bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Automatic subtyping of individuals with Primary Progressive Aphasia

Charalambos Themistocleous, Ph.D., Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Bronte Ficek, M.H.S, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Kimberly Webster, M.A., M.S., Johns Hopkins School of Medicine, Baltimore MD 21287, USA

Dirk-Bart den Ouden, Ph.D., Arnold School of Public Health, University of South Carolina, Columbia, SC, USA

Argye E. Hillis, M.D., Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Kyrana Tsapkini, Ph.D., Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Corresponding author:

Kyrana Tsapkini, Ph.D.

Assistant Professor of Neurology

Johns Hopkins University School of Medicine

600 N. Wolfe Street, Phipps 446,

Baltimore, MD 21287, USA

Cell: 1-410-736-2940 | tsapkini@jhmi.edu

Corresponding author and person responsible for statistical analysis:

Charalambos Themistocleous, Ph.D.

Johns Hopkins University School of Medicine

600 N. Wolfe Street, Phipps 488

Baltimore, MD 21287, USA

4433711207 | cthemis1@jhu.edu

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Publication history: This manuscript was previously published in bioRxiv:

Manuscript word count: 3,574

Abstract word count: 208

Title character count: 67

Study funding: This project was supported by the Science of Learning Institute at Johns Hopkins University grant to Dr. Kyrana Tsapkini (NIH/NIDCD R01 DC014475).

Search terms: primary progressive aphasia, deep neural networks, connected speech

Disclosures: All of the authors report no disclosures.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Objective

The classification of patients with Primary Progressive Aphasia into variants is time consuming, costly, and requires combined evaluations by clinical neurologists, neuropsychologists, speech pathologists, and radiologists. Therefore, our aim is to determine if acoustic and linguistic variables provide accurate classification of PPA patients into one of the three variants.

Methods

In this paper, we present a machine learning model based on Deep Neural Networks for the subtyping of patients with PPA into the three main variants using combined acoustic and linguistic information elicited automatically using acoustic and linguistic analysis. The performance of the Deep Neural Networks was compared to the classification accuracy of Random Forests, Support Vector Machines, and Decision Trees. It was also compared to the classification based on auditory scores provided by clinicians.

Results

The DNN model resulted in 80% classification accuracy providing reliable subtyping of patients with PPA into variants that outperformed other machine learning models and auditory classification of patients into variants by clinicians.

Conclusion

We show that combined measures of speech and language function as the patients' fingerprint and provide information about patients' symptoms and variant subtyping. This approach can enable clinicians and researchers to employ this fingerprint and provide an automatic classification of patients with PPA saving much time and money.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Primary Progressive Aphasia (PPA) is a progressive neurological condition affecting speech and language[1,2] with substantial symptom variability. Patients are classified into three main variants: nonfluent (nfvPPA), semantic (svPPA), and logopenic (lvPPA) PPA,[1] to facilitate PPA prognosis and evaluation and ultimately improve therapy decisions. The gold standard of PPA classification is the manually subtyping of patients with PPA by clinical experts, which requires combining MRI or PET scan reports with language and cognitive evaluations by clinical neurologists, neuropsychologists, and speech-language pathologists. Subtyping patients into variants is time consuming, arduous, and expensive but it provides clues to the most likely pathology, guides medical treatment, and explains potential subsequent symptoms. So, there is a critical need for an accurate, quick, and easy evaluation system, consistent with the established criteria[1], and sensitive to speech and language deficits that characterize patients by variant.

As patients with different PPA variants differ mainly in their language symptoms, the acoustics of their speech and grammar can function as a fingerprint enabling the variant identification of patients without requiring further tasks or measures as in earlier studies or at the very least as an efficient aid in further clinical classification by expert clinicians [3–5]. Our aim is to determine if combined acoustic and linguistic measures are able to provide an automated classification of patients with PPA into all three variants, using deep neural networks. Implemented as a web application, the automatic system provides a consultation tool that can expedite the opinion of the expert clinician and inform and guide the opinion of the less specialized clinician.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

METHODS

Participants

The 44 participants had a diagnosis of PPA from an expert neurologist, a history of at least two years of progressive language deficits with no other etiology (e.g., stroke, tumors, etc.), and relatively preserved memory as shown from the general Clinical Dementia Rating (CDR)[6]. All participants were right-handed and native speakers of English. Differential diagnosis of patients with PPA and PPA variant subtyping was based on Magnetic Resonance Imaging (MRI) results, clinical and neuropsychological examination, and speech and language evaluations following the consensus criteria by Gorno-Tempini et al., 2011[1]. Specifically, 9 participants were subtyped as svPPA, 16 as lvPPA, and 19 as nfvPPA. Table 1 provides baseline information for the study participants.

Table 1 Demographic information of the participants for each PPA variant (for age, education, onset of the condition in years, language severity and total severity, the mean and the standard deviation in parenthesis is provided).

Variant	svPPA	lvPPA	nfvPPA
Female	5	8	7
Male	4	8	12
Total Speakers	9	16	19
Age	66.59 (6.06)	67.93 (7.55)	69.07 (5.57)
Education	16.30 (1.92)	16.92 (2.24)	16.42 (1.37)
Onset years	6.48 (2.31)	3.88 (3.23)	3.49 (1.80)
Language severity	2.27 (0.56)	1.39 (0.75)	1.77 (0.48)

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Total severity	7.75 (4.36)	4.98 (2.82)	6.04 (3.10)

Table 1 provides biographical demographic information of the participants for each PPA variant. One-way ANOVA tests showed that there were no significant differences between variants for sex (F=5.82, df=2, p= 0.09), age (F=0.354, df=2, p=0.705), education (F=0.162, df=2, p=0.853), language severity (F=0.154, df=2, p=0.86) and total severity (F=1.162, df=2, p=0.33). We used the revised frontotemporal dementia clinical dementia rating (FTD-CDR) to rate language and total severity in PPA[7]. Data collection was conducted as part of a clinical trial on Transcranial Direct Current Stimulation for Primary Progressive Aphasia conducted at Johns Hopkins University (NCT:02606422). The Johns Hopkins Institutional Review Board approved this study. All participants provided written informed consent for research participation.

Materials

Data from connected speech productions were recorded during a simple and widely used assessment test, the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (BDAE) by trained clinicians and assistant clinicians[8]. A clinician presented the picture to the participant and prompted the participant following the standard BDAE instructions by saying: "tell me everything you see going on in this picture". The patient was instructed to describe the picture speaking in sentences and talk about the objects, people, activities shown in the picture. Clinicians did not interrupt the patient during the task. The picture description session was audio recorded.

Analyses

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Picture descriptions were converted into 16000 Hz mono format and were transcribed using the Themis[9,10]. The output of the transcription was evaluated twice by the first author and by comparing the output manually transcribing 1/5 of the sounds and comparing the output to that of two independent evaluators. The transcriptions were not modified, as we were primarily interested in seeing how well the transcript would perform in the absence of any modification. No pauses were coded in the transcript, but their duration was estimated from the acoustic signal during speech segmentation. Fillers such as 'um..' and 'uh..' were transcribed (and analyzed), but were not included in the total word count. Repetitions, false starts and repeated but incomplete attempts at a given word were transcribed in Roman alphabet; repetitions of words and false starts were included in the total word count. Neologisms were transcribed using standard orthography using the Roman alphabet. The following three preprocessing pipelines were developed to analyze the acoustic and linguistic (morphosyntactic) properties and generate the classification data.

Pipeline 1: Audio transcription and segmentation. The sounds were processed using 'Themis', a python library developed in house that provides a text file with the audio transcription of each word and segment—vowel, consonant, pause—and a table that contains the times (onset time and offset time) of each word and segment. All transcripts were evaluated manually by two independent evaluators to confirm the faithfulness of the speech-to-text system and the presence of incorrect transcriptions of words, for example due to environmental noise, low intensity speech production, etc. The text was converted into TextGrid text format files with time information about the beginning and end of vowels and consonants. Pause duration was calculated during segmentation from the automatic alignment system.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Pipeline 2: Audio processing. A second pipeline was employed for the extraction of acoustic information from the segmented vowels, namely the following acoustic properties were measured:

i. Vowel formants. Formant frequencies from *F*1...*F*5 were measured at three different locations across a vowel: 25%, 50%, and 75% mark of vowel duration.

iii. Vowel duration. Vowel duration was measured from the onset to the offset of the *F*1 and *F*2 format frequencies.

iv. Fundamental frequency. (*F*0). We calculated the mean *F*0, minimum *F*0, and maximum *F*0 for each vowel production. *F*0 calculation was conducted using pitch detection algorithm implemented in Praat[11].

v. H1–H2, H1–A1, H1–A2, H1–A3. Harmonic and spectral amplitude measures were extracted from the vowels.

We conducted acoustic analysis for frequency determination using Praat's standard algorithm for pitch detection and formant frequency identification[11]. Overall, we employed the following 40 predictors: vowel duration, pause duration, *F*1 … *F*5 measured at three locations inside the vowel at the 25%, 50%, and 75% mark of vowel total duration, voice quality features (H1-H2, H1-A1, H1-A2, H1-A3), measures of *F*0 (Minimum *F*0, Mean *F*0, Maximum *F*0).

Pipeline 3: A third pipeline was employed for conducting the automatic linguistic analysis. We employed the Natural Language Toolkit python library[12]. Measurements of characters, words, characters per word, etc., were calculated from the tokenized and parsed output, and the proportion of parts of speech: nouns, verbs, adjectives, adverbs, pronouns, and the ratio of each part of speech per total number of words were calculated, i.e., the noun-verb ratio, noun-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

adjective ratio, noun-adverb ratio, noun-pronoun ratio, verb-adjective ratio, verb-adverb ratio, verb-pronoun ratio, adjective-adverb ratio, adjective-pronoun ratio, and adverb-pronoun ratio. The outputs of the three pipelines were combined into a single comma-separated values (CSV) file that served as input for the machine learning models.

Deep Neural Network Architecture

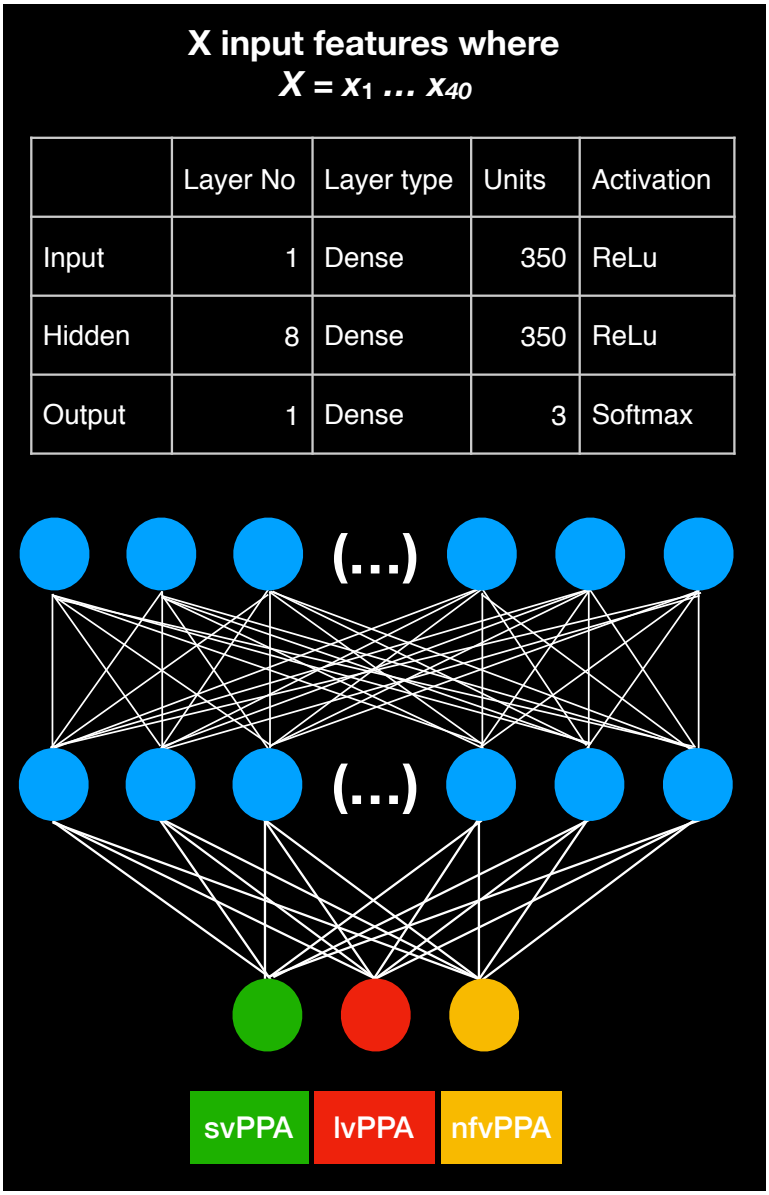
We randomized the data and then we standardized them using the StandardScaler function from Scikit-learn[13], which standardized the features by removing the mean and scaling to unit variance:

$$y = (x - \mu)/\sigma \tag{1}$$

where μ is the mean of the training samples; σ is the standard deviation of the training samples.

We standardized the training data and the test data separately to ensure that there is no information from the test set in the training set, which occurs when training and test data are transformed together. Standardization was shown to improve machine learning models[14].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.



	Layer No	Layer type	Units	Activation
Input	1	Dense	350	ReLu
Hidden	8	Dense	350	ReLu
Output	1	Dense	3	Softmax

Figure 1. **Neural Network Architecture.** Structure of the neural network designed for the study and feature properties, including the number of input features employed, the type and number of units and activation functions for the input, hidden, and output layer.

A feed-forward neural network (DNN) was designed for the classification (see Figure 1). The DNN presented here constitutes the final model, after tuning parameters. The DNN consists of an input layer with 350 dense units; the activation was a Rectified Linear Unit (ReLu). There were seven hidden layers all with 350 dense units and ReLu activation functions. The output

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

layer had 3 units and softmax activation function, which enables the classification of the three variants[14]. The ReLU activation computes the function $f(x) = max(0, x)$ and has the advantage that it computes and converges faster than other activation functions [15].

We compiled the model using a Root Mean Square Propagation (RMSProp) optimizer. In RMSProp, the learning rate was adapted for each of the parameters. This optimizer relies on dividing the learning rate for a weight by a running average of the magnitudes of recent gradients for that weight, its' mean square. Overall, RMSProp displays outstanding adaptation of learning rate.

The loss function was set to categorical cross-entropy. A higher value for the loss function implied a greater error for the predictions of the model. We fitted the network batch size set to 32. Within each fold, the neural net was trained for 30 epochs, using validation split to 20%.

Comparison support vector machines, random forests, and decision trees.

We compared three machine learning models to the DNN, which were selected because they are often employed in medical studies [16]: support vector machines (SVM) [17], random forests (RF) [18], and decision trees (DT).

i. DTs classify two categories by splitting the data using the best marker that can account for the data. One big advantage of DTs is that the trees can be visualized and provide an understanding of the structure of the data, and the exact decisions that are made by the model are overt and clear. Nevertheless, DTs are often prone to overfitting as they create long and complicated trees that generalize very well to unknown data. SVMs classify data categories using hyperspaces that best separate the classes. One advantage of SVMs is that they can provide good results with high dimensional spaces, which is often the case with acoustic data. SVMs can employ

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

both linear and non-linear Kernels for decision function. One disadvantage of SVMs is that the optimization of their hyperparameters can be complex and time consuming.

ii. RFs are similar to DTs, but unlike DTs they are ensemble models, i.e., they fit several DTs on the acoustic samples collected, and then they consider the mean of all trees to improve the accuracy of the model. RFs can address the overfitting that often takes place in the case of DTs.

Model optimization and hyperparameter tuning

For the selection of the final neural network architecture, we tested several neural network architectures by varying both the number of hidden layers, the number of units per layer, the dropout [19], the activation methods, and the batch size. DT models are provided here as a comparison model and their output is reported without optimizations. We evaluated the SVMs models with both linear and non-linear kernels and optimized the models for the number of kernels by running the SVM models with 1 - 300 kernels. The SVM model contains 14 non-linear kernels, which provided the best results in SVM optimization. We evaluated the RF models by optimizing for the number of trees from 1 - 300 trees. The best RF model has been the one with 14 trees. Note that the minimum split number was set to two.

Model comparison and evaluation

We compared the performance of the models using *eight-fold* grouped cross-validation. This validation method splits the randomized data into eight folds and trains and evaluates the machine learning models eight times. In the grouped cross-validation that we employed, the participant was set as the grouping factor; this ensures that, when data are randomized for splitting, there are always different participants in the training and test sets. For every training session, *seven folds* were employed for training and *one fold* was employed as an evaluation

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

set; so, the machine learning models were trained on different folds of data in the training set and evaluated on different test data from unknown participants during evaluation.

To evaluate the models, we employed the following metrics: *accuracy, precision,* and *recall*. *Accuracy* is the total sum of correct predictions divided by the total number of both correct and incorrect predictions:

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + \ false\ positive + \ false\ negative} \quad (2)$$

The true positive and true negative are the outcomes where the machine learning model correctly predicts the positive and negative class correspondingly. Also, a false positive or a false negative is an outcome where the model predicts the positive and negative classes incorrectly. Precision is the result of division of the true positives with the sum of true positives and false positives (see formula 3). Recall (a.k.a., sensitivity) is the result of dividing the true positives with the sum of true positives and true negatives (see formula 4).

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (3)$$

$$Precision = \frac{true\ positive}{true\ positive + false\ negative} \quad (4)$$

Finally, the *F1 score* is the weighted average of the precision and recall, and ranges between 0 and 1. The *F1 score* can offer a more balanced estimate of the outcome than the accuracy.

$$F1\ score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (5)$$

All models were implemented in Keras [20] running on top of TensorFlow [21] in Python 3.6.1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Comparison to Human Raters

Three trained speech-language pathologists were asked to provide the PPA variant of 9 patients—three from each variant—by listening to their Cookie Theft productions. The 9 participants were the same that we were employed for the evaluation of the machine learning model. No information was provided about the task, such as the aims of the task and how many sounds correspond to each variant. The clinicians had not previous interaction with the patients in the recordings. The recordings were provided in random order. To estimate the accuracy of their responses, we compared their responses to the information about the PPA variant from the clinical subtyping that employed the full battery of neurophysiological tests and imaging.

Data Availability

Anonymized data will be deposited in the ClinicalTrials.gov; identifier: NCT02606422.

RESULTS

The Cookie Theft picture description recordings were analyzed to elicit measures of speech and language from patients with PPA, then these measures were employed to train a Deep Neural Network and also three other machine learning models, namely a Random Forest, a Support Vector Machine, and a Decision Tree, that aim to provide comparative results for estimating the performance of the DNN. All machine learning models were trained and evaluated using an 8-fold cross-validation method. Table 2 shows the results from the 8-fold cross-validation method. Overall, the neural network model provided 80% classification accuracy and outperformed the other three machine learning methods that were employed for model comparison. That is, Random Forests (RFs) provided a 58% classification accuracy (see in Table 4 panel b), followed by the Decision Tree model (DT) with 57% classification

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

accuracy (see in Table 4 panel c). The Support Vector Machines had the worst performance in the cross-validation task with 45% classification accuracy (see in Table 4 panel a).

Table 2 Results from eight-fold cross-validation for the Deep Neural Network (DNN), Support Vector Machines (SVM), Random Forest (RF), and Decision Tree (DT). Shown is the mean cross-validation accuracy, the 95% Confidence Intervals (95% CI) and the standard error (SE).

Model	Mean	95% CI	SE
DNN	**80**	**[53, 100]**	**11**
SVM	45	[31, 59]	5
RF	58	[43, 73]	8
DT	57	[38, 75]	8

The confusion matrix shown in Table 3 and 4 was calculated by summing the 8 confusion matrices produced during cross-validation for the DNN. The neural network provided improved identification of patients with lvPPA and nfvPPA with respect to svPPA. The patients with lvPPA were identified 95% correctly; 5% of patients with lvPPA were identified as nfvPPA. Patients with svPPA was identified correctly as svPPA in 65% of the cases, 30% of a sample from patients with svPPA were misclassified as lvPPA, and 6% as nfvPPA; 90% of patients with nfvPPA were correctly identified and 10% were classified as svPPA.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

		Predicted Class		
		svPPA	lvPPA	nfvPPA
True Class	svPPA	64	30	6
	lvPPA	-	95	5
	nfvPPA	10	-	90

Table 3 DNN normalized confusion matrix created by summing scores across cross-validation scores from the 8-fold cross-validation test, showing the predicted vs. actual values from the DNN.

(a) SVM

		Predicted Class		
		svPPA	lvPPA	nfvPPA
True Class	SVM	5	10	3
	RF	13	12	6
	DT	12	16	23

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

(b) RF

Predicted Class

True Class	svPPA	lvPPA	nfvPPA
SVM	19	13	3
RF	10	10	6
DT	3	13	22

(c) DT

Predicted Class

True Class	svPPA	lvPPA	nfvPPA
SVM	15	14	1
RF	14	14	12
DT	1	9	19

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4 Normalized confusion matrix created by summing scores across cross-validation scores from the 8-fold cross-validation tests for SVM (Panel a), RF (Panel b), and DT (Panel c); matrices show the predicted vs. actual values from the evaluation of SVM (Panel a), RF (Panel b), and DT (Panel c).

To estimate the performance of the DNN, we also compared its accuracy with the classification performance of three trained speech-language pathologists who were blind to the gold standard diagnoses. Their responses were compared to the gold standard combined subtyping that employs neurophysiological tests, imaging, language evaluation, etc. Clinicians displayed significant variation in their classification scores of patients' variants with mean 67% (SD= 11). The lowest classification was just above average (5/9) 56%, followed by (6/9) 66%, and the highest classification reached (7/9) 77.77%. Overall, using the same evaluation data the DNN provided more accurate results than the clinicians.

DISCUSSION

Manual subtyping of patients with PPA is time-consuming and requires a high degree of expertise on PPA subtyping, costly scans, and lengthy evaluations. In this study, neural networks were trained on acoustic and linguistic predictors derived from descriptive-speech samples from three PPA variants. The gold-standard classification of PPA patients was based on expert clinical and neuropsychological examination, MRI imaging, and speech and language evaluations following the consensus criteria by Gorno-Tempini et al., 2011[1]. All models were trained 8 times in an eight-fold cross-validation. The output of the DNN was compared with the performance of three other machine learning models, namely Random Forests, Decision Trees, and Support Vector Machines, as well as with human auditory classification. The DNN achieved an 80% classification accuracy and outperformed the three

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

other machine learning models. In short, we showed that combining acoustic and linguistic information from a short picture description, the automated machine-learning model achieved a high classification accuracy of the PPA variant (80% correct including the difficult lvPPA variant) compared to the gold standard (expert clinician's diagnosis after neurological, neurolingistic and imaging evaluation). Importantly, the model outperformed clinicians when provided the same information (only Cookie Theft picture descriptions). These results illustrate three important conclusions: (a) a minimal amount of acoustic and linguistic information from connected speech has great discriminatory ability, providing an identification fingerprint of the PPA variants when used in a DNN model, (b) the DNN can simultaneously perform classification of all three PPA variants, and (c) the present automated end-to-end program may significantly help both the expert clinician by confirming the variant diagnosis as well as the novice or less expert clinician by guiding the variant diagnosis.

An unexpected finding was the improved classification results for the patients with lvPPA, as the DNN model performed better than other machine learning models for lvPPA, such as employed by Hoffman et al. (2017) and Maruta et al. (2017). Hoffman et al. (2017) used unsupervised classification methods and analyzed results from linguistic (e.g., hesitations, phonological errors, picture-naming scores, single-word comprehension, category fluency scores, written competence) and non-linguistic (cube analysis, paired associate learning, etc.) neuropsychological evaluations and found that participants with lvPPA were not identified as a separate group but were mixed with other participants in both linguistic and non-linguistic tasks[22]. Another study by Maruta et al. using a combination of measures from language and neurophysiological assessments in Portuguese discriminates individuals with svPPA from nfvPPA but not individuals with nfvPPA and svPPA from lvPPA[23].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Another important finding was that the DNN machine learning model provides superior results compared to human auditory classification. Specifically, three trained speech-language pathologists were asked to provide a classification by listening to Cookie Theft productions. Clinicians listened to the same recordings that were used for training the network and scored lower than the DNN. Also, the clinicians differed considerably in their judgments and often had to listen multiple times to the recordings to provide a judgment about the variant. The clinician with the highest classification accuracy reported that she had to listen several times for the speakers who had "mild effects." Overall, human PPA subtyping based on the same information provided to the machine learning model (Cookie Theft descriptions) can be a very difficult task for clinicians, including SLPs, because of different training and experience levels and the use of different criteria for PPA subtyping. Further, speech samples are often hard to hear in PPA patients requiring frequent playback and attentive listening.

The limited amount of data used could be considered as the main limitation of our model. Although 44 patients with PPA is a very substantial number for a rare syndrome such as PPA, increasing the training data will enable the neural network to identify patterns between acoustic and linguistic predictors that characterize each variant with increased confidence. In fact, during the prefinal evaluation of machine learning models, it became evident that the amount of data in the training set has a significant impact on model accuracy. So, by increasing the overall data sample and obtaining data from more patients, as we will make the code available to the community for clinicians' use, our model's performance will most likely be increasingly higher than the current overall accuracy. A second limitation is inherent to the task used for eliciting connected speech samples, i.e., the Cookie Theft picture description. This task constrains speech production both acoustically and with respect to the required grammar. Speakers provide primarily declarative intonational patterns, whereas questions, commands,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

etc. are not elicited. Another common criticism of picture-description tasks is that they are inclined to elicit labelling, actions in the present tense, and sentences with factual content rather than wishes, commands, embedded sentences and other more complex structures. By contrast, other tasks, such as personal story telling, conversation, etc. have the potential to provide richer speech and language output that can enable an improved classification of PPA variants. Future classification work is likely to benefit from the employment of machine learning models that aim to offer simultaneous classification of PPA variants using multifactorial predictors from a variety of discourse settings and conversations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendices

Appendix 1 Authors

Name	Location	Contribution
Charalambos Themistocleous, Ph.D.	Johns Hopkins School of Medicine, Baltimore, MD 21287, USA	Conceptualized study; designed the machine learning models, analyzed the data, and drafted the manuscript for intellectual content. Approved the final version.
Bronte Ficek, M.H.S.	Johns Hopkins School of Medicine, Baltimore, MD 21287, USA	Major role in the acquisition of data; revised the manuscript. Approved the final version.
Kimberly Webster, M.A., M.S.	Johns Hopkins University, Baltimore	Major role in the acquisition of data; revised the manuscript. Approved the final version.
Dirk-Bart den Ouden, Ph.D.	Arnold School of Public Health, University of South Carolina, Columbia, SC, USA	Revised the manuscript for intellectual content. Approved the final version.
Argye E. Hillis, M.D.	Johns Hopkins School of Medicine, Baltimore, MD 21287, USA	Revised the manuscript for intellectual content. Approved the final version.
Kyrana Tsapkini, Ph.D.	Johns Hopkins School of Medicine, Baltimore, MD 21287, USA	Conceptualized study; led the study trial where the data were acquired and the data collection process. Approved the final version.

References

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Gorno-Tempini, M. L. *et al.* Classification of primary progressive aphasia and its variants. *Neurology* **76**, 1006–1014 (2011).
2. Mesulam, M. M. Slowly progressive aphasia without generalized dementia. *Ann. Neurol.* **11**, 592--598 (1982).
3. Mesulam, M. *et al.* Quantitative template for subtyping primary progressive aphasia. *Arch. Neurol.* **66**, 1545–51 (2009).
4. Fraser, K. C. *et al.* Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex* **55**, 43--60 (2014).
5. Wilson, S. M. *et al.* Connected speech production in three variants of primary progressive aphasia. *Brain* **133**, 2069–2088 (2010).
6. Morris, J. C. The Clinical Dementia Rating (CDR): Current version and. *Young* **41**, 1588–1592 (1991).
7. Knopman, D. S. *et al.* Development of methodology for conducting clinical trials in frontotemporal lobar degeneration. *Brain* **131**, 2957–68 (2008).
8. Goodglass, H. & Kaplan, E. *Boston diagnostic aphasia examination (BDAE)*. vol. null (1983).
9. Themistocleous, C. & Kokkinakis, D. THEMIS-SV: Automatic clas- sification of language disorders from speech signals. in.
10. Themistocleous, C., Eckerstrom, M. & Kokkinakis, D. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Front. Neurol.* **9**, 975 (2018).
11. Boersma, P. & Weenink, D. Praat: doing phonetics by computer. (2016).
12. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. (O'Reilly Media, Inc., 2009).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.04.025593. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

14. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).

15. M. D. Zeiler *et al.* On rectified linear units for speech processing. in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* 3517–3521 (2013). doi:10.1109/ICASSP.2013.6638312.

16. Maroco, J. *et al.* Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res. Notes* **4**, 299 (2011).

17. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, (1995).

18. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

20. Chollet, F. & others. *Keras*. (GitHub, 2015).

21. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* **abs/1603.04467**, (2016).

22. Hoffman, P., Sajjadi, S. A., Patterson, K. & Nestor, P. J. Data-driven classification of patients with primary progressive aphasia. *Brain Lang.* **174**, 86–93 (2017).

23. Maruta, C., Maroco, J., de Mendonça, A. & Guerreiro, M. Behavior Symptoms in Primary Progressive Aphasia Variants. in *Neuropsychiatric Symptoms of Cognitive Impairment and Dementia* 27–43 (Springer, 2017).