ABSTRACT Parkinson’s disease (PD) is a multi-symptom neurodegenerative disease. There are no biomarkers; the diagnosis and monitoring of the disease progression require clinical and functional symptom observation. Voice impairment is an early symptom of PD, and computerized analysis of voice has been proposed for early detection and monitoring of the disease. However, there is poor reproducibility of many studies, which is attributed to the experimental data having been collected under controlled conditions. To overcome the limitations of earlier works, this study has investigated three sustained phonemes: /a/, /o/, and /m/, which were recorded using an iOS-based smartphone from 72 participants (36 people with PD and 36 healthy) in a typical clinical setting. A number of signal features were obtained, statistically investigated, and ranked to identify the suitable feature sets. These were classified using machine learning models. The results show that a combination of phonemes /a/+ /o/+ /m/ was most suited to differentiate the voice of PD people from healthy control participants, with an average accuracy, sensitivity, and specificity of 100%, 100%, 100%, respectively, using leave-one-out validation. The findings of this study could assist in the clinical assessments and remote telehealth monitoring for people with parkinsonian dysarthria using smartphones.

INDEX TERMS Dysarthria, Parkinson’s disease, smartphone, sustained phonemes, voice impairment.

I. INTRODUCTION Parkinson’s disease (PD) is the second most common neurodegenerative disorder [1], which is expected to increase with an ageing population. There are no biomarkers to diagnose the disease, which requires the observation of the complex set of symptoms of the patients. Acoustic speech abnormalities have been reported even in early-stage PD patients and even when there is no perceptible dysarthria [2]. Several investigators have found impaired speech parameters in early-stage PD using objective acoustic measures [3], [4]. Several studies have investigated the difference between the voice of PD and healthy control (HC) using different approaches [4], [5], [6], [7], [8], [9], [10], [11], [12].

Human speech requires fine-motor control, cognitive abilities, auditory feedback, and muscle strength. Parkinsonian dysarthria can be characterized by reduced vocal tract loudness, reduced speech prosody, imprecise articulation, significantly narrower pitch range, longer pauses, vocal tremor, breathy vocal quality, harsh voice quality, and disfluency [4]. The differences in the voice parameters of sustained phonemes have been examined for detecting and monitoring PD [4], [13], [14]. A number of works have considered the signal features previously used for speech studies, such as speaker recognition [15], [16]. The investigation of sustained phoneme and text-dependent speech...
modalities for PD screening is reported in [13]. However, such analysis has confounding factors such as language skills, vision and hearing [17]. Tsanas et al. [18] have extended this to associate these with the motor disability score of PD patients.

The use of non-linear and hybrid features such as the fractal dimension (FD), entropy [19], deep multivariate features [20], and linear predictive models [21], [22] has been proposed. Godino-Llorente et al. [12] proposed an articulatory biomarker based on the kinetic envelope trace of voice that had an accuracy of 85%. In [6], 132 features were extracted from phonemes recorded in a sound-treated booth with a head-mounted microphone to train a support vector machine (SVM) and random forest classifier which achieved an accuracy of 97.7% and 90.2%, respectively to identify PD from HC.

Signal features have often been selected based on the understanding of the disease [23], [24]. The difference between the voice of healthy people and those with PD has been observed in their pitch frequency, jitter, shimmer, and harmonics to noise ratio [25]. The pitch frequency or the fundamental frequency of the vocal cords, $f_0$, is the number of cycles of the glottal vibration. Jitter, the perturbation of the glottal vibration period, is influenced by the motor control, rigidity, and tremor of the larynx. Shimmer, the amplitude perturbation, is related to the glottal resistance and increases with a lack of fine muscle control. Harmonics to noise ratio (HNR) or noise to harmonic ratio (NHR) indicates the relative harmonic strength are the ratios between the periodic (voiced) and non-periodic (noise) components of the speech. These reduce with diminished glottal vibration and low HNR is an indicator of dysarthria. However, some of these parameters may also be affected by other factors such as age, gender, and ethnicity.

The above studies have shown that there are several signal features that show significant differences between the voice of PD and HC. However, most studies have not considered real-world conditions where there is background noise, and there are differences between recording devices and conditions [26], [27], [28]. There are only a few studies that used data recorded in real-life clinical setup [19], [29], [30]. Therefore, further work is required to validate the use of this for real-life scenarios, especially for remote monitoring of the patients and other telehealth applications.

The aim of this study was to identify the most suitable signal classification method that can differentiate between PD and HC when the recordings are made in real-world conditions. We investigated the phonatory parameters of three sustained phonemes and compared people with PD with HC. The data were recorded in a typical clinical setting to check for its real-world suitability using smartphones [31], [32]. Besides the statistical analysis, the SVM classifier was used to classify the voice in two classes: PD and HC. The proposed model provides the following advantages over the existing alternatives:

**FIGURE 1.** The block diagram of identifying PD from HC using sustained phonemes. The model is trained and tested using 72 PD and HC participants.

1. Data were recorded in a normal clinical setting and with background noise conditions.
2. The recordings were made using commercially available smartphone with default settings.
3. Only three phonemes were recorded and it was not dependent on language skills.
4. The performance was perfect, with 100% sensitivity and specificity, outperforming the state-of-the-art methods.

**II. MATERIALS AND METHODS**

**A. PARTICIPANTS**

Seventy-two age-matched volunteers comprising 36 people with PD and 36 healthy age-matched participants as the HC group participated in this study. The data can be found in our previously reported work [30]. All the people with PD had been diagnosed with PD within the last ten years based on procedures complying with the Queen Square Brain Bank criteria for idiopathic PD [33]. The presence of any advanced PD clinical symptoms such as visual hallucinations, frequent falling, cognitive disability, or need for institutional care was an exclusion criterion [34]. People with PD were recruited from the movement disorder clinic at Monash medical center and Dandenong neurological clinic while the HC group participants were recruited from several retirement centers. Table 1 presents participants’ demographics, cognitive stage, and health history. The UPDRS-III scores [35] of all the participants show a clear difference between the groups, while the MoCA score confirms that both PD and HC did not have cognitive impairment.

The study protocol was approved by the ethics committee of Monash Health, Melbourne, Australia (LNR/16/MonH/319) and RMIT University Human Research Ethics Committee, Melbourne, Australia (BSEHAPP22-15KUMAR). Before the experiments, written consent was obtained from all the participants.

**B. METHODS**

Figure 1 illustrates the block diagram of the proposed method of classifying PD from HC. As shown in Figure 1, three phonemes were recorded from PD and HC participants using a smartphone. Each phoneme was segmented before extracting features from it. Machine learning based classification was applied to identify PD from HC. The detail of each section is described below:
1) VOICE RECORDING
Three sustained phonemes /a/, /o/, and /u/ were recorded from each participant. The phonemes were selected to examine a range of voice production models [36]. The vowel /a/, as in “cat”, is an open-back or low vowel, produced while the jaw is wide open, with the tongue that is inactive and low in the mouth. In this, the vibration of the vocal folds dominates the sound of the vowel. The vowel /o/, as in “oh”, is a closed-mid-back vowel. The back of the tongue is positioned high towards the palate, and the lips are rounded. The phoneme /u/ is a nasal phoneme produced by the vibration of the vocal folds with the air flowing through the nasal cavity. Although all three phonemes require control of the respiratory and laryngeal vocal fold muscles, there are considerable differences in patterns of activation of the rostral muscles of articulation (of pharynx, tongue, jaw, and lips).

The participants were asked to speak the phonemes for as long as it was comfortable, in their natural pitch and loudness. During the recording, they held the smartphone as if they were talking a phone call. The voice of 72 participants (36 PD and 36 HC) was recorded using an iOS-based smartphone (iPhone 6S plus) with its built-in microphone and default settings, while the participants were located in typical Australian clinics or office settings. The recordings were saved into a single-channel uncompressed WAV format with a sampling frequency \( f_s \) of 48.1 kHz and a 16-bit resolution. Each file contained one single sustained phoneme with varied duration, as shown in Table 2. In between each recording, there was minimum 15 seconds rest time.

2) AUTOMATED SEGMENTATION AND FEATURE EXTRACTION
All computations, including pre-processing, automated segmentation, and statistical analysis, were performed using Matlab2018b (MathWorks) and Python. All the recorded phonemes were segmented using an envelope detection and thresholding approach. The signal features were computed from each segment. Recordings with the voice of the instructor were removed. In the original recordings, the signal-to-noise ratio was 16-24 dB (average 19.26 dB), similar to typical Australian clinical conditions. The first step for feature extraction was to locate the time instances \( t_i \) and the amplitude \( A_i \) of the pulses in the recording representing the glottal vibration. The instantaneous period of the glottal wave \( T_i \) was calculated as the difference between subsequent instances of the pulses, \( T_i = t_{i+1} - t_i \).

The first set of features were six jitter parameters: jitter absolute \( (\text{jitter abs}) \), jitter relative \( (\text{jitter rel}) \), period perturbation quotient-3 \( (\text{period ppq3}) \), period perturbation quotient-5 \( (\text{period ppq5}) \), period perturbation quotient-11 \( (\text{period ppq11}) \), and frequency modulation \( (\text{Jitter FM}) \). Here, \( \text{ppq3} \), \( \text{ppq5} \), and \( \text{ppq11} \) are the perturbation of the difference between \( T_i \) and the moving average of \( T_i \) with a window size of 3, 5, and 11, respectively. The equations to calculate jitter parameters [32] are shown in equations 1 to 6:

\[
\text{Jitter (abs)} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_{i+1} - T_i| \tag{1}
\]

\[
\text{Jitter (rel)} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{T_{i+1} - T_i}{T_i} \right| \tag{2}
\]

\[
\text{Jitter (ppq3)} = \frac{1}{N-2} \sum_{i=2}^{N-1} \left| T_i - \left( \frac{2}{3} \sum_{n=1}^{i+1} T_n \right) \right| \tag{3}
\]

\[
\text{Jitter (ppq5)} = \frac{1}{N-3} \sum_{i=3}^{N-1} \left| T_i - \left( \frac{2}{5} \sum_{n=1}^{i+2} T_n \right) \right| \tag{4}
\]

\[
\text{Jitter (ppq11)} = \frac{1}{N-10} \sum_{i=11}^{N-2} \left| T_i - \left( \frac{2}{11} \sum_{n=1}^{i+5} T_n \right) \right| \tag{5}
\]

\[
\text{Jitter (FM)} = \frac{\max(T_i)_{i=1}^{N} - \min(T_i)_{i=1}^{N}}{\max(T_i)_{i=1}^{N} + \min(T_i)_{i=1}^{N}} \tag{6}
\]

Six shimmer parameters that were extracted from the segments are the absolute shimmer \( (\text{shimmer abs}) \) in dB, the relative shimmer \( (\text{shimmer rel}) \), amplitude perturbation quotient-3 \( (\text{apq3}) \), amplitude perturbation quotient-5 \( (\text{apq5}) \), amplitude perturbation quotient-11 \( (\text{apq11}) \), and amplitude modulation \( (\text{Shimmer AM}) \). Here, \( \text{apq3} \), \( \text{apq5} \), and \( \text{apq11} \) represent the perturbation of the difference between \( A_i \) and the moving average of \( A_i \) with a window size of 3, 5, and 11, respectively. The calculations to compute shimmer parameters are described in equations 7 to 12:

\[
\text{Shimmer (abs, dB)} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| 20 \log \left( \frac{A_{i+1}}{A_i} \right) \right| \tag{7}
\]

\[
\text{Shimmer (rel)} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|A_{i+1} - A_i|}{A_i} \tag{8}
\]

\[
\text{Shimmer (apq3)} = \frac{1}{N-2} \sum_{i=2}^{N-1} \left| A_i - \left( \frac{2}{3} \sum_{n=1}^{i+1} A_n \right) \right| \tag{9}
\]

### Table 1. Participants’ demographics and clinical characteristics.

| Demographics        | HC | PD |
|---------------------|----|----|
| Number of subjects  | 36 | 36 |
| Age                 | 65.45 ± 7.35 | 68.75 ± 8.60 |
| Duration of disease (years) | N/A | 6.40 ± 2.75 |
| MDS-UPDRS-III score | N/A | 16.26 ± 17.28 |
| MoCA                | 28.47 ± 1.59 | 25.57 ± 4.16 |

Caption: PD=Parkinson’s disease, HC=Healthy control, MDS-UPDRS=Movement disorder society-unified Parkinson’s disease Rating scale, MoCA=Montreal cognitive assessment, N/A=Not applicable.

### Table 2. Duration of the recordings.

| Phoneme | HC min–max (mean ± std) | PD min–max (mean ± std) |
|---------|--------------------------|-------------------------|
| /a/     | 5.1–19.1 (11.3±2.7)      | 5.3–14.4 (9.8±2.5)      |
| /o/     | 6.7–23.3 (12.2±3.6)      | 5.2–16.8 (11.3±3.1)     |
| /u/     | 5.7–16.5 (11.7±2.9)      | 5.6–15.3 (10.2±2.4)     |
VOLUME 10, 2022

97603

Since the coordinators of articulator of vocal tract such as tongue, jaw, lips are also affected by PD [43], we incorporated those features that best characterize the vocal tract coordinators such as mel-frequency cepstral coefficients (MFCCs).

MFCCs measures the energy of speech signal in each frequency band (equation 15). Since the coordinators of articulators of the vocal tract such as the tongue, jaw, lips are also affected by PD [43], it is hypothesized that MFCC will be different for PD and HC.

Spectral analysis is used to understand the oscillatory trend of the signal but does not carry the temporal information. Wavelet transform (WT) is a technique that is based on the use of time limited waves, referred to as wavelets, and performs multi-resolution, time-frequency analysis. In this context, it converts the single dimension time domain signal to two-dimensional time-frequency domain without losing the temporal information. The discrete WT (DWT) decomposes the signal into different frequency bands into approximation and detail coefficients, with each scale corresponding to scaling of the frequency by half. In this study, the recordings were decomposed at level 10 which covers the entire audible range of the recordings. Daubechies 10 (Db10) mother wavelet was chosen as the vanishing moment. Energy, entropy, and TEKO features were computed from each DWT decomposed approximation and detail coefficients.

C. FEATURE SELECTION

A large number of features increases the risks of overfitting, can lead to higher error, and increases the computational complexity [45, 46]. That is why the exclusion of redundant features is necessary [46]. During feature selection, the first step was to identify those features that were tested to be statistically different (p < 0.0001) for the two groups using the Mann-Whitney U test. Next, feature selection algorithms were applied to identify the best features. For the removal of algorithm bias, four different feature selection algorithms were compared: i) infinite latent feature selection (ILFS), ii) least absolute shrinkage and selection operator (LASSO), iii) Relief-F, and iv) unsupervised discriminative feature selection (UDFS).

D. MODEL TRAINING AND CLASSIFICATION

Support vector machine (SVM)-based machine learning classifier was deployed to label the selected features into two classes: PD and HC. The details of the SVM classifier and cross-validation are described below.
1) SUPPORT VECTOR MACHINE

Support vector machine (SVM) is a widely used supervised machine learning technique for classification. The decision boundaries or hyperplanes are developed based on the support vectors during training.

Let, vector $x$ denotes the feature to be classified and its label is denoted by $y$ where $y \in \{+1, -1\}$. Now, for a given set of training data, $\{(x_i, y_i) , i = 1, 2, \ldots, n\}$, the separating hyperplanes can be obtained by maximizing the margin, which is the minimization of the following function:

$$J(w, \beta) = \frac{1}{2} w^T w + C \sum \beta_i$$

With the following constrain function

$$y_i(w^T x + b) \geq 1 - \beta_i \text{ where } \beta_i \geq 0$$

Here, $w$ is the weight vector, $b$ is a constant, $C$ is a positive regularization parameter, and $\beta_i$ is the slack variable. Applying the Lagrange multipliers $\alpha_i$, for vector $x$, the solution of the decision function can be expressed as:

$$w = \sum \alpha_i y_i x_i$$

$$f(x) = \sum \alpha_i y_i x_i^T x + b$$

For the nonlinear SVM, a nonlinear mapping function $\varphi(x)$ is used to map the input feature into a higher dimensional feature space, thus making the samples more separable:

$$f(x) = \sum \alpha_i y_i K(x_j, x) + b$$

where, $x_j$ are the support vectors and $K(x_j, x)$ is the kernel function, for the polynomial kernel $K(x_j, x) = (x_j \cdot x + 1)^d$ and radial basis function (RBF) kernel $K(x_j, x) = \exp\left(-\gamma \|x_j - x\|^2\right)$. SVM details can be found in [47].

In this study, SVM with linear, polynomial and RBF kernels were used.

2) CROSS VALIDATION

We evaluated the model performance using leave one out cross validation (LOOCV) techniques [48]. The LOOCV method uses N-1 subjects for model training, 1 for testing, and is repeated N times, so that each subject gets a chance to be tested. The final result is the mean of the individual evaluations. The detail of the model training and testing using LOOCV is illustrated in Figure 2. Accuracy, sensitivity, specificity, and F1-score were computed as performance metrics.

### III. RESULTS

#### A. STATISTICAL ANALYSIS AND PD CLASSIFICATION

Anderson-Darling test confirmed that the voice parameters of three different sustained phonemes for two groups were not normally distributed and thus unsuitable for the parametric test. So, the group differences and significance of each feature for PD vs. HC were computed using Mann-Whitney U test [49]. Features having $p$-value $\leq 0.0001$ were the input to the Relief-F feature selection algorithms to sort the most significant features from the pool of feature sets. The model performance with the variation of the sorted features of each phoneme using SVM classifier is shown in Fig 3. For all phonemes, the accuracy of the model increases with the
increasing number of features till 15. The model performance remained almost unchanged between feature number 15 to 40 and the performance decayed after any inclusion of feature beyond 40. Since, the non-significant features have a very low separable capability, the inclusion of large number of insignificant features may mislead the classifier and decreases the classification accuracy. As is observed from Fig. 3, inclusion of features above 40 reduces the model performance.

The accuracy of the proposed model with top 15 sorted features extracted from individual phoneme /a/, /m/, and /o/ using SVM with RBF kernel is 97.22%, 95.83%, and 98.66% respectively. Based on the combined features extracted from two phonemes, PD classification was 97.22%, 98.66%, and 100% for /a/ + /m/, /m/ + /o/, and /a/ + /o/, respectively. The proposed model accuracy became 100% when the features obtained from the three phonemes /a/ + /m/ + /o/ were combined. The detailed performance of the proposed model using the different combinations of phonemes is shown in Table 3. It is found that features extracted from phoneme /o/ identified PD from HC with higher accuracy compared to other phonemes and the inclusion of features from phonemes /a/ and /m/ improved the performance. It showed the highest performance when features from all three phonemes were combined to train the model. The confusion matrix is shown in Fig. 4. The confusion matrix summarises the predicted and actual classes, providing an accurate assessment of the performance by providing true positives, true negatives, false positives, and false negatives.

B. COMPUTING THE EFFECT SIZE AND SPEARMAN CORRELATION OF EACH SIGNIFICANT FEATURE

The statistically significant features of each phoneme were sorted and ranked by the ReliefF-based feature selection technique. The effect size computed by Cohen’s d and the Spearman correlation coefficient of each selected phonemes are shown in Table 4. Based on the Mann-Whitney U test, each feature was assessed for statistical significance, and the corresponding p-value is listed in Table 4. The two-dimensional representation of the top two features of each phoneme is demonstrated in Fig. 5.

C. ROBUSTNESS OF THE MODEL

A larger sample size is necessary for the training to represent modelled phenomena. However, with limited labelled data samples, which is often the case with medical data, the resultant model needs to be tested for robustness. Hence, the system performance as a function of the minimum number of data points (participants) was conducted and is presented in Fig. 6. The performance was obtained by increasing the number of participants from 8 to 50 at an increment of 6. For this purpose, the complete dataset was subdivided into two groups to construct the training set and they were randomly subdivided to get the training set by stratified random sampling. This ensured that class balance was maintained for the training set. Each step was iterated ten times and the
IV. DISCUSSION

People with PD often have dysarthria or speech impairment which may appear in phonatory, articulatory, prosodic, and linguistic aspects. The change is complex and characterized by reduced loudness, reduced speech prosody, imprecise articulation, significantly narrower pitch range, longer pauses, vocal tremors, breathy vocal quality, harsh voice quality, and dysfluency [4]. Speech disorders are related to several factors such as inability to perform habitual tasks, loss of fine control, weakness, tremor, and rigidity of the speech production muscles.

This study has investigated the use of the utterance of phonemes /a/, /o/, and /m/ for differentiating the voice of people with PD from HC. The classification results confirm that identifying the voice of HC from PD improves when the combination of phonemes /a/ + /m/ + /o/ are used. The results also indicate that among the single phonemes, /o/ is more effective in differentiating the two groups than phoneme /a/ and /m/. The phoneme /a/ is produced while the tongue is pressed towards the jaw and the lips are wide open. Similarly, the production of the phoneme /m/ does not require voice box muscles because the lips are closed, and the air is passed through the nasal cavity. On the other hand, the production of phoneme /o/ requires precise positioning of the tongue at a mid-height position and the small-rounded position of the lips [50] than /a/ and /m/. Since the production of the phoneme /a/ and /m/ does not require the precise control of the tongue and lips, the tremor or weakness in the tongue or lips positioning should be more prominent in the production of /o/ than /a/ and /m/. This supports our finding that PD and HC are better distinguished with /o/ compared to /a/ and /m/.

However, these are only logical deductions at this stage, and further research needs to be conducted to confirm these.

It was also found that the MFCC and the features from the first and second derivatives of MFCC of phonemes /a/, /m/, and /o/ were significantly different between PD and HC. The cepstral analysis identifies the changes to the source and vocal cord factors, and this observation confirms that Parkinsonian dysarthria is associated with these changes. The average log

---

### TABLE 5. The comparison of the proposed model with the existing studies in literature for two class (sleep-wake) classification problem.

| Reference     | Cross validation | Modality      | Dataset       | Recording Environment                  | Accuracy |
|---------------|------------------|---------------|---------------|----------------------------------------|----------|
| Little et al. [5] | Bootstrap with 50 replicates | Sustained vowel | 23 PD, 8 Con  | Sound-treated booth with head mounted microphone | 91.4%    |
| Tsanas et al. [6] | 10-fold          | Sustained phonemes | 33 PD, 10 Con | Sound-treated booth with head mounted microphone | 97.7%    |
| Orozco et al. [11] | 10-fold          | /a/, /e/, /i/, /o/, /u/ | 50 PD, 50 Con | Noise controlled soundproof booth with dynamic omnidirectional microphone | 91.3%    |
| Rekha et al. [19] | Leave one out   | /m/          | 18 PD, 22 Con  | Normal clinical setting                 | 93%      |
| Das et al. [26] | 35% of total samples selected randomly | Sustained vowel | 23 PD, 8 Con  | Sound-treated booth with head mounted microphone | 92.9%    |
| Guo et al. [27] | 10-fold          | Sustained vowel | 23 PD, 8 Con  | Sound-treated booth with head mounted microphone | 93.1%    |
| Psorakis et al. [28] | 10-fold          | Sustained vowel | 23 PD, 8 Con  | Sound-treated booth with head mounted microphone | 89.5%    |
| Nemuel et al. [29] | 30% of total samples selected randomly | /a/, /m/, /o/ | 22 PD, 22 Con  | Normal clinical setting                 | 81.3%    |
| Rekha et al. [30] | Leave one out   | /a/, /m/, /u/ | 24 PD, 22 Con  | Normal clinical setting                 | 81%      |
| Proposed       | Leave one out   | /a/          | 36 PD, 36 Con  | Normal clinical setting                 | 98.6%    |
| Proposed       | Leave one out   | /a/+o/+m/+o/ | 36 PD, 36 Con  | Normal clinical setting                 | 100%     |
energy of phoneme /a/ was found to be significantly different which also indicates the reduced source strength of PD.

The significant difference between PD and HC of HNR and GNER of phoneme /o/ indicates the weakened vocal cords, due to which the relative voiced noise compared to resonatory sound is higher in the voice of PD. The classification results show that the inclusion of these features improves the model performance. The classification accuracy was 100% when using these features from the three phonemes, /a/, /m/, and /o/. Since, PD is a multi-symptom disease with complex display of the symptoms, and while the analysis of each phoneme captures some of the symptoms, it is the combination of all the three that appears to be capturing all the symptoms of the disease. The study has also investigated the effect of sampling frequency in differentiating between PD and HC. For sampling frequencies, \( fs = 48.1 \text{ kHz} \) and \( 8 \text{ kHz} \), the model shows exactly similar results. This indicates that the relevant frequency of interest is less than 4kHz.

Further, this work explored the performance of the four feature selection algorithms for phoneme-based PD classification. Though ReliefF and ILFS slightly performed better than LASSO and UDFS, similar performance was noticed for the higher number of features. It was also observed that any top twenty features selected by any of the four-feature selection algorithms showed above 95% classification accuracy.

The performance comparison of our approach with the existing state-of-art techniques in the literature is summarized in Table 5. As shown in Table, the model performance for phonemes recorded in noise-free soundproof environment with a microphone varies from 89.5% to 97.7%. On the other hand, the model performance varies from 81% to 93.1% for phonemes recorded in a normal clinical setting. While the ambient noise resulted in a fall of performance of the models in literature by 5.6% to 8.4%, our proposed model was less prone to the ambient noise and capable of identifying PD from HC with 100% accuracy.

There are four major achievements of this study. Firstly, it has been found that people with PD and healthy age-matched have the most significant difference in the production of the phoneme /o/ which is differentiable even with background noise and recorded using handheld smartphone. The statistical analysis and classification results confirm that the voice features of phoneme /o/ can discriminate people with PD from HC participants more accurately than /a/ and /m/ but the combination of phonemes /a/, /m/, and /o/ is the most accurate. Secondly, it has shown that computerized assessment of the voice of people with PD is suitable for real-world, regular clinical settings with background noise and using low sampling rate smartphone. Thirdly, this model requires only phonemes and thus, it is language independent. Finally, the model is trained and tested without favoring hyperparameters that are tailored to a specific gender, so this is a gender independent model.

The limitation of this study is that we did not consider factors such as accents because all participants were of suburban Melbourne only. There is also the need to test the individual multiple times to check for the repeatability of the results and to use multiple devices while this study used one phone only. Another weakness of this study was that people with PD were more than two years post-diagnosis and not in the very early stage of the disease.

V. CONCLUSION

This study has investigated the use of sustained phonemes for computerized diagnosis of PD based on the utterance of three phonemes /a/, /o/, and /m/ recorded using a handheld smartphone in real-world clinical conditions with ambient noise conditions of about 20 dB. It has been found that there were number of features with significant differences between PD and HC. After feature selection from the three phonemes, /a/+ /m/+ /o/, the classifier differentiated between HC and PD with 100% accuracy. Two prominent differences between PD and HC based on the selected features are a decrease in voice energy and increase in relative voice-noise. The novelty of this study is the selection of the acoustic features that are suitable for differentiating between PD and HC while using a handheld smartphone and is not sensitive to clinical ambient noise conditions. This study shows the potential of using phoneme based computerised diagnosis of PD that can be performed remotely using a smartphone. It has applications for assisting in the clinic or for telehealth.

ACKNOWLEDGMENT

The authors acknowledge the team at Dandenong Neurology and RMIT University which collected the data, and its online availability. Special thanks to Dr. Sumit Bhowmik and Dr. Sumaiya Kabir for their support and helpful discussion.

REFERENCES

[1] L. M. de Lau and M. M. Breteler, “Epidemiology of Parkinson’s disease,” Lancet Neurol., vol. 5, no. 6, pp. 525–535, Jun. 2006.
[2] K. Forrest and G. Weismer, “Acoustic analysis of motor speech disorders,” in Clinical Management of Sensorimotor Speech Disorders, Taylor & Francis, 2009, pp. 46–63.
[3] C. Stewart, L. Winfield, A. Hunt, S. B. Bressman, S. Fahn, A. Blitzer, and M. F. Brin, “Speech dysfunction in early Parkinson’s disease.” Movement Disorders, Off. J. Movement Disorder Soc., vol. 10, no. 5, pp. 562–565, 1995.
[4] J. Rusz, R. Cnecka, T. Tykalova, H. Ruzickova, J. Klenpfr, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, “Imprecise vowel articulation as a potential early marker of Parkinson’s disease: Effect of speaking task,” J. Acoust. Soc. Amer., vol. 134, no. 3, pp. 2171–2181, Sep. 2013.
[5] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,” IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1015–1022, Apr. 2009.
[6] 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. Biomed. Eng., vol. 59, no. 5, pp. 1264–1271, May 2012.
[7] A. Tsanas, “Accurate telemonitoring of Parkinson’s disease symptom severity using nonlinear speech signal processing and statistical machine learning,” Ph.D. thesis, Dept. Comput. Sci., Univ. Oxford, Oxford U.K., 2012.
[8] A. Tsanas, “Automatic objective biomarkers of neurodegenerative disorders using nonlinear speech signal processing tools,” in Proc. 8th Int. Workshop Models Anal. Vocal Emissions Biomed. Appl. (MAVERA), 2013, pp. 37–40.
[9] J. Rusz, R. Čmejla, H. Rážičková, J. Klempíř, V. Majerová, J. Picmausová, J. Roth, and E. Rážička, “Evaluation of speech impairment in early stages of Parkinson’s disease: A prospective study with the role of pharmacotherapy,” J. Neural Transmiss., vol. 120, no. 2, pp. 319–329, Feb. 2013.

[10] J. Rusz, R. Čmejla, H. Rážičková, E. Chvojka, and E. Rážička, “Quantitative acoustic measurements for the characterization of speech and voice disorders in early untreated Parkinson’s disease,” J. Accoust. Soc. Amer., vol. 129, no. 1, pp. 350–367, 2011.

[11] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Ratıva, and E. Noth, “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease,” in Proc. 9th Int. Conf. Lang. Resour. Eval., 2014, pp. 342–347.

[12] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J. Y. Choi, L. Moro-Velázquez, and J. A. Gómez-García, “Towards the identification of idiopathic Parkinson’s disease from the speech. New articulatory kinetic biomarkers,” PLoS ONE, vol. 12, no. 12, Dec. 2017, Art. no. e0189583.

[13] E. Vaičiukynas, A. Verikas, A. Gelzinis, and M. Bacauskienė, “Detecting Parkinson’s disease from sustained phonation and speech signals,” PLoS One, vol. 12, no. 10, pp. 1–17, 2017.

[14] B. E. Sakar, M. E. Ibenkul, C. O. Sakar, A. Serbhas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, “Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings,” IEEE J. BioMed. Health Inform., vol. 17, no. 4, pp. 828–834, Jul. 2013.

[15] T. Khan, E. Lundgren, D. G. Anderson, I. Nowak, M. Dougherty, A. Verikas, M. Pavel, H. Jönsson, S. Nowaczyk, and V. Aharonson, “Assessing Parkinson’s disease severity using speech analysis in non-native speakers.” Comput. Speech Lang., vol. 61, May 2020, Art. no. 101047.

[16] J. Rusz, M. Novotná, J. Hlavnicka, T. Tykalova, and E. Rážička, “High-accuracy voice-based classification between patients with Parkinson’s disease and other neurological diseases may be an easy task with inappropriate experimental design,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 8, pp. 1319–1321, Aug. 2017.

[17] N. D. Pah, M. Motin, and D. Kumar, “Voice analysis for diagnosis and monitoring Parkinson’s disease,” in Techniques for Assessment of Parkinsonism for Diagnosis and Rehabilitation. Singapore: Springer, 2022, pp. 119–133.

[18] A. Tsanas, M. Little, P. E. McSharry, and L. O. Ramig, “Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests,” IEEE Trans. Bio-Med. Eng., vol. 57, no. 4, pp. 884–893, Nov. 2009.

[19] R. Viswanathan, S. P. Arjunan, A. Bingham, B. Jelfs, P. Kempster, S. RagHAV, and D. K. Kumar, “Complexity measures of voice recordings as a discriminative tool for Parkinson’s disease,” Biosensors, vol. 10, no. 1, p. 1, Dec. 2019.

[20] P. Khogesteh, R. Viswanathan, B. Aliahmad, S. RagHAV, P. Zharn, and D. K. Kumar, “Parkinson’s disease diagnosis based on multivariate deep features of speech signal,” in Proc. IEEE Life Sci. Conf. (LSC), Oct. 2018, pp. 187–190.

[21] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, “Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson’s disease,” Appl. Soft Comput., vol. 62, pp. 649–666, Jan. 2018.

[22] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, “Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection,” Biomed. Eng., vol. 19, pp. 1–19, Jan. 2007.

[23] V. S. Chaudhary, D. Kumar, G. P. Mishra, S. Sharma, and S. Kumar, “Plasmonic biosensor with gold and titanium dioxide immobilized on photonic crystal fiber for blood composition detection,” IEEE Sensors J., vol. 22, no. 9, pp. 8474–8481, May 2022.

[24] G. P. Mishra, D. Kumar, V. S. Chaudhary, and S. Kumar, “Design and sensitivity improvement of microstructured-core photonic crystal fiber based sensor for methane and hydrogen fluoride detection,” IEEE Sensors J., vol. 22, no. 2, pp. 1265–1272, Jan. 2022.

[25] J. Rusz, J. Hlavnicka, T. Tykalova, J. Buskova, O. Ulmanova, E. Ruzicka, and K. Sonka, “Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder,” Sleep Med., vol. 19, pp. 141–147, Mar. 2016.

[26] R. Das, “A comparison of multiple classification methods for diagnosis of Parkinson disease,” Expert Syst. Appl., vol. 37, no. 2, pp. 1568–1572, 2010.

[27] P.-F. Guo, P. Bhattacharya, and N. Khara, “Advances in detecting Parkinson’s disease,” in Proc. Int. Conf. Med. Biometrics, 2010, pp. 306–314.
[50] D. H. Whalen, A. M. Kang, H. S. Magen, R. K. Fulbright, and J. C. Gore, “Predicting midsagittal pharynx shape from tongue position during vowel production,” *J. Speech, Lang., Hearing Res.*, vol. 42, no. 3, pp. 592–603, Jun. 1999.

MOHAMMOD ABDUL MOTIN (Senior Member, IEEE) received the B.Sc.Eng. and M.Sc.Eng. degrees in electrical and electronic engineering from the Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, in 2011 and 2014, respectively, and the Ph.D. degree from the Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC, Australia, in 2020. He is currently working as an Assistant Professor at the Rajshahi University of Engineering and Technology. His research interests include biomedical signal processing and modeling, time series analysis, machine learning, embedded systems, monitoring neurological disorders, and sensor devices for wearable health care monitoring.

NEMUEL DANIEL PAH (Member, IEEE) received the B.E. degree in electronics from the Institut of Teknologi, Sepuluh Nopember, Indonesia, in 1992, and the Ph.D. degree from RMIT University, Melbourne, VIC, Australia, in 2003. He is currently an Academic at the University of Surabaya, Indonesia, and working with RMIT University. His research interests include signal processing of bioelectrical signals and applications of wavelet transforms for signal analysis.

SANJAY RAGHAV is currently an Associate Professor at RMIT University and a Senior Lecturer at Monash University. He is a Senior Consultant Neurologist at the Monash Medical Centre and Frankston Hospital. He is an Assigned Supervisor and a Mentor to RACP trainees. He has a special interest in movement disorders which include tremors, Parkinson’s disease, tics, dystonia, and Huntington’s disease. In addition, he is involved in several research projects related to movement disorders at RMIT and Monash University. Several of his research papers and articles have been published in esteemed peer-reviewed medical journals and presented at national and international conferences. He is actively involved in teaching and training of Neurology Registrars and Fellows.

DINESH KANT KUMAR (Senior Member, IEEE) received the B.E. degree from the Indian Institute of Technology (IIT), Madras, India, and the Ph.D. degree from the IIT, Delhi, India, in 1990. He has been a Professor at RMIT University, Melbourne, since 1997, where he leads research in biosignals, and his passion is to develop devices for affordable diagnostics. He has experience in two start-ups. He has authored more than 420 articles and seven books. He has been a Senior Professorial Fellow in Brazil, Vajra Fellow in India, and is on the TGA (Australia) advisory panel for medical devices, and a fellow of the Australasian Institute of Digital Health (AIDH).