Supplementary material for **'Investigating Voice as a Biomarker for leucine-rich repeat kinase 2-Associated Parkinson's Disease'**

Supplementary Analysis

Feature extraction

For each recording, we extracted 292 summary measures (also referred to as *features*). These features can broadly be characterized as: (1) *Descriptive features*: statistical characteristics of voice. (2) *Vocal fold vibration-based features*: variation in speech frequency and amplitude. (3) *Cepstral coefficients*: subtle changes in the placement of the articulators (mouth, tongue, teeth and lips). (4) *Aeroacoustics-based features*: degree of turbulent noise in speech due to incomplete vocal fold closure. (5) *Wavelet features*: extended time-frequency domain properties. Details regarding these feature categories are provided in Supplementary Table 1.

Feature selection

For each pairwise comparison, salient features were identified using the following 5 feature selection algorithms that help enhance the explanatory power of the analysis by removing redundant and less informative features: (1) Minimum redundancy maximum relevance (mRMR) [27], (2) Gram-Schmidt orthogonalization (GSO) [25], (3) RELIEF [26], (4) Local learning-based feature selection (LLBFS) [28], and (5) Least absolute shrinkage and selection operator (LASSO) [29].

Validation

In this study, we used two model validation schemes: (1) 10-fold cross-validation: this scheme involves randomly splitting the data into two non-overlapping parts, the first part of

the data (comprising 90% of the recordings) are used to train the model (i.e., learn the underlying differences in patterns of voice-based features for each pairwise comparison), while the remaining 10% of the recordings are used for validation (i.e., evaluate the accuracy of model predictions). This process of randomized selection of training and validation sets was repeated multiple times, and the discrimination accuracies (quantified using sensitivity and specificity) were calculated on each repetition. (2) Leave-one-subject-out: this scheme involves splitting the data such that all recordings from only one participant are used for model validation, while all remaining recordings are used for training. This process is repeated multiple times. Note that for both schemes, the accuracy of model predictions are computed using only the validation set. Basically, the model is blinded to the validation set during the training process, which helps gauge generalizability of the model to previously unseen similar datasets.

In this study, we analysed all available/suitable voice recordings leading to a mismatch in the five group sizes (Table 1). Whilst there are fewer *LRRK2*-associated PD participants compared to iPD, it should be noted that we used a 'balanced cross-validation scheme' that results in an equal number of samples across different classes for each pairwise comparison. This scheme helps mitigate the issues associated with imbalanced datasets/differences in group sizes.

***LRRP2-PD vs idiopathic PD* (excluding 1 *LRRK2*-PD participant)**

LRRP2-PD participants had longer mean disease duration (n = 7, mean = 10.3 years, SD = 11.8 years) compared to participants with iPD (n = 17, mean = 5.4 years, SD = 5.8 years). One *LRRK2*-PD participant had a disease duration of 36 years, and without this participant, the mean disease duration for the remaining 6 *LRRK2*-PD participants was 6 years (SD = 3.7

years), which is similar to the mean disease duration for iPD participants (5.4 years). Statistical analyses were thus performed separately by excluding this *LRRK2*-PD participant (disease duration 36 years, female, age 85 years, one voice recording available). The rationale for this analysis was to investigate disease duration as a potential confounding factor in discriminating *LRRK2*-PD versus iPD. Using 10 voice recordings collected from the remaining 6 *LRRK2*-PD participants, we recomputed the accuracy in discriminating *LRRK2*-PD versus iPD.

Using only the 10 top-ranked features for this pairwise comparison, the mean sensitivity and mean specificity was 97.0% (SD 15.5%) and 87.2% (SD 31.1%) in discriminating *LRRK2*-PD from iPD using all recordings, 99.0% (SD 10.0%) and 82.4% (SD 38.1%) in discriminating *LRRK2*-PD from iPD using only female recordings, and 100% (SD 0%) and 88.9% (SD 31.6%) in discriminating *LRRK2*-PD from iPD using only male recordings. These accuracies were obtained using 10-fold cross-validation. Note that the excluded *LRRK2*-PD recording was collected from a female participant; hence the accuracy in discriminating *LRRK2*-PD versus iPD using only male recordings was the same as those reported in Table 1. For leave-one-subject-out cross-validation, mean sensitivity and mean specificity were 88.5% (SD 7.8%) and 81.3% (SD 12.3%) respectively, in discriminating *LRRK2*-PD from iPD using all recordings. The sample size was too small to draw any reliable inference based on discrimination accuracies for subgroup analysis stratified by sex. These sensitivity and specificity values (obtained using recordings from n = 6 *LRRK2*-PD participants) are in close agreement with the discrimination accuracies obtained using all available recordings for *LRRK2*-PD (n = 7), as reported in Table 1 and Supplementary Table 2. Encouragingly, for all pairwise comparisons reported above, the sensitivity and specificity results differed statistically significantly from comparable results obtained from completely randomized predictions about which

participants had a *LRRK2* mutation or were iPD (these predictions are akin to outcomes of an unbiased coin flip and are based on chance alone). Moreover, the differences in *LRRK2*-PD and iPD voice recordings is also evident from the scatterplot of salient features that shows two distinct clusters, as presented in Supplementary Figure 4.

Detecting and characterising identity confounding

Digital recordings of voice and other sensor data from individuals can capture properties of these data which are unique to particular individuals, in the following way. For example, in voice, the combination of various vocal features such as vocal pitch and spectral envelope may occupy an approximately unique region in feature space, distinct from all other individuals in the study. This uniqueness can interact with highly nonlinear classifiers to produce 'identity confounding' whereby the classifier finds a relationship between the individual and their specific clinical grouping, rather than a relationship between clinical symptoms and clinical grouping. This inadvertent relationship can confound predictions, which means that it is necessary to quantify the extent to which the classifier is making predictions which would generalize to individuals not in the study. To quantify this potential confound, we counted the number of observations and individuals per unique predicted value in the classification tree (across 500 trees used); whereby each tree was built using a bootstrap sample of the training data. Averaged over all cross-validation repetitions (10-fold with 100 repetitions), we found that the average number of observations per unique predicted value and corresponding number of participants were 8.2 and 4.5, respectively, for *LRRK2*-PD versus iPD. The average number of participants in the training set (after balancing the data and bootstrapping) in each tree for the above pairwise comparison was 10.9. While training the model, the unique predicted values thus had observations from around 41% of the individuals that were in the training set (4.5 individuals on average), thus

indicating that this form of identity confounding is unlikely to be a significant factor in the results presented in this study.

Prodromal versus Nonprodromal

We investigated if two non-manifesting carriers classified as being in the prodromal state were more similar to their non-prodromal counterparts compared to participants with *LRRK2*-PD. One of the two participants meeting prodromal criteria was more similar to the manifesting LRRK2 carriers on the basis of the two most salient voice features, (Supplementary Figure 3). However, the sample size was too small to draw any reliable inferences.

LRRK2 carriers vs idiopathic PD

For the *LRRK2* carrier group, we analyzed 50 recordings, collected from 27 individuals (mean age: 61.9 years (SD 15.3); % female: 48.2%; mean UPDRSIII: 8.7 (SD 12.4)), whereas for the iPD group, we analyzed 32 recordings, collected from 17 individuals (mean age: 63.4 years (SD 8.7); % female: 53.0%; mean UPDRSIII: 22.8 (SD 10.0)). Age for the *LRRK2* carrier and iPD groups were similar at 5% significance level (unpaired t-test). As expected, UPDRS III for the *LRRK2* carrier group was significantly lower compared to the iPD group (Mann–Whitney U test). In distinguishing *LRRK2* carriers and iPD, using the 10 most salient features, the mean sensitivity was 74.8% (SD 26.5%) and mean specificity was 83.0% (SD 22.3%). We used 10-fold CV with 100 repetitions. The differences in the features for *LRRK2*-carriers versus iPD were thus less pronounced, compared to the case when features for *LRRK2*-PD were compared against iPD. Further investigations are using larger cohorts are needed to investigate if non-manifesting and manifesting *LRRK2* carriers can be treated as belonging to the same clinical group.

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FIGURES

SUPPLEMENTARY TABLE 1: Brief description of five categories of features extracted from the voice recordings.

Category	Brief description
Category 1: Descriptive features:	
Mean, median, standard deviation, skewness, interquartile range etc.	Quantifies statistical characteristics of the voice signal
Category 2: Vocal fold vibration-based features:	
Jitter	Quantifies the instabilities of the oscillating pattern of the vocal folds by measuring cycle-to-cycle changes in the fundamental frequency (measure of roughness in voice)
Shimmer	Quantifies the instabilities of the oscillating pattern of the vocal folds by measuring cycle-to-cycle changes in the amplitude (measure of roughness in voice)
Teager-Kaiser Energy Operator (TKEO)	Measures the instantaneous changes in voice energy (takes into account both amplitude and frequency)
Category 3: Cepstral coefficients based features:	
Mel Frequency Cepstral Coefficients (MFCCs)	Computes the contribution of the energy of the speech signal at each frequency band (are aimed at detecting subtle changes in the motion of the articulators)
Category 4: Aeroacoustics, aperiodicity, and frequency based features:	
Recurrence Period Density Entropy (RPDE)	Quantifies any ambiguity in fundamental pitch (RPDE is zero for perfectly periodic signals and one for purely stochastic signals). Higher RPDE has been associated with voice impairment
Detrended Fluctuation Analysis (DFA)	Characterizes the changing detail of aero-acoustic breath noise
Pitch Period Entropy (PPE)	Measures the impaired control of stable pitch, a property common in PD
Harmonics-to-Noise Ratio (HNR)	Quantifies noise in the speech signal, caused mainly due to incomplete vocal fold closure
Glottal to Noise Excitation (GNE) ratio	Quantifies the extent of noise in speech using linear and nonlinear energy measures
Vocal Fold Excitation Ratios (VFER)	Quantifies the extent of noise in speech using energy (linear and nonlinear) and entropy-based measures
Perturbation Quotient (PQ)	Quantifies variations in speech signal
Glottis Quotient (GQ)	Quantifies properties of the vocal folds (when glottis is open and closed)
F0 contour features	Measures based the summary statistics of the fundamental frequency
Category 5: Wavelet-based features:	
Wavelet related measures	Variants of above-discussed summary measures applied to wavelet coefficients of the speech signal

SUPPLEMENTARY TABLE 2. Discrimination accuracy for the leave-one-subject-out (LOSO) cross-validation (CV) scheme for the three pairwise comparisons: *LRRK2*-associated Parkinson's disease (*LRRK2*-PD) versus idiopathic PD (iPD), non-manifesting *LRRK2* mutation carriers (NMC) versus related non-carriers (RNC), and NMC versus healthy controls, computed using a machine learning algorithm (random forest) and a naïve benchmark (randomized predictions).

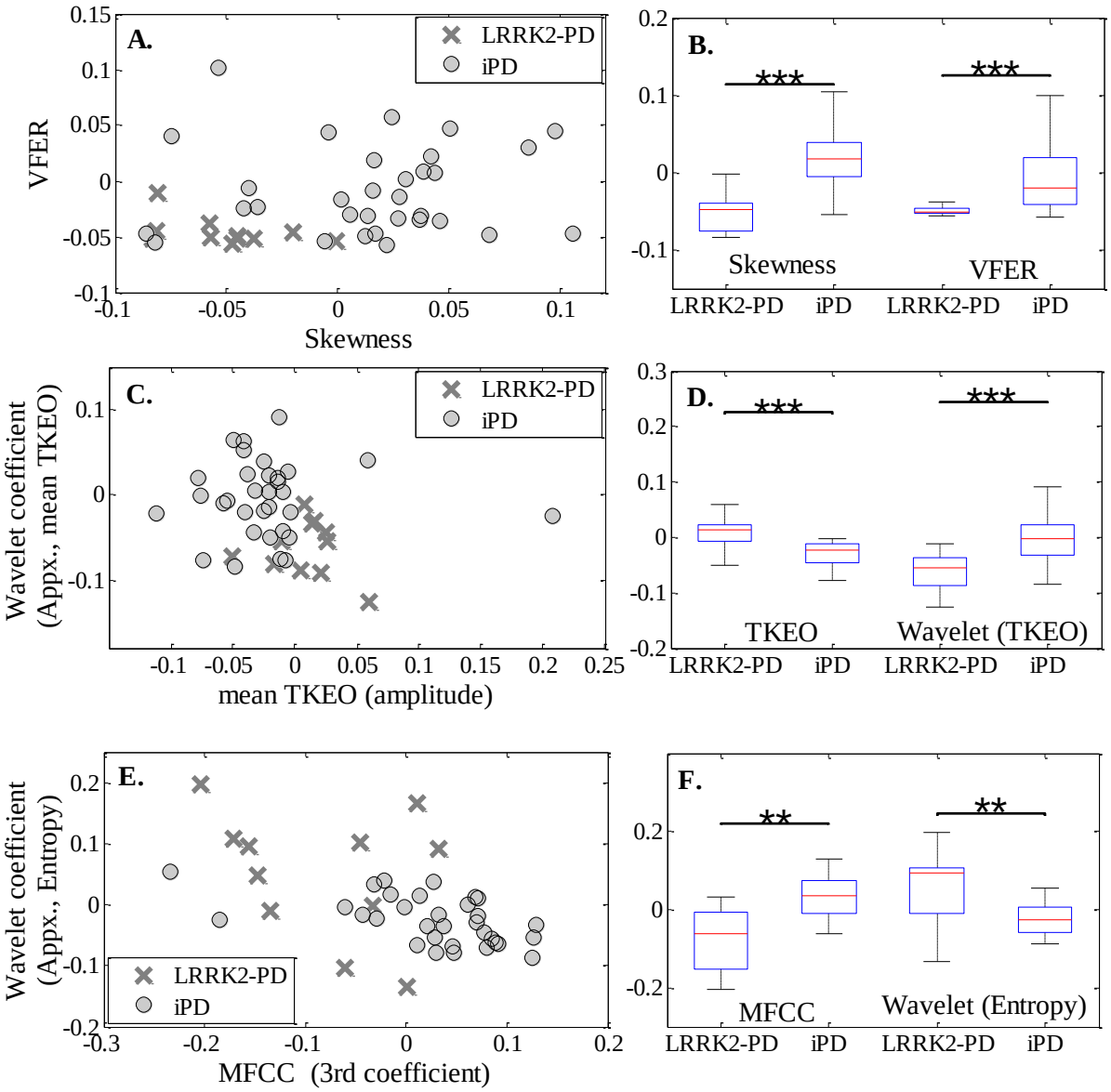
Discrimination accuracy	Sensitivity (%) Mean (SD)	Specificity (%) Mean (SD)
LRRK2-PD vs iPD *(ALL)*		
Random forest	83.7% (7.1%)	88.5% (8.4%)
Randomized predictions	51.6% (20.2%)	47.0% (18.3%)
NMC vs RNC (ALL)		
Random forest	67.3% (3.2%)	69.8% (4.9%)
Randomized predictions	48.6% (13.1%)	47.3% (11.1%)
NMC vs Healthy (ALL)		
Random forest	72.4% (3.2%)	69.9% (5.0%)
Randomized predictions	50.4% (10.0%)	48.9% (11.6%)

The above sensitivity and specificity values were computed separately for each of the three priority pairwise comparisons (1. *LRRK2*-PD vs iPD, 2. NMC vs RNC, and, 3. NMC vs Healthy) using a leave-one-subject-out (LOSO) cross-validation (CV) scheme, employing 10 most salient voice features. Validation scheme involved repetitive splitting of the data such that at a given CV iteration, all voice tests from only one randomly selected participant were employed for model validation, while voice tests from all remaining participants were used for training. We used LOSO CV scheme with 100 repetitions. The data was balanced to account for differences in number of participants in each clinical group. Accuracies are reported for a machine learning classifier (random forest) and a naïve benchmark based on randomized predictions (expected accuracy around 50%), using all available voice tests from the five clinical groups (1. *LRRK2*-PD, 2. iPD, 3. NMC, 4. RNC, and, 5. Healthy controls). The sensitivity and specificity values were presented in percentage (%) as mean (and standard deviation, in brackets), whereby the standard deviation denotes the variability in the accuracy across multiple CV repetitions. The rankings of the most salient features were obtained separately for each of the three pairwise comparisons, using a majority voting scheme (using 5 feature selection algorithms). Abbreviations used: iPD, idiopathic Parkinson's disease; *LRRK2*-PD, *LRRK2*-associated Parkinson's disease; NMC, non-manifesting carriers; RNC, related non-carriers; SD, standard deviation.

SUPPLEMENTARY TABLE 3: List of 10 salient features selected for the three pairwise comparisons.

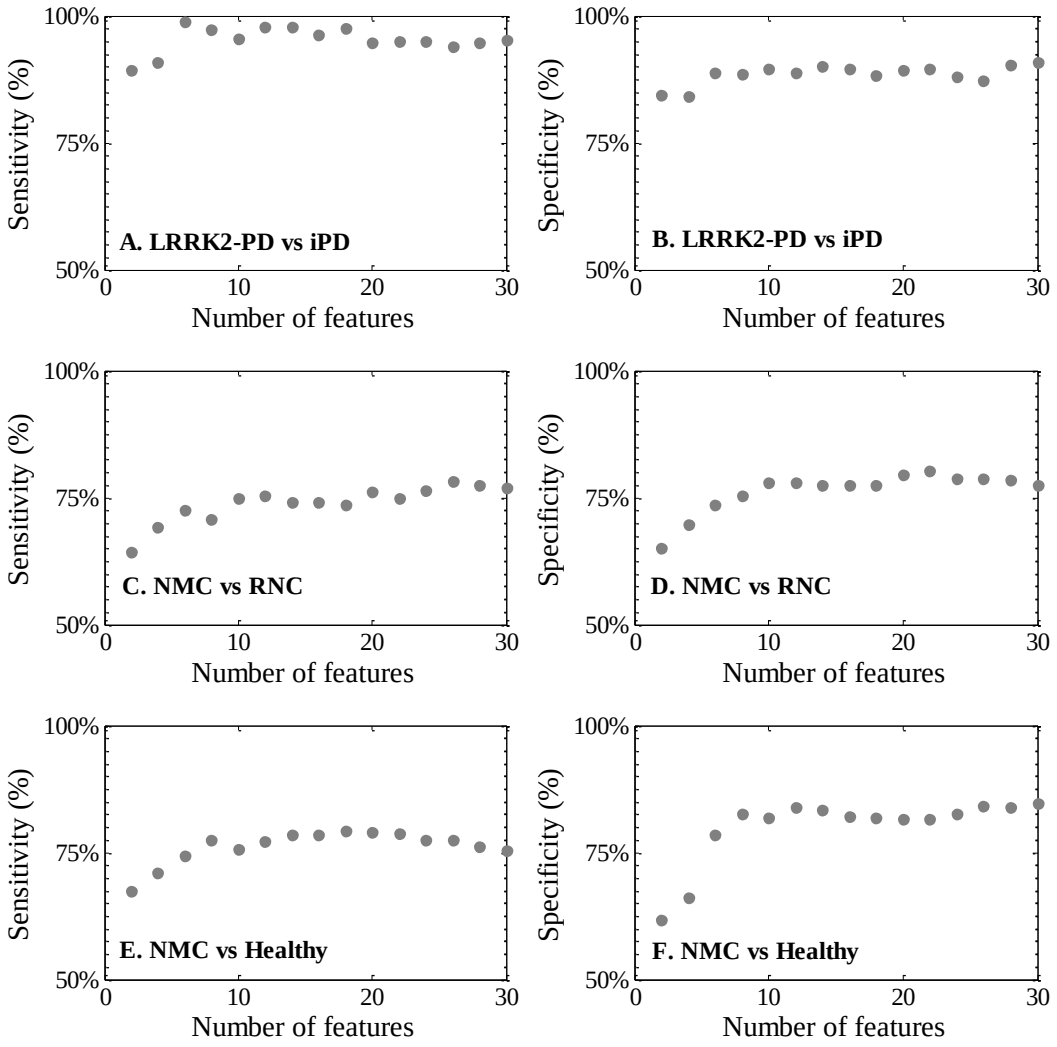
Feature Name	Brief description
Comparison 1*: LRRK2*-associated PD vs iPD	
Skewness	Quantifies asymmetry of the distribution
det_entropy_log_6_coef	Wavelet log-entropy of the 6th detail coefficient of F0, quantifies subtle changes in the details of F0 fluctuations
GNE-SEO	Glottal to Noise Excitation Squared Energy Operator, quantifies excessive noise and turbulence in the voice
VFER-LF-TKEO	Vocal Fold Excitation Ratio, quantifies incomplete vocal fold closure which creates vortices and inconsistencies across frequency bands in terms of energy
prctile50TKEO_A0	Median of the Teager-Kaiser Energy of amplitude
app_det_TKEO_mean_4_coef	Mean Teager-Kaiser Energy of the 4th wavelet decomposiition coefficient decomposing F0, quantifies subtle changes in the energy of F0
medMFCC3	Median value of the 3rd MFCC coefficient, quantifies envelope structure fluctuations
app_entropy_log_8_coef	Entropy of the 8th approximation wavelet decomposition coefficient, quantifies changes in F0
det_entropy_shannon_6_coef	Shannon entropy of the 6th detail wavelet decomposition coefficient, quantifies changes in F0
Q1	25th quartile
Comparison 2*:* Non-manifesting carriers (NMC) versus related controls	
det_LT_TKEO_mean_4_coef	Mean Teager-Kaiser Energy of the 4th detail wavelet decomposition coefficient, quantifies changes in F0
HNR(1)	Harmonics to Noise Ratio, quantifies signal to noise, i.e. the extent of vocal noise using standard autocorrelation
Mean(A0)	Mean amplitude
medShimmer	Quantifies amplitude perturbations
PQ11.class_Schoentgen	Amplitude perturbation using a 11-sample window
muDiffMFCC5	5th MFCC
medMFCC10	Median of 10th MFCC, quantifies mostly higher harmonic components in the signal
medJitter	Quantifies frequency perturbations
mode_F0	Dominating F0 value
det_LT_TKEO_mean_8_coef	Mean Teager-Kaiser Energy of the 8th detail wavelet decomposition coefficient
Comparison 3*:* Non-manifesting carriers (NMC) versus unrelated controls	
muDiffMFCC13	13th MFCC
muDiffMFCC8	Quantifies mostly higher harmonic components in the signal
medShimmer	Quantifies amplitude perturbations
Ed2_8_coef	Wavelet energy of the 8th wavelet coefficient
PQ11.class_Schoentgen	Amplitude perturbation using a 11-sample window
medMFCC10	Quantifies mostly higher harmonic components in the signal
Ed2_7_coef	Wavelet energy of the 7th wavelet coefficient
det_entropy_log_6_coef	Wavelet log energy the 6th detail wavelet decomposition coefficient, quantifies changes in F0
P0	Perturbation quotient (zeroth order)
HNR(1)	Harmonics to Noise Ratio, quantifies signal to noise, i.e. the extent of vocal noise using standard autocorrelation

SUPPLEMENTARY FIGURE 1. Scatterplots and boxplots of salient features for the pairwise comparison: *LRRK2*-associated Parkinson's disease (*LRRK2*-PD) versus idiopathic PD (iPD).



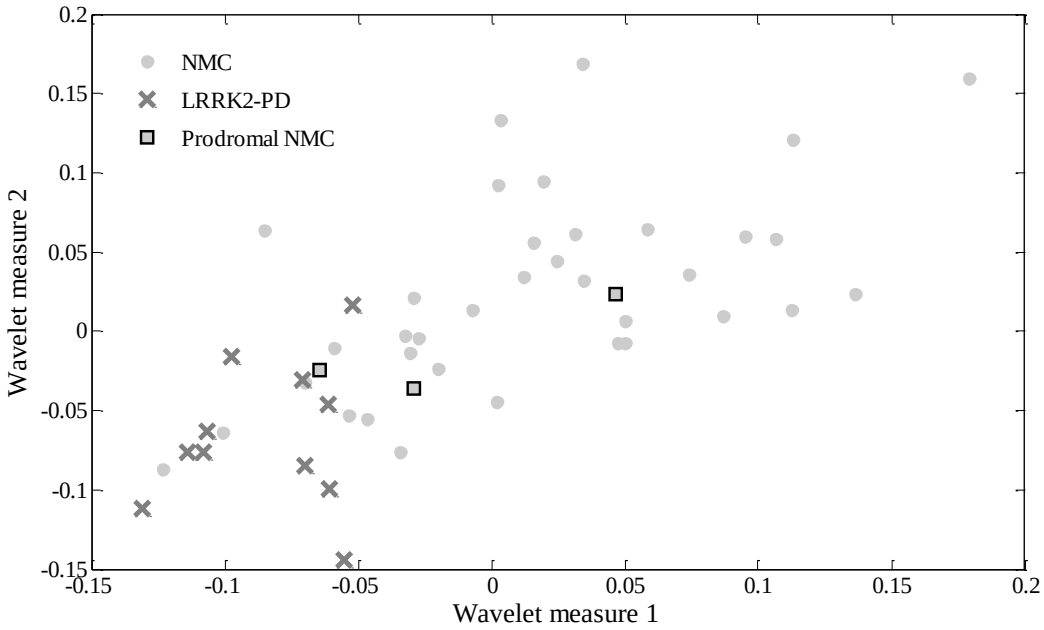
Panel A plots two salient features, Skewness (amplitude) and Vocal Fold Excitation Ratio (VFER, the degree of signal strength over noise resulting from incomplete vocal fold closure), both skewness and VFER were significantly different between the two groups, (p<0.001, denoted by ***) (Panel B). Panel C plots mean Teager Kaiser Energy Operator (TKEO, quantifies instantaneous changes in voice energy) and a Wavelet coefficient (based on TKEO), while Panel D shows that these features were significantly different between *LRRK2*-PD and iPD. Panel E plots the Mel Frequency Cepstral Coefficients (MFCC, quantifies vocal fold dynamics taking into account the properties of the articulators) and Entropy (entropy computed after wavelet decomposition, computes the extent of randomness in a signal), while Panel F shows that these features were statistically significantly different (p<0.01). Features with high discriminatory power were identified using five different feature selection algorithms. The above plots were generated using all voice recordings collected from participants with *LRRK2*-PD and iPD. p values reported above were computed using the nonparametric two-sided Kolmogorov-Smirnov (KS) test.

SUPPLEMENTARY FIGURE 2. Discrimination accuracies as a function of the number of salient features used in the machine learning discrimination analysis, for the three pairwise comparisons: *LRRK2*-associated Parkinson's disease (*LRRK2*-PD) versus idiopathic PD (iPD) (Panels A and B), non-manifesting *LRRK2* mutation carriers (NMC) versus related non-carriers (RNC) (Panels C and D), and NMC versus healthy controls (Panels E and F).



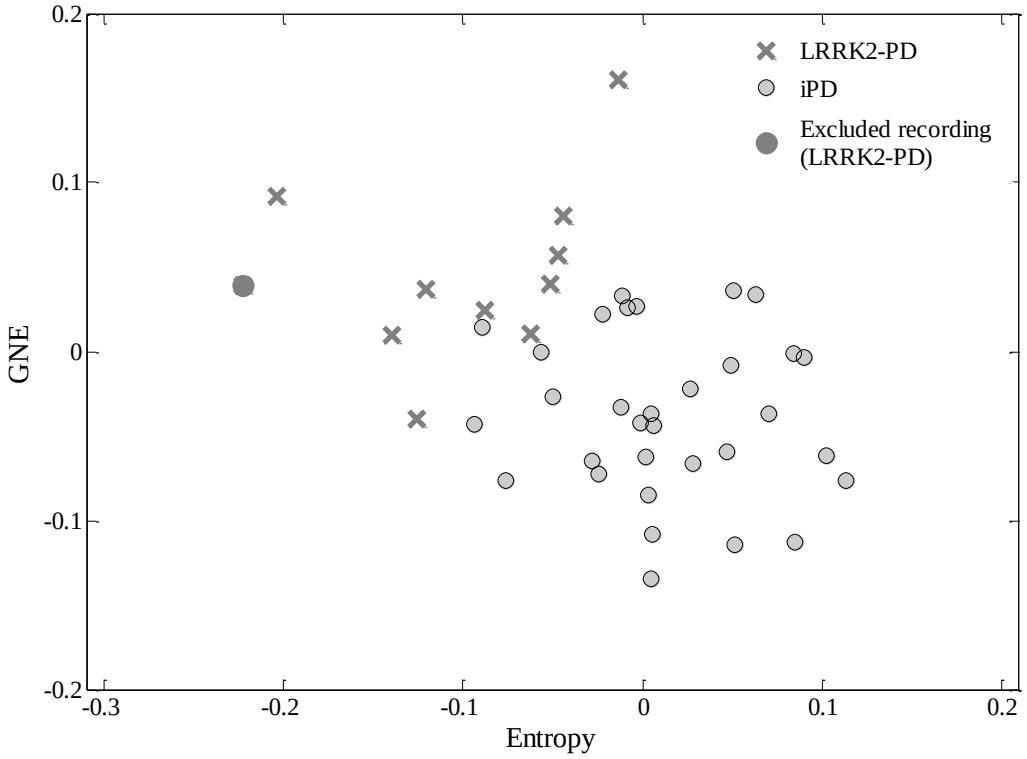
The above accuracies were computed using all available voice recordings from the five clinical groups (1. *LRRK2*-PD, 2. iPD, 3. NMC, 4. RNC, and, 5. Healthy controls), using 10-fold cross-validation (100 repetitions). The rankings of the most salient features were obtained using a majority voting scheme (using 5 feature selection algorithms). The feature rankings were obtained separately for each of the above 3 pairwise comparisons (1. *LRRK2*-PD vs iPD, 2. NMC vs RNC, and, 3. NMC vs Healthy). Features were added into the machine learning classifier (random forest) in increments of 2 (starting from 2, and going up to 30), whereby higher ranked features were added first. The whole process of training and validation was repeated each time two new features were included. Mean sensitivity and specificity values are denoted as grey circles and reported in percentage (%).

SUPPLEMENTARY FIGURE 3. Scatterplot of two most salient features that help discriminate *LRRK2*-associated Parkinson's disease (*LRRK2*-PD) versus non-manifesting *LRRK2* mutation carriers (NMC), plotted along with features from prodromal participants.



The feature rankings were obtained separately for the above pairwise comparison: *LRRK2*-PD versus NMC. We analysed three voice recordings from two prodromal participants (denoted as a grey square).

SUPPLEMENTARY FIGURE 4. Scatterplot of two most salient features that help discriminate *LRRK2*-associated Parkinson's disease (*LRRK2*-PD) versus idiopathic PD (iPD), plotted along with voice features from the excluded *LRRK2*-PD participant.



We analysed 10 voice recordings collected from six *LRRK2*-PD participants (denoted as grey crosses) and 32 voice recordings obtained from seventeen iPD participants (denoted as light grey circles). Analysis excluded one *LRRK2*-PD participant who had disease duration of 36 years, note that we only had one decent quality voice recording for this participant (denoted as a dark grey circle).