapping segments of the original waveform. Each segment is stored as a separate one-dimensional array, forming a list of chunks that collectively represent the full recording while preserving continuity and context for subsequent transcription. The transcription of audio recordings begins by loading metadata containing the file paths of all recordings into a `pandas DataFrame`. After each audio file is segmented into overlapping chunks, as described previously, the Whisper model performs automatic ASR on each chunk to produce a Spanish transcript. Non-empty transcripts from all chunks of a single recording are concatenated to generate a complete transcript for that file. This procedure is applied iteratively to all recordings. The resulting transcripts are stored in a new column of the `DataFrame` and saved to a `CSV` file, ensuring that the full audio content is preserved and ready for text embedding generation.

D. Text Embedding Generation and Speaker-Level Feature Aggregation

Text embeddings were generated using the pre-trained model `mixedbread-ai/mxbai-embed-large-v1` [10] of size 1024 based on a Transformer encoder architecture for generating contextual sentence embeddings. Each transcript was first checked to ensure that it was a non-empty string of sufficient length, and a contextual prefix: "Parkinson speech severity prediction", was added to guide the embedding model toward the clinical prediction task. Analytically, each token in the transcript interacts with the prefix tokens during self-attention, altering the hidden representations throughout the network. This results in the final 1024-dimensional embedding being slightly biased toward speech patterns associated with severity, such as slurring, pauses, or monotone delivery. Although all recordings are Parkinson-related, the prefix helps highlight subtle variations that are most informative for predicting UPDRS scores, producing embeddings that are more discriminative for downstream regression modeling. The cleaned transcripts were then encoded using this pre-trained model in batches of 16, producing 1024-dimensional dense vectors that capture semantic and contextual information, enabling the integration of linguistic information into machine learning models to predict clinical scores. After generating embeddings for each transcript, speaker-level feature vectors were constructed to represent an individual's recordings. The dataset was filtered to include only PD patients. For each speaker, all their transcript embeddings were retrieved and aggregated by taking the median across embeddings, yielding a robust representation that reduces the influence of outlier segments. The aggregated vector was then ℓ_2-normalized, ensuring the resulting feature vector had unit length, thereby standardizing the input scale for downstream models. Each speaker-level embedding was stored in a feature matrix $\boldsymbol{X}$, and the corresponding UPDRS score was stored in the target vector $\boldsymbol{y}$. This process resulted in a structured dataset of speaker-level embeddings and labels suitable for regression analysis, with $\boldsymbol{X}$ having shape (50×1024) and $\boldsymbol{y}$ of size 50 containing the associated UPDRS scores. Let $t_{i,j}$ denote the jth transcript of speaker i. Each transcript is augmented by appending the task-specific prefix p:

$$\tilde{t}_{i,j} = [p; t_{i,j}]. \quad (1)$$

The embedding model [10] $f_\theta(\cdot)$ converts each input sequence to a 1024-dimensional vector:

$$\mathbf{e}_{i,j} = f_\theta(\tilde{t}_{i,j}) \in \mathbb{R}^{1024}. \quad (2)$$

For each speaker, the embeddings are aggregated component-wise using the median:

$$\mathbf{s}_i = \text{median}\left(\{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \ldots, \mathbf{e}_{i,n_i}\}\right). \quad (3)$$

The resulting speaker representation is then ℓ_2-normalized:

$$\hat{\mathbf{s}}_i = \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|_2} \quad (4)$$

and the final dataset is defined as:

$$\mathcal{D} = \{(\hat{\mathbf{s}}_i, y_i)\}_{i=1}^{N}, \quad N = 50. \quad (5)$$

E. Regression and UPDRS Prediction

To predict the UPDRS units from speaker-level embeddings, SVR was applied within a leave-one-out cross-validation (LOOCV) framework combined with hyperparameter tuning (`GridSearchCV`). In each fold, one speaker from the 50 PD speakers is held out for testing, while the remaining speakers are used for training, ensuring the model does not see the test embeddings during training. The pipeline for this task included feature standardization via `StandardScaler()` [14] to prevent scale issues in SVR, optional dimensionality reduction via PCA [15], and SVR. Hyperparameters for both PCA and SVR were optimized using a 5-fold grid search on the training set. For PCA, the number of components tested was 5, 10, 20, and 40. For SVR, the grid search performed for choosing the kernel type $\in$ {linear, RBF, sigmoid, and polynomial}, the regularization parameter `C` $\in \{0.1, 1, 10, 100, 200\}$, and the scale parameter `gamma` $\in \{10^{-3}, 10^{-2}, 10^{-4}\}$. For the sigmoid kernel, `coef0` values $\in \{0, 0.1, 1\}$ were tested. For the polynomial kernel, degrees 2 and 3 were evaluated using `scikit-learn`. Precisely, inside each LOO iteration, the training set (49 speakers) is further split into 5 folds. The kernels used for this grid search were linear, RBF, and sigmoid. The grid search result corresponding to the best model was used to predict the UPDRS score for the held-out speaker, and the RMSE was recorded. This procedure was repeated for each speaker, resulting in a vector of RMSE scores. The mean RMSE is computed as well as its standard error (SE) defined as $\frac{std}{\sqrt{N}}$ where std denotes the standard deviation and N is the sample size (i.e., 50). The SE provides confidence in the estimated mean RMSE, i.e., how accurate the estimate is.

III. RESULTS

Table II provides insight into SVR for the linear and the RBF kernels. The average RMSE across all LOOCV folds is expressed in raw UPDRS units. The sigmoid kernel is excluded from Table II, because it was not selected in any fold. Among the evaluated kernels, the RBF kernel achieved the lowest mean RMSE of 14.94 with hyperparameters PCA $n_{\text{components}} = 5$, $C = 100$, and $\gamma = 0.001$, indicating better average predictive performance than the linear kernel, which yielded a mean RMSE of 24.837. The total RMSE was calculated across all folds and incorporated the predictions from all selected kernels. The total RMSE among all folds was 15.71, the MAE was 11.04, and the R^2 was 0.25. This result suggests that the relationship between the embeddings and UPDRS scores is probably nonlinear, which explains why the RBF kernel performs better.

TABLE II. PERFORMANCE COMPARISON OF SVR KERNELS EXCLUDING THE POLYNOMIAL.

Kernel	Mean RMSE	SE
linear	24.837	10.962
RBF	14.94	1.4851

The lowest mean RMSE was achieved with the RBF kernel (i.e., 14.94), compared to the linear kernel (i.e., 24.837), highlighting its potential to capture complex nonlinear relationships in certain cases. The SE of the RBF (1.4851) is lower than that of the linear kernel (10.962), indicating that the RBF kernel exhibits more stable behavior and reflects its ability to model complex patterns in the embedding space. When the polynomial kernel is excluded from the grid search, as demonstrated in Table II, the overall performance improves, with the RBF kernel achieving the lowest mean RMSE (i.e., 14.94).

When the polynomial kernel was frequently selected across folds during the grid search in inner cross-validation, as shown in Table III, it resulted in worse generalization performance. The polynomial kernel was selected 39 times during the grid search. The mean RMSE achieved was 17.7 from the polynomial kernel only, which is higher than the best mean RMSE of 14.94 from the RBF kernel in Table II. The total RMSE was also higher, at 16.752, with total SE of 1.623 and R^2 = 0.146, indicating that it does not capture the linguistic characteristics well. Also, the linear kernel was not selected in this case, not even once. These RMSE values differ because the first refers to the RMSE obtained using only the polynomial kernel. In contrast, the total RMSE/SE is calculated across all folds and incorporates predictions from all selected kernels. Polynomial kernels are highly flexible and can easily overfit small, high-dimensional datasets, especially in clinical datasets with a limited number of speakers. While the polynomial kernel may fit the training folds very closely, this increased complexity often leads to poorer performance on the unseen data. The exclusion of the polynomial kernel reduces the risk of overfitting by limiting the model's flexibility, thereby allowing the RBF kernel to better capture nonlinear relationships in the data.

However, the RBF kernel still exhibits variability, as indicated by its relatively high SE of 1.91, suggesting that performance remains sensitive to the specific training folds. However, the SE of the polynomial kernel was higher, i.e., 2.80, as demonstrated in Table III. A larger standard error indicates greater variability in model performance across cross-validation folds and, therefore, lower confidence in the estimated mean performance. Overall, these results indicate that the RBF kernel yields a smoother nonlinear decision function that captures complex relationships without overfitting noise or speaker-specific variations.

TABLE III. KERNEL PERFORMANCE COMPARISON.

Kernel	Mean RMSE	MAE	SE
polynomial	17.7	13.15	2.80
RBF	12.8558	9.3	1.91

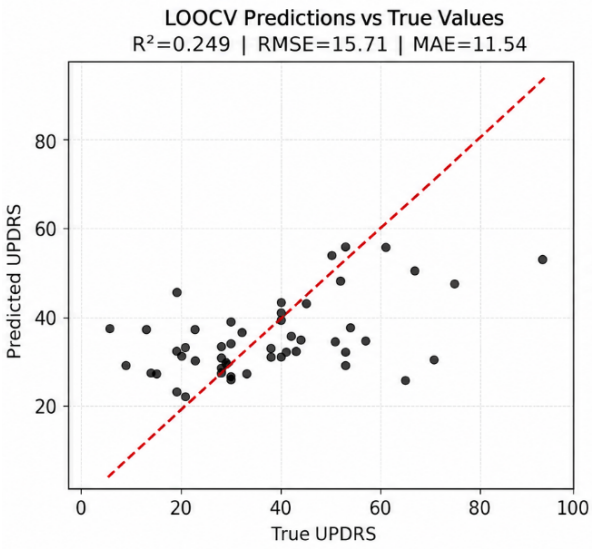

Figure 2. Predicted vs True values of the target variable across all LOOCV folds.

Figure 2 shows predicted versus true target values across LOOCV folds from the grid search procedure, with results summarized in Table II. The points are colored by the SVR kernel used (linear or RBF). The black dashed line represents perfect prediction. Points close to the line indicate accurate predictions, while deviations show errors. The majority of predictions, particularly for the RBF kernel, cluster around this diagonal line, but noticeable deviations indicate prediction errors. Notably, only a single point corresponds to the linear kernel. This occurs because model selection is performed independently within each fold via a grid search, and the RBF kernel outperformed the linear kernel in all other folds. Also, the plot shows that the model underestimates higher UPDRS scores and overestimates lower ones, as many points are compressed toward the center of the range. This effect is more pronounced with the RBF kernel, which, despite achieving a lower average RMSE, exhibits significant variability and scattered predictions across the range. In contrast, the linear kernel produces fewer predictions, with points clustering in a more consistent region, albeit with less flexibility in capturing extreme values. Overall, the plot confirms that while the RBF kernel can model nonlinear relationships, it is somewhat unstable, whereas the linear kernel provides more consistent but less expressive predictions.

Figure 3 shows the per-fold RMSE plot illustrating the model's performance variation across speakers under LOOCV. Each point represents the prediction error for a single test speaker across folds, while the dashed horizontal line indicates the mean RMSE of 15.71. The coefficient of determination R^2 was 0.25 when the best model was selected independently in

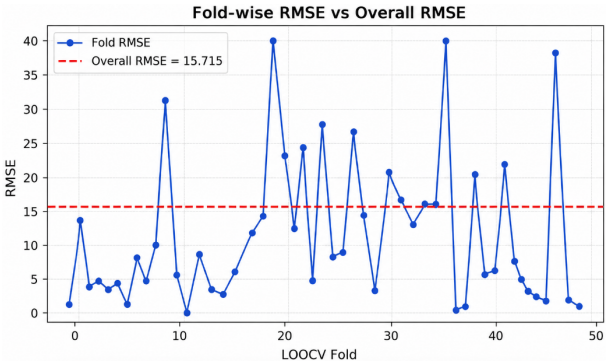

Figure 3. RMSE per fold corresponding to a speaker left out in cross-validation.

TABLE IV. MODEL PERFORMANCE WITH AND WITHOUT THE USE OF TEXTUAL PREFIX.

Setup	Dominant Kernel	RMSE
With prefix	RBF	15.71
Without prefix	RBF (29 times selected)	14.8628
Without prefix	linear (17 times selected)	18.4772
Without prefix	sigmoid (4 times selected)	24.6858

each fold, with a SE of 1.52, indicating moderate predictive accuracy and variability across train–test splits. The relatively low R^2 highlights the complexity of PD severity prediction and the influence of clinical and behavioral factors not fully captured by speech transcriptions. As can be seen, there is considerable variation in error across folds, with RMSE values ranging from very low to above 40. This suggests that the model performs quite accurately for some speakers but struggles with others, resulting in noticeably higher error rates. This variability indicates a strong dependence on speaker-specific characteristics, such as speech patterns, recording conditions, or disease severity, and highlights limited generalization across samples.

The textual prefix ("Parkinson speech severity prediction:") was introduced to guide the embedding model toward capturing clinically relevant semantic information. To verify its contribution to the results, two implementations were conducted: one using the prefix and one without it. The results presented in Table IV show that without the prefix, during the nested cross-validation procedure, all three kernel functions (linear, RBF, and sigmoid) were selected as the optimal SVR kernel in at least one fold and achieved an RMSE of 17.115, R^2 = 0.11 indicating no good generalization, whereas with the prefix, a non-linear RBF kernel was selected in most folds, leading to a lower total RMSE of 15.71. This indicates that the prefix is useful in this experiment because it alters the structure of the embedding space, enabling the model to capture more complex, non-linear relationships between speech-derived features and PD severity. These findings support the inclusion of the prefix as a simple yet effective strategy for improving representation quality and improving prediction.

The mean RMSE of the proposed method is compared with the mean RMSE disclosed in [16], where linguistic embeddings combined with SVR achieved mean RMSE values

between 13.5 for Bidirectional Encoder Representations from Transformers (BERT) [17] and 14.6 for XLNet [18], with the best performance of (12.9 $\pm$ 1.68) observed using the *text-embedding-3-large* model [19], employing embeddings of 768 dimensions and the sigmoid kernel. In this proposed approach, the model achieves a mean RMSE of (15.71 $\pm$ 1.52) with a coefficient of determination R^2 = 0.25, indicating improved predictive accuracy. However, the results also show good variability, particularly for the linear kernel, indicating less stable performance across folds, likely due to the small dataset and the choice of kernel. Additionally, hyperparameter tuning is performed separately within each fold, which can lead to different model configurations and further increase variability.

IV. CONCLUSION AND FUTURE WORK

The experimental results suggest that audio recordings transcribed to text can provide useful information about PD severity. The pipeline used for predicting UPDRS scores in PD included audio transcription, transformer-based embeddings, and SVR. On average, the predictions were approximately 16 points off the UPDRS score, with a minimum of 6 and a maximum of 93. The results indicate that linguistic representations extracted from transcribed speech contain relevant information for estimating clinical severity. Although this model has not achieved the highest precision among the previously described models, the coefficient of determination indicates a good prediction potential. It is not sufficient on its own for clinical estimation due to a 16-unit offset relative to an accurate estimate. The comparison between SVR kernels showed that the RBF kernel achieved a lower average RMSE. However, this improvement came with increased variability, indicating reduced robustness across folds. The linear kernel didn't produce more stable but less accurate predictions, highlighting a trade-off between model flexibility and generalization. Based on the results of the LOOCV folds, a tendency toward regression to the mean is relieved, with higher UPDRS scores being underestimated and lower scores overestimated. This behavior, along with the variability observed in per-fold RMSE, suggests that the model struggles to capture extreme cases and that performance is sensitive to individual speaker characteristics.

Some limitations include the small dataset (50 PD speakers), the use of text-only features, the aggregation of embeddings across recordings from the same speaker (which may obscure variability between recordings), and prediction errors of about 12 UPDRS points, which are not precise enough for detailed clinical use. Despite these challenges, the proposed approach demonstrates the potential of combining pre-trained models for transcription and embedding generation in clinical prediction tasks. The use of a contextual prefix further enhances the relevance of the extracted features with minimal complexity. Future work should focus on expanding the dataset through data augmentation, combining text with acoustic embeddings, and exploring more robust modeling approaches to improve performance and generalization.

ACKNOWLEDGMENT

The authors would like to thank Professor Juan Rafael Orozco Arroyave for sharing the PC-GITA dataset with them.

REFERENCES

[1] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," *Research and Reviews in Parkinsonism*, vol. 9, pp. 9–16, 2019.

[2] Y. Liu *et al.*, "Automatic assessment of Parkinson's Disease using speech representations of phonation and articulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 242–255, 2022.

[3] B. Karana, S. S. Sahua, J. R. Orozco-Arroyave, and K. Mahtoa, "Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction," *Computer Speech & Language*, vol. 69, p. 101216, 2021.

[4] B. Karan and S. S. Sahu, "An improved framework for Parkinson's Disease prediction using variational mode decomposition-Hilbert spectrum of speech signal," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 717–732, 2021.

[5] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's Disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.

[6] Scikit-learn Developers, "Randomforestregressor," https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, 2026, [retrieved: May, 2026].

[7] ——, "Mlpregressor," https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html, 2026, [retrieved: May, 2026].

[8] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[9] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," 2022, [retrieved: May, 2026]. [Online]. Available: https://arxiv.org/abs/2212.04356

[10] S. Lee, A. Shakir, D. Koenig, and J. Lipp, "Open source strikes bread - new fluffy embeddings model," 2024, [retrieved: May, 2026]. [Online]. Available: https://www.mixedbread.ai/blog/mxbai-embed-large-v1

[11] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 342–347.

[12] American Parkinson Disease Association, "Stages in Parkinson's Disease," 2013, [retrieved: May, 2026]. [Online]. Available: https://www.apdaparkinson.org/article/stages-in-parkinsons

[13] Y.-Y. Yang *et al.*, "TorchAudio: Building blocks for audio and speech processing," 2021, [retrieved: May, 2026]. [Online]. Available: https://arxiv.org/abs/2110.15018

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[16] J. Crawford, "Linguistic changes in spontaneous speech for detecting parkinson's disease using large language models," 2024, [retrieved: May, 2026]. [Online]. Available: https://arxiv.org/abs/2404.05160

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, [retrieved: May, 2026]. [Online]. Available: https://arxiv.org/abs/1810.04805

[18] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," 2020, [retrieved: May, 2026]. [Online]. Available: https://arxiv.org/abs/1906.08237

[19] OpenAI, "OpenAI. new embedding models and api updates," 2020, [retrieved: May, 2026]. [Online]. Available: https://developers.openai.com/api/docs/guides/embeddings