Parkinson's Disease Detection Based on Running Speech Data From Phone Calls

Christos Laganas, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sofia B. Dias, Sevasti Bostantzopoulou, Zoe Katsarou, Lisa Klingelhoefer, Heinz Reichmann, Dhaval Trivedi, K. Ray Chaudhuri, and Leontios J. Hadjileontiadis, *Senior Member, IEEE*

***Abstract*— *Objective*: Parkinson's Disease (PD) is a progressive neurodegenerative disorder, manifesting with subtle early signs, which, often hinder timely and early diagnosis and treatment. The development of accessible, technology-based methods for longitudinal PD symptoms tracking in daily living, offers the potential for transforming disease assessment and accelerating diagnosis. *Methods*: A privacy-aware method for classifying patients and healthy controls (HC), on the grounds of speech impairment present in PD, is proposed. Voice features from running speech signals were extracted from passively-recordings over voice calls. Language-aware training of multiple- and single-instance learning classifiers was employed to fuse and predict on voice features and demographic data from a multilingual cohort of 498 subjects (392/106 self-reported HC/PD patients). *Results*: By means of leave-one-subject-out cross-validation, the best-performing models yielded 0.69/0.68/0.63/0.83 area under the Receiver Operating Characteristic curve (AUC) for the binary classification of PD patient vs. HC in sub-cohorts of English/Greek/German/Portuguese-speaking subjects, respectively. Out-of sample testing of the best performing models was conducted in an additional dataset, generated by 63 clinically-assessed subjects (24/39 HC/early PD patients). Testing has resulted in 0.84/0.93/0.83 AUC for the English/Greek/German-speaking sub-cohorts, respectively. *Conclusions*: The proposed approach outperforms other methods proposed for language-aware PD detection considering the ecological validity of the voice data. *Significance*: This paper introduces for the first time a high-frequency, privacy-aware and unobtrusive PD screening tool based on analysis of voice samples captured during routine phone calls.**

***Index Terms*— Parkinson's Disease, digital biomarkers, voice impairment, machine learning, speech processing**

Manuscript received September 13, 2020; revised March 8, 2021, July 7, 2021 and September 15, 2021; accepted September 22, 2021. Date of publication xxxx x, 2021; date of current version September 28, 2021. This work has been funded by the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement No 690494—i-PROGNOSIS: Intelligent Parkinson early detection guiding novel supportive interventions. *(Christos Laganas and and Dimitrios Iakovakis contributed equally to this work.) (Corresponding author: Leontios J. Hadjileontiadis.)*

Christos Laganas, Dimitrios Iakovakis, Stelios Hadjidimitriou, and Vasileios Charisis are with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR 54124 Thessaloniki, Greece (e-mail: chr.laganas@gmail.com; dimiiako12@gmail.com; stelios.hadjidimitriou@gmail.com; vcharisis@ee.auth.gr)

Sofia B. Dias is with the Faculty of Human Kinetics, University of Lisbon, 1495-688 Cruz Quebrada, Lisbon, Portugal (e-mail: sbalula@fmh.ulisboa.pt)

Sevasti Bostantzopoulou is with the 3rd Department of Neurology, G. Papanikolaou Hospital, School of Medicine, Aristotle University of Thessaloniki, GR 570 10 Thessaloniki, Greece (e-mail: bostkamb@otenet.gr).

Zoe Katsarou is with the Department of Neurology, Hippokration Hospital, GR 54642 Thessaloniki, Greece (e-mail: katsarouzoe@gmail.com).

Lisa Klingelhoefer, and Heinz Reichmann, are with the Department of Neurology, Technical University of Dresden, 01307 Dresden, Germany (e-mail: lisa.klingelhoefer@uniklinikum-dresden.de; Heinz.Reichmann@uniklinikum-dresden.de)

Dhaval Trivedi and K. Ray Chaudhuri, are with the King's College Hospital NHS Foundation Trust, SE5 9RS London, United Kingdom (e-mail: dhaval.trivedi1@nhs.net, Ray.chaudhuri@nhs.net).

Leontios J. Hadjileontiadis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR 54124 Thessaloniki, Greece and also with the Department of Electrical Engineering and Computer Science/Biomedical Engineering, Khalifa University, 127788 Abu Dhabi, UAE (e-mail: leontios@auth.gr; leontios.hadjileontiadis@ku.ac.ae)

I. Introduction

PARKINSON'S Disease (PD) is a progressive neurodegenerative disorder of high prevalence rate, with 1% of people above the age of 60 being affected [1]. PD mainly impacts people over 50 years old and, considering the continuously growing population, the number of affected individuals will only increase [2]. The depletion of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies and accumulations of alpha-synuclein protein are the main pathological hallmarks of PD [3]. PD manifests with motor symptoms, including tremor, bradykinesia, muscle rigidity and speech impairment, as well as non-motor signs, i.e., sleep disorders, cognitive impairment and constipation [4], [5]. Timely diagnosis of PD is often delayed, since symptoms are subtle at the early stages [6], [7] and their assessment usually requires an in-clinic evaluation of the subject's condition by a movement disorders expert. The latter usually takes place based on standardized scales and questionnaires, such as the Unified Parkinson's Disease Rating Scale (UPDRS) [8]. In particular, UPDRS Part III involves 14 items, dedicated to the assessment of the subject's motor performance [9]. Overall, the standard medical practice regarding PD diagnosis is of subjective nature; its effectiveness depends on years of expertise and its accuracy is considered low [10]. Thus, objectivity is required to improve the diagnostic procedure and, consequently, improve treatment outcomes [11].

Information and Communication Technology (ICT)-based solutions have provided a wide variety of tools to quantify the motor status of PD patients. In recent years, human mobile interaction is part of everyday life, yielding large-scale data streams. Mobile devices with their built-in sensors have been exploited for capturing and extracting information related to PD, and the biggest smartphone-based PD study for PD [12], recruited participants remotely, to self-report their demographics and perform structured tests for collecting data over a six-month period. The study initiated a promising remote data collection pathway with 9000 participants, the revealed associated requirement of users' active data contribution resulted in high dropout rates ($\geq$90%). Prominent examples of passive capturing of the data with high data fidelity, have used the touchscreen input [13] during typing activity or the accelerometer signals [14] during voice calls, to estimate motor-related impairment of PD in the context of the i-PROGNOSIS study [15] (`http://www.i-prognosis.eu/`). Another example of smartphone technology is [16], where authors have introduced a mobile application that can be used for evaluation and monitoring of motor impairments in PD patients through task-based speech, gait and hand movements. Preliminary results show that most of the evaluated features indicate significant differences between PD patients and healthy controls, paving the way for further research of smartphone-based applications in evaluation of PD.

The speech flaws that PD causes consist of reduced intensity of voice [17], [18], monopitch [19] and incorrect articulation of consonants [20], among several others [21]. Furthermore, 90% of PD patients face voice impairment [22]. This has made the acoustic analysis of speech signals for early PD detection an intensive research area. Previous methods has mainly focused on the vocal impairment estimation using sustained vowel phonations; 33 PD and 10 Healthy Controls (HC) were classified based on analysis of the phonation /a/ [18], while 100 subjects (50 PD and 50 HC) were classified based on the analysis of /i/ [23]. Running speech analysis from data (155 PD and 150 HC) captured in-the-clinic [24] has demonstrated a performance of 98% accuracy for PD patients classification versus HC. The features used relate with the vocal fold vibration changes seen between voice (i.e., vibrating vocal folds) and unvoiced (i.e., non-vibrating vocal folds) sounds, as reflected in the voice frequency content. The strong research contribution on the field and the high accuracy results presented so far in the lab-setting is a solid stepping stone towards a remote speech-based PD marker. However, reaching to the ecologically valid capturing of vocal flaws caused by PD in running-speech across different spoken languages, genders and ages make the running speech analysis challenging in a real life PD detection scenario.

So far, machine learning methods have demonstrated the potential for early PD prognosis, based on passive data collection through dedicated apps (e.g., iPrognosis app `https://play.google.com/store/apps/details?id=com.iprognosis.gdatasuite`). Specifically, touchscreen typing pattern analysis has been used for early detection of fine-motor impairment [13], whereas accelerometer-based analysis was employed for tremor detection [14], expressing the effect of PD to upper-extremity motor function of PD patients. Use of speech recordings through smartphone devices of participants enrolled in conducted studies have been used on [25], [26], where authors used a feature-space that represents the key aspects of hypokinetic dysarthria in the early stages of PD, indicating that early screening is possible via acoustic segment analysis. The current work acts as an extension to the previous approaches by focusing on voice-related passive data collection captured during phone calls via the iPrognosis app. The design of the latter is enhanced to capture 75 seconds of running speech and provide on-device voice-based features that are used in the machine learning framework. An end-to-end system is introduced for the first time that effectively combines: (a) a scalable smartphone-based capturing system during voice calls, used during a remote human study for voice data collection; (b) privacy-aware processing and effective fusion of voice features with demographic variables, and (c) language aware machine learning models that allow voice-based early PD detection from data captured in-the-wild.

The rest of the paper is organized as follows: Section II describes the datasets used for training and validation of the adopted methodologies and experiments. Section III presents the results obtained, while Section IV discusses the findings. Finally, Section V concludes the paper.

II. Materials and Methods

The overall concept of the study pipeline is depicted in Fig. 1. Initially, voice calls signals (GData in Fig. 1(a)) are passively captured by the smartphone's microphone via the iPrognosis app (`https://play.google.com/store/apps/details?id=com.iprognosis.gdatasuite`). Then, voice-related features are locally extracted on each subjects' smartphone and used for the training of language-aware models that classify PD vs. HC (Fig. 1(b)). Finally, the trained models are further tested on a dataset still captured in-the-wild (SData in Fig. 1(c)), yet clinically evaluated, providing, thus, labelling ground truth by the physicians. In the context of the current work, sensitivity measures the proportion of actual PD patients that are correctly identified as having the condition, whereas specificity measures the proportion of actual healthy controls who are correctly identified as not having the condition. Analytical description of the employed steps follows.

A. In-the-Wild Data Acquisition

The voice data captured in-the-wild were collected from subjects coming from seven countries across EU, who remotely contributed de-identified data via the iPrognosis app. During the installation process of the app, it provided information regarding the study details. Moreover, an electronic informed consent was obtained from the subjects enrolled via digital signing. Subjects held the right to withdraw from the procedure at any time via the available option within the application with the option to delete their collected data so far.

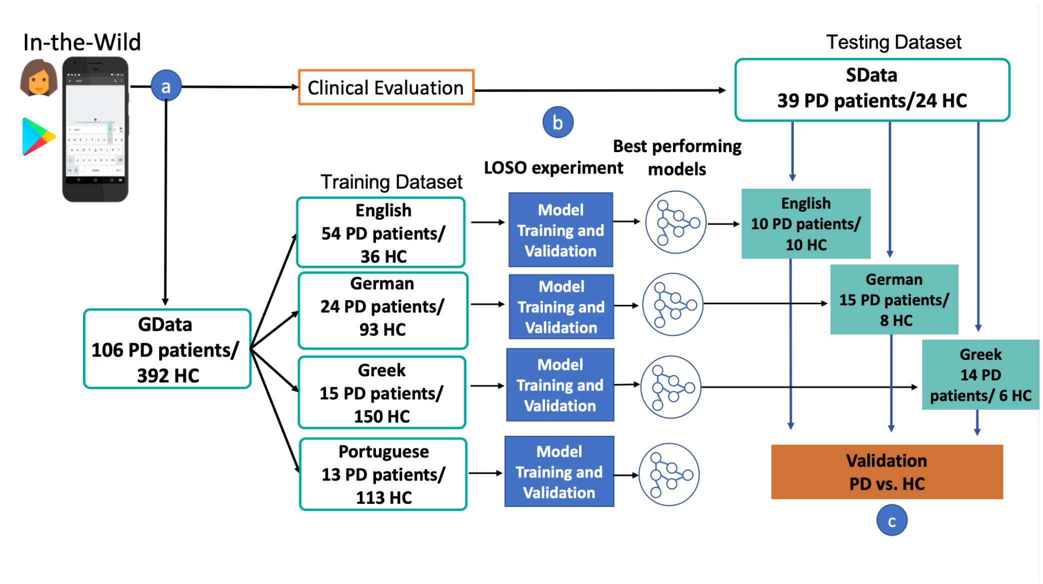


Fig. 1. Overall pipeline for the development and evaluation of the proposed method. (a) Voice features collected in-the-wild during phone calls via a remote data collection study: data (GData) from 498 subjects, who self-reported their health status, are used for training, whereas a separate set of 63 clinically-assessed subjects (39 PD patients/24 HC) and their voice features are used for testing (SData). Note that the training (GData) and testing (SData) cohorts are totally independent of each other. (b) The complete training dataset is split based on subjects' native language and each language-specific dataset is used for training based on a Leave-One-Subject-Out (LOSO) cross validation scheme, in order to produce the best performing model per language. (c) The best performing models of stage (b) are then evaluated in terms of their ability to discriminate between Parkinson's Disease (PD) patients and Healthy Controls (HC), using the unseen SData set.

The iPrognosis application only recorded the subject's input from the smartphone microphone for consent purposes; hence the voice of the callee on the other side of the phone was not captured. All analyses were performed using the speech data from calling person only; no speech utterances from distant speakers were presented at final recordings used for acoustic analysis. Subjects were speaking over voice calls for the duration of their will and the application captured the first 15-75 seconds of their call in the background. Hence, no speech-source separation was needed. Each voice call with duration at least 15 seconds triggered the capturing process and the first S seconds ($S \in [15, 75]$s) of the related voice signal were acquired (see Section II-C) from the smartphone microphone and locally stored for feature extraction analysis. The 15s threshold was used to discard non-informative calls, whereas the 75s one was chosen as a trade-off between optimization of the smartphone memory usage and sufficient voice data windowing, informatively expressing the PD-related voice symptoms [27]. After the end of the feature extraction analysis, the raw voice data were deleted from the smartphone memory. Then, the extracted features were stored in a `.JSON` format file, and indexed as an entry to a local `SQLite` database. The accumulated entries were daily uploaded to a Microsoft Azure Cloud-based database, accompanied by a unique coded User-ID, to sustain data privacy and ensure General Data Protection Regulation [28] compliance. After the Cloud-based uploading, the `.JSON` file was deleted from the smartphone memory for efficiency purposes.

1) Ethics Approval: All the experimental and ethical protocols were approved by Ethik-Kommission an der Technischen Universität Dresden, Dresden, Germany (EK 44022017), Greece, Bioethics Committee of the Aristotle University of Thessaloniki Medical School, Thessaloniki, Greece (359/3.4.17), Portugal Conselho de Ética, Faculdade de Motricidade Humana, Lisbon, Portugal (CEFMH 17/2017), United Kingdom London, Dulwich Research Ethics Committee (17/LO/0909),Comité de Ética de la Investigación Biomédica de Andalucia, Spain (60854c5dbc58dda37b4730edb590a503edbd3572), Sunshine Coast Hospital and Health Service, Australia (41562 HREC/18/QPCH/266) and Studienzentrum der Prosenex AmbulatoriumbetriebsgesmbH an der Privatklinik Confraternitaet, Wien, Austria (002/2018).

B. Data Cohorts

For the purposes of this study, 29,048 voice calls (II-A) were gathered in-the-wild from 835 total users (both self-reported Healthy Controls (HC) and PD patients) via the iPrognosis app. Speech impairment has been suggested to produce subtle effects even five years prior to diagnosis [29], [30], thus PD patients that generated the test set were selected to examine the potential of the proposed method for the early detection of PD. The different cohorts used for the purpose of model training and testing, following specific cohort inclusion/exclusion criteria, as explained below.

1) Training Dataset (GData): The training dataset, namely General Data (GData) (see Fig. 1), consists of subjects who self-reported demographic characteristics (i.e., age, country, gender, PD patient or not). Recruitment of the subjects that generated the training set has taken place via a remote data crowdsourcing study. As a result, detailed clinical information could not be retrieved and only high-level labelling of subjects as PD or HC could be collected by the subjects themselves, providing the bare minimum information for training the binary classification algorithms. Subjects provided pseudoanonymized voice data remotely via the mobile application and those younger than 40 and older than 85 years were excluded from the training subset to remove the age-covariate. In total, the GData set includes 18,284 voice feature sets (mean/std recordings per subject: 36.71/72.74), from 498 users (392 HC, 106 PD). Specifically, 90 English subjects contributed 1,907 voice feature sets (21.19/32.7), 165 Greek subjects contributed 8,982 voice feature sets (54.44/88.19), 117 German subjects contributed 2,044 voice feature sets (17.47/24.75) and 126 Portuguese subjects contributed 5,351 voice feature sets (42.47/92.40). Supplementary Material (Suppl. Mat.) Table I summarizes the GData subjects' self-reported health status and demographic characteristics per country.

2) Testing Dataset (SData): For the purpose of testing, a subset of voice data from subjects that met the inclusion criteria of GData, yet were independent of the GData cohort, was formed, namely Specific Data (SData) (see Fig. 1). The two cohorts GData and SData are composed from separate speakers. SData dataset did not participate in any of GData training procedures making it totally independent of the GData cohort, in order to have unbiased results. In particular, SData consisted of data from 63 subjects (24 HC, 39 PD), all clinically assessed by specialised neurologists in three medical centers (Aristotle University of Thessaloniki Medical School, King's College London Hospital, Technischen Universitat Dresden). In total, 20 English subjects contributed 403 voice calls (20.2/22.9), 20 Greek subjects contributed 1857 voice calls (92.9/111.8) and 23 German subjects contributed 794 voice calls (34.5/43.5), respectively. Years since diagnosis and medication in the levodopa equivalent dose (LEDD) along with the UPDRS Part III Item 18 are tabulated, along with the other demographic and clinical characteristics for the SData group, in Suppl. Mat. Table II.

C. Data Conditioning

Prior to data feature extraction analysis, a data conditioning step was adopted. This pre-processing was used for: a) resampling, b) digital zero removal, c) Direct Current (DC) offset removal, and d) normalization of the voice signal. Modern android smartphones (OS >5.0) are equipped with noise-canceling microphones picking up sound from different directions. They have a primary microphone that faces speakers' mouth, and a secondary microphone that picks up any background noise and the other side speaker. iPrognosis application has been designed to capture the sounds from the microphone from call transmission that suppresses the interference from the other callee and environment noise. In particular, according to the Android Operating System (AOS) documentation, the $f_s = 44.1$ kHz sampling frequency that ensures high audio quality is not supported by all devices. As iPrognosis app is running in a variety of smartphone devices with at least AOS 5.0 and a minimum of $f_s = 16kHz$ (A/D PCM 16-bit), resampling of the captured voice data at $f_s = 16kHz$ was adopted where needed, ensuring wide device compatibility. As the main spectral content affected by PD is concentrated below 8kHz [31], the adopted resampling did not affect the underlying PD-related information of the voice data. Furthermore, digital zero removal was adopted to replace zero data values with Gaussian noise random variable ϵ $(0 < \epsilon \ll 10^{-3})$ to avoid divisions by zero and/or Not-a-Number exceptions in feature processing. Digital zeros can occur at the start and end of recordings, when the device provides audio buffers that are only filled partially with the sound pressure measurements. DC offset can cause erroneous calculations, e.g., estimation of the power of the audio signal, which would be distorted by inaudible low frequency signals with large amplitude. Usually, DC offset is removed by the sound card on the smartphone; yet, this is not always the case. To ensure this, a DC offset removal was applied by using a second order Infinite Impulse Response (IIR) highpass filter with a cutoff frequency of 10Hz. Finally, an amplitude normalisation within [-1,1] was performed, to compensate for the amplitude variability in the acquired voice data, due to their collection from heterogeneous smartphones.

D. Feature Extraction

The smartphone-based feature extraction process adopted here was motivated by previous works [24], [32]. At first, the fundamental frequency, also known as pitch was extracted. Then, the voice signal was divided into 40ms-long frames with a hop size of 10ms. For each windowed voice signal, a threshold of the normalized voice signal amplitude (± 0.05) was used to discard silent frames. In the remaining ones, the pitch ($F0$) was extracted using the auto-correlation combined with peak pruning [32]. If the estimated $F0$ lied outside [70,600] Hz, the candidate frame was discarded. Then, the time points of the onsets $\widehat{O(t)}$ (i.e., when the pitch starts to be active) and the offsets $\check{O}(t)$ (i.e., when the pitch starts to be inactive) were derived from the (smoothed) pitch curve. In fact, these time points correspond to signal locations, where the vocal fold vibration is started and stopped [24], respectively. Next, at the estimated onsets and offset locations, an overall time period of $\pm$40ms around each onset/offset location was selected. In the latter, two sets of well-known spectral descriptors were calculated for signal frames of length 25 ms window size with 10 ms hop size. Mel Frequency Cepstral Coefficients (MFCCs) [33] were extracted with a Hamming window size of 20ms for 26 filters, covering the frequency range of [0-8kHz]. Each filter in the filter bank is triangular, having a response of 1 at the center frequency, linearly decreasing towards 0 until it reaches the center frequencies of the two adjacent filters where the response is 0. To decorrelate the filter bank coefficients, the discrete cosine transform was applied and the first 13 MFCCs were used to characterize the spectral description, along with 22 Bark-band Energies (BBE) [34]. The mean (μ), standard deviation (σ), kurtosis ($kurt$) and skewness ($skew$) of the estimated $F0$, MFCCs and BBE across all windows were computed and the latter two are used as features. Furthermore, we have used gender and age as features of the training set to the priors as users self-reported. This was intended to capture the prevalence-related priors of PD. However, as the existence of such bias could skew the models' results, we formulated the

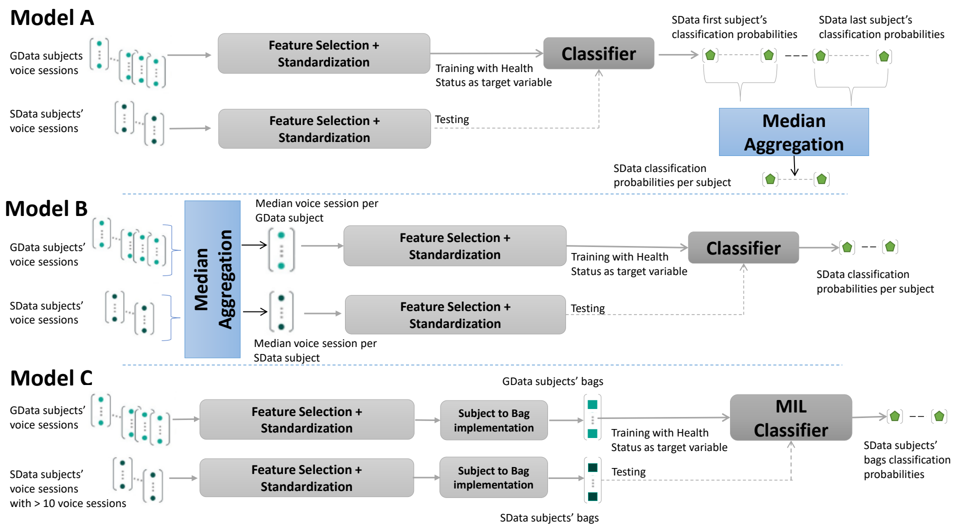

Model A
GData subjects voice sessions
Feature Selection + Standardization
Training with Health Status as target variable
Classifier
SData first subject's classification probabilities
SData last subject's classification probabilities
SData subjects' voice sessions
Feature Selection + Standardization
Testing
Median Aggregation
SData classification probabilities per subject
Model B
GData subjects' voice sessions
Median Aggregation
Median voice session per GData subject
Feature Selection + Standardization
Training with Health Status as target variable
Classifier
SData classification probabilities per subject
SData subjects' voice sessions
Feature Selection + Standardization
Testing
Median voice session per SData subject
Model C
GData subjects' bags
GData subjects' voice sessions
Feature Selection + Standardization
Subject to Bag implementation
Training with Health Status as target variable
MIL Classifier
SData subjects' bags classification probabilities
SData subjects' voice sessions with > 10 voice sessions
Feature Selection + Standardization
Subject to Bag implementation
Testing
SData subjects' bags

Fig. 2. Model A. Feature selection, classifier training and evaluation on a single-instance level (i.e., individual feature vectors of single voice recordings). Each subject is characterized by the median of the probabilities reflecting the prediction for the produced classification score. Model B. The median of the feature vectors per subject is computed before training and evaluation of the feature selection and classification pipeline on a subject-level. Model C. Multiple Instance Learning (MIL) pipeline is trained and evaluated on a bag of feature vectors for each subject.

test-scenarios in demographically-matched cohorts. Therefore, the total length of the feature vector was 2x4x13 (MFCC) + 2x4x22 (BBE) + 2 (metadata) = 282.

E. Proposed Models

Three different modeling approaches were examined here (see Fig. 2) to learn a language-aware PD vs. HC classification pipeline. In particular:

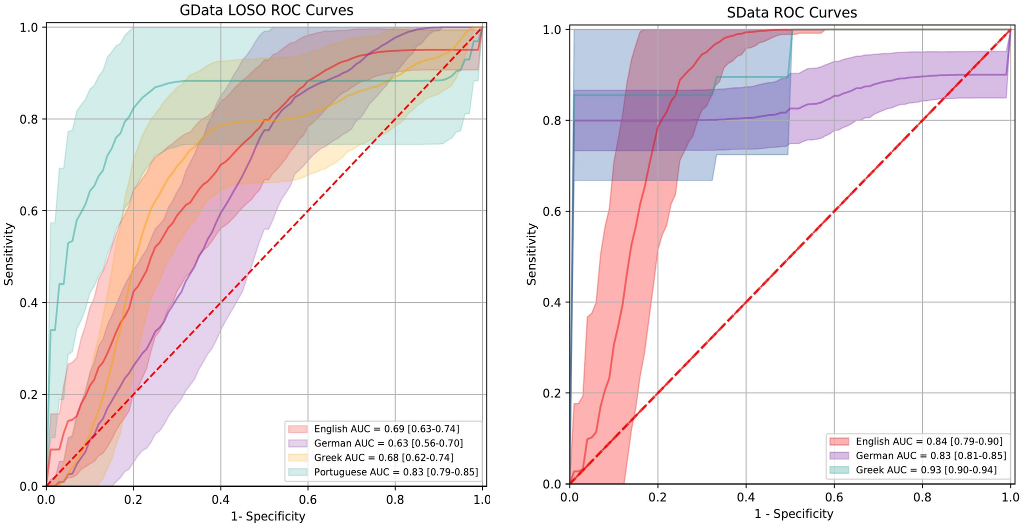
- Model A (see Fig. 2-top panel), involves feature selection with four feature ranking-selection methods, i.e., Lasso [35], Ridge [36], Gini Impurity [37] and ANOVA-based selection [38], and standardization process in the GData set, in order to extract the highest ranking features and reduce the dimension of the feature vector. Further fitting of the GData set in a single-instance classifier, i.e., Linear Support-Vector Machine [39], Logistic Regression [40] and Random Forest [41], was used and the subject under test was characterized by the median of the prediction probabilities.
- Model B (see Fig. 2-middle panel) initially involved the computation of the median feature vector for each subject. Then, the same feature selection process as in Model A was applied, and, finally, the single-instance classifiers, as in Model A, were trained. The hypothesis of representing each subject by the median of his/her voice features was adopted in order to minimize the effect of variations, due to different number of voice calls per subject.
- Model C (see Fig. 2-bottom panel) was based on Multiple Instance Learning (MIL) [42] machine learning algorithms, with the MIL classifiers under evaluation consisting of Normalized Set Kernel (NSK) [43], Statistic Kernel (STK) [43], sparse MIL (sMIL) [44] and multiple instance SVM (mi-SVM) [45]. In MIL the decision was made on multiple instances and not single ones. This is clearly suitable for the examined case, since not all voice calls capture equal amount of information for the PD detection. Specifically, the classifier receives a set of bags of feature vectors under the same label that corresponds to subject's self-reported label (HC or PD).

A LOSO cross validation scheme was used in the GData set for the different languages (see Fig. 1(b)), to identify which classification methods (Models A, B, C) performs better in terms of classifying PD vs. HC.Regarding the class imbalance problem that the training set has, a penalty has been applied to the cost function of each model. The cost term weights each class to be proportional to the inverse of its frequency, to avoid possible model skew towards one class in case of unbalanced classes, such as in GData (Suppl. Mat. Table I). The best performing models outputted from the LOSO cross validation were then tested on the SData set (see Fig. 1(c)). For the feature selection algorithms and classification models implementation, Python's scikit-learn library was used [46], [47].

Moreover, an evaluation about the possible transfer of the aforementioned language-aware modelling across different languages was examined. In this vein, cross-language model-testing was tested to evaluate the models' performance for classifying PD vs. HC subjects using other languages (e.g., training in the English dataset-tested in Greek and German datasets). For this cross-language testing scenario, the optimized models per language group were validated in the other language groups. The additional hypothesis that a single model (language-unaware) can effectively discriminate among PD vs. HC without language-specific modelling was tested as well.

F. Model Optimization

The results presented hereby were from the best performing classification pipelines as evaluated in the GData set, based on the LOSO cross validation scheme. In the training phase, prior to the classifiers' training and optimization procedures,



(a) (b)

Fig. 3. Receiver operating characteristic (ROC) curves of the best-performing language-specific models for (a) the Leave-One-Subject-Out (LOSO) cross validation experiment and the four different cohorts (i.e., English, German, Greek, and Portuguese), and (b) the testing on the three cohorts of clinically-assessed subjects (English, German, and Greek).

training subsets were subjected to three feature elimination techniques. Specifically, a tree-based feature selection technique is used to compute the impurity-based feature importances, which in turn are used to discard irrelevant features with the threshold being the mean feature importance value. The second feature selection technique is around Lasso[48] and Ridge[36] regularization with features whose coefficients are less than the mean coefficient value being discarded, while the third involves a grid search of the number of selected features used in the range of 1-100 using K-best features selection procedure based on ANOVA for selecting the most discriminant features. In this case, the feature selection process considers a score for each feature expressing ”how well this feature discriminates between two classes”. This is translated by measuring the distance between means of class distributions over the variance of each single class via the ANOVA feature selection process. An F-statistic, or F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples or the explained and unexplained variance by a statistical test, like ANOVA. The ANOVA method is a type of F-statistic referred to here as an ANOVA f-test [49]. The ANOVA feature selection technique was not used in any of the language groups models. For English and German language groups, tree-based feature selection techniques were ultimately used, where a Ridge regularization feature selection was performed for the Greek language group. The experiments were repeated for each set of the classifier hyperparameters. The range of the examined hyperparameters was: C $\in [10^{-4} : 10^{4}]$ for the linear kernel, C,$\gamma \in [10^{-3} : 10^{3}]$ for the Radial Basis Function (RBF) kernels of the Multiple Instance SVMs, and C $\in [10^{-4} : 10^{4}]$ for the Logistic Regression classifiers. For each classification method, grid search was performed via LOSO cross-validation for hyperparameter optimization, in order to ensure that the models were optimised.

G. Statistical Performance Evaluation

The two groups PD patients and HC are tested in terms of statistical significance. Chi-squared tests were adopted for testing the p-value of categorical demographic variables, whereas a two-sided Mann-Whitney U-Test [50] is used for the age variables. Statistically significant difference was set at the level of $p < 0.05$. For German and English SData language groups all demographic information is balanced between Healthy Controls and PD patients. For Greek SData test set, 10 random age-matched subsets are chosen and used as testing sets. The age-matched subsets consist of 85% of the original set. GData training groups are not demographically balanced, as the data were gathered from an in-the-wild environment with remote collection of data, happening through the user’s smartphone via the iPrognosis app.

III. Results

A. Hyperparameter Optimization Results

Results from the hyperparameter optimization process are depicted in Suppl. Mat. Table III. From the latter it is clear that Multiple instance SVM and Logistic Regression are optimized with C=100, showing a consistency of the model parameter setting across all languages.

B. Training and Evaluation (GData)-Related Results

Results from the initial LOSO experiment applied on GData are presented in Fig. 3(a). In the latter, the ROC curves of the best classification models per language are depicted, corresponding to AUC scores (95% Confidence Interval (CI)/best model) of: {English Cohort: AUC 0.69 (CI: 0.63-0.74/Model C-STK)}; {German Cohort: AUC 0.63 (CI: 0.56-0.70/Model C-NSK)}; {Greek Cohort: AUC 0.68 (CI: 0.62-0.74/Model A-Logistic Regression with Ridge)} and {Portuguese Cohort: AUC 0.83 (CI: 0.79-0.85/Model C-NSK)}. From these results

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2021.3116935, IEEE Transactions on Biomedical Engineering

it is clear that, in most cases, Model C was the most efficiently performing model, resulting in AUC $\geq$0.63. The sensitivity values of the aforementioned models at different specificity levels are tabulated in Table I. In all four cases gender and age variables were present in the feature vector that describe the best performing models.

C. Testing (SData)-Related Results

1) Language-Aware Results: The optimized models per language cohort from the training phase (see Section III-B) were tested in the SData set (Suppl. Mat. Table II). In the latter, the age-distribution of PD vs. HC in the Greek cohort, unlike all others, exhibited statistically significance difference (p<0.05). In response to that, the resulted classification metrics for the Greek cohort were extracted from 100 random bootstraps of age-matched subgroups. In Fig. 3(b), the ROC curves of the classification models are depicted with the corresponding AUC scores (CI) of: {English Cohort: AUC 0.84 (CI: 0.79-0.90)}; {German Cohort: AUC 0.83 (CI: 0.81-0.85}; and {Greek Cohort: AUC 0.93 (CI: 0.90-0.94}. When collating the predictions of the three cohorts, the overall performance was found equal to AUC=0.82 (CI:0.78-0.89). The sensitivity values, at different levels of specificity, are tabulated in Table I, where it is noticeable that even at the specificity level of 80%, sensitivity is $\geq$ 77% in all cohorts. Confusion matrices for all language groups, as well as for the whole SData group are tabulated in Suppl. Mat. Table IV. In the latter, results are tabulated for subjects that had more than one, as well as more than 10 voice sessions. Moreover, as Fig. 4 illustrates, the performance efficiency of the proposed models increases, as the number of sessions per subject under test (SData set) increases, as well. Clearly, both the AUC scores (Fig. 4(a)) and F-score (Fig. 4(b)) surpass the threshold of 0.80 when at least 50 sessions per subject have been accumulated. Beyond the cost-sensitive optimization process, we also tested a resampling strategy to mitigate the imbalance between the two populations. We randomly chose 50 balanced subsets from each language in the GData training set and were tested on the language appropriate SData test sets. This approach significantly reduced the total number of training subjects from 165/117/90 to 30/48/72 for the GR/GER/EN group, respectively. The effect of this reduction had impact on the features that were induced in the optimization problem, yielding a significant reduction in the mean AUC scores, i.e., 0.73 [0.68 - 0.79], 0.50 [0.47 -0.54], and 0.45 [0.40-0.50] for the Greek, German and English language groups, respectively

2) Cross-Language-Aware Results: The analysis of the cross-language testing of the proposed models, when training them with GData set from one cohort (i.e., {Greek (GR), English (EN), German (GE)}; Portuguese cohort was omitted from this analysis) and testing on SData set from the remaining two cohorts, has resulted in AUC values of {(GR,GE|EN): (0.70,0.54)}, {(GR,EN|GE): (0.72,0.38)}, and {(EN,GE|GR): (0.50,0.43)}, respectively.

3) Language-Unaware Results: When the language-unaware modelling approach was tested, the Model A outperformed the other two, resulting in 0.71 AUC score on discriminating HC from PD via the SData set testing. When compared with the results in Section III-C.1, it is clear that language consideration enhances the discrimination performance of the proposed approach.

TABLE I
ESTIMATED SENSITIVITY AT VARIOUS LEVELS OF SPECIFICITY FOR THE CASE OF LANGUAGE-AWARE ANALYSIS

	Specificity		
	0.70	0.80	0.90
GData			
English	0.59	0.41	0.21
German	0.42	0.27	0.10
Greek	0.69	0.46	0.12
Portuguese	0.85	0.80	0.56
SData			
English	0.94	0.77	0.31
German	0.79	0.80	0.80
Greek	0.86	0.86	0.86

IV. DISCUSSION

The current clinical practice for the characterization of PD presents with several limitations that highlight the clear need for objective measurement of symptoms. In the context of this study, an on-device based feature extraction process was adopted from the launch of the study, which limits the continuous generation of features over time. The extracted representation focused on the extraction of spectral description for the onset and offset of the speech, which was the state-of-the-art in running speech analysis for PD [24] when the study was designed. The latter language-unaware feature extraction assisted to have a uniform and privacy aware feature extraction process. Ranking the selected features per language group yielded high diversity, validating that language-aware analysis captures the subtle PD-related degradation. The novelty of the paper is reflected in the engineering for the first time of an end-to-end remote tool that is adapted to sustain a longitudinal monitoring tool that preserves privacy and long-term adherence. In the current study, authors contribute a method for analysis of running-speech data captured in-the-wild. It is evident that passive capturing sustains adherence, while the combination of the classification results of the voice-based analysis could be useful for a passive-based screening tool for early PD patients. This was enabled from a privacy-aware remote-data acquisition that has the potential to support diagnosis with objective measures, which is a limitation of the current clinical practice that heavily misses the timely early PD detection[10]. Moreover, in order to expand the feature extraction pipeline, inclusion of the cepstral separation in the speech processing part can improve the feature-representation space, in order to capture dysphonia and dysprosody in running speech [51].

The current study focused on a) privacy-aware feature-only transmission and b) light-weight processing for maximizing of long-term adherence. That being said authors employed a representation focused on the extraction of spectral description for the onset and offset of the speech, which was the state-of-the-art in running speech analysis for PD[24], [52] when the study was designed. Hence, the system-design criteria,

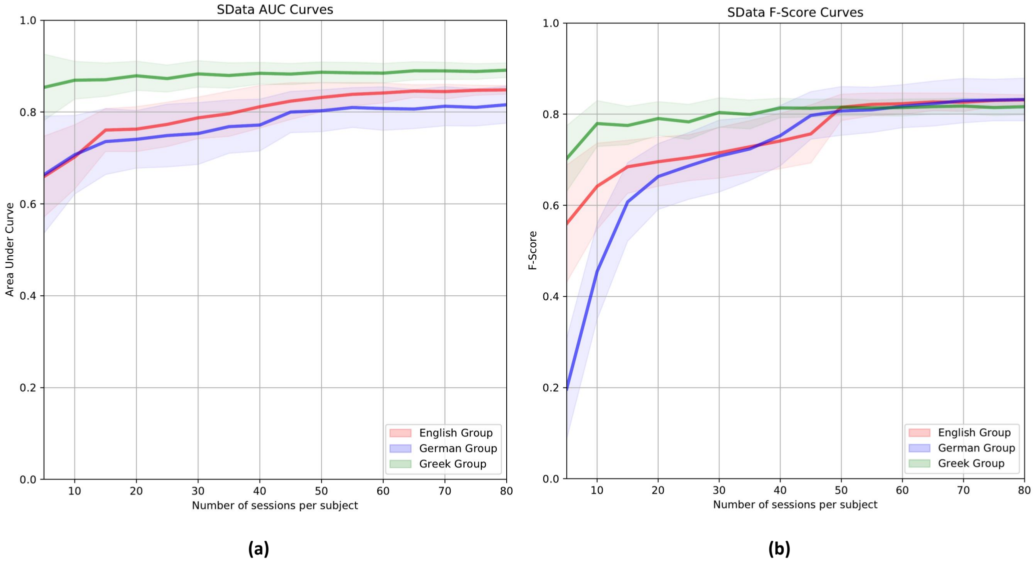


Fig. 4. Evolution of (a) AUC and (b) F-score with respect to the number of phone call session-extracted input feature vectors available to the models. Models are evaluated in three language-specific test sets by incrementing the number of input feature vector per subject. A clear improvement of the classification performance is observed as more voice-related input data become available for each test subject.

restricting raw data transmission, do not allow for novelty in the area of representation learning via deep neural networks, currently used in large-scale automatic speech recognition. On the contrary, having amalgamated the requirements of privacy and ethics via this engineering design for longitudinal behavioral data collection, the results show that for the first time a high-accuracy PD classification is feasible by processing of passive recordings from running speech segments with a machine learning model.

The motivation behind the use of MFCCs as features in the current study is prompted by their wide use on speech recognition systems, as well as being the primary voice feature of previous studies on the research field, such as in [52] and [24]. In addition to MFCCs, BBE are also extracted and used as features for the classification models. Since the data acquisition and feature extraction process were performed by the user's smartphone, the authors focused on privacy-by-design to perform this study, in order not to reflect any information related to content of spoken speech, which can be possibly done by extracting windows from all the speech segments captured. Future research that can combine federated learning approaches (that avoid data transfer) with more granular description of speech subsystems [53] could pave the way for a more descriptive model for running speech. In this work, we showcase that machine learning-based tools fusing demographic and voice features can be used for remote, in-the-wild PD detection for different spoken languages. The ecological validity of our method lies in the fact that users do not need to carry any additional hardware or perform a structured test, since data capturing takes place unobtrusively, during the natural use of the smartphone. In particular, data collection is automatically evoked during a phone call, preserving the user's privacy by extracting features locally on the smartphone and deleting the raw voice data afterwards. Remote research data crowdsourcing studies have the innate risk of data labelling inaccuracy due to the self-reporting processes involved, impacted further by concerns on data privacy, which may lead certain subjects not to report their true health status. In this vein, the recruitment process of the SData cohort has taken place by sampling from the pool of remote study participants from three different countries. Through this process, labels of SData participants who underwent detailed clinical assessment were 100% correct. The authors acknowledge that the same cannot be claimed for the labels reported by the GData subjects. Nevertheless, compliance with all related ethical guidelines and data protection regulations (GDPR[28]), as well as the reliability of EU-based dissemination channels employed for remote recruitment may have enhanced GData study participants' trust regarding the preservation of anonymity, data privacy and transparency of the processes, which in turn may have led to more accurate self-reporting. The latter can be a highlighting remark for future research data crowdsourcing studies, aiming at collecting longitudinal behavioural data. Finally, the number of subjects that were clinically assessed and used for evaluation could be potentially used for training of the severity of speech impairment. However, this approach was not followed here due to the limited size of clinical cohort/language and the noise in the clinically-evaluation process to assess speech impairment (this is shown in its low diagnostic properties over PD). However, this is left in future extension of this study, when more clinically evaluated subjects become available.

A. Comparative Analysis

All PD patients are considered early, with H/Y stages of 1 or 2 and an average total UPDRS Part III score (avg. 15.8/15.0/22.8 for the EN/GE/GR cohorts, respectively), representing motor status to be low as tabulated in Suppl. Mat. Table II. The univariate classification performance (early PD vs. HC) of the UPDRS Part III Item 18 score, representing the

TABLE II
COMPARATIVE ANALYSIS RESULTS

Language	UPDRS Item 18 AUC [CI]	Proposed approach AUC [CI]
English	0.43 [0.37-0.50]	**0.84 [0.79-0.90]**
German	0.79 [0.76-0.81]	**0.83 [0.81-0.85]**
Greek	0.67 [0.62-0.71]	**0.93 [0.90-0.94]**
All	0.66 [0.62-0.71]	**0.82 [0.78-0.89]**

Comparison of the univariate classification performance of the UPDRS Part III Item 18 score and that of the models of the proposed machine learning approach, per language-specific dataset. UPDRS: Unified Parkinson's Disease Rating Scale; CI: 95% Confidence Interval.

TABLE III
RESULTS ON SDATA SET FROM [24]

Language	Testing AUC (CI)	LOSO Testing AUC (CI)
English Onset	0.42 (0.37-0.48)	0.69 (0.64-0.76)
English Offset	0.42 (0.38-0.45)	0.73 (0.68-0.79)
German Onset	0.65 (0.59-0.69)	0.38 (0.31-0.45)
German Offset	0.44 (0.37-0.50)	0.25 (0.20-0.30)
Greek Onset	0.53 (0.44-0.62)	0.37 (0.33-0.41)
Greek Offset	0.58 (0.49-0.67)	0.33 (0.31-0.36)

Testing corresponds to GData set-based training and SData set-based testing, whereas LOSO Testing corresponds to SData set-based training and testing under the LOSO cross-validation scheme. AUC: Area Under the ROC Curve; CI: 95% Confidence Interval.

physician-based evaluation of speech impairment, is provided in Table II, along with the performance of the proposed machine learning approach. Results suggest that the subtleness of PD-related speech impairment may not be easily detected in the cohort, while the proposed approach yields better PD detection performance.

Vocal folds-related features and their connection to speech impairment in PD patients are reflected by the selected lower and higher order statistics of the MFCCs and Bark On-Off offsets, as presented before [52]. This is further explored here, as these features are used to train the proposed machine learning models with voice data obtained within a noisy environment met at a real life setting. For a comparative analysis between the proposed approach and the previous most promising running speech-based methodology regarding PD detection [52], a validation of the latter was performed on the SData set for all of its language cohorts. The results of such analysis are tabulated in Table III. There, two types of testing are reported, i.e., GData set-based training and SData set-based testing (second column) and SData set-based training and testing under the LOSO cross-validation scheme (third column). Apparently, by comparing the results presented in Sections III-B, III-C and in Table III, it is clear that the proposed approach exhibits higher performance than [52] in both testing scenarios. The enhanced classification approach over the previous methods is a result of two factors. Firstly, the in-the-wild data distribution arising from data collection over voice calls is different from the in-the-clinic environment, where the previous approaches were trained and validated; this, directly affects the distributions of some of the features. In the proposed approach, an improvement over the ones based in-the-clinic data is achieved, as noisy features are discarded in a language-aware fashion through the combination of the feature selection process and the multiple-instance-based learning techniques. These construct a pipeline that focuses on the consistent features over the classification approach. Secondly, the fusion of the demographic information enhances the performance via injecting prior information of the subject's characteristics; the validation in demographically matched subjects allows for the use of such prior.

B. Language Dependence

1) On Cumulative SData Set Testing: As shown in Section III, when using the whole SData set for testing, each language cohort exhibited different feature distributions, with lower discrimination power compared to the multivariate analysis. The language-aware training outperformed the cross-language-aware experiment, validating in this way the initial hypothesis that running speech signals, unlike sustained vowel phonations, diverge due to different pronunciations and phonetics among different languages. The latter hypothesis is also validated when the language-unaware model is used, with the classification performance being lower when compared to the language-aware training. The latter implies that running speech-based PD detection may need language-specific data collection towards a multilingual voice-based PD detection. Similar studies [54], [55] have also shown that language generalization between training and testing subsets is achieving lower classification results than language-aware. More specifically, both studies suggest that when data from target language are added to the training set, the classification results are highly improved, as opposed to the drop in the diagnostic performance during the cross-language experiments. The current study aims to minimize the Hawrthone effect that can skew the data distribution and sustain long-term adherence. The speech segments of vocal onsets and offsets can be modified during the articulation of the language if a subject is a non-native speaker. However, in the acquired data, there is not available information reported in the self-reported demographics and the test set only consisted of native speakers. Nevertheless, the results forming the language-unaware classification showcase the lower bound of the expected performance of the classification to a non-native dataset, allowing the transferability of the proposed modeling approach to classification subgroups of non-native speakers, when the corresponding labelled data are available. The latter results are also supported by recent findings in speech analysis performed in seven languages for idiopathic REM sleep behavior disorder[56], indicating that cross-language indicators could be used as leading indicators for PD symptoms detection.

2) On Daily SData Set Testing: Figure Suppl. Mat. 1 shows an example of the classification output for six participants (3 HC and 3 PD), when daily voice bags of SData set are used. As Fig. Suppl. Mat. 1 illustrates, Greek and German HCs are constantly classified correctly with low variability of predictions, whereas the predictions of English HC are more variant. When PD subjects are examined regarding in terms of their weekly data contribution, all three patients exhibit high variability, probably due to Levodopa intake; yet, still subjects were classified correctly via the aggregation of their sessions. The study was performed in an in-the-wild environment with remote collection of data, happening through the user's smartphone via the iPrognosis app for a longitudinal period of time, without any instruction or daily-tasks. The aim was to understand the ability to detect the early-PD cases in an unobtrusive design in order to be able to remote screen for PD. Data coming from patients alternating between their On-Off states indeed could be contaminated with noise; that was, however, the motivation of using the MIL model, that takes into account the whole subject-based set, and not only single observations. Future semi-controlled data collection of reports of levodopa intake would better assist the model to infer for the intra-day fluctuations and to take into account patient's On and Off states. The authors acknowledge the noisy data that are contained within, but, nonetheless, the models achieve high classification results, despite the noisy environment that are trained and tested at. Moreover, the continuous accumulation of voice sessions improves the diagnostic performance per language, increasing both the AUC and the F-score of the models for optimizing the Youden index, as depicted in Fig. 4. Additionally, the overall diagnostic performance of the developed models improves as information from a larger number of voice calls is available. The latter highlights that non-conventional aggregation methods, as is the case of MIL, mitigate the within-subject data variability and noise of data. In order for the proposed method to be integrated in a multi-modal screening tool that will aggregate different data modalities, the minimum number of voice calls (50) needed should be considered, so to provide a confident prediction regarding PD (Fig. 4).

C. Interpretation of the Results

The interpretation of the results is based on the following pillars. Firstly, the examination of the initial hypothesis that PD-related degradation on the onset and offset of vocal folds is being made via observing which is the intersection of selected features for the different language settings. During the LOSO training process, the feature selection lowered the initial dimensionality (282) space based on univariate statistical criteria for feature selection. The discarded features set per language could not meet a satisfactory univariate difference for distinguishing the two groups, probably due to the noise-injection of the data collection framework. The selected features (69/78/114 selected features for GR/EN/DE, respectively) though, surpassed a univariate performance and, therefore, the representation space has been of lower dimensions, avoiding noise that mixes the two classes. The latter formed a representation space that, in all language groups, the resulting models had the parameter C=100 with reduced margins for the classification process. The difference of the performance of more wide margins from the hyperplane (i.e., smaller C values) was small (for C=1: -3%/-6%/-0.01% for GR/EN/DE, respectively). Accent differences imposed from the variations from intonations are projected in the frequency distribution and compressed in the mean values of MFCCs. In order to interpret the impact of the linguistic differences in the features representing the spectrum, the selected mean MFCCs coefficients were retrospectively analyzed as descriptors of the spectral distribution. The resulted distribution shows that in Greek language, being a language with more variant phonemes and syllables, all (100%) of the 13 mean spectral coefficients are exploit for both onsets (coefficient order (co):[10,4,5,3,11,7,6,8,2,9,12,13,1]) and offsets (co: [8,2,4,13,5,6,7,3,9,19,12,1]). Moreover, in English language, the corresponding percentages of 92% for the offsets (co: [11,13,4,12,6,5,8,7,9,3,10,2]) and 85% for the onsets (co: [5,6,11,13,12,3,8,7,2,9,10]) were found, whereas in Portuguese language a 85% of the 13 mean spectral coefficients was used for both onsets (co: [10,4,7,8,11,3,2,6,12,5,9]) and offsets(co: [4,8,3,7,10,2,5,11,6,12,9]). These results showcase that tonal frequency in these three languages is widely distributed in the spectrum, expected hence these languages have a soft accent. Moreover, there is no shared pattern in the spectrum coefficients order over languages or onset/offsets. However, in the case of German language, a hard accent language, tones are more concentrated in particular frequencies over the spectrum; hence, only 70% of mean offsets (co: [1,9,13,7,5,4,2,6,3]) and onsets (co: [9,7,11,5,13,12,6,3,4]) are used. The latter findings are in agreement with previous research works in the spectral distribution of speech analysis [51]. The language unaware classification process, resulted in lower classification score of 0.70 AUC than the language-aware modelling. However, the usefulness of the latter model is not only that can provide a moderate classification score, but also can be transferred to new languages without the need of new labelled data and retraining time, due to its language unaware approach. The universal set of features for the language-unaware classifiers, showcase that all 86 are MFCC-only related features and all (100%) of the cepstral descriptors are included to be selected in the process. The two most dominant feature-groups have been shown to be standard deviation of the MFCCs captured in the offset and the mean MFCCs captured in the onset of the voice signals. Interestingly, the intersection of features in all language groups consists of the mean values of MFCCs [2,3,5-7,9] vocal offsets, mean values of MFCCs [3,5,7,9,11,12] for vocal onsets, and MFCCs skewness of the 2nd MFCC. The resulted intersection mainly consists of the mean of spectral descriptors, expressed via MFCCs, which are designed to transform the frequency bands on the Mel scale, approximating the human auditory system's response. Moreover, the mean, as a first-order representator of the distribution, can possibly reflect the human-observable differences between the two classes of HC and early PD. No single feature has shown enough discrimination power alone to classify the two populations. This explains the high dimensionality (>69 features) of the selected features in each language, to best classify the subjects. Moreover, ranking of language-dependent features yielded diversely selected features across the languages, validating that language-aware analysis captures the subtle PD degradation, as it can be expressed in the different running speech styles. The lower order MFCCs contain most of the information about the overall spectral shape of the transfer function, whereas PD affects the articulation over time and the higher-order MFCCs capture the latter variability. Inclusion of the grouped features sets per language group is tabulated in Suppl. Mat. Table V, along with their univariate AUC on the GData. The grouped features set indicate that mean and variance of MFCCs and

BBE features in the onset and the offset are the most dominant in the classification process, for the different languages for the two cohorts (training and testing). The latter facts, plausibly associate how clinical experts form their decision based on the acoustic signal on the degradation-related differences and resulted from an objective in-the-wild approach presented hereby, with the standardized clinical setting.

Finally, the hypothesis of the current study, that demographic information (age and gender) can positively impact the performance has been examined. Specifically, the impact is measured by excluding the two demographic features from the process, showcasing that mutual dependencies between age, gender and voice characteristics can boost the performance to 14/32/8% in ROC AUC scores for English/Greek/German demographically-matched language cohorts, respectively. The latter showcases that latent variables that inform the input signal to the machine learning models can significantly boost the performance of the classification process.

D. Limitations

Despite the promising results, our study has certain limitations that need to be considered prior to future adoption. GData were gathered in an in-the-wild environment with remote collection of data, happening through the user's smartphone. The authors acknowledge that this method of data gathering is prone to containing noisy data which make the accuracy on training data lower than the testing which were clinically evaluated by specialized neurologists. The training was performed in a cohort that self-reported the health status, i.e., subjects' labels were not clinically validated in the GData set. As a consequence, errors in labels may explain the lower predictive performance observed in the LOSO experiment, which is influenced by both the feature noise (affects the observed values of the feature during measurement) and label noise (since other speech-related impairment may mislabel the instances) in the dataset with self-reported ground truth [57]. Despite the latter, the proposed method is optimized to learn a robust predictive mechanism from noisy data [58], based on the largest running speech dataset in the literature. Misclassification of certain training samples is not directly aimed at dealing with the label noise, but robustifies boosting[59], while the approach can be used to find difficult or informative patterns on the data with self-reported labels. A second limitation is that the proposed method is based on the *a priori* assumption that the locally (on the smartphone) extracted voice features are informative enough to detect PD. The local feature extraction approach was adopted in order to avoid raw data transmission and render the system privacy-aware. Nevertheless, this approach reduces the available range of signal representation methods to be used. However, the obtained results show that the extracted voice features, which were used in previous research [24], [52], have satisfactory discrimination potential, at least in the case of the binary PD vs. HC classification. Finally, due to the nature of data collection, a dataset imbalance is observed between the populations of Greek and Portuguese speakers. The cost function of each model during the training phase was optimized considering the weight of each class to be proportional to the inverse of its frequency, to avoid possible model skew towards one class. That being said, the Portuguese model that was trained by the LOSO experiment on the GData group was not used in the clinically assessed test group (SData) due to the absence of any Portuguese clinical center in the study. Moreover, for the Greek group, a sensitivity of 0.86 with a level of 0.90 for specificity is achieved, proving that the models are correctly classifying both the PD and the HC.

V. Conclusion

In this paper a machine learning-based approach for voice-based smartphone data from subjects' running speech recordings is presented. The results indicate that voice-based detection of PD can be done by language-dependent training and evaluation of multiple/single-instance based machine learning models using an in-the-wild dataset. These showcase that PD can be detected from voice calls and features that were already validated in the literature. The latter is further validated in a test dataset from different countries, where subjects were clinically examined, showing that the extracted metrics can discriminate the early PD from the HC and can be further used in the everyday life for the monitoring of voice degradation.

ACKNOWLEDGMENT

The authors acknowledge the donations of time and data from all of the i-PROGNOSIS study participants, and the engineering contributions from George Ntakakis, Fotis Karayiannis (Microsoft Innovation Center, Greece), Konstantinos Kyritsis, (Aristotle University of Thessaloniki, Greece), Hagen Jaeger, Michael Stadtschnitzer (Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany) for their support of the iPrognosis app and data-collection system.

References

[1] S. Campenhausen *et al.*, "Prevalence and incidence of Parkinson's disease in Europe," *European Neuropsychopharmacology*, vol. 15, no. 4, pp. 473–490, 2005.

[2] K. Stephen, M. Caroline, L. Allan, and D. Robin, "Incidence of Parkinson's Disease: Variation by age, gender, and race/ethnicity," *American Journal of Epidemiology*, vol. 157, no. 11, pp. 1015–1022, 2003.

[3] L. V. Kalia and A. E. Lang, "Parkinson disease in 2015: Evolving basic, pathological and clinical concepts in PD," *Nature Reviews Neurology*, vol. 12, no. 2, p. 65, 2016.

[4] B. Thomas and M. F. Beal, "Parkinson's disease," *Human Molecular Genetics*, vol. 16, no. R2, R183–R194, 2007.

[5] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[6] A. Schrag, Y. Ben-Shlomo, and N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?" *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. 5, pp. 529–534, 2002.

[7] A. Schrag, L. Horsfall, K. Walters, A. Noyce, and I. Petersen, "Prediagnostic presentations of Parkinson's disease in primary care: A case-control study," *The Lancet Neurology*, vol. 14, no. 1, pp. 57–64, 2015.

[8] S. Fahn, "Unified Parkinson's disease rating scale," *Recent development in Parkinson's disease*, 1987.

[9] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[10] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of parkinson disease: A systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[11] P. Farzanehfar *et al.*, "Objective measurement in routine care of people with Parkinson's disease improves outcomes," *npj Parkinson's Disease*, vol. 4, no. 1, p. 10, 2018. DOI: 10.1038/s41531-018-0046-4.

[12] B. M. Bot *et al.*, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[13] D. Iakovakis *et al.*, "Motor impairment estimates via touchscreen typing dynamics toward Parkinson's disease detection from data harvested in-the-wild," *Frontiers in ICT*, vol. 5, p. 28, 2018.

[14] A. Papadopoulos, K. Kyritsis, L. Klingelhoefer, S. Bostanjopoulou, K. R. Chaudhuri, and A. Delopoulos, "Detecting Parkinsonian tremor from IMU data collected in-the-wild using deep multiple-instance learning," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[15] A. Burton, "Smartphones versus Parkinson's disease: i-PROGNOSIS," *The Lancet Neurology*, vol. 19, no. 5, pp. 385–386, 2020.

[16] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, P. Klumpp, P. A. Pérez-Toro, D. Escobar-Grisales, N. Roth, C. D. Rios-Urrego, M. Strauss, H. A. Carvajal-Castaño, S. Bayerl, *et al.*, "Apkinson: The smartphone application for telemonitoring parkinson's patients through speech, gait and hands movement," *Neurodegenerative Disease Management*, vol. 10, no. 3, pp. 137–157, 2020.

[17] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[18] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[19] Z. Galaz *et al.*, "Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 301–317, 2016.

[20] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, and E. Ruzicka, "Distinct patterns of imprecise consonant articulation among parkinson's disease, progressive supranuclear palsy and multiple system atrophy," *Brain and language*, vol. 165, pp. 1–9, 2017.

[21] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation," *Neural Transmission*, vol. 124, no. 3, pp. 303–334, 2017.

[22] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioural Neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[23] T. Villa-Cañas, J. Orozco-Arroyave, J. Vargas-Bonilla, and J. Arias-Londoño, "Modulation spectra for automatic detection of Parkinson's disease," in *Proc. of the Image Signal Processing and Artificial Vision (STSIVA) 2014 XIX Symposium*, IEEE, 2014, pp. 1–5.

[24] Orozco-Arroyave *et al.*, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *Sixteenth Annual Conf. of the International Speech Communication Association*, 2015, pp. 95–99.

[25] J. Rusz, J. Hlavnička, T. Tykalova, M. Novotn, P. Dušek, K. Šonka, and E. Ržička, "Smartphone allows capture of speech abnormalities associated with high risk of developing parkinson's disease," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 26, no. 8, pp. 1495–1507, 2018.

[26] S. Arora, C. Lo, M. Hu, and A. Tsanas, "Smartphone speech testing for symptom assessment in rapid eye movement sleep behavior disorder and parkinson's disease," *IEEE Access*, vol. 9, pp. 44 813–44 824, 2021.

[27] R. J. Holmes, J. M. Oates, D. J. Phyland, and A. J. Hughes, "Voice characteristics in the progression of Parkinson's disease," *International J. of Language & Communication Disorders*, vol. 35, no. 3, pp. 407–418, 2000.

[28] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[29] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain and Cognition*, vol. 56, no. 1, pp. 24–29, 2004.

[30] S.-M. Fereshtehnejad, C. Yao, A. Pelletier, J. Y. Montplaisir, J.-F. Gagnon, and R. B. Postuma, "Evolution of prodromal parkinson's disease and dementia with lewy bodies: A prospective study," *Brain*, vol. 142, no. 7, pp. 2051–2067, 2019.

[31] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of parkinson's disease using telephone-quality voice," *Acoustical Society of America*, vol. 145, no. 5, pp. 2871–2884, 2019.

[32] H. Jaeger, M. Stadtschnitzer, A. Rizos, F. Karayiannis, G. Ntakakis, and L. Hadjileontiadis, *i-Prognosis: Verwendung von Sprachmerkmalen als biomarker zur Detektion der Parkinson-erkrankung*, Mar. 2018. DOI: 10 . 5281 / zenodo . 3678669. [Online]. Available: https://doi.org/10.5281/zenodo.3678669.

[33] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling.," in *Proc. Int. Symp. Music Information Retrieval*, vol. 270, 2000, pp. 1–11.

[34] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[35] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale L1-regularized least squares," *IEEE Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[36] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proc. of the 21st International Conf. on Machine learning*, 2004, pp. 78–85.

[37] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[38] S. Lars *et al.*, "Analysis of variance (ANOVA)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, 1989.

[39] N. Cristianini *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[40] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1-2, pp. 167–179, 1967.

[41] A. Liaw *et al.*, "Classification and regression by Random Forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[42] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. of Neural Information Processing Systems*, 1998, pp. 570–576.

[43] T. Gartner *et al.*, "Multi-instance kernels," in *Proc. of the Nineteenth International Conf. on Machine Learning*, 2002, pp. 176–186.

[44] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. of the 24th International Conf. on Machine Learning*, 2007, pp. 105–112.

[45] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. of Neural Information Processing Systems*, vol. 15, 2003, pp. 577–584.

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

[47] G. Doran and S. Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," *Machine learning*, vol. 97, no. 1-2, pp. 79–102, 2014.

[48] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1–25, 2017.

[49] K. A. Fox and T. K. Kaul, "Intermediate economic statistics," Wiley New York, Tech. Rep., 1968.

[50] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

[51] T. Khan, L. E. Lundgren, D. G. Anderson, I. Nowak, M. Dougherty, A. Verikas, M. Pavel, H. Jimison, S. Nowaczyk, and V. Aharonson, "Assessing parkinson's disease severity using speech analysis in non-native speakers," *Computer Speech & Language*, vol. 61, p. 101 047, 2020.

[52] Orozco-Arroyave *et al.*, "Phonation and articulation analysis of spanish vowels for automatic detection of Parkinson's disease," in *International Conf. on Text, Speech, and Dialogue*, Springer, 2014, pp. 374–381.

[53] T. Khan, J. Westin, and M. Dougherty, "Classification of speech intelligibility in parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 35–45, 2014.

[54] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Noeth, "Gender-dependent gmm-ubm for tracking parkinson's disease progression from speech," in *Speech Communication; 12. ITG Symposium*, VDE, 2016, pp. 1–5.

[55] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: A deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1618–1630, 2018.

[56] J. Rusz, J. Hlavnička, M. Novotn, T. Tykalová, A. Pelletier, J. Montplaisir, J.-F. Gagnon, P. Dušek, A. Galbiati, S. Marelli, *et al.*, "Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease," *Annals of Neurology*, 2021.

[57] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2013.

[58] R. J. Hickey, "Noise modelling and evaluating learning from examples," *Artificial Intelligence*, vol. 82, no. 1-2, pp. 157–179, 1996.

[59] G. Rätsch, B. Schölkopf, A. J. Smola, S. Mika, T. Onoda, and K.-R. Müller, "Robust ensemble learning for data mining," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2000, pp. 341–344.