RESEARCH ARTICLE

Linguistic changes in spontaneous speech for detecting Parkinson's disease using large language models

Jonathan L. Crawford *

Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, United States of America

* jcraw@bu.edu

Abstract

Parkinson's disease is the second most prevalent neurodegenerative disorder with over ten million active cases worldwide and one million new diagnoses per year. Detecting and subsequently diagnosing the disease is challenging because of symptom heterogeneity with respect to complexity, as well as the type and timing of phenotypic manifestations. Typically, language impairment can present in the prodromal phase and precede motor symptoms suggesting that a linguistic-based approach could serve as a diagnostic method for incipient Parkinson's disease. Additionally, improved linguistic models may enhance other approaches through fusion techniques. The field of large language models is advancing rapidly, presenting the opportunity to explore the use of these new models for detecting Parkinson's disease and to improve on current linguistic approaches with high-dimensional representations of linguistics. We evaluate the application of state-of-the-art large language models to detect Parkinson's disease automatically from spontaneous speech with up to 78% accuracy. We also demonstrate that large language models can be used to predict the severity of PD in a regression task. We further demonstrate that the better performance of large language models is due to their ability to extract more relevant linguistic features and not due to increased dimensionality of the feature space.

Author summary

Our work explores the use of large language models to detect Parkinson's disease from spontaneous speech. Parkinson's disease is the second most common neurodegenerative disorder, affecting over 10 million people worldwide. However, it is often diagnosed late due to the complex nature of symptoms and the limitations of existing detection methods. Language impairment can precede the physical symptoms currently used for diagnosis, which suggests speech analysis could allow for earlier detection. We evaluate several state-of-the-art large language models to automatically analyze transcripts of spontaneous speech from Parkinson's patients and healthy controls. The large language models extract linguistic features such as word usage, grammar, meaning, and context into high-dimensional representations. We demonstrate that the models can be used to detect Parkinson's

OPEN ACCESS

Citation: Crawford JL (2025) Linguistic changes in spontaneous speech for detecting Parkinson's disease using large language models. PLOS Digit Health 4(2): e0000757. https://doi.org/10.1371/journal.pdig.0000757

Editor: Hadi Shafiee, Brigham and Women's Hospital, Harvard Medical School, UNITED STATES OF AMERICA

Received: April 21, 2024

Accepted: January 21, 2025

Published: February 10, 2025

Data Availability Statement: The PC-GITA dataset is used in this study. A description of the dataset can be found at: https://aclanthology.org/L14-1549. The dataset consists of the spontaneous speech of 50 subjects with PD and 50 health controls for a total of 100 samples in the dataset. The authors have not received any special privileges in accessing the data that other researchers would not have. The dataset is available by request using the contact form at the following site: https://gita.udea.edu.co/contact.

Funding: The author(s) received no specific funding for this work.

Competing interests: The author has declared that no competing interests exist.

disease and outperform prior methods. This demonstrates the potential for artificial intelligence to aid in screening for Parkinson's disease based on subtle speech patterns. The work paves the way for further research into leveraging these powerful language models as affordable, non-invasive screening tools. Ultimately, the use of large language models could allow for earlier diagnosis and treatment intervention, improving quality of life for millions affected by Parkinson's disease.

Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder with over ten million active cases worldwide, one million new diagnoses per year, and an exponentially growing incidence rate [1]. The global prevalence of PD continues to rise due to increased life expectancy and industrialization [2]. PD is a chronic and progressive neurodegenerative disorder that induces physical and cognitive impairment. The pathogenesis and pathophysiology of the disease are poorly understood. Current research indicates that the pathogenesis of PD involves an interplay of unknown genetic susceptibilities and environmental exposures [3]. PD is characterized by the progressive loss of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies, leading to central nervous system degradations. The disease is also characterized by motor symptoms, namely bradykinesia, resting tremor, rigidity, and postural instability. Non-motor symptoms also manifest heterogeneously [4].

PD is clinically diagnosed late relative to the pathogenesis and with low accuracy [5,6]. This can be attributed to multiple confounding factors including the absence of early-stage biomarkers and screening methods, the complex symptomatology of the disease, and the limitations of diagnostic methods in timely detection and differentiation. The pathogenesis of the disease is estimated to begin decades before the manifestation of the phenotypic symptoms necessary for clinical diagnosis [7]. PD is currently diagnosed with clinical evaluations. The clinical evaluation utilizes phenotypic symptoms augmented with neuroimaging to exclude other conditions. The primary symptoms used for PD diagnosis are tremor, rigidity, bradykinesia, and postural instability. These fine motor-skill deteriorations are considered the cardinal and first observable signs of PD. Dependence on these symptoms is problematic given their variability, non-specificity, and potential overlap with other diseases [8]. Inconsistent symptom onset and presentations across populations further add to the complex symptomatology. Reliance on these physical symptoms leads to late detection because these symptoms do not present until around 80% neural degradation [9].

There is a need for additional biomarkers and new methods to detect PD. Language impairment can present in the prodromal phase and precede motor symptoms suggesting that a linguistic-based approach could serve as a diagnostic method for incipient PD [10]. Linguistics models may also be used to detect PD across all stages and enhance other approaches through fusion techniques.

The architecture, parameter structure, and training of the large language models allow the models to extract and encode into text embeddings a unique linguistic feature space representing the morphology, syntax, semantics, and pragmatics of the spontaneous speech signals [11]. The architecture of a large language model is defined by the transformers, attention mechanisms, and depth of layers, along with specific parameter configurations and the nature of its training data. The differing architecture of each model results in embeddings that capture distinct linguistic patterns. Specific dimensions in this space include the usage of syntactic structures, the frequency of certain words or phrases, the presence of semantic themes, the distribution of parts of speech, and the occurrence of named entities [12]. For PD detection, language impairments may align with dimensions in the feature space including semantic richness, grammatical usage, fluency patterns, and syntactic complexity. For instance, reduced information content may appear as low semantic richness, impaired grammaticality may be evident in deviations from standard grammar, disrupted fluency may be characterized by pauses and self-corrections, and reduced syntactic complexity may be reflected in simpler sentence structures. Mapping these deficits may allow the models to detect deviations from typical language patterns associated with PD.

For example, Bidirectional Encoder Representations from Transformers (BERT) has been used to detect PD [13]. The field of large language models is advancing rapidly, which presents the opportunity to explore the use of the new models for detecting PD and to improve on current linguistic approaches with high-dimensional representations of linguistics [11].

The purpose of this work is to demonstrate the potential of state-of-the-art large language models to improve the field of detection of PD using linguistics. This may eventually lead to a detection method for PD in the prodromal phase. We evaluate the application of state-of-the-art large language models to detect PD from spontaneous speech. The state-of-the-art large language models lead to improved performance over the prior methods using our implementation. We also demonstrate that large language models can be used to predict the severity of PD. We expect that our work will be built upon by us and future researchers to eventually develop a method to detect early onset PD in clinical applications.

Results

We report on comparing multiple methods to detect PD using state-of-the-art large language models. A high-level description of a system to detect PD is detailed in Fig 1. The system inputs digitized spontaneous speech from participants. The speech either belongs to a control group without PD or a group that has PD of varying degrees of severity. The speech is transcribed automatically with an automated speech recognition model. A large language model then generates a linguistic feature space from the transcription. A classification algorithm subsequently processes the feature space to make a PD/Non-PD diagnosis. Low-level details on each step are presented below. The computer code used to create the results shown in this paper is available upon request from the author.

Biomarker generation

Speech-to-text via automated speech recognition. The audio files are automatically transcribed into textual form using Whisper, an automated, multilingual speech recognition model from Open AI [14]. The translation endpoint is used to transcribe the audio into English.

Text embeddings via large language models. Text embeddings are generated from the transcriptions using large language models. We evaluate the efficacy of multiple state-of-the-art large language models. Each model generates a high dimensional linguistic feature space

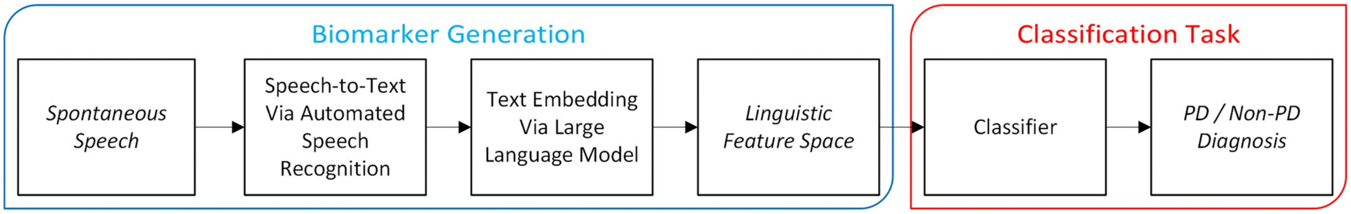

Fig 1. A high-level representation of the methodology. Italicized boxes signify inputs and outputs.

https://doi.org/10.1371/journal.pdig.0000757.g001 with each dimension representing different linguistic features determined by the architecture and training of the models. The models were chosen because they are widely applied and considered to be state-of-the-art as of the writing of this paper.

We present results from: Bidirectional Encoder Representations from Transformers (BERT) [15]; XLNet [16]; Generative Pre-trained Transformer 2 (GPT-2) [17]; text-embedding-ada-002 [18,19]; and text-embedding-3-small and text-embedding-3-large [20]. We note that BERT has been previously applied for PD detection [13]. We also implement Word2Vec [21], which is a word embedding technique, for comparison with prior linguistic methods.

The dimensionality of the text-embedding-3-small and text-embedding-3-large outputs can be reduced through an application programming interface (API) endpoint parameter. This represents a trade-off between performance and the cost of using embeddings. Specifically, embeddings are shortened internally by the model without losing their concept-representing properties [22]. We present results from the two models at both their default (maximum) dimensionality and at a reduced dimensionality, adjusted to match that of the other evaluated models. This approach enables a direct comparison of performance, standardized based on dimensionality, between these models and the other models evaluated.

Classification

The classification task presents challenges due to the high dimensionality of the feature spaces compared to the small size of the dataset. A support vector machine (SVM) is utilized for its robustness when handling such high-dimensional data and for its effectiveness in classification tasks through the creation of optimal hyperplanes in a transformed feature space. Specifically, SVM models are resistant to overfitting when regularized [23]. The embeddings are standardized to allow the SVM model to process the feature space [24]. All dimensions of the embeddings are used to train the SVM.

Assessing performance is challenging due to the limited dataset sizes. To address this, 10-fold stratified cross-validation is employed to tune hyperparameters and provide a stochastic estimate of performance. We follow the procedure of [25] to allow comparison of results with others in the field who also employ this standardized procedure. The dataset consists of 100 samples, so 90 are used as the training set and 10 are treated as an unseen and validation set in each fold. A grid search is used to optimize the hyperparameters based on accuracy. The hyperparameters considered are the kernel type, the regularization parameter (C), and the kernel coefficient gamma. The kernel types considered are Polynomial, Radial Basis, and Sigmoid functions, which are denoted in the results section below as poly, rbf, and sig, respectively. The regularization parameter ranges over the values $[10^{-5}, 10^{-4}, \ldots, 10^{5}]$, which helps in controlling the trade-off between achieving a low training error and a low testing error, thereby avoiding overfitting. The kernel coefficient, which influences the decision boundary, also spans across the same range of values: $[10^{-5}, 10^{-4}, \ldots, 10^{5}]$.

After each fold in the cross-validation process, the evaluation metrics are recorded, and their mean values are calculated upon completion of all folds. The evaluation metrics are accuracy, precision, recall, and the area under the curve (AUC). Positives are defined as the group with PD and negatives are defined as the healthy control group. Accuracy is defined as correct predictions over all predictions. Precision is defined as true positives over true positives plus false positives (positive predictions). Recall is defined as true positives over true positives plus false negatives (all positives). Area under the curve is defined as area under the receiver operating characteristic curve, where a perfect area has a value of one. Standard Error is defined as the standard deviation of the metric across folds divided by the square root of the number of samples, ten in this case.

Table 1. Results of the application of state-of-the-art large language models. Performance metrics and standard errors are reported as percentages, expressed as mean (μ) ± standard error (SE). AUC is reported as a fraction.

Model	Dimen-sion	Accuracy	Precision	Recall	AUC	Kernel	C	Gamma
Word2Vec	300	54 ± 4.0	58 ± 5.9	60 ± 6.7	0.44 ± 0.05	linear	10^{-3}	10^{-4}
BERT	768	66 ± 3.7	64 ± 3.0	80 ± 6.0	0.69 ± 0.06	linear	10^{-4}	10^{-4}
XLNet	768	61 ± 4.1	65 ± 6.9	56 ± 7.7	0.61 ± 0.03	linear	10^{-4}	10^{-4}
GPT-2	768	70 ± 4.9	73 ± 5.0	72 ± 8.0	0.76 ± 0.06	linear	10^{-3}	10^{-4}
text-embedding-ada-002	1536	70 ± 4.7	72 ± 4.4	70 ± 6.8	0.70 ± 0.06	sig	10^{3}	10^{-3}
text-embedding-3-small	1536	78 ± 3.5	80 ± 4.8	80 ± 7.3	0.78 ± 0.05	sig	10^{1}	10^{2}
text-embedding-3-small	768	76 ± 4.3	85 ± 5.3	68 ± 7.4	0.80 ± 0.04	sig	10^{1}	10^{-2}
text-embedding-3-large	3072	73 ± 3.8	72 ± 5.5	80 ± 6.0	0.75 ± 0.05	linear	10^{-2}	10^{-4}
text-embedding-3-large	768	71 ± 4.8	74 ± 6.0	70 ± 5.4	0.74 ± 0.06	linear	10^{-2}	10^{-4}

https://doi.org/10.1371/journal.pdig.0000757.t001

The results of our methods are reported in Table 1.

In addition to the classification task, a regression model is developed to predict Movement Disorder Society-Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS-III, hereinafter denoted UPDRS) scores using the embeddings. This can be interpreted as predicting the severity of the disease based on the speech samples of the PD group. A Support Vector Regression (SVR) model is selected for its ability to manage high-dimensional data and prevent overfitting through regularization. The embeddings are standardized [24]. Samples without UPDRS scores, representing healthy controls, are excluded from the analysis. The grid search from the classification task is used to optimize the SVR hyperparameters. Leave-One-Out Cross-Validation (LOOCV) is used to tune the hyperparameters and evaluate the performance of the model. The hyperparameters are determined based on the negative mean squared error. LOOCV is then performed to calculate the Root Mean Squared Error (RMSE) for each fold and the mean values and standard errors are reported in Table 2.

Discussion

We demonstrate that the state-of-the-art large language models can detect PD with up to 78% accuracy using a linguistic feature space generated with large language models. We show that the text-embedding-3 models outperform the other models. This finding is consistent with the benchmarked performance of all of the models across a variety of tasks [27]. The previous research for PD detection with large language models, specifically BERT, is only 66% accurate

Table 2. Results of the Support Vector Regression for predicting UPDRS scores. Metrics are Root Mean Squared Error (RMSE) with standard errors. We note that the MFCC+EGEMAPS (acoustic) method uses 10-fold cross validation [26].

Model	Dimension	RMSE	Kernel	C	Gamma
Word2Vec	300	14.5 ± 1.6	sig	10^{1}	10^{-2}
BERT	768	13.5 ± 1.77	sig	10^{2}	10^{-3}
XLNet	768	14.6 ± 1.65	sig	10^{1}	10^{-2}
GPT-2	768	14.3 ± 1.64	sig	10^{-3}	10^{-4}
text-embedding-ada-002	1536	14.7 ± 1.58	sig	10^{2}	10^{-2}
text-embedding-3-small	1536	15.3 ± 1.57	sig	10^{-3}	10^{-4}
text-embedding-3-small	768	14.2 ± 1.47	sig	10^{1}	10^{3}
text-embedding-3-large	3072	13.4 ± 1.62	sig	10^{1}	10^{-2}
text-embedding-3-large	768	12.9 ± 1.68	sig	10^{1}	10^{-1}
MFCC+EGEMAPS (acoustic) [26]	N/A	16.4	N/A	N/A	N/A

https://doi.org/10.1371/journal.pdig.0000757.t002 with our implementation and with the dataset that we used [13]. We demonstrate that the text-embedding-3 models surpass BERT across performance benchmarks.

The performance metrics for text-embedding-3 are largely independent of the dimensionality of the embedding output. In particular, even with the dimensionality reduced to 768 to match that of BERT and of the other large language models, the performance metrics remain better than the other large language models. We conclude that the better performance of large language models is due to their ability to extract more relevant linguistic features and not due to increased dimensionality of the feature space. This counters the notion that higher-dimensional output, with more features, automatically leads to better feature representation and performance metrics. We show that text-embedding-3 models with the same dimensionality as other LLMs outperform these models. This indicates that the architecture of the text-embedding-3 models captures a better representation of the linguistic differences between the control and PD groups, rather than merely outputting more features. The architecture of a large language model refers to the design, engineering, and organization of the model's layers, structures, and mechanisms, which define how it processes and represents the information.

Additionally, we evaluate the performance of the models using precision and recall metrics, which are algebraically related to false positive and false negative rates. Our best results are a precision of 80% and a recall of 80%. These values are approximately ten points better than the prior art. While we are not clinicians, we surmise that 80% is insufficient for clinical applications. High recall is important to detect as many actual PD cases as possible and reduce missed cases (false negatives). In contrast, high precision is important to avoid false positives and minimize unnecessary follow-up actions. Our precision and recall scores show potential, but achieving near-perfect metrics is essential for these models to move from research to clinical use. The purpose of this paper is to demonstrate the potential of large language models to improve linguistic-based detection. We subjectively feel that a six-point increase in accuracy with our implementation and a ten-point increase in precision and recall merits reporting. This statement is based on the exponential difficulty in improving the accuracy (and area under the ROC) as 100% precision and 100% recall are approached [28].

The regression analysis shows that the linguistic feature spaces generated by large language models can predict the severity of Parkinson's disease, as quantified by UPDRS scores. Notably, the text-embedding-3 method with a reduced dimensionality of 768 achieves the lowest Root Mean Squared Error (RMSE) of 12.9 ± 1.68. We note that the error bars between the large language models overlap and there may not be a statistical difference between the performances. However, the performance of all the models indicates that the embeddings capture linguistic patterns for all levels of disease severity. The text-embedding-3 method outperforms the results of acoustic methods such as MFCC+EGEMAPS within experimental uncertainty [29]. We note that MFCC+EGEMAPS uses 10-fold cross validation whereas we use LOOCV. LOOCV generally exhibits lower bias but higher variance, while 10-fold CV has higher bias but lower variance [30,31]. Despite these differences, both methods estimate the same underlying out-of-sample performance [30,31]. As a result, we can still compare overall trends. However, absolute values should be interpreted with caution due to the distinct bias–variance trade-offs. The findings suggest that large language models and linguistics can serve as a method for monitoring disease progression and assessing the effectiveness of therapeutic interventions.

We note that the presentation of PD symptoms across populations is not homogenous. The heterogeneity of PD symptoms may limit the ability of detection methods to generalize to all patients and presentations of PD. Dysarthria (acoustic impairment) is estimated to occur in 90% of PD patients [32]. The percentage of PD patients who have language impairment is a subject of ongoing research, and the etiology of language deficits in PD is not definitive

[33,34]. Research has hypothesized and correlated cognitive impairment to language impairment in PD, and has estimated cognitive impairment to occur in 80–100% of cases of PD [35]. Language impairments in PD have been characterized by reduced information content, impaired grammaticality, disrupted fluency, and reduced syntactic complexity [33]. For example, individuals with PD produce language that contains fewer correct information units, more grammatical errors, and less complex syntactic structures compared to healthy adults. They may also exhibit difficulties in generating verbs, producing longer sentences due to listing events, and including more pauses and hesitations in speech.

The processes for distilling and interpreting dimensions within the feature spaces generated by large language models are not yet fully understood and are the subject of ongoing research [36]. Specific language impairments in PD may be correlated to specific dimensions in the linguistic feature space produced by a large language model [37–39]. Reduced information content may be captured by analyzing semantic richness and coherence within the text. Impaired grammaticality may be detected through features representing grammatical usage and syntactic correctness, which identify deviations from standard grammar rules. Disrupted fluency may be reflected in features that capture fluency patterns including the frequency and duration of pauses, hesitations, and self-corrections. Reduced syntactic complexity may be observed by analyzing sentence structures and the use of complex grammatical constructions, noting a preference for simpler sentences. Mapping these deficits to corresponding dimensions in the linguistic feature space enables large language models to identify deviations from typical language patterns associated with PD.

Comparison between different detection methods is difficult due to the use of proprietary datasets, differing and insufficient implementation details, differing validation methods, and differing classification algorithms. We aim to overcome these limitations by using a dataset that is in the public domain, providing implementation details, and making our source code available upon request. This enables other researchers to replicate our methods and implementation.

Past work

Motor-speech impairment is estimated to occur in over 90% of PD cases [40]. Past research has shown that acoustic features extracted from speech signals including prosodic, vocal, and lexical elements can be used to detect PD [41]. Research has also been conducted on the processing of acoustic speech signals for detecting Alzheimer's disease [42].

Acoustic models of speech have demonstrated high accuracy on performance tasks [43]. However, acoustic features are phonetic and arise from physical changes in the vocal tract [44]. They therefore may not manifest until late into the pathogenesis, which limits the utility of acoustic models for screening and asymptomatic detection [45]. The results from the regression task demonstrates the strengths of LLMs over acoustic features for PD screening, as they can more reliably predict disease severity across different stages of the disease and thus support earlier detection.

Linguistic models have also been proposed for both PD and Alzheimer's disease. Large language models have been utilized to distinguish Alzheimer's disease from spontaneous speech with 80.3% accuracy [12]. However, the research is conducted in the context of an aphasia exam, and Alzheimer's disease and PD have different manifestations of language impairment. BERT has been implemented to detect PD from spontaneous speech [13]. A framework for automated semantic analyses of action stories capturing action-concept markers was developed to distinguish PD [46]. Morphological analysis tools were used to study cognitive impairment and utterance alterations in PD on a Japanese dataset [47]. The research showed

Table 3. Performance results of comparable studies to detect PD. All studies listed use 10-fold cross validation.

Ref.	Features Generation	Dataset	Classifier	Accuracy
[29]	Mel spectrograms (acoustic)	PC-GITA	2D-CNNs & LSTM	98.6
[48]	Bag of Words	PC-GITA	Random Forest	70.0
[48]	Term Frequency-Inverse Document Frequency	PC-GITA	Random Forest	67.0
[48]	Word2Vec	PC-GITA	SVM	72.0
[13]	Wav2Vec 2.0	See ref.	SVM	69.7
			CNN	73.8
[13]	BERT	See ref.	SVM	61.8
			CNN	74.2
[13]	BETO [52]	See ref.	SVM	62.5
			CNN	77.9

https://doi.org/10.1371/journal.pdig.0000757.t003

that cognitively unimpaired PD patients exhibited different usage rates of morphological language components when compared to the healthy control group. Morphological analysis tools encompass only one of the five linguistic components (morphology), whereas text embeddings capture four, including morphology, syntax, semantics, and pragmatics.

Research reports accuracy of up to 72% accuracy with the PC-GITA dataset using Word2-Vec [21] word embeddings on a manual transcription of the PD monologues [48]. Word embedding models are context-independent, meaning each word has the same embedding regardless of its context in a sentence. Therefore, word embeddings primarily extract features at the word level. The study also eliminates punctuation and stop words and performs lexicon normalization. This process results in the loss of linguistic information beyond the semantic component of each isolated word [49]. Manual transcription requires domain- and task-specific knowledge from the transcriber. The transcriber and their domain-experience therefore become a variable in the experiment and may introduce implicit bias during the transcription process [50]. In contrast, automatic speech recognition models do not introduce the variable of the transcriber and provide a uniform and scalable approach. However, automatic speech recognition models may struggle with domain-specific jargon, accents, and speech nuances. Specifically, they may fail to identify, either by commission or omission, the explicit incorrect use of language that may be present in individuals with language impairments. Automated transcription may lower accuracy in performance tasks, specifically in the classification of neurodegenerative diseases [51]. We report the performance metrics of our text embeddings approach, applied to an automatically transcribed version of the monologues. We suspect that the accuracy does not fully represent the capabilities of the approach. If the embedding models were applied to a manually transcribed version of the monologues, we may achieve even better performance.

A summary of results of other methods to detect PD is shown in Table 3. We note that while acoustic methods provide the highest accuracy, they may be limited in application and for detecting PD in the prodromal stage.

Limitations

The dataset for training and testing the classifier is small relative to the high dimensionality of the feature spaces, which may restrict the generalizability and scalability of the results. The small dataset size also increases the risk of overtraining the model and makes it challenging to create representative validation and testing sets. However, the Support Vector Machine with regularization may prevent overtraining. Additionally, the small number of samples in test sets limits the implementation of statistical methods to assess significance. We note that our cross-validation strategy may provide an overly optimistic estimate of the performance of the models. This is because the hyper-parameters were optimized on the test set. Nested cross-validation can address this issue. However, we present results using a non-nested cross-validation strategy to allow benchmarking with similar studies in the field with the same validation. We note that text-embedding-3-small outperforms text-emebedding-3-large across performance metrics in our use case. This contradicts the results of standardized performance benchmarks, which show that text-embedding-3-large outperforms text-embedding-3-small [20]. We note that there is a risk of larger LLMs overfitting with a small dataset. We surmise that the performance of the LLMs will improve with more data given the large dimensional output and large internal architecture of the models [12].

The dataset for training and testing the classifier may have the following limitations. The dataset may not be representative of all ethnic diversities and cultures, which may limit scalability of the model [53]. The potential for misdiagnosis in the PD patient group and undiagnosed neurological conditions in the control group introduces uncontrolled variables that could impact the findings of the study. The spontaneous speech data may have biases related to language dependence, dialect, and accent [54]. The speech is of only a single language, which may limit the ability of the model to generalize to other languages. For generating the monologues, participants were asked to describe their daily routine. The nature of the prompt may not induce the participants to exhibit all forms of language impairment present.

The time since disease diagnosis (time post diagnosis) for PD patients ranges from 0.4 to 43 years. The high variability in disease duration may lower the application of the method for early disease detection. However, we note that the mean average time post diagnosis for the 50 PD patients is 11.2 ± 9.9 years. This indicates a skew towards the earlier phase of the disease. Additionally, 46% of the UPDRS scores within the PD cohort fall below the 32-point threshold [55]. This indicates only mild motor impairment. Of all the scores, 76% are beneath the 59-point threshold for severe motor impairment. We also note that 70% of the PD group have H&Y scores below two. A H&Y score below two corresponds to the early stage of PD [56]. The distribution of UPDRS and H&Y scores indicate that the majority of the cohort exhibits only mild to moderate symptoms and are in the early stage of PD.

The publication of "Attention is All You Need" [57] and the release of OpenAI's GPT-3 model has led to rapid growth in the fields of generative AI and large language models. Large language models may be implemented in a clinical setting to help detect PD in the future. However, this raises several concerns and limitations. One concern is the implementation of a test based on a non-transparent and non-open-source model. There are issues of interpretability when relying on a decision made by a black box in a clinical setting, especially when the model is not 100% accurate. The result of a misdiagnosis may be catastrophic, which raises the question of who is accountable for the decision of the model. Research has shown that large language models may reflect societal biases [58]. These biases can include those related to gender and race. Large language models may perpetuate biases present in their training data. They may also suffer from a lack of training data for specific demographics. In our case, large language models may not have sufficient training data on certain vernaculars and cultures. This lack of data may lead to misdiagnosis. These concerns highlight a larger issue as large language models continue to be implemented in the healthcare field and society [59]. More research is needed to ensure that the models and tests are developed and implemented responsibly.

Future directions

We recommend that future data collection efforts consider various forms of participant prompting. Examples of different prompts include memory-dependent tasks, narrative construction, abstract thinking, and problem-solving scenarios. Additionally, we suggest considering different mediums for conversational tasks, including monologue, dialogue, and multilogue formats. We recommend that future data collection incorporate a dimension of time. By tracking patients longitudinally, research could capture the progression of the disease and its linguistic markers. This could offer information on how early these markers appear and how they evolve. We also recommend the procurement of a standardized dataset with multiple neurodegenerative disorders and languages.

This research highlights spontaneous speech as a classifiable biomarker through linguistic representation in text embeddings. Based on this observation, several questions and future directions emerge. Previous research has demonstrated high accuracy in using speech signals to distinguish various neurodegenerative diseases [60]. However, most of the research uses healthy controls and positives of the disease for binary classification. It remains unclear whether each neurodegenerative disease has a unique signature in its speech patterns, whether acoustic or linguistic. This raises the question of how these models will perform in the real world, where individuals may have other conditions with similar or overlapping symptoms. This also raises the question of whether a single model, either binary or multimodal, can distinguish between multiple neurodegenerative diseases or diseases with similar presentations.

The manifestation of language impairments across various languages is still unclear. The Spanish-based model may be extended and applied to other languages. Techniques such as transfer learning and zero-shot learning may offer effective adaptation strategies for new languages. There also remains the possibility of increasing the accuracy and performance of our approach in this classification task by utilizing more complex classification methods such as computational Neural Networks and Deep learning. Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) are considered effective models for distilling and classifying text embedding feature spaces and may be applied to increase accuracy [61]. Feature spaces may be fused together to improve the accuracy and robustness of a classification method if the features are orthogonal [62]. Acoustic and linguistic impairments stem from different pathophysiological mechanisms and therefore may be orthogonal. Therefore, a fusion method incorporating acoustic and linguistic features may also be developed to increase performance [63].

There has been extensive research in developing new biomarkers and detection methods for PD [41,64,65]. Recent advances include the identification of blood-based protein markers through proteomic phenotyping, progress in neuroimaging techniques for assessing dopamine system function, recognition of REM sleep behavior disorder as a preclinical marker, the use of wearable devices and smartphones for collecting digital biomarkers, and the application of artificial intelligence to analyze nocturnal respiratory signals for PD assessment in home settings [41,65,65]. Linguistic features may be combined with these methods to enhance overall detection accuracy and provide a more robust view of disease markers.

Materials and methods

Dataset (spontaneous speech)

The PC-GITA dataset is used in this study [66]. The dataset consists of the spontaneous speech of 50 subjects with PD and 50 health controls (HC) for a total of 100 samples in the dataset. The data are age and gender matched. The p-value for age, calculated with a two-sided Mann-Whitney U test, is 0.99. The data consist of monologues where each subject is asked to speak about what they do on a normal day. The average duration of the monologues is 44.86 seconds. The speech is of native Colombian Spanish speakers. The recordings were taken with the PD patients in the ON state, which refers to a period of time when medication effectively alleviates

Table 4. Patient demographics and calculated statistics derived from the dataset. μ: average, σ: standard deviation.

	PD Patients		HC Subjects	
	Male	**Female**	**Male**	**Female**
Number of subjects	25	25	25	25
Age [years] (μ±σ)	61.3 ± 11.4	60.7 ± 7.3	60.5 ± 11.6	61.4 ± 7.0
Range of age [years]	33–81	49–75	31–86	49–76
Time post diagnosis [years] (μ±σ)	8.7 ± 5.8	13.8 ± 12.4		
Range of time post diagnosis [years]	0.4–20	1–43		
MDS-UPDRS-III (μ±σ)	37.8 ± 22.1	37.6 ± 14.0		
Range of MDS-UPDRS-III	6–93	19–71		

https://doi.org/10.1371/journal.pdig.0000757.t004

the motor symptoms of the disease. Recordings were conducted no more than three hours after medication was taken. The healthy controls do not have symptoms associated with PD or any other neurological disease. Data for each PD participant is labeled with Movement Disorder Society-Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS-III) [55], Hoehn & Yahr (H&Y) [56], and time post diagnosis. MDS-UPDRS-III is a clinician-scored monitored motor examination. Scores range from 0 to 132 points with scores below 32 points indicating mild motor impairment, scores from 33 to 58 points indicating moderate motor impairment, and scores of 59 or higher indicating severe motor impairment. The H&Y scale is a clinical tool used to describe the progression of PD. It ranges from Stage 1 (mild symptoms, minimal disability) to Stage 5 (severe disability, requiring full-time care). An H&Y score less than or equal to two corresponds to the disease being in the early stage. The time post diagnosis refers to the duration since each PD patient was diagnosed. Data for both the HC and PD groups are labeled with sex and age. The demographic and statistical details are summarized in Table 4.

Ethics statement

The dataset used for this study is in the public domain and was provided to the authors de-identified. The dataset complies with the Helsinki Declaration. The collection of the dataset was approved by the Ethics Committee of the Clínica Noel, in Medellín, Colombia. Each participant signed a written informed consent.

Computational approaches

The details of the implementation of the methods are shown in this section.

The code was written in Python using Google Colab.

The details of the Speech-to-Text Via Automated Speech Recognition step are as follows.

- Model: Whisper-1
- API Endpoint: https://api.openai.com/v1/audio/transcriptions
- Response_format parameter: set as "text"
- Other Parameters: none set or invoked

The details of the Text Embedding Via Large Language Model step are as follows.

OpenAI Endpoint Models:

- text-embedding-ada-002
    - API Endpoint: discontinued

- ○ Parameters: model = "text-embedding-ada-002"
- text-embedding-3-small
  - ○ API Endpoint: https://api.openai.com/v1/embeddings
  - ○ Parameters: model = "text-embedding-3-small"
  - ○ dimensions = 768 (for text-embedding-3-small with reduced dimensions; otherwise, the parameter is not invoked, and the dimension is automatically set to the default maximum length).
- text-embedding-3-large
  - ○ API Endpoint: https://api.openai.com/v1/embeddings
  - ○ Parameters: model = "text-embedding-3-large"
  - ○ dimensions = 768 (for text-embedding-3-large with reduced dimensions; otherwise, the parameter is not invoked, and the dimension is automatically set to the default maximum length).

Hugging face Transformers Endpoint Models: https://huggingface.co/docs/transformers/index

- BERT
  - ○ Tokenizer: BertTokenizer.from_pretrained('bert-base-uncased')
  - ○ Model: BertModel.from_pretrained('bert-base-uncased')
- XLNet
  - ○ Tokenizer: XLNetTokenizer.from_pretrained('xlnet-base-cased')
  - ○ Model: XLNetModel.from_pretrained('xlnet-base-cased')
- GPT-2
  - ○ Tokenizer: GPT2Tokenizer.from_pretrained('gpt2')
  - ○ Model: GPT2Model.from_pretrained('gpt2')

GENSIM Endpoint Models: https://radimrehurek.com/gensim/models/word2vec.html

- Word2Vec:
  - ○ Model: glove-wiki-gigaword-300

The performance metrics, validation, state vector machine (SVM) classifier, and Support Vector Regression (SVR) are implemented with scikit-learn [67]

Acknowledgments

The author thanks the following people for supporting this project: Professor Juan Rafael Orozco-Arroyave for providing access to the PC-GITA dataset; Dr. Pippa Simpson, Dr. Rebecca Kinraide, Dr. Henry Medeiros, Dr. Malcolm Slaney, and Dr. David Nahamoo for their technical expertise and guidance in developing the methodology; Dr. Greg Marks, Mr. Eliot Scheuer, and Mr. Bryan Pack for their continual support; and Dr. Michael Bartl for teaching the mathematics and problem-solving skills necessary to tackle this problem. Finally, the author is especially grateful for and thankful to Mr. Robert Juranitch for his persisting mentorship, and guidance.

Author Contributions

Conceptualization: Jonathan L. Crawford.

Formal analysis: Jonathan L. Crawford.

Investigation: Jonathan L. Crawford.

Methodology: Jonathan L. Crawford.

Project administration: Jonathan L. Crawford.

Writing – original draft: Jonathan L. Crawford.

References

1. Ou Z, Pan J, Tang S, Duan D, Yu D, Nong H, et al. Global Trends in the Incidence, Prevalence, and Years Lived With Disability of Parkinson's Disease in 204 Countries/Territories From 1990 to 2019. Front Public Health. 2021; 9: 776847. https://doi.org/10.3389/fpubh.2021.776847 PMID: 34950630
2. Dorsey ER, Sherer T, Okun MS, Bloem BR. The Emerging Evidence of the Parkinson Pandemic. Brundin P, Langston JW, Bloem BR, editors. J Park Dis. 2018; 8: S3–S8. https://doi.org/10.3233/JPD-181474 PMID: 30584159
3. Kouli A, Torsney KM, Kuan W-L. Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. In: John Van Geest Centre for Brain Repair, Department of Clinical Neurosciences, University of Cambridge, UK, Stoker TB, Greenland JC, editors. Parkinson's Disease: Pathogenesis and Clinical Aspects. Codon Publications; 2018. pp. 3–26. https://doi.org/10.15586/codonpublications.parkinsonsdisease.2018.ch1
4. Simuni T, Sethi K. Nonmotor manifestations of Parkinson's disease. Ann Neurol. 2009; 64: S65–S80. https://doi.org/10.1002/ana.21472 PMID: 19127582
5. Beach TG, Adler CH. Importance of low diagnostic Accuracy for early Parkinson's disease. Mov Disord. 2018; 33: 1551–1554. https://doi.org/10.1002/mds.27485 PMID: 30288780
6. Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. Neurology. 2016; 86: 566–576. https://doi.org/10.1212/WNL.0000000000002350 PMID: 26764028
7. Kilzheimer A, Hentrich T, Burkhardt S, Schulze-Hentrich JM. The Challenge and Opportunity to Diagnose Parkinson's Disease in Midlife. Front Neurol. 2019; 10: 1328. https://doi.org/10.3389/fneur.2019.01328 PMID: 31920948
8. Adler CH, Beach TG, Hentz JG, Shill HA, Caviness JN, Driver-Dunckley E, et al. Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study. Neurology. 2014; 83: 406–412. https://doi.org/10.1212/WNL.0000000000000641 PMID: 24975862
9. Bernheimer H, Birkmayer W, Hornykiewicz O, Jellinger K, Seitelberger F. Brain dopamine and the syndromes of Parkinson and Huntington Clinical, morphological and neurochemical correlations. J Neurol Sci. 1973; 20: 415–455. https://doi.org/10.1016/0022-510X(73)90175-5 PMID: 4272516
10. Postuma RB, Berg D. Prodromal Parkinson's Disease: The Decade Past, the Decade to Come. Mov Disord. 2019; 34: 665–675. https://doi.org/10.1002/mds.27670 PMID: 30919499
11. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A Comprehensive Overview of Large Language Models. arXiv; 2024. Available from: http://arxiv.org/abs/2307.06435
12. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. Geisler BP, editor. PLOS Digit Health. 2022; 1: e0000168. https://doi.org/10.1371/journal.pdig.0000168 PMID: 36812634
13. Escobar-Grisales D, Ríos-Urrego CD, Orozco-Arroyave JR. Deep Learning and Artificial Intelligence Applied to Model Speech and Language in Parkinson's Disease. Diagnostics. 2023; 13: 2163. https://doi.org/10.3390/diagnostics13132163 PMID: 37443557
14. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. 2022 [cited 27 Feb 2024]. https://doi.org/10.48550/ARXIV.2212.04356
15. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018 [cited 27 Feb 2024]. https://doi.org/10.48550/ARXIV.1810.04805
16. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019 [cited 26 Feb 2024]. https://doi.org/10.48550/ARXIV.1906.08237

17. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019; 1: 9. Available from: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

18. Greene R, Sanders T, Wang L, Neelakantan A. New and Improved Embedding Model. 2022. Available from: https://openai.com/blog/new-and-improved-embedding-model

19. Neelakantan A, Xu T, Puri R, Radford A, Han JM, Tworek J, et al. Text and Code Embeddings by Contrastive Pre-Training. 2022 [cited 26 Feb 2024]. https://doi.org/10.48550/ARXIV.2201.10005

20. OpenAI. New Embedding Models and API Updates. 2024. Available from: https://openai.com/blog/new-embedding-models-and-api-updates

21. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 [cited 29 Feb 2024]. https://doi.org/10.48550/ARXIV.1301.3781

22. Kusupati A, Bhatt G, Rege A, Wallingford M, Sinha A, Ramanujan V, et al. Matryoshka Representation Learning. 2022 [cited 22 Mar 2024]. https://doi.org/10.48550/ARXIV.2205.13147

23. Xu H, Caramanis C, Mannor S. Robustness and Regularization of Support Vector Machines. 2008 [cited 23 Mar 2024]. https://doi.org/10.48550/ARXIV.0803.3490

24. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013. Available from: https://www.springer.com/gp/book/9781461471370

25. Narendra NP, Schuller B, Alku P. The Detection of Parkinson's Disease From Speech Using Voice Source Information. IEEEACM Trans Audio Speech Lang Process. 2021; 29: 1925–1936. https://doi.org/10.1109/TASLP.2021.3078364

26. Liu Y, Reddy MK, Penttila N, Ihalainen T, Alku P, Rasanen O. Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation. IEEEACM Trans Audio Speech Lang Process. 2023; 31: 242–255. https://doi.org/10.1109/TASLP.2022.3212829

27. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive Text Embedding Benchmark. 2022 [cited 1 Mar 2024]. https://doi.org/10.48550/ARXIV.2210.07316

28. Van Trees HL. Detection, Estimation, and Modulation Theory. New York: Wiley; 2001.

29. Er MB, Isik E, Isik I. Parkinson's Detection Based On Combined CNN And LSTM Using Enhanced Speech Signals With Variational Mode Decomposition. 2021. https://doi.org/10.21203/rs.3.rs-305818/v1

30. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2009. Available from: https://books.google.com/books?id=eBSgoAEACAAJ

31. Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. Int Stat Rev Int Stat. 1992; 291–319.

32. Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S. Speech Impairment in a Large Sample of Patients with Parkinson's Disease. Behav Neurol. 1999; 11: 131–137. https://doi.org/10.1155/1999/327643 PMID: 22387592

33. Altmann LJP, Troche MS. High-level language production in Parkinson's disease: a review. Park Dis. 2011; 2011: 238956. https://doi.org/10.4061/2011/238956 PMID: 21860777

34. Liu L, Luo X-G, Dy C-L, Ren Y, Feng Y, Yu H-M, et al. Characteristics of language impairment in Parkinson's disease and its influencing factors. Transl Neurodegener. 2015; 4: 2. https://doi.org/10.1186/2047-9158-4-2 PMID: 25685335

35. Berg E, Björnram C, Hartelius L, Laakso K, Johnels B. High-level language difficulties in Parkinson's disease. Clin Linguist Phon. 2003; 17: 63–80. https://doi.org/10.1080/0269920021000055540 PMID: 12737055

36. Paulo G, Mallen A, Juang C, Belrose N. Automatically Interpreting Millions of Features in Large Language Models. arXiv; 2024. https://doi.org/10.48550/ARXIV.2410.13928

37. Bjerva J, Östling R, Veiga MH, Tiedemann J, Augenstein I. What Do Language Representations Really Represent? Comput Linguist. 2019; 45: 381–389. https://doi.org/10.1162/coli_a_00351

38. Du X, Tanaka-Ishii K. Correlation dimension of natural language in a statistical manifold. Phys Rev Res. 2024; 6: L022028. https://doi.org/10.1103/PhysRevResearch.6.L022028

39. Antonello R, Turek J, Vo V, Huth A. Low-Dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. arXiv; 2021. https://doi.org/10.48550/ARXIV.2106.05426

40. Ramig L, Fox C, Sapir S. Speech and Voice Disorders in Parkinson's Disease. 1st ed. In: Olanow CW, Stocchi F, Lang AE, editors. Parkinson's Disease. 1st ed. Wiley; 2011. pp. 346–360. https://doi.org/10.1002/9781444397970.ch31

41. Dixit S, Bohre K, Singh Y, Himeur Y, Mansoor W, Atalla S, et al. A Comprehensive Review on AI-Enabled Models for Parkinson's Disease Diagnosis. Electronics. 2023; 12: 783. https://doi.org/10.3390/electronics12040783

42. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: The ADReSSo Challenge. 2021 [cited 27 Feb 2024]. https://doi.org/10.48550/ARXIV.2104.09356

43. Chiaramonte R, Bonfiglio M. Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies. Rev Neurol. 2020; 70: 393–405. https://doi.org/10.33588/rn.7011.2019414 PMID: 32436206

44. Rusz J, Cmejla R, Ruzickova H, Ruzicka E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J Acoust Soc Am. 2011; 129: 350–367. https://doi.org/10.1121/1.3514381 PMID: 21303016

45. Holmes R J., Oates J M., Phyland D J., Hughes A J. Voice characteristics in the progression of Parkinson's disease. Int J Lang Commun Disord. 2000; 35: 407–418. https://doi.org/10.1080/136828200410654 PMID: 10963022

46. García AM, Escobar-Grisales D, Vásquez Correa JC, Bocanegra Y, Moreno L, Carmona J, et al. Detecting Parkinson's disease and its cognitive phenotypes via automated semantic analyses of action stories. Npj Park Dis. 2022; 8: 163. https://doi.org/10.1038/s41531-022-00422-8 PMID: 36434017

47. Yokoi K, Iribe Y, Kitaoka N, Tsuboi T, Hiraga K, Satake Y, et al. Analysis of spontaneous speech in Parkinson's disease by natural language processing. Parkinsonism Relat Disord. 2023; 113: 105411. https://doi.org/10.1016/j.parkreldis.2023.105411 PMID: 37179151

48. Pérez-Toro PA, Vásquez-Correa JC, Strauss M, Orozco-Arroyave JR, Nöth E. Natural Language Analysis to Detect Parkinson's Disease. In: Ekštein K, editor. Text, Speech, and Dialogue. Cham: Springer International Publishing; 2019. pp. 82–90. https://doi.org/10.1007/978-3-030-27947-9_7

49. Jurafsky D, Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd ed. 2024. Available from: https://web.stanford.edu/~jurafsky/slp3.

50. Poland BD. Transcription Quality as an Aspect of Rigor in Qualitative Research. Qual Inq. 1995; 1: 290–310. https://doi.org/10.1177/107780049500100302

51. Soroski T, Da Cunha Vasco T, Newton-Mason S, Granby S, Lewis C, Harisinghani A, et al. Evaluating Web-Based Automatic Transcription for Alzheimer Speech Data: Transcript Comparison and Machine Learning Analysis. JMIR Aging. 2022; 5: e33460. https://doi.org/10.2196/33460 PMID: 36129754

52. Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J. Spanish Pre-trained BERT Model and Evaluation Data. arXiv; 2023. Available from: http://arxiv.org/abs/2308.02976

53. Zhao D, Andrews JTA, Papakyriakopoulos O, Xiang A. Position: Measure Dataset Diversity, Don't Just Claim It. arXiv; 2024. https://doi.org/10.48550/ARXIV.2407.08188

54. Paul S, Maindarkar M, Saxena S, Saba L, Turk M, Kalra M, et al. Bias Investigation in Artificial Intelligence Systems for Early Detection of Parkinson's Disease: A Narrative Review. Diagnostics. 2022; 12: 166. https://doi.org/10.3390/diagnostics12010166 PMID: 35054333

55. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov Disord. 2008; 23: 2129–2170. https://doi.org/10.1002/mds.22340 PMID: 19025984

56. Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Counsell C, et al. *Movement* Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations The *Movement* Disorder Society Task Force on rating scales for Parkinson's disease. Mov Disord. 2004; 19: 1020–1028. https://doi.org/10.1002/mds.20213 PMID: 15372591

57. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2017 [cited 22 Mar 2024]. https://doi.org/10.48550/ARXIV.1706.03762

58. Ayoub NF, Balakrishnan K, Ayoub MS, Barrett TF, David AP, Gray ST. Inherent Bias in Large Language Models: A Random Sampling Analysis. Mayo Clin Proc Digit Health. 2024; 2: 186–191. https://doi.org/10.1016/j.mcpdig.2024.03.003

59. Ghassemi M, Birhane A, Bilal M, Kankaria S, Malone C, Mollick E, et al. ChatGPT one year on: who is using it, how and why? Nature. 2023; 624: 39–41. https://doi.org/10.1038/d41586-023-03798-6 PMID: 38036860

60. Hecker P, Steckhan N, Eyben F, Schuller BW, Arnrich B. Voice Analysis for Neurological Disorder Recognition–A Systematic Review and Perspective on Emerging Trends. Front Digit Health. 2022; 4: 842301. https://doi.org/10.3389/fdgth.2022.842301 PMID: 35899034

61. Bai X. Text classification based on LSTM and attention. 2018 Thirteenth International Conference on Digital Information Management (ICDIM). Berlin, Germany: IEEE; 2018. pp. 29–32. https://doi.org/10.1109/ICDIM.2018.8847061

62. Rodgers JL, Nicewander WA, Toothaker L. Linearly Independent, Orthogonal, and Uncorrelated Variables. Am Stat. 1984; 38: 133–134. https://doi.org/10.1080/00031305.1984.10483183

63. Amato F, Borzì L, Olmo G, Orozco-Arroyave JR. An algorithm for Parkinson's disease speech classification based on isolated words analysis. Health Inf Sci Syst. 2021; 9: 32. https://doi.org/10.1007/s13755-021-00162-8 PMID: 34422258

64. Palmirotta C, Aresta S, Battista P, Tagliente S, Lagravinese G, Mongelli D, et al. Unveiling the Diagnostic Potential of Linguistic Markers in Identifying Individuals with Parkinson's Disease through Artificial Intelligence: A Systematic Review. Brain Sci. 2024; 14: 137. https://doi.org/10.3390/brainsci14020137 PMID: 38391712

65. Miller DB, O'Callaghan JP. Biomarkers of Parkinson's disease: present and future. Metabolism. 2015; 64: S40–46. https://doi.org/10.1016/j.metabol.2014.10.030 PMID: 25510818

66. Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, Gonzalez-Rátiva MC, Nöth E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. LREC. 2014. pp. 342–347. Available from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/7_Paper.pdf

67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. 2012 [cited 23 Mar 2024]. https://doi.org/10.48550/ARXIV.1201.0490