scientific reports

OPEN

Application of machine learning and temporal response function modeling of EEG data for differential diagnosis in primary progressive aphasia

Heather Dial[1,2,8]✉, Lokesha S. Pugalenthi[3,4,8], G. Nike Gnanateja[5], Junyi Jessy Li[6] & Maya L. Henry[2,7]

Primary progressive aphasia (PPA) is a neurodegenerative syndrome characterized by progressive decline in speech and/or language. There are three PPA subtypes with distinct speech-language profiles. Early diagnosis is essential for optimal provision of care but differential diagnosis by PPA subtype can be difficult and time consuming. We investigated the diagnostic utility of a novel electroencephalography (EEG)-based biomarker in conjunction with machine learning. Individuals with semantic, logopenic, or nonfluent/agrammatic variant PPA and healthy controls (*n* = 10 per group) listened to a continuous narrative while EEG responses were recorded. The speech envelope and linguistic features representing core language processes were extracted from the narrative speech and temporal response function (TRF) modeling was used to estimate the neural responses to these features. Although TRF modeling has shown promise for clinical applications, research is lacking regarding its diagnostic utility in populations like PPA. This study sought to provide preliminary evidence to address this gap. The resulting TRFs for channel Cz were used as input to machine learning algorithms for classification of PPA vs. healthy controls, three-way classification by PPA subtype, classification of a single PPA subtype relative to the other two (e.g., semantic vs. logopenic/nonfluent variant), and pairwise classification by PPA subtype. F1 scores were highest for the latter tasks (F1's from 0.73 to 0.74), with better-than-chance classification in all tasks. Additional analyses determined that the TRF beta weights significantly improved classification over preprocessed EEG waveforms alone for all but one task (PPA vs. healthy controls). Our preliminary findings demonstrate the potential utility of this approach for differential diagnosis of PPA, warranting further investigation.

Keywords Primary progressive aphasia, Temporal response function, Differential diagnosis, Electroencephalography, Machine learning, Natural language processing

Primary progressive aphasia (PPA) is a neurodegenerative syndrome characterized by the progressive deterioration of language and/or speech[1,2]. Although additional cognitive, behavioral, and motoric deficits emerge over time, speech and language deficits are the primary contributors to impaired activities of daily living in early stages of disease. There are three PPA subtypes, each with a distinct speech-language profile[1]. The semantic variant (svPPA) is associated with a loss of core semantic knowledge, leading to deficits in word retrieval and word comprehension. The logopenic variant (lvPPA) is associated with impaired phonological

[1]Department of Communication Sciences and Disorders, University of Houston, 3871 Holman St, Houston, TX 77204, USA. [2]Department of Speech, Language, and Hearing Sciences, The University of Texas at Austin, 2504a Whitis Ave a1100, Austin, TX 78712, USA. [3]Department of Electrical and Computer Engineering, Rice University, 6100 Main Street MS 366, Houston, TX 77005, USA. [4]Department of Computer Science, The University of Texas at Austin, 2317 Speedway, Austin, TX 78712, USA. [5]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, 1975 Willow Drive, Madison, WI 53706, USA. [6]Department of Linguistics, The University of Texas at Austin Dell Medical School, 305 E 23rd St Suite 4.304, Austin, TX 78712, USA. [7]Department of Neurology, The University of Texas at Austin Dell Medical School, 1501 Red River St, Austin, TX 78712, USA. [8]Heather Dial and Lokesha S. Pugalenthi contributed equally to this work. ✉email: hrdial@central.uh.edu

processing, with associated deficits in word retrieval and repetition. Lastly, the non-fluent variant (nfvPPA) is characterized by impaired expressive grammar and/or motor speech impairment.

Early and accurate diagnosis is essential for optimal provision of care for individuals with PPA, both in terms of speech-language services and potential forthcoming pharmacological interventions. With regard to speech-language intervention, the most appropriate restitutive interventions differ by PPA subtype[3–5]. Interventions targeting word retrieval may be most relevant for svPPA and lvPPA, whereas interventions targeting motoric aspects of speech and/or grammar may be of most benefit in nfvPPA. For disease-modifying treatments, it is important to note that PPA subtypes are associated with distinct underlying pathological profiles[6]. As a consequence, early clinical diagnosis contributes to pathological prediction which, in turn, may facilitate identification of appropriate pharmacological interventions as they become available. However, differential diagnosis by PPA subtype can be challenging, even for experienced speech-language clinicians.

Differential diagnosis by PPA subtype

In standard clinical care, differential diagnosis by PPA subtype requires comprehensive cognitive-linguistic assessment[7]. Diagnostic assessment typically requires hours of testing with tasks requiring overt responses (e.g., naming pictures, yes/no responses, repeating words and phrases), which may lead to fatigue and potentially compromise the validity and reliability of the results. Perhaps more importantly, even after comprehensive cognitive-linguistic assessment, a definitive diagnosis may be elusive. Whereas svPPA and nfvPPA are typically straightforward to differentiate behaviorally, distinguishing lvPPA from nfvPPA can be challenging due to overlapping clinical features, including reduced speech fluency in both subtypes[8,9]. Fluency is a multidimensional construct, reflecting motor speech, grammar, word finding, and prosody. Thus, although the source of impaired fluency in lvPPA and nfvPPA differs (deficits in phonological processing vs. motor speech and/or grammar, respectively), the two PPA subtypes may present similarly, particularly in mild, early stages[10]. Moreover, phonological paraphasias, which are common in lvPPA, can be difficult to distinguish from apraxic speech sound errors, which are common in nfvPPA. Differentiating lvPPA from svPPA also presents challenges, as anomia is a core feature for both subtypes. Moreover, additional overlapping clinical features emerge over time; for example, in lvPPA, semantic deficits may become apparent with progression[11].

Differential diagnosis using biomarkers and machine learning

Given the challenges of differential diagnosis based on behavioral assessment, clinicians and researchers seek alternative or complementary tools for confirming a diagnosis[12]. Blood, cerebrospinal fluid (CSF), and neuroimaging (e.g., magnetic resonance imaging [MRI] and positron emission tomography) biomarkers have shown promise for identifying the underlying etiology of PPA[13–27]. To further improve diagnostic accuracy and efficiency, researchers have used neuroimaging biomarkers with machine learning (ML)[13,14,19,26]. Most studies using neuroimaging with ML have focused on structural MRI[14,19,26], although resting-state electroencephalography (EEG)/magnetoencephalography (MEG), which reflects network dynamics, has also been used with ML for PPA subtype classification[28–30]. Although each of these studies achieved high classification accuracy for some diagnostic tasks (e.g., differentiating lvPPA vs. controls), poorer classification accuracy was achieved for other tasks (e.g., differentiating nfvPPA vs. lvPPA).

EEG has fewer contraindications relative to MRI and MEG (which exclude patients with implanted metal, for example) and is significantly less expensive (cost to record EEG data is negligible compared to the hundreds of dollars per hour for MRI and MEG). However, only one study has used ML with EEG for classification of PPA. Moral-Rubio et al.[28] used resting-state EEG data as input into seven ML classification algorithms (random forest, decision tree, k-nearest neighbors (kNN), support vector machine (SVM), elastic net, Gaussian Naive Bayes, and multinomial Naive Bayes). They achieved good classification of controls vs. PPA ($F1 = 0.83$), and relatively worse, but still better-than-chance, four-way classification of controls vs. lvPPA vs. nfvPPA vs. svPPA ($F1 = 0.60$).

In sum, although the use of neuroimaging biomarkers with ML classification algorithms has proven useful for differential diagnosis, the identification of novel, reliable biomarkers and accompanying analytical approaches will continue to benefit the field. Biomarkers derived using techniques that are non-invasive and affordable, such as EEG, are particularly valuable. Despite the language-based nature of PPA syndromes, the utility of neuroimaging data obtained during language processing tasks has yet to be evaluated. Considering the nature of PPA and the distinct language phenotypes associated with each PPA subtype, a language-based EEG biomarker could prove particularly effective for differential diagnosis.

In recent years, temporal response function (TRF) modeling has gained traction as an ecologically-valid approach for characterizing neural processing of acoustic and linguistic features of continuous speech[31,32]. In TRF modeling, a linear function is estimated to map acoustic and/or linguistic features of speech to neurophysiological data. The accuracy of the resulting TRF can be tested by comparing the observed neurophysiological data with the TRF-predicted response, providing a measure of the fidelity of the neural representation in the brain. The TRF itself provides additional information about the time course of processing that specific feature. Researchers have argued that the TRF approach has potential as a tool for improving clinical diagnosis[33,34],but TRF-derived measures have not been evaluated as diagnostic tools. In the current study, we sought to provide preliminary evidence regarding the diagnostic utility of TRF modeling and ML algorithms for differential diagnosis of PPA subtypes.

Current study

In this proof-of-concept study, we examined the utility of ML classification algorithms for diagnosis of PPA using EEG data collected while participants listened to 30 one-minute segments of a continuous speech narrative (15 minutes each from two audiobooks). TRF modeling was used to derive a linear function to map acoustic and linguistic features of the audiobook onto each participant's EEG data. TRFs were estimated separately for the delta (1–4 Hz) and theta (4–8 Hz) EEG frequency bands, as they have been argued to support different levels of speech processing (e.g., delta band: word- and phrase-level representations, theta band: syllable-level representations[32]). Our first research question was whether the TRF holds promise for classifying participants by clinical subtype. Our second research question examined whether TRFs provided additional benefit compared to using the (preprocessed) EEG data alone. In other words, do the TRF-derived beta weights improve classification compared to the EEG data alone (without TRF mapping to the acoustic and linguistic features)? We predicted that the TRF beta weights would outperform the EEG-only data because they reflect processing of the acoustic and linguistic features of the continuous narrative. EEG waveforms, on the other hand, contain neural activity both related and unrelated to processing the narrative. The study workflow is presented in Figure 1.

Method

Participants

Participants included 10 healthy, age-, education-, and hearing-matched control participants, 10 individuals with svPPA, 10 individuals with nfvPPA, and 10 individuals with lvPPA (Table 1; note that control participants and participants with lvPPA are also presented in[33]). Participants with PPA were recruited as part of a speech-language intervention trial conducted by the Aphasia Research and Treatment Lab at the University of Texas at Austin[35–38]. Individuals with PPA were required to have a Mini-Mental State Exam[39] score greater than 15 and to meet criteria for one of the canonical subtypes of PPA based on international consensus criteria[1]. Clinical diagnosis was based on comprehensive neurological and cognitive-linguistic assessment. Exclusion criteria for controls included a history of stroke, neurodegenerative disease, severe psychiatric disturbance, or developmental speech and language deficits. Due to the acoustic nature of the stimuli, hearing thresholds at 500, 1000, 2000, and 4000 Hz were collected for both ears. The pure tone average across frequencies and ears is reported in Table 1 for each participant group. The study was approved by the Institutional Review Board of the University of Texas at Austin and participants provided written informed consent. The study was conducted in accordance with relevant guidelines and regulations. Because control participants were not recruited as part of the larger clinical trial, they were paid $15/hour for their participation. All participants were native English speakers who spoke English as their primary language.

Stimuli and task

Stimuli consisted of 15-minute segments from each of two audiobooks, *Alice's Adventures in Wonderland*[50], and *Who Was Albert Einstein?*[51], the latter of which has been validated for use in stroke-induced aphasia[52]. Each audiobook was divided into 15 one-minute tracks, ensuring that each track started and ended with a complete sentence. Stimuli were presented binaurally using insert earphones (ER-3A, Etymotic Research, Elk Grove Village, IL). After listening to each track, participants were asked two multiple choice questions to encourage close attention to the audiobook (accuracy presented in Table 1). These questions were not evaluated for their validity in assessing story comprehension, though we note that an analysis of variance revealed significant differences across the groups (F (3, 26) = 8.21, $p < 0.001$); post hoc comparisons performed using Tukey's Honestly Significant Difference test indicated that individuals with lvPPA and svPPA performed significantly worse than control participants, and individuals with svPPA also performed significantly worse than individuals with nfvPPA. To mitigate fatigue, participants were given the opportunity to take a break between tracks and were instructed to press the spacebar when they were ready to move on. For two participants with svPPA and five participants with nfvPPA, data were only available for the 15 tracks from *Alice's Adventures in Wonderland* (see Supplementary Materials, Supplementary Table 1), creating an imbalance in samples between subtypes.

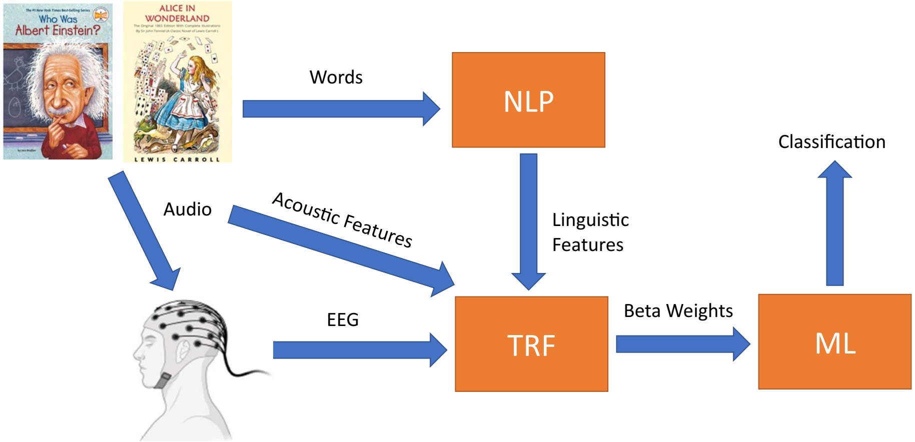

Figure 1. Study workflow. EEG data were acquired while participants listened to 30 one-minute tracks of a continuous narrative. Acoustic features were derived from the audio. Additionally, for each word in the stimulus, linguistic feature values were derived using natural language processing (NLP). Acoustic and linguistic features were used to estimate a TRF to map feature values to a participant's EEG responses. The resulting TRF beta weights were then used as input to a ML-based classifier.

	Healthy controls	svPPA	lvPPA	nfvPPA
Demographic information				
Age (years)	65.9 (s = 6.4, range = 58–79)	65.1 (s = 7.6, range = 54–76)	68.6 (s = 8.8, range = 55–81)	70.2 (s = 6, range = 61–78)
Sex (F/M)	9/1	4/6	5/5	4/6
Handedness (Right/Left)	10R	10 R	10 R	10 R
Education (years)	15.2 (s = 2.3, range = 12–18)	17.4 (s = 3.2, range = 12–22)	16.1 (s = 2.1, range = 12–18)	16 (s = 3, range = 12–22)
Hearing threshold, pure tone average**	18.6 (s = 11.5, range = 8.8–45.0)	22.9 (s = 15.1, range = 6.9–54.4)	18.4 (s = 8.2, range = 3.1–31.9)	26.8 (s = 10.0, range = 10.6–41.2)
Race, ethnicity	10 White, non-Hispanic	10 White, non-Hispanic	10 White, non-Hispanic	10 White, non-Hispanic
General cognition				
MMSE (30)[a]	25.8/26, 29.5/30*	24.0 (s = 4.0, range = 16–28)[+]	23.7 (s = 3.3, range = 18–27)[+]	26.6 (s = 2.8, range = 22–30)[+]
Verbal memory				
CVLT total (36)[a]	–	16.7 (s = 6.5, range = 10–30)	15.1 (s = 6.7, range = 7–28)	24.9 (s = 5.7, range = 16–32)
CVLT recall (9)[a]	–	1.7 (s = 2.2, range = 0–6)	2.8 (s = 2.9, range = 0–9)	7.1 (s = 1.9, range = 3–9)
Visuospatial processing				
Benson figure copy (17)[a]	–	16.5 (s = 0.7, range = 15–17)	15.5 (s = 1.8, range = 11–17)	14.4 (s = 3.6, range = 6–17)
Benson figure recall (17)[a]	–	8.3 (s = 3.5, range = 2–14)	9.1 (s = 3.8, range = 2–14)	10.9 (s = 4.3, range = 6–17)
Benson figure recognition (Correct / Incorrect)[a]	–	8 Correct/ 2 Incorrect	10 Correct/ 0 Incorrect	7 Correct/ 3 Incorrect
Phonological working memory				
Forward digit span[a]	–	6.5 (s = 1.3, range = 5–9)	4.5 (s = 0.8, range = 3–6)	4.4 (s = 1.5, range = 1–6)
Backward digit span[a]	–	4.6 (s = 2.2, range = 0–8)	3.1 (s = 0.6, range = 2–4)	2.9 (s = 1.1, range = 1–4)
Western Aphasia Battery Repetition (100)[b]	–	92.3 (s = 6.1, range = 82–100)	74.6 (s = 10.3, range = 61–94)	87.5 (s = 8.7, range = 66–98)
Auditory word comprehension				
PPVT short (16)[a]	–	9.3 (s = 3.2, range = 5–13)	14.9 (s = 1.2, range = 13–16)	15 (s = 1.4, range = 12–16)
Verbal fluency				
Letter fluency (d)[a]	–	6.8 (s = 6, range = 0–22)	7.7 (s = 2.8, range = 4–13)	4.8 (s = 2.7, range = 2–9)
Category fluency (Animals)[a]	–	7.2 (s = 4.7, range = 3–18)	10.3 (s = 3.7, range = 5–16)	11.2 (s = 4.9, range = 2–18)
Aphasia severity				
Western Aphasia Battery Aphasia Quotient (100)[b]	–	84.5 (s = 8.4, range = 70.6–93.6)	84.2 (s = 4.2, range = 78.7–90.9)	83.3 (s = 9.1, range = 65–96)
Object knowledge				
Pyramids & Palm Trees: Pictures[c]	–	76.9 (s = 14.4, range = 55.8–98.1)	92.7 (s = 4.1, range = 82.7–98.1)	99.3 (s = 2.3, range = 92.9–100)
Naming				
Boston Naming Test[a]	–	22.2 (s = 15.1, range = 5–53.3)	59.0 (s = 25.9, range = 15–93.3)	86.7 (s = 17.5, range = 40–100)
Syntactic processing				
Auditory sentence-picture matching (%)[d]	–	95.3 (s = 9.9, range = 68.8–100)	89.4 (s = 8.2, range = 72.9–100)	90.9 (s = 7.5, range = 79.2–100)
NAT (%)[e]	–	89.8 (s = 13.7, range = 66.7–100)	71.7 (s = 24.3, range = 8.3–91.7)	64 (s = 19.9, range = 16.7–86.7)
Reading				
Regular word reading (%)[f]	–	92.2 (s = 15.3, range = 44.4–100)	95.0 (s = 11.1, range = 66.7–100)	92.2 (s = 11.5, range = 55.6–100)
Irregular word reading (%)[f]	–	60.0 (s = 26.6, range = 0–100)	74.4 (s = 21.1, range = 22.2–100)	82.8 (s = 19.6, range = 33.3–100)
Spelling				
Regular word spelling (%)[f]	–	74.0 (s = 28.4, range = 0–100)	76.0 (s = 24.8, range = 20–100)	80.0 (s = 15.9, range = 40–100)
Continued				

	Healthy controls	svPPA	lvPPA	nfvPPA
Irregular word spelling (%)[f]	–	30.0 (s = 31.5, range = 0–100)	38.0 (s = 27.5, range = 0–80)	61.0 (s = 24.7, range = 20–100)
Apraxia of speech severity				
Apraxia of speech severity rating[g]	–	0 (s = 0, range = 0–0)	0.1 (s = 0.3, range = 0–1)	2.5 (s = 1.2, range = 1–4)
Multiple choice question accuracy				
Multiple choice question accuracy (%)	90.8 (s = 5.4, range = 81.7–96.7)	58.2 (s = 22.9, range = 20.0–81.7)	63.8 (s = 15.1, range = 36.7–90.0)	78.8 (s = 16.8, range = 48.3–98.3)

Table 1. Demographic characteristics, results of neuropsychological assessments of cognitive and linguistic processing, and performance on comprehension questions used in the current study. Mean, standard deviation (s), and range are reported. *F* = female; *M* = male; *R* = right; *L* = left; *MMSE* = Mini Mental State Exam; *CVLT* = California Verbal Learning Test; *PPVT* = Peabody Picture Vocabulary Test; *NAT* = Northwestern Anagram Test. *Six controls were tested with the 26 item telephone-modified MMSE[40] and four were tested on the traditional MMSE[39]. **Analysis of variance indicated that there were no significant differences across groups on pure tone average, $F\ (3, 36) = 1.16$, $p = 0.338$. Note that the pure tone average was not available for one participant with lvPPA. [+]Analysis of variance indicated that there were no significant differences across PPA subtypes on the MMSE, $F\ (2, 28) = 2.62$, $p = 0.091$. [a]Assessments from neuropsychological battery described in[41–43]. For the Boston Naming Test, the full 60-item version was only administered to individuals with svPPA and lvPPA. Individuals with nfvPPA received a shortened 30-item version. Percent correct is reported. [b]From[44]. [c]The full 52-item Pyramids and Palm Trees test[45] was only given to individuals with svPPA or lvPPA. Individuals with nfvPPA received a shorter version, developed by[46] from the standard 52-item version. Percent correct is reported. [d]From[10]. [e]From[47]. The full 30-item version was only administered to individuals with nfvPPA. Individuals with lvPPA and svPPA received a shortened 12-item version. Percent correct is reported. [f]Adapted from the Arizona Battery for Reading and Spelling[48]. [g]Subjective clinician ratings from the Motor Speech Examination[49].

We chose to use F1 for ranking classifier performance in order to minimize issues related to this imbalance (see Analyzing model performance for definition and justification for using F1).

EEG data collection and preprocessing

While participants listened to the audiobooks, EEG data and audio were sampled at 25,000 Hz using a 32-channel (10–20 system) BrainVision actiCHamp active electrode system and BrainVision StimTrak, respectively (Brain Products, Gilching, Germany). The data were re-referenced offline using the common average reference. EEG data were preprocessed using EEGLAB 2019.1[53] in MATLAB 2016b (MathWorks Inc., Natick, MA, USA). Data were downsampled to 128 Hz, then filtered from 1 to 15 Hz using a non-causal, Hamming windowed-sinc FIR filter (high pass filter cut-off = 1 Hz, filter order = 846; low pass filter cut-off = 15 Hz, filter order = 212). Channels whose activity was > 3 standard deviations from surrounding channels were rejected and replaced via spherical spline interpolation. Large artifacts were suppressed using artifact subspace reconstruction[54], with sixty seconds of manually-defined clean data used as calibration data. Lastly, independent component analysis using the infomax algorithm was performed to correct for eye movement, muscle, and electrocardiographic artifacts, with components manually identified for correction. The cleaned EEG data were further filtered into the delta (1–4 Hz) and theta (4–8 Hz) bands, as these two frequency bands have been identified as important for speech processing but may support different aspects of processing. Specifically, the delta band has been linked to processing longer speech units (e.g., words and phrases) and the theta band has been linked to processing shorter speech units (e.g. syllables)[32].

Acoustic feature derivation

Cortical tracking of the speech envelope has proven sensitive to hearing impairment in neurotypical older adults[55] and multiband envelope tracking has been shown to differ significantly between individuals with lvPPA and neurotypical older adults[33]. Thus, we investigated whether TRFs reflecting cortical tracking of acoustic features would be successful in PPA classification. Two acoustic features, the multiband speech envelope and broadband envelope derivative, were calculated for each of the audio tracks to be used for TRF modeling.

Multiband speech envelope The multiband speech envelope reflects syllable, word, and phrase boundaries as well as prosodic cues[56,57]. To derive the multiband speech envelope, auditory stimuli from the audiobooks were first filtered through 16 gammatone filters to produce 16 bands[58]. The absolute value of the Hilbert transform in each of the 16 bands comprised the multiband stimulus envelope, which was then raised to a power of 0.6 to mimic the compression characteristics of the inner ear[59]. This resulted in 16 band-specific speech envelopes. TRFs were estimated for each of the 16 bands. The TRF beta weights were averaged across the 16 bands for ML classification.

Broadband envelope derivative The broadband envelope derivative reflects acoustic onsets and offsets critical for identifying syllable, word, and phrase boundaries[60]. The auditory cortex, including the superior temporal gyrus, has been shown to be particularly sensitive to acoustic edges[61]. Considering that the superior temporal gyrus is a site of prominent atrophy in lvPPA[62], we sought to determine whether cortical tracking of the broadband envelope derivative would be useful for PPA classification. Thus, we took the first temporal derivative of the broadband envelope to be used for TRF estimation. Only the positive values of the derivative were used.

Linguistic feature derivation

Linguistic features were selected that correspond to the core language domains implicated in PPA and used for PPA subtype classification. Specifically, we selected features reflecting phonological processing (significantly impaired in lvPPA), semantic processing (significantly impaired in svPPA), and syntactic processing (significantly impaired in nfvPPA). Critically, the specific linguistic features we selected to represent each of these levels of processing have been demonstrated to have better-than-chance prediction accuracy in previous studies utilizing TRF modeling[63–65]. Prosodylab-Aligner[66] was used to temporally align phonemes and words with the audio tracks (i.e., for identification of phoneme and word onsets and offsets), with manual correction by expert linguists and highly trained research assistants. [Because of coarticulation, there is no "ground truth" for where one phoneme/word begins and another ends, and so we thus emphasized consistency in alignment by having the first author review each track, making edits as needed. We note that although "errors" in alignment would impact the accuracy of TRF modeling, this would be consistent across participants and therefore should not impact classification performance.] Phoneme and word onsets were subsequently used to temporally align linguistic features with the EEG responses.

Phonological feature: cohort entropy Cohort entropy quantifies the degree of uncertainty regarding word identity at the current phoneme based on competition among words in the cohort (the list of words with the same phonemes up to that point in the word). It was derived at the phoneme level and mapped to phoneme onsets for TRF estimation. Notably, the first phoneme in each word lacks a feature value. A phoneme's cohort entropy is defined as the Shannon entropy for the cohort of words consistent with the phonemic makeup up to that phoneme[64]. Each word's entropy is defined as its word frequency multiplied by the natural log of its word frequency. To derive word frequency, the frequency count of the word was determined based on the SUBTLEX_us_2007 corpus[67] and then divided by the total number of words in the corpus, forming a probability distribution among the words; frequency is then defined as the natural logarithm of each word's probability. For the *i*th phoneme in a word, the following formula was used to compute cohort entropy.

$$\sum_{word}^{cohort} freq\left(word\right) \cdot ln\left(freq\left(word\right)\right)$$

Semantic features These features were derived at the word level and were subsequently mapped to word onsets for TRF estimation.

Word frequency

Word frequency represents how frequently a word appears in the English language. As previously indicated, to derive word frequency, the frequency count of the word was determined based on the SUBTLEX_us_2007 corpus[67] and then divided by the total number of words in the corpus, forming a probability distribution among the words; frequency is defined as the natural logarithm of each word's probability, assuming no prior context[64]. For any word *w*, its word frequency can be mathematically formulated as its natural log probability, $ln\left(p\left(w\right)\right)$, where p represents probability as defined above, independent of context.

Semantic dissimilarity

Semantic dissimilarity represents how semantically dissimilar a word is compared to the preceding words in a sentence[63]. To calculate semantic dissimilarity, we first used the well-established NLP model GPT2 to derive a semantic feature vector for each word[68]. GPT2 was chosen because it is a widely used neural language model yielding contextualized word representations (i.e., "feature vector")[69] that are sensitive and accurate to preceding context. Computations were run on Google Colab Pro's GPUs and TPUs. Semantic dissimilarity was then derived by taking each word's GPT2 feature vector and obtaining 1 minus the correlation coefficient between that vector and the mean of the vectors for all previous words in the sentence. As such, the first word for each sentence does not have a feature value. Dissimilarity values ranged from 0 to 2, with larger values reflecting larger dissimilarity. SciPy was used to compute the mean feature vector across words and NumPy was used to compute the correlations across feature vectors. For the *i*th word in a text, its semantic dissimilarity is mathematically formulated as

$$1 - r\left[f\left(w_i\right), \, mean\left[f\left(w_{i-1}\right), f\left(w_{i-2}\right), \ldots, f\left(w_2\right), f\left(w_1\right)\right]\right]$$

where *r* represents Pearson's correlation and *f*(w) represents a word's feature vector.

Syntactic feature: syntactic surprisal Syntactic surprisal was derived at the word level and subsequently mapped to word onsets for TRF estimation. Syntactic surprisal represents how surprising the part of speech (POS) tag of the current word is given the preceding words. A word's syntactic surprisal is defined as the log probability of its POS tag conditioned on previous text[65], where the next-word probability distribution was extracted using GPT2[70]. As with semantic dissimilarity, GPT2 was chosen because of its contextualized word representations. To form the next-word probability distribution with GPT2, the text preceding the current word was fed into GPT2, which outputted logits. A softmax was applied to the logits to form a probability distribution. From this distribution, we decoded using the nucleus sampling algorithm[70] with $p = 0.9$ (i.e., the smallest set of next-word predictions such that the cumulative probability was 0.9). Each word in this nucleus sample was then tagged with the POS tagger from SpaCy's *en_core_web_lg* model (https://spacy.io/models/en#en_core_web_lg). From this, counts of each POS tag were computed and then normalized to form the POS tag probability distribution. For the *i*th word of a text, its syntactic surprisal can be mathematically formulated as

$$ln\left[p\left[pos\left(w_i\right)|w_{i-1},w_{i-2},\ldots,w_2,w_1\right]\right]$$

where $p(pos\left(w_i\right)|w_{i-1},w_{i-2},\ldots,w_2,w_1\,)$ is computed from the nucleus sampling outlined above.

TRF modeling

TRF estimation was conducted using EEG data that were *z*-scored to each participant's mean across channels. TRFs were constructed to map each track's acoustic or linguistic features to a participant's corresponding EEG data, with separate TRFs estimated for each acoustic and linguistic feature. For the multiband envelope, TRFs were estimated separately for each of the 16 frequency bands, then averaged across those bands. For the broadband envelope and linguistic features, a single TRF was estimated. Each TRF was estimated by minimizing the least-squares distance between EEG values predicted from a given feature and the participant's observed EEG data. Time lags of – 500 to 1000 ms were used. TRFs were derived using regularized linear ridge regression and validated using leave-one-out cross-validation, implemented in the mTRF Toolbox[71]. The resulting TRFs represented a vector of beta weights that were then used as input to the ML algorithms described below.

Classification

Classification tasks

The broadest task was to classify each participant as either a control participant or an individual with PPA (controls vs. PPA). Differential classification across participant groups (four-way classification, controls vs. svPPA vs. lvPPA vs. nfvPPA) and by PPA subtype (three-way classification, svPPA vs. lvPPA vs. nfvPPA) was also pursued. Additionally, we sought to classify one type of PPA by ruling out the other two types of PPA (svPPA vs. nfvPPA and lvPPA; lvPPA vs. svPPA and nfvPPA; nfvPPA vs. svPPA and lvPPA), which would be clinically useful in cases where overall PPA diagnosis is conferred and one PPA subtype is suspected. This is also a common methodology for multiclass classification that enables superior performance by ML classification algorithms. Lastly, we sought pairwise (two-way) classification by PPA subtype (svPPA vs. lvPPA; svPPA vs. nfvPPA; and lvPPA vs. nfvPPA), which would be useful in cases where PPA diagnosis is conferred and narrowed down to one of two possible subtypes.

Reading in EEG and TRF data

All participants had EEG data from 30 EEG channels, but only data from channel Cz (10–20 electrode system placement[72]) were fed into ML classification algorithms (see ML classification algorithms) because a single vector concatenating all channels (i.e., 30 channels × 8307 timestamps) would be too large for our computational constraints. Channel Cz was selected based on its common use for analysis and display purposes in previous TRF literature[73,74]. Further, it is not as susceptible to bias by hemispheric differences, which is particularly important in a population like PPA, where there is asymmetric neurodegeneration. Lastly, Cz has also been linked to language-related ERPs, such as the N400[75,76]. Participant-level data were reorganized into track-level data, resulting in 1095 tracks (33 participants with 30 tracks and 7 participants with 15 tracks) that were used for training and evaluating the ML classification algorithms. The number of data points (1095) used for training exceeds any in the literature on automated approaches to PPA classification. The number of data points for each subgroup overall is presented in Supplementary Table 1. The results reported in the main text reflect classification performance at the track-level. In the Supplementary Materials, we also report the classification performance when track-level predictions are merged into individual-level predictions (Supplementary Table 2).

TRF beta weights were available for every audio track. As with EEG data, each participant's channel Cz TRF beta weights were used to build a ML-based classifier. Standardization of both TRF and EEG data is discussed in the "ML classification algorithms" section. We note that the input to the model was a single vector, both for each classification task's TRF-based model and the EEG-based model. The "Hyperparameter tuning" section describes the process used to select the single acoustic/linguistic feature and the single ML classification algorithm used in each classification task's model.

ML classification algorithms

It is common practice to test a variety of classification algorithms to achieve the best classification performance[77–82]. In this study, we evaluated nine ML classification algorithms from the Python ScikitLearn package[83]: decision tree, random forest, extremely randomized trees (aka ExtraTrees), SVM, kNN, logistic regression, Gaussian Naive Bayes, Adaboost, and Multilayer Perceptron (MLP). This is similar to the seven ML classification algorithms used by[28] for PPA classification. Note that kNN, SVM, and MLP required prior scaling as these algorithms are based on the notion of distance between data points; scaling here refers to standardizing all input TRF/EEG by subtracting the mean and scaling to unit variance. The other six ML classification algorithms did not require any preprocessing of the TRF beta weights or EEG data as they are not based on distance between data points.

Cross validation

At the participant level, the data were split into 5 stratified outer folds, where 80% of each fold was designated for training and 20% was designated for testing. Special care was taken to ensure this was done at the participant level instead of the track level so that results generalize across individuals. In other words, all tracks for a given participant were either in the training set or the test set (not both). The classifier's predictions on each outer fold's test set were merged to form a set of predictions for all data points, which were then compared to ground truth (see "Analyzing model performance"). This use of cross validation ensures the reported results are applicable across all participants in our sample. This is in contrast to the 80–20 train-test split, where the classifier would be trained on 80% of the data and only evaluated on 20% of the data (i.e., results only reflect a fifth of the dataset). Our decision to use cross validation instead of train-test split is motivated by the small N of our dataset.

Hyperparameter tuning

For each classification task's model, we built a classifier for each possible combination of EEG frequency band, single acoustic/linguistic feature used to derive TRF weights, and single classification algorithm into which the TRFs were fed. The combination that resulted in the best performance on the nested cross-validation per classification task is reported in Tables 2, 3, 4, 5, 6. The classification performance for all classifiers constructed (i.e., each combination of frequency band, acoustic/linguistic feature, and classification algorithm) is reported for each classification task in Supplementary Tables 3–12, and the best classification performance for delta and theta bands, specifically, is reported in Supplementary Table 13. The percentage of classifiers outperforming the random sampling baseline is reported in Supplementary Table 14 (see "Analyzing model performance"). For each classifier built, we used 5-fold nested cross validation to determine the internal hyperparameters of the ML classification algorithm. For each of the five outer folds, its training set is split into five stratified inner folds (i.e., running a 5-fold cross validation on an outer fold's training set, where 80% of each fold's training set is designated for training and 20% is designated for validation). When evaluating a particular set of hyperparameters, classification performance was computed for each inner fold (i.e., trained on the inner fold's training set and evaluated on the inner fold's validation set) and then averaged. This process was repeated for several sets of hyperparameters, from which the best performing hyperparameters were identified. Note that only the outer fold's training set was used to determine the best hyperparameter combination. Then, only the best performing hyperparameters were used to train a model on all of the outer fold's training data, which was evaluated on the outer fold's test set (which was not seen/used in the hyperparameter tuning process, thus giving an unbiased estimate of the hyperparameter's true performance). This process was then repeated for the 2nd outer fold and so on, where each outer fold may select different hyperparameters from its training set, which was then evaluated on its test set. Finally, each outer fold's test set predictions were merged to form a set of predictions for all data points, which were then compared to ground truth (see "Analyzing model performance"). This nested cross validation process allows us to optimize each classifier's hyperparameters without compromising the validity of its evaluation and generalization to new patients. In sum, for each classification task, the inner folds were used for selecting the model's best hyperparameter combination and the outer folds were used for final evaluation of the model itself.

Analyzing model performance

Recall, precision, and F1 score were metrics of interest. A class' recall reflects the proportion of true positive cases predicted as positive relative to all true positive cases (e.g., how many individuals with PPA were classified as having PPA). Precision reflects the proportion of true positive cases predicted as positive relative to all predicted positive cases (e.g., for all samples classified as PPA, how many actually had PPA). Lastly, F1 score reflects the harmonic mean of its precision and recall, ranging from 0 to 1, where 1 reflects perfect classification. F1 was used to evaluate each model's performance in lieu of accuracy for two reasons. First, for many of the selected classification tasks, there was an uneven class distribution; for example, for the classification task of svPPA vs. lv/nfvPPA, there were twice as many lv/nfvPPA samples as svPPA samples. Using the macro (i.e., unweighted) average of each class' F1 is ideal for use in situations where there is class imbalance because it gives equal weighting to both the dominant and non-dominant class, avoiding artificial inflation of the F1 score by the dominant class (which could potentially have a higher F1 score). Using accuracy, as many previous studies have done, can result in a classifier achieving seemingly good performance by always predicting the dominant class; for example, given that there are three times as many PPA samples as controls, our classifier for controls vs. PPA would achieve 75% accuracy by classifying every sample as PPA. For F1, however, this would correspond to a much lower score. Unlike accuracy, F1 also balances the need for simultaneously good precision and recall. To show that our classifiers achieved meaningful, above-chance performance, F1 scores were derived by randomly sampling each prediction using the uniform distribution and the sample-label distribution (Supplementary Table 14). These baselines were computed through ScikitLearn's DummyClassifier model, where its strategy parameter was set to either "uniform" or "stratified".

For all classification tasks, McNemar tests from the mlxtend package[84] were used to compare the best EEG-only model that used EEG waveforms as input against the best model that used TRF beta weights as input in order to determine whether the derivation of the TRF beta weights provided additional benefit to classification performance. From the predictions of the best TRF-based and the best EEG-only classifiers, a 2 × 2 contingency table was formed using the *mcnemar_table* function from the mlx_extend package. From this contingency table, the McNemar test statistic and corresponding *p*-value was computed using the *mcnemar* function from the mlextend package.

Results

Classification using TRF beta weights

Our first research question was whether TRF beta weights can be used to successfully classify individuals with PPA across the classification tasks described above. First, for classification of samples as healthy controls or PPA, we achieved an F1 score of 0.60 (Table 2), which outperformed random sampling predictions by either a uniform or sample-label distribution by 0.14 (Supplementary Table 4). Based on precision and recall, PPA samples were more likely to be accurately classified than control samples. Second, for four-way classification by participant group (controls vs. svPPA vs. lvPPA vs. nfvPPA), we achieved an F1 score of 0.34 (Table 3), which outperformed

Feature	Frequency band	Algorithm	Participant group	Precision	Recall	F1	Accuracy
Broadband envelope derivative	Delta	Decision tree	Healthy controls	0.45	0.35	0.60	0.71
			PPA	0.77	0.84		

Table 2. Differentiation of healthy controls from individuals with PPA with TRF beta weights as input. F1 refers to the macro average of both class' F1 scores.

Feature	Frequency band	Algorithm	Participant group	Precision	Recall	F1	Accuracy
Semantic dissimilarity	Delta	Logistic regression	Healthy controls	0.41	0.39	0.34	0.34
			svPPA	0.40	0.35		
			lvPPA	0.30	0.36		
			nfvPPA	0.28	0.28		

Table 3. Four-way classification by participant group (controls vs. svPPA vs. lvPPA vs. nfvPPA) with TRF beta weights as input. F1 refers to the macro average of all class' F1 scores.

Feature	Frequency band	Algorithm	Participant group	Precision	Recall	F1	Accuracy
Word frequency	Delta	Extremely randomized trees	svPPA	0.45	0.51	0.48	0.51
			lvPPA	0.60	0.66		
			nfvPPA	0.42	0.22		

Table 4. Three-way classification by PPA subtype (svPPA vs. lvPPA vs. nfvPPA) with TRF beta weights as input. F1 refers to the macro average of all class' F1 scores.

Task	Feature	Frequency band	Algorithm	Participant group	Precision	Recall	F1	Accuracy
svPPA vs. lvPPA and nfvPPA	Cohort entropy	Theta	Adaboost	svPPA	0.53	0.67	0.67	0.68
				lvPPA and nfvPPA	0.80	0.69		
lvPPA vs. svPPA and nfvPPA	Semantic dissimilarity	Delta	Naive Bayes	lvPPA	0.72	0.60	0.73	0.76
				svPPA and nfvPPA	0.78	0.86		
nfvPPA vs. lvPPA and svPPA	Word frequency	Delta	Naive Bayes	nfvPPA	0.54	0.54	0.68	0.74
				lvPPA and svPPA	0.82	0.82		

Table 5. Classification of a single PPA subtype relative to the other two PPA subtypes with TRF beta weights as input. F1 refers to the macro average of both class' F1 scores.

Task	Feature	Frequency band	Algorithm	Participant group	Precision	Recall	F1	Accuracy
nfvPPA vs. lvPPA	Semantic dissimilarity/Word frequency (tied)	Delta	Decision tree	nfvPPA	0.67/0.79	0.72/0.56	0.73	0.73/0.75
				lvPPA	0.78/0.73	0.76/0.89		
svPPA vs. nfvPPA	Multiband envelope	Theta	Decision tree	svPPA	0.75	0.66	0.74	0.75
				nfvPPA	0.74	0.81		
svPPA vs. lvPPA	Semantic dissimilarity	Delta	Naive Bayes	svPPA	0.69	0.83	0.74	0.74
				lvPPA	0.81	0.66		

Table 6. Pairwise classification by PPA subtype with TRF beta weights as input. F1 refers to the macro average of both class' F1 scores.

our baseline of randomly sampling predictions by 0.10 (Supplementary Table 14). Based on precision and recall, control (Precision = 0.41, Recall = 0.39) samples were more likely to be accurately classified than the other groups (Precision and Recall ranging from 0.28 to 0.40). Next, for differential classification of samples by PPA subtype, we achieved an F1 score of 0.48 (Table 4), which outperformed our baseline of randomly sampling predictions by more than 0.16 (Supplementary Table 14). Based on precision and recall, however, confidence in this model's classification would be relatively low, regardless of how a sample was classified. Subsequently, we sought to classify one PPA subtype by ruling out the other two PPA subtypes (Table 4). For classification of samples as svPPA or lvPPA/nfvPPA, we achieved an F1 score of 0.67; for classification of samples as lvPPA or svPPA/nfvPPA,

Task	Best EEG-based model's F1	Best TRF-based model's F1	*p*-value
Controls vs. PPA	0.56	0.60	0.026
Controls vs. svPPA vs. lvPPA vs. nfvPPA	0.27	0.34	< 0.001*
svPPA vs. lvPPA vs. nfvPPA	0.40	0.48	< 0.001*
svPPA vs. lvPPA and nfvPPA	0.51	0.67	< 0.001*
lvPPA vs. svPPA and nfvPPA	0.61	0.73	< 0.001*
nfvPPA vs. lvPPA and svPPA	0.54	0.68	< 0.001*
lvPPA vs. nfvPPA	0.58	0.73	< 0.001*
svPPA vs. nfvPPA	0.49	0.74	< 0.001*
lvPPA vs. svPPA	0.62	0.74	< 0.001*

Table 7. Comparison of the best TRF and EEG models for all classification tasks. The best model is defined as the model with the highest F1 score (see "Hyperparameter tuning"), where F1 refers to the macro average of both class' F1 scores ("Analyzing model performance"). *Indicates significant after Bonferroni correction, with alpha set at 0.05 (new threshold for significance = 0.0056).

we achieved an F1 score 0.73; and for classification of samples as nfvPPA or lvPPA/svPPA, we achieved an F1 score 0.68. Each of these three classification tasks outperformed baselines by more than 0.15 (Supplementary Table 14). Based on precision and recall, our classifiers did a better job at ruling out one PPA subtype relative to the other two subtypes than it did at diagnosing that subtype (e.g., for the classification task of svPPA vs. lv/nfvPPA, the model had a much higher precision score and a slightly higher recall score for classifying a case as belonging to the lv/nfvPPA class than for classifying a case as belonging to the svPPA class). Lastly, we conducted pairwise classification by PPA subtype (Table 5). For differentiating nfvPPA from lvPPA, we achieved an F1 score of 0.73. Differentiation of nfvPPA from svPPA had an F1 score of 0.74, as did the differentiation of lvPPA and svPPA. Classifiers for pairwise classification by PPA subtype outperformed baselines by more than 0.22 (Supplementary Table 14). Notably, although a relation between PPA subtypes and linguistic features most relevant for classification might be anticipated, this was not the case, as no clear pattern emerged regarding classification accuracy and the specific linguistic features used to derive TRF beta weights. Further, the different EEG frequency bands used as input to the models had no clear effect on classification accuracy and no single classification algorithm had the best performance across a majority of classification tasks.

Classification performance for TRF beta weights versus EEG

Our second research question was whether the use of TRF beta weights would improve classification performance over the use of (preprocessed) EEG waveforms alone. Accordingly, for each classification task, channel Cz of the EEG data was fed into ML classification algorithms. The outcomes from the best EEG-based classifier were then compared to the best TRF-based classifier (Tables 2, 3, 4, 5, 6). Equivalent or superior performance of EEG data relative to TRF beta weights for PPA classification would indicate that TRF modeling is not necessary. For every classification task except the broad classification of controls vs. PPA, the best TRF-based model outperformed the best EEG-based model at the 99.9% confidence level (Table 7). This provides preliminary evidence that TRF modeling is worth the time and expertise required to extract TRF beta weights because it improved predictive accuracy relative to EEG alone.

Discussion

In the current study, we explored the potential utility of temporal response function (TRF) modeling for classification of primary progressive aphasia (PPA) using electroencephalography (EEG) and machine learning (ML) classification algorithms in order to provide initial demonstration of the feasibility of the approach. Individuals with PPA and healthy controls listened to 30 minutes of continuous speech while EEG responses were recorded. TRF modeling was used to derive a linear function to map acoustic and linguistic features of the continuous speech onto the EEG data. Either the resulting TRF beta weights or (preprocessed) EEG data constituted input to the ML classification algorithms, which were used to perform a number of different classification tasks. We addressed two research questions in the current study.

Our first research question was whether TRF beta weights hold promise for use in PPA classification. The findings of the current study indicate that TRF beta weights may be useful for PPA classification, with better-than-chance classification performance observed for all tasks, although success varied across classification tasks. The most successful models were pairwise classification of PPA subtypes, with the best classification performance observed for svPPA vs. nfvPPA and nfvPPA vs. svPPA (F1s = 0.74), followed by nfvPPA vs. lvPPA (F1 = 0.73). Relatively good classification performance was also observed for classifying lvPPA vs. svPPA/nfvPPA (F1 = 0.73), with poorer classification performance observed for nfvPPA vs. svPPA/lvPPA (F1 = 0.68), svPPA vs. nfvPPA/lvPPA (F1 = 0.67), PPA vs. controls (F1 = 0.60), three-way classification by PPA subtype (F1 = 0.48), and four-way classification (controls vs. svPPA vs. lvPPA vs. nfvPPA, F1 = 0.34). However, we would note that, clinically, a PPA diagnosis must be conferred before differential diagnosis by PPA subtype. Considering the hierarchical approach to diagnosis (general PPA diagnosis to specific subtype diagnosis), it is not as clinically relevant to be able to perform four-way classification. The poor classification of PPA vs. controls could potentially emerge from the heterogeneity in TRFs across the PPA subtypes, precluding clear differentiation from controls. The F1 score of 0.73 for classification of nfvPPA vs. lvPPA is especially notable, given that differential diagnosis of nfvPPA vs.

lvPPA can be challenging for clinicians[8,9]. Taken together, the findings are particularly promising for situations where a diagnosis of PPA has been established, but differential diagnosis by subtype remains elusive, particularly if diagnosis has been narrowed to one of two subtypes. These results provide preliminary evidence regarding the potential value of TRF-based biomarkers for facilitating differential diagnosis in PPA.

Our second research question was whether there was an added benefit of incorporating TRF beta weights compared to utilizing preprocessed EEG waveforms alone. The findings of the current study indicate that use of TRF beta weights leads to significantly better classification performance over EEG alone, except in the classification of PPA vs. controls, where performance was similar between TRF- and EEG-derived classifications. Overall, we provide preliminary evidence that TRF modeling is worth the additional effort compared to EEG data alone, although future work should focus on how to make TRF modeling accessible within clinical practice settings since the current methods require access to proprietary software and technical expertise.

Previous research on automated approaches to diagnosis of PPA with neuroimaging data have utilized a variety of different inputs to the models, including structural magnetic resonance imaging (MRI)[14,26], functional connectivity from magnetoencephalography (MEG)[29], power spectral density from resting-state EEG[30], and graph theory-derived measures from resting-state EEG[28]. Of most relevance to the current study is the work of Moral-Rubio and colleagues[28], in which two classification tasks were performed (PPA vs. controls and four-way classification of controls, svPPA, nfvPPA, and lvPPA). In that study, classification of PPA vs. controls was superior to our study (F1 = 0.83 vs. F1 = 0.60), as was four-way classification of controls, svPPA, nfvPPA, and lvPPA (F1 = 0.60 vs. F1 = 0.39). Although Moral-Rubio et al.[28] achieved better classification performance for PPA vs. controls and for four-way classification for some ML algorithms, we extend their work by performing a larger number of classification tasks and using EEG data collected while participants engaged with language stimuli. Differences in the number of data points used for model training and in the model architectures themselves preclude direct comparison of classification performance across these studies. However, our results are largely consistent with previous research in supporting a potential role for automated approaches to PPA diagnosis.

Overall, we demonstrate that ML utilizing TRF-based biomarkers derived from EEG data holds promise as a means to support diagnostic decision-making in PPA. In contrast to automated approaches using MRI or MEG as input, EEG has the benefit of being affordable, with no contraindications for use. Further, EEG is non-invasive, as opposed to positron emission tomography or cerebrospinal fluid-based biomarkers that are currently used in standard clinical practice, making it a safer approach to informing diagnosis. These findings in PPA add to the evidence base suggesting a role of TRF modeling in improving diagnostic decision-making in clinical populations more broadly. Automated approaches developed to aid diagnosis hold potential for addressing health disparities associated with diagnosis/misdiagnosis as a function of race/ethnicity (see[85] for discussion) or English-speaking status. For example, only ~8% of America's speech-language pathologists speak a language other than English (ASHA 2023[86]) and many standard assessment materials are developed in English only. The development of automated approaches to diagnosis in languages other than English could mitigate the influence of these factors.

Limitations and future directions

The current study marks an important step toward use of automated approaches to diagnosis of PPA, and the exploratory nature of this study presents multiple avenues for further research. The current study included a relatively small number of participants ($n = 10$ per participant group) that were not perfectly matched for demographic characteristics (e.g., there is a larger proportion of female participants in the control group than in the PPA groups), limiting the generalizability of findings (although we would note that the 1095 data points included in the ML classification is at least one order of magnitude larger than all previous research on automated classification of PPA). Future research should be conducted with a larger number of participants to further improve classification performance and generalizability to new samples. It will also be important to consider whether and to what extent the current approach improves upon the current gold standard cognitive-linguistic assessments used for diagnosis.

It was somewhat surprising that one of the poorest performing classification tasks was for classification of PPA vs. controls. This is also the only classification task where TRF beta weights did not outperform the EEG-only classification. As indicated previously, it is possible that this is a consequence of the heterogeneity of TRFs across PPA subtypes, making it difficult to clearly identify a TRF profile that distinguishes all PPA subtypes from controls. However, distinguishing neurotypical older adults from persons with PPA is likely to be the least relevant for standard clinical practice, as individuals with PPA are more likely to be misdiagnosed with a different neurodegenerative syndrome or psychiatric condition[87,88], rather than classified as healthy. In other words, the potential utility of a TRF-based classifier for differentiating PPA vs. controls is likely limited. Instead, classification of PPA vs. Alzheimer's dementia or PPA vs. severe clinical depression, for example, would be more clinically useful. Thus, future research may focus on the development of automated tools for differential diagnosis across neurodegenerative syndromes and/or other neurological or psychiatric conditions.

There is a great deal to be learned regarding factors contributing to the relative success of one TRF model over another. For example, in the current study, analyses were restricted to electrode Cz in order to determine whether the approach was useful for classification of PPA and PPA subtypes. Given the modest success in the current study, future work should seek to identify optimal electrode configurations that maximize classification success. Along these lines, an important next step is to apply more advanced deep learning approaches, such as convolutional neural networks, to PPA classification[30]. Applying more advanced deep learning approaches has the potential to improve classification performance while providing more interpretability, allowing for the identification of features of the input that most strongly contribute to classification accuracy. Contrary to the ML classification algorithms used in the current study, all channels of EEG data can be fed into deep learning classification algorithms (compared to only channel Cz in this paper); thus, it will be possible to identify which channels (i.e., electrodes) are most useful for classification. Relatedly, due to the lack of interpretability offered by the ML models paired with the TRF beta weights in the current study, there are a number of questions that remain unanswered. For example, why was there no clear relation between classification accuracy and the specific linguistic features used to derive TRF beta weights, and why did certain features perform better than others? Future work should focus on developing a better understanding of the factors that influence classification, with a particular emphasis on identifying acoustic and linguistic features that maximize classification accuracy. The results of such work may provide valuable insights into nature and diagnosis of PPA syndromes as well as our understanding of the neural processing of the specific acoustic and linguistic features being modeled.

Conclusion

In the current study, we showed that TRF-derived beta weights for acoustic and linguistic features of a continuous narrative hold promise for use in PPA classification. In doing so, we demonstrate the potential clinical utility of this automated approach using a TRF-based biomarker derived from EEG. With recent efforts to draw attention to the amount of testing required of individuals with PPA[89], automated approaches to diagnosis will likely continue to gain traction. The current study marks an important first step toward more automated approaches to diagnosis, particularly those using TRF modeling. It provides proof-of-concept for the utility of TRF modeling for use in clinical diagnostic decision-making, motivating future research seeking to fine-tune the specific parameters used for classification. Future work should seek to make these automated approaches more accessible to clinicians, moving this research a step closer to use in clinical practice.

Data availability

The data supporting the findings of this study are subject to HIPAA regulations. However, these data may be made available upon request and with the execution of appropriate data use agreements. Please contact the corresponding author, Heather Dial, for data use inquiries at hrdial@central.uh.edu.

Received: 5 December 2024; Accepted: 21 July 2025
Published online: 12 August 2025

References

1. Gorno-Tempini, M. L. et al. Classification of primary progressive aphasia and its variants. *Neurology* **76**(11), 1006–1014 (2011).
2. Mesulam, M. M. Primary progressive aphasia. *Ann. Neurol.* **49**(4), 425–432 (2001).
3. Cadório, I., Lousada, M., Martins, P. & Figueiredo, D. Generalization and maintenance of treatment gains in primary progressive aphasia (PPA): A systematic review. *Int. J. Lang. Commun. Disord.* **52**(5), 543–560 (2017).
4. Volkmer, A. et al. Speech and language therapy approaches to managing primary progressive aphasia. *Pract. Neurol.* **20**(2), 154–161 (2020).
5. Wauters, L. D., Croot, K., Dial, H. R., Duffy, J. R., Grasso, S. M., Kim, E. & Henry, M. L. Behavioral treatment for speech and language in primary progressive Aphasia and primary progressive Apraxia of speech: A systematic review. *Neuropsychol. Rev.* 1–42 (2023).
6. Spinelli, E. G. et al. Typical and atypical pathology in primary progressive aphasia variants. *Ann. Neurol.* **81**(3), 430–443 (2017).
7. Henry, M. & Grasso, S. Assessment of individuals with primary progressive aphasia. *Semin. Speech Lang.* **39**(3), 231–241. https://doi.org/10.1055/s-0038-1660782 (2018).
8. Croot, K., Ballard, K., Leyton, C. E. & Hodges, J. R. Apraxia of speech and phonological errors in the diagnosis of nonfluent/agrammatic and logopenic variants of primary progressive aphasia (2012).
9. Leyton, C. E. & Hodges, J. R. Differential diagnosis of primary progressive aphasia variants using the international criteria. *Aphasiology* **28**(8–9), 909–921 (2014).
10. Wilson, S. M. et al. Connected speech production in three variants of primary progressive aphasia. *Brain* **133**(7), 2069–2088 (2010).
11. Ramanan, S. et al. Understanding the multidimensional cognitive deficits of logopenic variant primary progressive aphasia. *Brain* **145**(9), 2955–2966 (2022).
12. Europa, E. et al. Diagnostic assessment in primary progressive aphasia: An illustrative case example. *Am. J. Speech Lang. Pathol.* **29**(4), 1833–1849. https://doi.org/10.1044/2020_AJSLP-20-00007 (2020).
13. Agosta, F. et al. Differentiation between subtypes of primary progressive aphasia by using cortical thickness and diffusion-tensor MR imaging measures. *Radiology* **276**(1), 219–227 (2015).
14. Bisenius, S. et al. Predicting primary progressive aphasias with support vector machine approaches in structural MRI data. *NeuroImage Clin.* **14**, 334–343 (2017).
15. Conca, F., Esposito, V., Giusto, G., Cappa, S. F. & Catricalà, E. Characterization of the logopenic variant of primary progressive aphasia: A systematic review and meta-analysis. *Ageing Res. Rev.* **82**, 101760 (2022).
16. Gonzalez-Gomez, R., Ibañez, A. & Moguilner, S. Multiclass characterization of frontotemporal dementia variants via multimodal brain network computational inference. *Netw. Neurosci.* **7**(1), 322–350 (2023).
17. Grossman, M. Biomarkers in the primary progressive aphasias. *Aphasiology* **28**(8–9), 922–940 (2014).
18. Hüper, L. et al. Neurofilaments and progranulin are related to atrophy in frontotemporal lobar degeneration–a transdiagnostic study cross-validating atrophy and fluid biomarkers. *Alzheimer's Dement.* **20**, 4461–4475 (2024).
19. Kim, J. P. et al. Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer's disease. *NeuroImage Clin.* **23**, 101811 (2019).
20. Paraskevas, G. P. et al. Cerebrospinal fluid biomarkers as a diagnostic tool of the underlying pathology of primary progressive aphasia. *J. Alzheimer's Dis.* **55**(4), 1453–1461 (2017).
21. Perini, G. et al. Role of cerebrospinal fluid biomarkers and (18) F-florbetapir PET imaging in the diagnosis of primary progressive aphasia: A retrospective analysis. *Alzheimer Dis. Assoc. Disord.* **33**(3), 282–284 (2019).
22. Rabinovici, G. D. et al. Aβ amyloid and glucose metabolism in three variants of primary progressive aphasia. *Ann. Neurol.* **64**(4), 388–401 (2008).
23. Rohrer, J. D., Rossor, M. N. & Warren, J. D. Syndromes of nonfluent primary progressive aphasia: A clinical and neurolinguistic analysis. *Neurology* **75**(7), 603–610 (2010).
24. Teichmann, M. et al. Deciphering logopenic primary progressive aphasia: A clinical, imaging and biomarker investigation. *Brain* **136**(11), 3474–3488 (2013).

25. Whitwell, J. L. et al. Working memory and language network dysfunctions in logopenic aphasia: A task-free fMRI comparison with Alzheimer's dementia. *Neurobiol. Aging* **36**(3), 1245–1252 (2015).
26. Wilson, S. M. et al. Automated MRI-based classification of primary progressive aphasia variants. *Neuroimage* **47**(4), 1558–1567 (2009).
27. Xu, J. et al. Clinical features and biomarkers of semantic variant primary progressive aphasia with MAPT mutation. *Alzheimer's Res. Ther.* **15**(1), 21 (2023).
28. Moral-Rubio, C. et al. Application of machine learning to electroencephalography for the diagnosis of primary progressive aphasia: A pilot study. *Brain Sci.* **11**(10), 1262 (2021).
29. Ranasinghe, K. G. et al. Distinct spatiotemporal patterns of neuronal functional connectivity in primary progressive aphasia variants. *Brain* **140**(10), 2737–2751 (2017).
30. Quinn, C., Craik, A., Tessmer, R., Henry, M. L. & Dial, H. Utilization of resting-state electroencephalography spectral power in convolutional neural networks for classification of primary progressive aphasia. *NeuroImage Rep.* **5**(1), 100242 (2025).
31. Crosse, M. J. et al. Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Front. Neurosci.* **15**, 705621 (2021).
32. Gnanateja, G. N. et al. On the role of neural oscillations across timescales in speech and music processing. *Front. Comput. Neurosci.* **16**, 872093 (2022).
33. Dial, H. R. et al. Cortical tracking of the speech envelope in logopenic variant primary progressive aphasia. *Front. Hum. Neurosci.* **14**, 597694 (2021).
34. Di Liberto, G. M. & Lalor, E. C. Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hear. Res.* (2017).
35. Dial, H. R., Hinshelwood, H. A., Grasso, S. M., Hubbard, H. I., Gorno-Tempini, M. L. & Henry, M. L. Investigating the utility of teletherapy in individuals with primary progressive aphasia. *Clin. Interv. Aging 453–471* (2019).
36. Dial, H. R. et al. Baseline structural imaging correlates of treatment outcomes in semantic variant primary progressive aphasia. *Cortex* **158**, 158–175 (2023).
37. Henry, M. L. et al. Treatment for word retrieval in semantic and logopenic variants of primary progressive aphasia: Immediate and long-term outcomes. *J. Speech Lang. Hear. Res.* **62**(8), 2723–2749 (2019).
38. Henry, M. L. et al. Retraining speech production and fluency in non-fluent/agrammatic primary progressive aphasia. *Brain* **141**(6), 1799–1814 (2018).
39. Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**(3), 189–198 (1975).
40. Newkirk, L. A. et al. Validation of a 26-point telephone version of the mini-mental state examination. *J. Geriatr. Psychiatry Neurol.* **17**(2), 81–87 (2004).
41. Knopman, D. S. et al. Development of methodology for conducting clinical trials in frontotemporal lobar degeneration. *Brain* **131**(11), 2957–2968 (2008).
42. Kramer, J. H. et al. Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. *Cogn. Behav. Neurol.* **16**(4), 211–218 (2003).
43. Staffaroni, A. M. et al. Longitudinal multimodal imaging and clinical endpoints for frontotemporal dementia clinical trials. *Brain* **142**(2), 443–459 (2019).
44. Kertesz, A. *Western Aphasia Battery–Revised* (PsychCorp, 2007).
45. Howard, D. & Patterson, K. E. The pyramids and palm trees test (1992).
46. Breining, B. L. et al. A brief assessment of object semantics in primary progressive aphasia. *Aphasiology* **29**(4), 488–505 (2015).
47. Weintraub, S. et al. The Northwestern Anagram test: Measuring sentence production in primary progressive aphasia. *Am. J. Alzheimers Dis. Other Dement.* **24**, 408–416. https://doi.org/10.1177/1533317509343104 (2009).
48. Beeson, P. M. & Rising, K. *Arizona Battery for Reading and Spelling* (Canyonlands Publishing, 2010).
49. Wertz, R. T., LaPointe, L. L. & Rosenbek, J. C. *Apraxia of Speech in Adults: The Disorder and its Management* (Grune and Stratton, 1984).
50. Carrol, L. Alice's Adventures in Wonderland (1865). Available online at: https://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-5.
51. Brallier, J. *Who was Albert Einstein?* (Penguin, 2002).
52. Wilson, S. M., Yen, M. & Eriksson, D. K. An adaptive semantic matching paradigm for reliable and valid language mapping in individuals with aphasia. *Hum. Brain Mapp.* **39**(8), 3285–3307 (2018).
53. Delorme, A. & Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**(1), 9–21 (2004).
54. Mullen, T. R. et al. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Trans. Biomed. Eng.* **62**(11), 2553–2567 (2015).
55. Fuglsang, S. A., Märcher-Rørsted, J., Dau, T. & Hjortkjær, J. Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *J. Neurosci.* **40**(12), 2562–2572 (2020).
56. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* **28**(24), 3976–3983 (2018).
57. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**(6174), 1006–1010 (2014).
58. Slaney, M. Auditory toolbox. *Interval Res. Corp. Tech. Rep.* **10**(1998), 1194 (1998).
59. Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z. & Francart, T. Speech intelligibility predicted from neural entrainment of the speech envelope. *J. Assoc. Res. Otolaryngol.* **19**, 181–191 (2018).
60. Zuk, N. J., Murphy, J. W., Reilly, R. B. & Lalor, E. C. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLoS Comput. Biol.* **17**(9), e1009358 (2021).
61. Oganian, Y. & Chang, E. F. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.* **5**(11), eaay6279 (2019).
62. Gorno-Tempini, M. L. et al. The logopenic/phonological variant of primary progressive aphasia. *Neurology* **71**(16), 1227–1234 (2008).
63. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**(5), 803–809 (2018).
64. Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T. & Brodbeck, C. Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *J. Neurosci.* **41**(50), 10316–10329 (2021).
65. Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P. & De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci.* **119**(32), e2201968119 (2022).
66. Gorman, K., Howell, J. & Wagner, M. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Can. Acoust.* **39**(3), 192–193 (2011).
67. Brysbaert, M. & New, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* **41**(4), 977–990 (2009).
68. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019).
69. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst*. (2017).

70. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. In: *International Conference on Learning Representations* (2020).
71. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
72. Jasper, H. H. The Ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* **10**, 371–375 (1958).
73. Bai, F., Meyer, A. S. & Martin, A. E. Neural dynamics differentially encode phrases and sentences during spoken language comprehension. *PLoS Biol.* **20**(7), e3001713 (2022).
74. Sassenhagen, J. How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Lang. Cogn. Neurosci.* **34**(4), 474–490 (2019).
75. Alday, P. M., Schlesewsky, M. & Bornkessel-Schlesewsky, I. Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: Event-related potentials reflect continuous model updates. *eneuro,* **4**(6), (2017).
76. Kutas, M. & Federmeier, K. D. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**(1), 621–647 (2011).
77. Celik, E. & Omurca, S. I. Improving Parkinson's disease diagnosis with machine learning methods. In: *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1–4). IEEE (2019).
78. Govindu, A. & Palwe, S. Early detection of Parkinson's disease using machine learning. *Procedia Comput. Sci.* **218**, 249–261 (2023).
79. Järvelin, A. & Juhola, M. Comparison of machine learning methods for classifying aphasic and non-aphasic speakers. *Comput. Methods Programs Biomed.* **104**(3), 349–357 (2011).
80. Joshi, S., Shenoy, D., Simha, G. V., Rrashmi, P. L., Venugopal, K. R. & Patnaik, L. M. Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods. In: *2010 Second International Conference on Machine Learning and Computing* (pp. 218–222). IEEE (2010).
81. Mahmoud, S. S., Kumar, A., Li, Y., Tang, Y. & Fang, Q. Performance evaluation of machine learning frameworks for aphasia assessment. *Sensors* **21**(8), 2582 (2021).
82. Matias-Guiu, J. A. et al. Machine learning in the clinical and language characterisation of primary progressive aphasia variants. *Cortex* **119**, 312–323 (2019).
83. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
84. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* **3**(24), 638 (2018).
85. Gianattasio, K. Z., Prather, C., Glymour, M. M., Ciarleglio, A. & Power, M. C. Racial disparities and temporal trends in dementia misdiagnosis risk in the United States. *Alzheimer's Dementia Transl. Res. Clin. Interv.* **5**, 891–898 (2019).
86. American Speech-Language-Hearing Association. (2023). 2022 profile of multilingual service providers. https://www.asha.org/siteassets/surveys/2022-profile-of-multilingual-service-providers.pdf.
87. Hussain, M., Kumar, P., Khan, S., Gordon, D. K. & Khan, S. Similarities between depression and neurodegenerative diseases: Pathophysiology, challenges in diagnosis and treatment options. *Cureus,* **12**(11), (2020).
88. Voros, V. et al. Untreated depressive symptoms significantly worsen quality of life in old age and may lead to the misdiagnosis of dementia: A cross-sectional study. *Ann. Gen. Psychiatry* **19**(1), 1–6 (2020).
89. Gallée, J., Cartwright, J., Volkmer, A., Whitworth, A. & Hersh, D. "Please don't assess him to destruction": The RAISE assessment framework for primary progressive aphasia. *Am. J. Speech Lang. Pathol.* **32**(2), 391–410 (2023).

Acknowledgements

We wish to thank members of the Aphasia Research and Treatment Lab at the University of Texas at Austin for assisting with data collection. We also wish to thank our participants with PPA and their families for the time and effort that they have devoted to our research. This work was funded by grants from the National Institutes of Health (NIDCD F32 DC016812-01 to HRD; NIDCD R01 DC016291 and NIDCD R03 DC013403 to MLH).

Author contributions

Heather Dial: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Curation, Writing-Original Draft, Writing-Review & Editing, Funding acquisition, Project administration; Lokesha Pugalenthi: Conceptualization, Data curation, Formal analysis, Methodology, Writing-Original Draft, Writing-Review & Editing; G. Nike Gnanateja: Conceptualization, Formal analysis, Methodology, Writing-Review & Editing; Junyi Jessy Li: Conceptualization, Methodology, Supervision, Writing-Review & Editing; Maya Henry: Conceptualization, Resources, Writing-Review & Editing, Supervision, Funding acquisition.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-13000-8.

Correspondence and requests for materials should be addressed to H.D.