Audio-Based Neural Network to Classify Patients With Early Parkinson's Disease

Hou, Weikang ; Quan, Changqin ; Chen, Zhonglue ; Cao, Sheng ; Ren, Kang ; Su, Wen ; Luo, Zhiwei

(Citation)
IEEE Access,13:154283-154294

(Issue Date)
2025-09-02

(Resource Type)
journal article

(Version)
Version of Record

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2025, accepted 25 August 2025, date of publication 2 September 2025, date of current version 8 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3604877

METHODS

Audio-Based Neural Network to Classify Patients With Early Parkinson's Disease

WEIKANG HOU[1], CHANGQIN QUAN[1], ZHONGLUE CHEN[2], SHENG CAO[1], (Member, IEEE), KANG REN[2], WEN SU[3], AND ZHIWEI LUO[1]

[1]Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan
[2]GYENNO Technologies Company Ltd., Shenzhen 518000, China
[3]Department of Neurology, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing 100700, China

Corresponding authors: Changqin Quan (quanchqin@gold.kobe-u.ac.jp) and Wen Su (suwendy@126.com)

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP25K15078.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Beijing Hospital.

ABSTRACT Parkinson's disease is a progressive neurodegenerative disorder, and identifying patients at the premotor stage enables early intervention and improved treatment outcomes. Dysarthria affects over 90% of Parkinson's patients, making speech a valuable biomarker. In this study, we propose an end-to-end deep learning model to detect early-stage Parkinson's disease based on speech signals. The model was trained and evaluated using recordings from 131 early-stage PD patients and 42 healthy controls, including sustained vowels (e.g., /A/, /O/) and repetitive syllables (e.g., /pa/, /ta/). Experimental results demonstrate that our model outperforms various deep learning and ensemble learning classifiers in terms of detection accuracy and F1-score, Among indicators, ACC reached 0.78, and F1-score reached 0.831. Furthermore, we explore the temporal dynamics of speech sequences to reveal their correlation with disease progression.

INDEX TERMS Parkinson disease, deep learning, spectrum, speech recognition.

I. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease that occurs in middle-aged and elderly people. The gradual loss of dopaminergic neurons in the substantia nigra of the midbrain is one of the main pathological features of PD, which ultimately leads to dysfunction of the basal ganglia and a series of movement disorders, such as tremor, bradykinesia, and postural instability, and seriously affects the quality of life of patients. Consequently, basal ganglia dysfunction gives rise to bradykinesia, and postural instability, markedly reducing quality of life. In Europe, the prevalence of Parkinson's disease has reached 108-257/100,000 per year in 2020, which is a high level among neurodegenerative diseases [1]. In addition to movement disorders, Parkinson's disease leads to a range of non-motor disorders such as olfactory disturbances, dementia, sleep disorders, and language disorders [2], with patients with REM sleep behavior disorders (RBD) having a risk of PD of up to 80-90%. In early studies related to the pathology of Parkinson's disease, these non-motor symptoms were often attributed to other diseases or ignored. With the deeper exploration of studies related to neurodegenerative diseases, these non-motor symptoms have been shown to frequently appear several years - or even more than a decade - before the classical motor symptoms [3]. Therefore, tracking and detecting non-motor symptoms has become crucial in optimizing the early diagnosis of Parkinson's disease. In addition, improving the timeliness of intervention has attracted increasing attention from the medical community due to its critical role.

In general, the diagnosis of Parkinson's disease focuses on the clinical assessment of movement status, but this diagnostic approach based on movement tracking has some shortcomings. On the one hand, such diagnosis requires sufficient clinical experience of clinicians, because early PD

The associate editor coordinating the review of this manuscript and approving it for publication was Ding Xu.

patients may have very mild motor symptoms, and traditional clinical movement tests (e.g., UPDRS score) may not be sensitive enough to recognize mild symptoms. On the other hand, movement monitoring requires high-precision motion sensors and relatively closed experimental environments, and its results are affected by a series of external factors such as equipment quality. This leads to the fact that the accuracy and sensitivity of the results of movement analysis are very dependent on professional equipment and manual evaluation with rich clinical experience, the complexity of the data is high and the degree of automation is low. In addition, complex motion detection devices and sensors make it difficult to follow the patient's condition remotely for long periods. Although voice analysis methods also have certain requirements for recording devices (such as microphones), compared to motion sensors, the process of obtaining voice data is more natural and non-contact, and is more accessible and scalable in a variety of practical application scenarios. Therefore, in the task of assisting in the screening of mild cases, voice-based analysis methods are expected to serve as a more convenient and universally applicable supplementary measure.

Modern Parkinson's disease detection systems are often based on multiple types of instruments to improve early detection accuracy and diagnostic efficiency in terms of imaging, movement, and biomarkers. The most common of these symptoms are abnormal voice symptoms such as vocalization, resonance, and intonation, and these motor dysarthria caused by PD can significantly affect the speech ability of the patients. More importantly, some existing studies have found that most patients develop voice defects in the early stage of the disease [4], [5], [6], and the movement data in the same period often do not have the sensitivity required for early diagnosis. Signal data also has the advantages of being non-invasive, convenient, and less costly to obtain and analyze. The non-invasive evaluation capability provides a good basis for remote and continuous disease tracking, which opens up better ways for PD healthcare optimization and personalized treatments, and the low cost also makes large-scale comprehensive screening of PD possible.

The general prodromal symptoms of patients with Parkinson's disease include hyposmia, impaired color vision, constipation, and erectile dysfunction; in the early stage, dysarthria and constipation may also occur [7]. And the severity of dysarthria tends to increase as the disease progresses. With the gradual deterioration of the disease, dysarthria becomes more serious and the patient's pronunciation becomes more and more difficult to hear. UPDRS (Unified Parkinson's Disease Rating Scale) is a measure of the severity of Parkinson's disease, proposed by the American Movement Disorder Society (AMDS) in the 1980s. The unified parkinson's disease rating scale was developed by the American Movement Disorders Society in the 1980s to quantitatively assess the severity of Parkinson's disease in patients by quantifying symptoms such as hypophonia, slurred and monotonous speech, and rapid or irregular speech rate, rapid and irregular speech rate are mainly based on self-reporting by the patient and subjective assessment by the physician. However, the reproducibility of this method is poor and the ability to dynamically monitor is limited; more importantly, because of the lack of analysis of specific acoustic characteristics, in this era of emphasis on precision medicine, qualitative evaluation may not be able to comprehensively and objectively assess language impairment in Parkinson's disease [8].

In recent years, the rapid development of signal processing, speech recognition, and deep learning technologies has significantly advanced the field of artificial intelligence (AI). AI has shown great potential in assessing speech disorders in Parkinson's disease (PD) [9]. It enables high-throughput, automated analysis of patient characteristics and supports dynamic, personalized medical decision-making by leveraging large-scale data processing and time-series modeling. Karaman et al. developed a deep convolutional neural network model based on speech signals and migration learning to recognize PD using biomarker-derived speech signals, achieving 89.75% accuracy and 91.50% sensitivity [10]. Khaskhoussy and Ayed used CNN to extract deep features from raw speech signals from patients with Parkinson's disease and used a multilayer perceptron to detect Parkinson's disease, achieving precision 75% with 50-50 cross-validation [11]. Asmae Ouhmida et al. used two databases of the UCI machine learning repositories to use convolutional neural networks (CNNs) and artificial neural networks (ANNs) to categorize healthy patients and PD based on voice characteristics, with an accuracy of up to 93.10%. Speech analysis in Parkinson's disease mainly focuses on raw features extracted manually by signal processing techniques [12] and deep features automatically learned and extracted from speech signals by deep neural networks with high-dimensional abstract representations [13], which include fundamental frequency, volume, and jitter, and have the advantages of high interpretability and low computational complexity, and are good for clinical quantitative analysis. They have good usability in clinical quantitative analysis, however, the reliance on raw features on manual design and development limits their application to the task of analyzing large-scale multimodal datasets, where deep features are generally used to capture complex speech patterns and nonlinear variations. Past studies have shown that both types of features have the potential to show the best classification results in different multilingual dataset-based classification tasks [14], but in most of the studies, these two types of features have been considered and processed separately [15], [16], [17]. In clinical applications, if we can break through the limitation of a single feature, we can improve the accuracy of the model prediction and the comprehensiveness of the model, while maintaining good interpretability. This can provide strong support for personalized long-term prediction of Parkinson's disease process and treatment evaluation.

Section II of this paper describes in detail the related work and the methodology used in this study, Section III discusses the dataset and the specific setup of the experiments, Section IV discusses the possibilities of exploratory analyses and summarizes the experimental results, and Section V discusses the nature of the research in this paper and presents the relevant conclusions.

II. MATERIALS AND METHODS

A. SPEECH CHARACTERIZATION PLAYS A ROLE IN THE DETECTION AND DIAGNOSIS OF PARKINSON'S DISEASE

Motion detection based on motor function assessment has been one of the main clinical diagnostic tools for Parkinson's disease for a long time [18]. Neuronal degeneration in the substantia nigra region of the brain in PD patients leads to the destruction of the basal ganglia, which regulates motor function, thus causing uncontrollable limb tremor in patients. Currently, the most popular detection methods are advanced motion sensors [19] and handwriting analysis [20]. The rapid development of given signal recording techniques and wearable devices has reinvigorated the subject in recent years, and skills are now capable of realizing the distinction between PD patients and healthy individuals, and even distinguishing PD patients taking and not taking specific medications by their notes [21]. However, in the diagnosis of early-stage PD, motion detection has some insurmountable disadvantages, such as the fact that motor symptoms are usually not obvious in early-stage patients [2], and the cost of data detection tracking is generally high.

In patients with PD, ambiguous speech and dysarthria usually manifest at the early stage of the disease, and decreased coordination of the oral muscles and respiratory control difficulties caused by PD manifest themselves in the form of slower speech speed, disorganized intonation, and reduced volume. In detail, motor dysfunction caused by neurodegenerative changes in Parkinson's disease patients leads to muscle stiffness and weakened respiratory support, resulting in a narrower fundamental frequency range, unstable pitch, and weaker and unstable loudness when PD patients speak. In contrast, healthy individuals exhibit natural and stable changes in fundamental frequency and loudness when speaking. Furthermore, from the perspective of speech data characteristics, the speech spectrum of healthy individuals is typically evenly distributed with minimal or no fluctuations. In contrast, PD patients exhibit significant abnormalities in spectral characteristics due to the severe fluctuations in voice timbre, reflecting changes in phonation and resonance of the vocal fold during speech.

Traditional PD speech analysis relies on a manual qualitative assessment, which is inefficient and highly subjective. The intervention of machine learning and deep learning technologies has made the quantitative assessment of speech features possible. With the help of these tools, it is possible to identify appropriate attributes that have not traditionally been applied in the medical diagnosis of Parkinson's disease and to rely on these alternative metrics in the preclinical stage of Parkinson's disease [22], [23], [24]. For example, Nagasubramanian and Sankayya utilized heterogeneous datasets specially developed with absolute multi-variate speech attribute processing algorithm for effective value creation, to construct a clear data model for improved Parkinson's disease detection [25]. Thanks to DL techniques represented by CNN, LSTM, and Transformer, in addition to the original speech features, neural networks can be used to extract deep nonlinear features that capture subtle spatial patterns in the spectrum and large rhythmic pauses in the time series, which can be exploited to optimize the classification performance of the model. Deep Convolutional Neural Networks (DCNN) have been used to differentiate between the voices of Parkinson's disease patients and those of healthy individuals, which, unfortunately, initially provided an accuracy of only 75% [26]. Some recently proposed related studies aim to focus on improving the effectiveness of PD classification and diagnosis tasks based on speech signal features using more advanced optimization strategies and model algorithms, which include hyperparameter fine-tuning, feature correlation analysis, synthetic minority over-sampling technique (SMOTE) to solve the category imbalance problem and migration learning based on a publicly available large database [27], [28]. These deep neural network-based methods for extracting deep features from PD speech data have shown considerable discriminative power and can capture correlations and potentially complex patterns that are difficult to distinguish in the original conventional speech features. However, DL-based deep feature studies often lack interpretability and thus have low acceptance for clinical applications, and model training is demanding in terms of both data quality and computational resources. Therefore, synthesizing raw speech and deep neural features and striking a balance between performance and usability may be one of the best solutions at present.

B. DATA SET

The speech dataset used in the study related to this paper was obtained from Department of Neurology, National Center for Geriatrics, Beijing Hospital, Institute of Geriatrics, Chinese Academy of Medical Sciences, Beijing, China.Total 173 participants, of which 131 were individuals with early-stage Parkinson's disease and the remaining 42 were healthy participant controls. For each participant, 11 different speech sounds were recorded using three different devices, and the distribution of clinical characteristics of all participants is shown in Table 1. In this paper, three different devices were used for data collection, including the RØDE Wireless GO II RX using a 2.4 GHz digital wireless transmission (Series IV 2.4 GHz), the Xbox NUI Sensor using four linear array microphones with a sampling rate of 16 kHz, and the Intel® Smart Sound Technology uses a digital microphone with a sampling rate of 16 kHz and a sensitivity of -42 dBV/Pa. We saved all the speech data collected by the devices as WAV files, and the whole process of speech

data collection was carried out in a quiet soundproof room with an average noise level of no more than 25 dB, and all microphones were located at a distance of no more than 10 cm from the patient's mouth. The entries of these speech data are shown in Table 2. Informed consent was signed by all the subjects from whom data were collected in this study. According to the relevant laws and regulations of the data collection region, the dataset used in this study does not belong to the public dataset.

In this study, each participant's speech segment was recorded using three different microphone models, generating three independent but corresponding data subsets. To assess the robustness of the model across different devices, we performed five-fold cross-validation on the data recorded by each microphone and calculated the average of the evaluation results of the three models in the final validation experiment as an indicator of the overall performance of the model.

TABLE 1. Distribution of data for PD and non-PD.

	Healthy Individuals	Patients with Early PD	Total
Persons counted	42	131	173
Age (years)	57.6 ± 6.22	54.5 ± 7.56	55.2 ± 7.32
Gender (M/F)	16/26	76/55	92/81
Hoehn-Yahr staging	–	1.606 ± 0.44	1.606 ± 0.44
MDS-UPDRS scores	–	42.2 ± 17.7	42.2 ± 17.7

Age (years),Hoehn-Yahr staging and MDS-UPDRS scores take the mean and standard deviation.

TABLE 2. Voice data entry.

No	Transcription
A	/a/
B	/ɐ/
C	/i/
D	/u/
E	/pa/
F	/ta/
G	/ka/
H	/pataka/
I	A 40-second fairy tale read in Mandarin
J	/sishisizhishishizi/
K	/igedahuawankaozheyizhidahuohama/

All participants were native speakers of Mandarin Chinese and residents of mainland China. The study included patients with Parkinson's disease (PD) and healthy controls. Additionally, to control for speech variability, participants with other known speech or cognitive impairments were excluded to ensure the voice features analyzed were primarily attributable to PD. About Ethics approval, the Parkinson's disease voice recordings were collected at Beijing Hospital with approval from the hospital's Ethics Committee. All participants provided written informed consent. The present study was conducted using de-identified data in accordance with institutional and international ethical guidelines.

Inclusion criteria for PD patients were:

1) A confirmed diagnosis of idiopathic PD following the MDS Clinical Diagnostic Criteria.
2) Classification as early-stage (Hoehn and Yahr ≤ 2).
3) Native Mandarin proficiency.
4) Ability to provide informed consent.

Exclusion criteria included:

1) Presence of other neurodegenerative diseases.
2) History of psychiatric or speech-language disorders unrelated to PD (e.g., aphasia, vocal fold pathology).
3) Non-native speakers of Mandarin.
4) Individuals with severe hearing impairment.
5) Missing or incomplete voice data.

C. PREPROCESSING AND RAW FEATURE EXTRACTION

For the extracted raw speech data, we firstly resampled all the audios at a standardized sampling rate of 48000, and then we used the python based Librosa to convert the speech files into Mel-spectrogram as shown in Figure1, and we used Librosa to extract the Mel Frequency Cepstrum Coefficient(MFCC), Root Mean Square energy (Root Mean Square), Zero Crossing Rate, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff and Pitch. In addition, we also extracted some original speech features using parselmouth (Praat) that are associated with PD pathology in previous studies [29], including fundamental frequency cyclic fluctuations (Jitter), amplitude fluctuations (Shimmer), and resonance peaks (Formants).

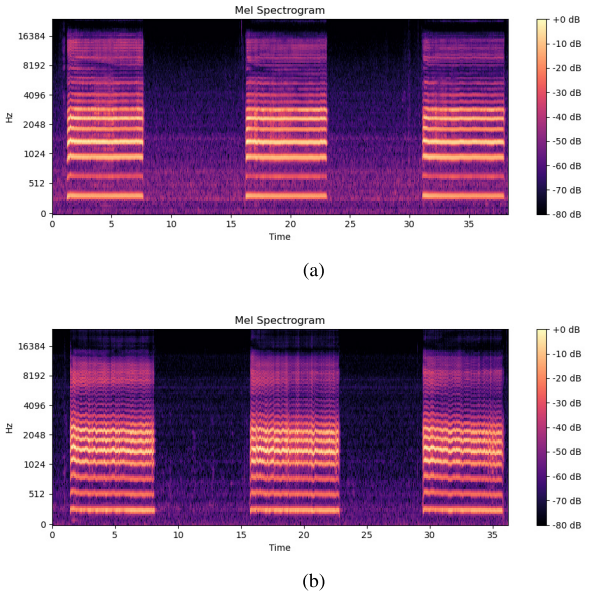


FIGURE 1. MEL spectrograms of /i/ vowels for healthy (a) and PD patients (b),In PD voice, the resonance peak position may be unclear or have low energy, appearing as blurred or discontinuous horizontal stripes in the spectrum diagram.

For each audio file, we extracted a total of 16 original speech features and then used principal component analysis (PCA) to filter out 9 features from these features with higher

correlation with the classification prediction results [30]. The benefit of feature dimensionality reduction is that removing the redundant noise from the features helps to optimize the efficiency of the model, reduce the complexity, and improve the accuracy and performance.

In addition, the nine original speech features included in this study are as follows:

1) Duration: Refers to the duration of the voice data.
2) RMS: Measures the magnitude of the energy of the speech signal, reflecting the volume.
3) SpectralCentroid: Indicates the ''center of mass'' of the spectrum, a weighted mean of frequencies present in the signal, reflecting the brightness of a sound.
4) SpectralRolloff: Is the frequency below which a specified percentage of the total spectral energy is contained.
5) MFCC_3: MFCCs (Mel frequency cepstral coefficients) are parameters that represent the characteristics of the vocal tract. The 3rd coefficient captures acoustic information in a certain frequency range.
6) MFCC_5: Similar to MFCC_3, the 5th coefficient captures acoustic info in another frequency band, providing complementary spectral features.
7) jitter: Measures tiny variations in the fundamental frequency (pitch) period, reflecting vocal fold vibration stability.
8) Shimmer: measures small variations in amplitude between consecutive vocal cycles, indicating loudness stability.
9) F0: or fundamental frequency, is the rate of vocal fold vibration and determines the pitch of the voice.

D. ARCHITECTURE OF THE PROPOSED MODEL

The method proposed in this paper aims to use speech data from patients with PD to determine whether an individual suffers from early-stage PD or not. Figure 2 describes its workflow. It is mainly divided into four steps: speech data preprocessing, raw feature extraction, feature splicing, and fusion, and prediction of classification results by neural network. To examine the deep feature representation extracted by the neural network more comprehensively, three different pre-trained neural network architectures are used in the experiments of this paper, which are Temporal convolutional neural network based on time, Two-Level Ensemble Network, and Time- Warped Input Echo State Network. Warped Input Echo State Network. The Temporal convolutional neural network based on time is characterized by a series of time-distributed 2D-CNN (Two-Dimensional Convolutional Neural Network) modules that transform the inputs into time-series dynamic features. The obtained time series dynamic features are then passed to a second module containing 1D-CNN (1D Convolutional Neural Network) modules to learn the dependencies between them. This method combines time series encoding and spectrum-based feature extraction to effectively capture the dynamics of speech signals. It is the first end-to-end deep learning model that combines the time series features and local spatial information of speech signals for Parkinson's disease detection. The principle of a Two-Level Ensemble Network is to train the data in parallel using multiple base models and to aggregate the prediction results of the individual models at subsequent layers of the network using methods such as weighted tie-breaking and meta-modeling. Models' prediction results, this architecture makes good use of the diversity of parallel models and is especially resistant to overfitting in small and medium data sets. Time-Warped Input Echo State Network is an improved echo state network (ESN) model, that deals with nonlinearities and temporal deformations of time-series data by introducing time-warping techniques, The network structure can efficiently capture the complex patterns and dynamic features of temporal variations, and thanks to the unique ability to analyze data with temporal irregularities and complex time-dependent data, this network performs very well, especially in the task of speech data analysis [31]. TleNet effectively captures both temporal and local articulation features of Parkinson's speech, even under limited computational resources, facilitating early detection of subtle motor impairments in speech. TWESEN incorporates timewarping edit similarity to enhance robustness against rhythm variations and misalignments in speech, enabling more accurate differentiation between normal and pathological speech patterns.Both of these methods are commonly used and highly efficient in PD speech analysis tasks. In this study, they serve as the control group, while Time-CNN is the temporal neural network we propose, which will be briefly introduced in the next subsection.

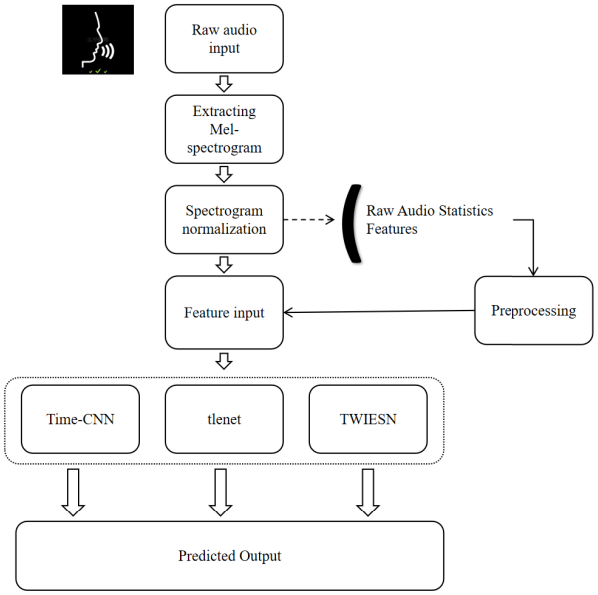


FIGURE 2. System architecture flowchart.

In the experimental phase, we cross-compare the performance of hybrid and single features in three different network architectures, while in the performance evaluation phase, we use five-fold cross-validation to weaken the impact of the evaluation bias caused by data segmentation chance on the results and to keep the complexity and computational overhead of the network within a manageable range.

Additionally, to reduce the bias caused by the initial weight values, we trained these models using 10 different runs and took the average, i.e. the average of these 10 runs on the test set. Since excessive training cycles may lead to overfitting of the training dataset, we determine the optimal model by monitoring model performance on the validation dataset and then use it for testing. The number of epochs for all models is set to 200. It is important to note that, due to model performance monitoring in the validation set, the number of epochs for the optimal model is sometimes less than 200.

E. END-TO-END TIME-SERIES CONVOLUTIONAL NEURAL NETWORK APPROACH

The temporal neural network used in this paper consists of two main layers [32], the first one consists mainly of 2D-CNNs about temporal distributions, which convert the Mayer spectrograms into dynamic features about time series, and these features are then passed to the second layer consisting of a one-dimensional convolutional neural network to learn the dependencies of the features.

Specifically, the modules in the first layer acquire feature blocks through a fixed time window slide along the time axis, and extract features from these fixed-length feature blocks using 2D-CNNs, which consist of a stack of convolutional layers, a batch normalization layer, an average pooling layer, and a dropout layer, and the main purpose of this layer is to capture local spatial features of the spectrograms and to mine the temporal structure of the temporal distribution of time-dependent features. The second layer consists of a one-dimensional convolutional neural network and the average pooling layer, where the time series features output from the first layer are flattened to perform the convolution operation, and the final results are output to the fully connected layer for prediction.As shown in the Figure 3.

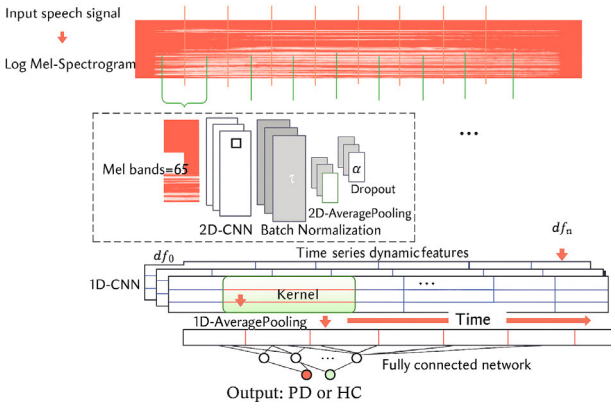
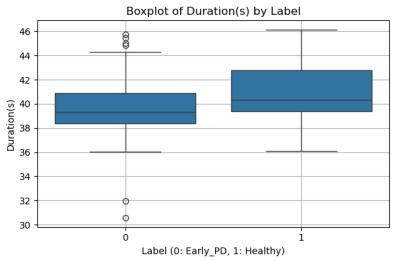
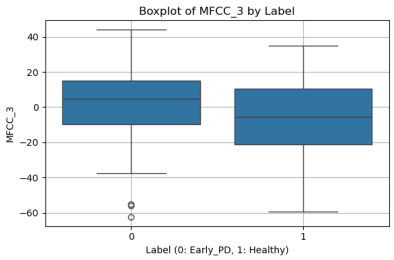


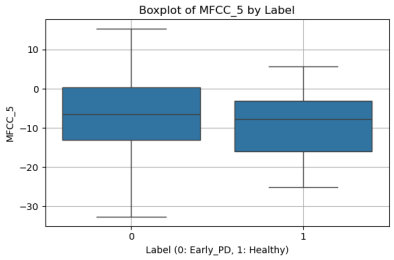
FIGURE 3. End-to-end temporal neural network architecture.



(a) Duration



(b) MFCC-3



(c) MFCC-5

FIGURE 4. Significant differences in the distributions of (a) duration and the two Meier spectral features (b) MFCC-3, (c) MFCC-5 according to patient group($p < 0.05$).

III. IMPLEMENTATION

A. EXPERIMENTAL ENVIRONMENT

The network models used in this paper were performed on an identical workstation with Ubuntu 20.04.6 LTS, configured with 32Gib RAM, Intel(R) Core(TM) i7-10700 CPU@2.90GHz and NVIDIA RTX 3080 GPUs, and a deep learning architecture optimized with the GeForce CUDA library. The deep learning architecture is optimized with the GeForce CUDA library, and the programming environment is based on Python 3.8.0 and Pytorch 2.4.0.

B. FEATURE CORRELATION ANALYSIS

To transform the original audio files into Mel spectrograms, the Python-based Librosa 0.11.0 library was used. Based on the spectrogram results, we extracted a series of original speech features including Mel frequency cepstrum coefficients (MFCC), and to optimize the classification reliability and complexity of the model and to improve the computational efficiency, we used principal component analysis to downsize the 16 standardized original speech features [33]. In this method, the principal components with

the highest correlation with the original feature distribution are removed by maximizing the variance of the data features, and the high-dimensional features are projected to the low-dimensional space, to reduce the feature dimensionality and remove the redundancy. Based on cumulative explained variance, we select 9 features with high correlation with PD distribution from 16 original features, as shown in Figure 4, the features extracted by PCA exhibit substantial distributional differences across distinct labels, suggesting a strong correlation between the selected features and the classification labels.

During model training, we employ regularization to enable the network to better adapt to differences in feature distributions and to enhance its generalization capability. In addition, We calculated the Pearson correlation between each of the 16 voice features and the binary early-PD label. Several features, including jitter and MFCC3, showed moderate positive correlations ($|r| > 0.4$, $p < 0.05$), indicating their potential as discriminative markers for early PD, as shown in Figure 5.

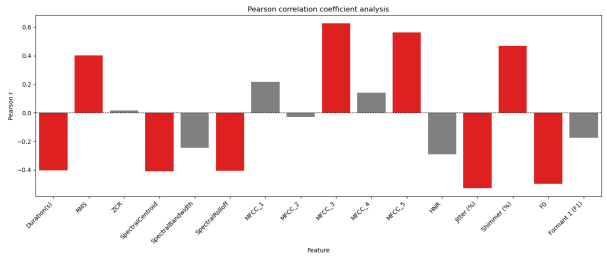


FIGURE 5. Pearson correlation coefficient between voice characteristics and early PD labels, among them, speech features with p less than 0.05 are marked in red.

C. RESULTS AND EVALUATION

The experiments in this paper are conducted using five-fold cross-validation (CV), which means that the dataset is divided into five subsets of the same size in each round of iteration. Four of them are used randomly for training, the remaining one subset i.e., is used to validate the prediction of the model, and the average of five predictions is averaged as the result after five rounds of iteration.

In order to calculate the performance of hybrid features in different neural network architectures, the experimental evaluation metrics include accuracy, F-score, specificity, sensitivity, and Mathews correlation coefficient (MCC), which are calculated using the following formulas:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.$$

$$F\_\text{score} = \frac{2 \times \text{specificity} \times \text{sensitivity}}{\text{specificity} + \text{sensitivity}}.$$

$$\text{specificity} = \frac{TP}{TP + FP}.$$

$$\text{sensitivity} = \frac{TP}{TP + FN}.$$

$$MCC = \frac{TP \times TN\text{-}FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (1)$$

Among them, TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively, and F-score is used to comprehensively measure the model precision and recall. Especially for the scenarios of labeled unbalanced datasets like the one studied in this paper, the reconciled mean of the F-score can better evaluate the classification effect of the model, and the value usually fluctuates within the range of 0.5-1, with 0.5 denoting immediate prediction and 1 representing perfect prediction. 0.5 means immediate prediction, 1 means perfect prediction, MCC (Matthews Correlation Coefficient) is a specific measure of the overall performance of the dichotomous classification model is a widely used evaluation index, its value is in the range of -1-1, -1 means the prediction is completely wrong, 0 means the result is consistent with random prediction, 1 means the prediction is completely correct, 0 means the result is consistent with random prediction, and 1 means the prediction is completely correct. The value is in the range of -1-1, -1 means the prediction is completely wrong, 0 means the result is consistent with the random prediction, and 1 means the prediction is completely correct, which comprehensively reflects the correlation between the model prediction result and the real label. Sensitivity refers to the model's ability to identify ''positive'' individuals correctly. In this study, high sensitivity refers to the model's ability to effectively identify individuals with the disease, particularly in the early stages, which is crucial for timely intervention and treatment planning. While specificity refers to the model's ability to correctly identify ''negative'' individuals, in this study, the high-specificity explanatory model demonstrated high accuracy in excluding individuals without early-stage PD, thereby helping to reduce misdiagnosis and lower the risk of unnecessary anxiety or overtreatment.

In a comparative experimental setup with different data branches, we utilize log-Meier spectrograms as the input for the time-frequency representation of the deep learning model. Log-Mel spectrograms were extracted from all recordings (resampled to 24000 Hz) using librosa, with a window size of 2048, a hop length of 512, and 65 Mel bands. Subsequently, the effects of the extracted raw features and the hybrid features spliced with neural network features were tested on three different neural network architectures, for each network architecture we used Grid Search to determine the optimal hyperparameters based on five-fold cross-validation, and in addition, for the control group using only a single neural network feature, we also made a reference comparison on three different neural network architectures. Experiments, as shown in Table 3.

It can be noticed that the classification performance of hybrid features in independently different network models always outperforms or equalizes the control group with single neural network features. This demonstrates that hybrid

TABLE 3. Performance of hybrid and single features in different networks.

Feature Type	Model	Accuracy	F1 Score	MCC	Sensitivity	Specificity
Hybrid Mixed Features	Time-CNN	0.780 ± 0.018	0.831 ± 0.024	0.470 ± 0.027	0.813 ± 0.022	0.747 ± 0.019
	Tlenet	0.748 ± 0.046	0.802 ± 0.051	0.422 ± 0.038	0.786 ± 0.049	0.710 ± 0.043
	TWIESN	0.750 ± 0.020	0.819 ± 0.031	0.436 ± 0.025	0.803 ± 0.029	0.697 ± 0.024
Only Neural Networks	Time-CNN	0.765 ± 0.009	0.793 ± 0.015	0.455 ± 0.013	0.778 ± 0.014	0.752 ± 0.011
	Tlenet	0.730 ± 0.029	0.805 ± 0.030	0.401 ± 0.037	0.787 ± 0.028	0.673 ± 0.026
	TWIESN	0.730 ± 0.016	0.808 ± 0.022	0.406 ± 0.018	0.790 ± 0.021	0.670 ± 0.018

features possess superior competitiveness in early PD diagnosis classification tasks. Although neural networks cannot provide detailed interpretability analysis like traditional methods, the fusion strategy reduces human intervention while enhancing the stability and accuracy of classification predictions. In the experiments, the Hybrid feature achieved an ACC metric 0.015 higher than using neural networks alone in Time-CNN, and also saw improvements of 0.018 and 0.02 in this metric in tlenet and TWIESN, respectively. These results suggest that combining manually extracted features with deep learning features may provide a more comprehensive representation of speech signals in patients with early-stage Parkinson's disease. To assess the stability of the model under different data partitions, we used 5-fold cross-validation and calculated the standard deviation of the performance metrics for each fold. In our 5-fold cross-validation experiment, the performance rankings of different models across folds were not entirely consistent. In most folds, Time-CNN showed better results. Tlenet exhibited greater variability, possibly due to higher sensitivity to data partitioning in cross-validation. Of course, this inconsistency may also stem from the heterogeneity of the data itself, with minor differences in category distribution, feature variability, or noise levels across folds. We also noticed that after introducing hybrid features, the standard deviation of each fold increased across different models. This may be due to the expansion of feature dimensions relative to the insufficient number of samples, causing differences in data structure to have a significant impact on the prediction results. The hyperparameter settings for the three different network models are shown in Table 4.

In addition, Time-CNN always reported the best results of the three different methods, whether in experiments using hybrid or single features, reaching 0.831 and 0.47 for F1 and MCC in the hybrid model experiments, which are significantly better than the rest of the tlenet and TWIESN, the relationships between accuracy and loss across different training epochs are shown in the figure 6. This may be since Time-CNN uses a 2D-CNN structure to process the log-Meier spectrograms compared to the tlenet's multi-scale convolutional kernel temporal features and TWIESN's projection space processing, Time-CNN better preserves the structural properties of the spectrograms, which results in stronger sensitivity to specific anomalous patterns. Besides, Time-CNN takes a direct spreading between 2D-CNN and ordinary CNN and subsequently uses the output of ordinary CNN to input the fully-connected layer, and this modular construction is easier to tune, while the multiple-scale convolutional kernel fusion paths of TLENET make the optimal hyper-parameter tuning more complicated, and TWIESN is because most of the parameters are initialized by the reservoir (e.g. spectral radius, leak rate) defined and controlled by the reservoir initialization [34], which makes the hyper-parameter tuning very sensitive to the classification results and difficult to tune compared to Time-CNN.

TABLE 4. Hyperparameter search space of neural networks.

Deep learning model	Hyperparameter search space
Time-CNN	Time-distributed 2D-CNNs layer:Number of filters: {16, 32, 64, 128}, kenel size: {(2, 2), (3, 3), (5, 5)} , Time-distributed 2D-AveragePooling layer: pool size: {(2, 2), (3, 3), (5, 5)} , Kernel size of 1D-CNN layer: {2, 3, 5} , Number of filters of 1D-CNN layer: {16, 32, 64, 128} , Pool size of each 1D-AveragePooling layer: {2, 3} , Number of frames = {5, 10, 15, 25, 35, 45, 55, 65, 75, 85, 95, 115, 135, 145}
tlenet	Number of filters: {32, 64, 128} , Kernel size: {3, 5, 7} , Activation: {ReLU, LeakyReLU} , Stride: {1} , Number of blocks: {2, 3, 4} , Pool size: {2, 3} , Dropout rate: {0.1, 0.3, 0.5} , Learning rate: {1e-4, 5e-4, 1e-3} , Batch size: {16, 32, 64}
TWIESN	Reservoir size: {200, 300, 500, 700} , Spectral radius: {0.7, 0.8, 0.9, 0.99} , Input scaling: {0.01, 0.05, 0.1, 0.5} , Leak rate: {0.1, 0.2, 0.3, 0.5} , Washout (initial discard): {10, 20, 30} , Embedding dimension: {5, 10, 20} , Delay (tau): {1, 2, 3, 4} , Regularization: {1e-6, 1e-5, 1e-4, 1e-3}

In addition, we analyzed the log-Meier spectrogram and linear scale of the speech data in the speech data sustained vowel /a/. Figure 7 shows an example of log-Meier spectrogram heatmap image of a 52-year-old male healthy control and a 57-year-old female early PD patient. The heatmap is generated by back-propagating the gradient of the last convolutional layer based on the classification output and mapping it back to the original image size superimposed. We note that the energy of the speech signal of the early patient is still concentrated in the low-frequency region ($<$1000 Hz) compared to that of the healthy controls, but the energy in the high-frequency region (2000 Hz and above) is unevenly attenuated, as indicated by the appearance of slight ripples and irregularities over time, which resembles

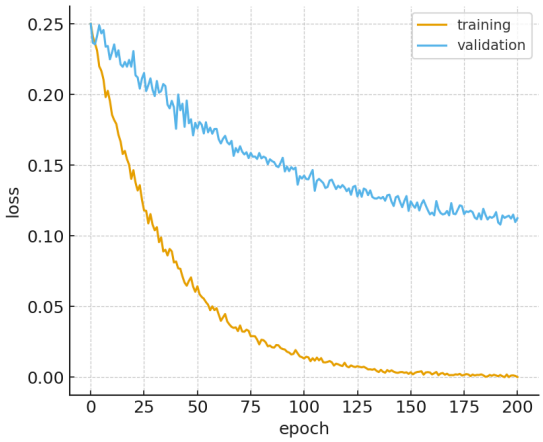


FIGURE 6. The training and validation process of Time-CNN model shows the relationship between accuracy and loss metrics across different epochs.

the pattern that has been frequency perturbations of speech in PD patients found in published studies [35]. In addition, the horizontal bright lines show periodic interruptions or frequency drifts, which may reflect the deterioration of vocal fold modulation, especially in the early stages of Parkinson's disease when this trend has been observed [12], which is also consistent with the model classification results. In contrast to the early-stage patients, the speech energy of the healthy individuals was concentrated in a consistent low-frequency band, showing a very stable frequency pattern.

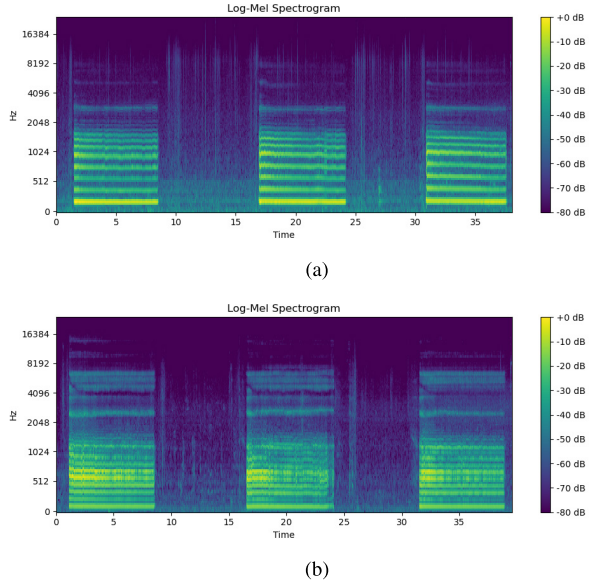


FIGURE 7. Log-Meier spectral thermograms in 52-year-old healthy men (a) and 57-year-old female patients (b) with early PD.

IV. DISCUSSION

In this study, we demonstrate the performance of using raw speech features and neural network temporal features together in the task of detecting speech patterns in Parkinson's disease and classifying patients with early-stage Parkinson's disease. We conducted experiments on three different pre-trained neural networks using speech data with 11 different entries, and the results show that the hybrid features based on spectrograms exhibit impressive classification performance in different neural network architectures. The results show that the hybrid features based on spectrograms exhibit considerable classification performance in different neural network architectures, and it also demonstrates that the fusion strategy is feasible for early PD diagnosis tasks. In both the hybrid feature experiment and the comparison group using ordinary neural network temporal features, the Time-CNN model consistently reported superior results compared to the other two network models. In the hybrid feature experiment, the ACC was 0.78 and the F1-score was 0.83; when using only neural network temporal features, the ACC was 0.765 and the F1 score was 0.79. The experimental results suggest that the Time-CNN model identifies learned dynamic features that capture the narrowed overall frequency range and reduced variability of Parkinson's disease-related sounds, which are important clinical indicators for detecting patients with Parkinson's disease. In addition, compared to the direct use of speech data, spectrograms can retain and highlight subtle features such as frequency anomalies, jitter, and pitch changes in the articulation process of patients with PD, leading to better classification performance. Moreover, because the redundancy of the original speech waveforms is generally high, it is difficult to converge in a conventional neural network [36], and the spectrogram data also have better training efficiency. In traditional clinical research on Parkinson's disease, feature extraction has been highly dependent on domain knowledge and expert experience for feature design and selection. In the experiments studied in this paper, we employed a fusion strategy combining traditional features with automatically extracted features from neural networks to train the neural network model. In addition, we used Principal Component Analysis (PCA) to select and reduce the dimensionality of traditional speech features. This method has been proven to retain high-frequency information in speech representations efficiently [37]. In particular, speech data with a high degree of redundant noise have the dual characteristics of high dimensionality and low sample size, which makes the selection of initial candidate variables particularly important [38], [39].

With the widespread application of deep learning, end-to-end methods are increasingly being used in clinical research to assist in the diagnosis of movement disorders and cognitive disorders related to dementia [40], [41]. Compared with traditional machine learning methods, deep learning techniques have a stronger representation of sensitive features, while the performance will continue to enhance with the accumulation of data and have a higher generalization ability. In addition, several studies have shown [42], [43] that deep models can significantly outperform traditional machine learning methods in early PD vs HC classification tasks, especially in multicenter and multispecies speech tasks.

Chandrasekaran et al. used recurrent neural networks and fuzzy KNN based on magnetic resonance imaging of the brain to achieve early Parkinson's disease detection [44], and the results outperformed the general decision tree and random forest algorithms. Srinivasan et al. employed a variety of techniques including the synthetic minority oversampling technique (SMOTE) for solving the category imbalance problem, feature selection for identifying the most relevant features, and the use of RandomizedSearchCV for the detection of Parkinson's disease, based on a small dataset of 195 speech recordings from 31 patients. Hyper-parameter tuning using strategies such as RandomizedSearchCV, and achieved 99.11% and 95.89% accuracy using KNN and feed-forward neural network (FNN) models, respectively [28]. Rahul Nijhawan et al. developed a Transformer-based approach to retrieve dysarthria indicators from subjects' speech recordings to detect PD, and Nijhawan et al. developed a Transformer-based approach to retrieve dysarthria indicators from subjects' speech recordings to detect PD. Metrics to detect PD, in addition to providing an XgBoost-based feature selection method and a fully connected neural network layer technique for incorporating continuous dysphonia measures. Their proposed method comprehensively outperforms traditional machine learning techniques (e.g., Multilayer Perceptron (MLP), Support Vector Machines (SVMs), and Random Forests (RFs)) in all conventional metrics, and their study also found that The accuracy of partial discharge (PD) detection can be further improved by using a constant-length vector representation generated by the Transformer. The solution they used can also be used in a setup similar to the Siamese network with a triple-state contrast loss function to bring vector representations of similar classes closer together through direct supervision [16], which confirms the feasibility and frontiers of neural networks in complex sequence modeling and feature representation.

Although making precise judgments about important spectral regions for CNN classification decisions remains one of the greater challenges in the field of PD diagnosis, it is still feasible to emphasize some of the routine differences between PD patients and healthy controls [45]. In this paper, the average length of the original recordings was more than 60 seconds, and we uniformly cut them to 40 seconds to standardize them for neural network input. Interpretability has always been one of the limitations of AI application in clinical medicine, we have conducted some analyses of different branching features to explore the independent value and potential connection of different modal features, for example, in the analysis of the data in this paper, we found that in the narrow band of PD patients at 500-2000 HZ, the healthy controls appeared to be significantly different, which may be due to the peak frequency of the articulatory resonance close to the vowels. The original speech features used in this paper have been proven to be directly related to the physiological mechanisms of PD in previous studies [2], and these features have been clearly defined and mature quantitative standards, which have significant advantages in terms of interpretability, and can be used in conjunction with temporal features extracted from neural networks to complement the ''black-box'' characteristics of neural networks. The combination of these features with the time-series features extracted from neural networks may be able to complement, in a sense, the lack of interpretability of neural networks due to their ''black box'' characteristics, and thus improve the performance and clinical confidence of the early Parkinson's disease detection system.

Additionally, although this study collected speech data in a controlled studio environment to minimize background noise interference, we recognize that this setting differs from actual remote or home monitoring scenarios. Future research will focus on evaluating the robustness of the model in everyday environments with different types of devices and background noise to verify the feasibility and cost-effectiveness of this method in a wider range of practical applications.This study used five-fold cross-validation to evaluate model performance, but this method cannot cover cross-device generalization issues completely that may arise in practical applications. Future work should further verify the generality of this method on different microphones, different noise environments, or multi-center datasets. Considering the interpretability of enhanced deep learning models in clinical settings, we plan to introduce interpretable artificial intelligence methods such as saliency maps and layer-wise relevance propagation (LRP) in subsequent studies. This will help clinicians intuitively understand the speech features that the model focuses on, thereby improving their understanding and trust in the model's decision-making basis.

V. CONCLUSION AND PROSPECTS

In this paper, we propose a multimodal diagnostic model based on speech recording data for the diagnosis of patients with early Parkinson's disease and determine the early Parkinson's disease through the original speech features and the temporal features extracted by the neural network. We combined spectral-based raw speech features and time-series-based deep features extracted by end-to-end neural networks in different pre-trained deep learning architectures to classify healthy and early PD patients. Unlike most previous studies that treated these two types of features separately, the unified framework we propose can simultaneously capture complementary information in the frequency domain and time domain, thereby improving the prediction accuracy and robustness of the model. Specifically, Our model takes the raw features as branching inputs to an intermediate layer and then extracts time-series dynamics deep features through a series of temporally distributed two-dimensional convolutional neural networks. The time-series dynamics deep features are extracted through a fully-connected layer pre concatenate splicing the outputs of the two types of features to achieve feature fusion. Cross-validation is performed to ensure the reliability of the experimental results, we compared our Time-CNN model against two existing

published methods (TLENet and TWIESN), ensuring a fair and representative evaluation, and the results demonstrate that this feature fusion method achieves equal or better results compared to a single feature in several different neural network architectures. In addition, we discussed the relevance of the features and the patterns exhibited by the spectrum charts of different patient groups. The experimental results prove that the hybrid features in different pre-training network architectures have a better effect than a single temporal feature, and also complement the problem of lack of interpretability that the deep learning technology has always had in the clinical application.

Although our work has achieved some results, it can be extended in at least two directions: first, this method can be tested on larger and more diverse datasets, which not only refers to the increase in the number of speech entries and patients, but also implies that there are more kinds of speech data in different languages, and the validity of the method can be verified by expanding the dataset. Secondly, there is still room for more extensions to this feature fusion approach, and additional techniques (e.g., cross-modal attentional fusion and multiscale adaptive fusion) have a lot of room to be utilized in the Parkinson's disease speech diagnosis task. These improvements are promising to further improve the robustness and classification performance of the model.

REFERENCES

[1] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A. Schrag, and A. E. Lang, "Parkinson disease," *Eur. J. Neurol.*, vol. 3, no. 1, pp. 27–42, 2017.

[2] M. J. Armstrong and M. S. Okun, "Diagnosis and treatment of Parkinson disease: A review," *Jama*, vol. 323, no. 6, pp. 548–560, 2020.

[3] E. Majda-Zdancewicz, A. Potulska-Chromik, M. Nojszewska, and A. Kostera-Pruszczyk, "Speech signal analysis in patients with Parkinson's disease, taking into account phonation, articulation, and prosody of speech," *Appl. Sci.*, vol. 14, no. 23, p. 11085, Nov. 2024.

[4] H. Gunduz, "Deep learning-based Parkinson's disease classification using vocal feature sets," *IEEE Access*, vol. 7, pp. 115540–115551, 2019.

[5] 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, Jul. 2019.

[6] L. Ramig, A. Halpern, J. Spielman, C. Fox, and K. Freeman, "Speech treatment in Parkinson's disease: Randomized controlled trial (RCT)," *Movement Disorders*, vol. 33, no. 11, pp. 1777–1791, Nov. 2018.

[7] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[8] W. Pawlukowska, A. Szylińska, D. Kotlęga, I. Rotter, and P. Nowacki, "Differences between subjective and objective assessment of speech deficiency in Parkinson disease," *J. Voice*, vol. 32, no. 6, pp. 715–722, Nov. 2018.

[9] I. Nissar, W. A. Mir, Izharuddin, and T. A. Shaikh, "Machine learning approaches for detection and diagnosis of Parkinson's disease—A review," in *Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2021, pp. 898–905.

[10] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated Parkinson disease detection based on voice signals with transfer learning," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 115013.

[11] R. Khaskhoussy and Y. B. Ayed, "Improving Parkinson's disease recognition through voice analysis using deep learning," *Pattern Recognit. Lett.*, vol. 168, pp. 64–70, Apr. 2023.

[12] N. P. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1925–1936, 2021.

[13] A. Mahmood, M. Mehroz Khan, M. Imran, O. Alhajlah, H. Dhahri, and T. Karamat, "End-to-end deep learning method for detection of invasive Parkinson's disease," *Diagnostics*, vol. 13, no. 6, p. 1088, Mar. 2023.

[14] Z. Galaz, P. Drotar, J. Mekyska, M. Gazda, J. Mucha, V. Zvoncak, Z. Smekal, M. Faundez-Zanuy, R. Castrillon, J. R. Orozco-Arroyave, S. Rapcsak, T. Kincses, L. Brabenec, and I. Rektorova, "Comparison of CNN-learned vs. Handcrafted features for detection of Parkinson's disease dysgraphia in a multilingual dataset," *Frontiers Neuroinform.*, vol. 16, May 2022, Art. no. 877139.

[15] P. Valarmathi, Y. Suganya, K. R. Saranya, and S. S. Priya, "Enhancing Parkinson disease detection through feature based deep learning with autoencoders and neural networks," *Sci. Rep.*, vol. 15, no. 1, p. 8624, Mar. 2025.

[16] R. Nijhawan, M. Kumar, S. Arya, N. Mendirtta, S. Kumar, S. K. Towfek, D. S. Khafaga, H. K. Alkahtani, and A. A. Abdelhamid, "A novel artificial-intelligence-based approach for classification of Parkinson's disease using complex and large vocal features," *Biomimetics*, vol. 8, no. 4, p. 351, Aug. 2023.

[17] I. Ahmed, S. Aljahdali, M. Shakeel Khan, and S. Kaddoura, "Classification of Parkinson disease based on patient's voice signal using machine learning," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 705–722, 2022.

[18] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, and P. Bonato, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jun. 2016, pp. 655–658.

[19] A. C. Albán-Cadena, F. Villalba-Meneses, K. O. Pila-Varela, A. Moreno-Calvo, C. P. Villalba-Meneses, and D. A. Almeida-Galárraga, "Wearable sensors in the diagnosis and study of Parkinson's disease symptoms: A systematic review," *J. Med. Eng. Technol.*, vol. 45, no. 7, pp. 532–545, 2021.

[20] I. Aouraghe, G. Khaissidi, and M. Mrabti, "A literature review of online handwriting analysis to detect Parkinson's disease at an early stage," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11923–11948, Mar. 2023.

[21] A. Letanneux, J. Danna, J.-L. Velay, F. Viallet, and S. Pinto, "From micrographia to Parkinson's disease dysgraphia," *Movement Disorders*, vol. 29, no. 12, pp. 1467–1475, Oct. 2014.

[22] A. Rana, A. Dumka, R. Singh, M. K. Panda, N. Priyadarshi, and B. Twala, "Imperative role of machine learning algorithm for detection of Parkinson's disease: Review, challenges and recommendations," *Diagnostics*, vol. 12, no. 8, p. 2003, Aug. 2022.

[23] C. Quan, K. Ren, and Z. Luo, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," *IEEE Access*, vol. 9, pp. 10239–10252, 2021.

[24] T. Fujita, Z. Luo, C. Quan, K. Mori, and S. Cao, "Performance evaluation of RNN with hyperbolic secant in gate structure through application of Parkinson's disease detection," *Appl. Sci.*, vol. 11, no. 10, p. 4361, May 2021.

[25] G. Nagasubramanian and M. Sankayya, "Multi-variate vocal data analysis for detection of Parkinson disease using deep learning," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 4849–4864, May 2021.

[26] P. Khojasteh, R. Viswanathan, B. Aliahmad, S. Ragnav, P. Zham, 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.

[27] J. D. Arias-Londoño and J. A. Gómez-García, "Predicting updrs scores in Parkinson's disease using voice signals: A deep learning/transfer-learning-based approach," in *Proc. 1st Workshop Autom. Assessment Parkinsonian Speech*, Cambridge, MA, USA. Cham, Switzerland: Springer, 2019, pp. 100–123.

[28] S. Srinivasan, P. Ramadass, S. K. Mathivanan, K. P. Selvam, B. D. Shivahare, and M. A. Shah, "Detection of Parkinson disease using multiclass machine learning approach," *Sci. Rep.*, vol. 14, no. 1, p. 13813, Jun. 2024.

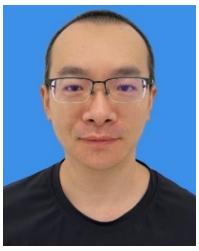
[29] T. Gonçalves, J. Reis, G. Gonçalves, M. Calejo, and M. Seco, "Predictive models in the diagnosis of Parkinson's disease through voice analysis," in *Proc. Intell. Syst. Conf.* Cham, Switzerland: Springer, 2024, pp. 591–610.

[30] V. Mittal and R. K. Sharma, "Machine learning approach for classification of Parkinson disease using acoustic features," *J. Reliable Intell. Environments*, vol. 7, no. 3, pp. 233–239, Sep. 2021.

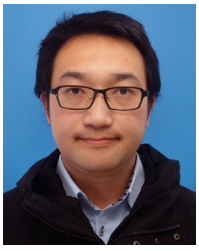
[31] C. Sun, M. Song, D. Cai, B. Zhang, S. Hong, and H. Li, "A systematic review of echo state networks from design to application," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 23–37, Jan. 2024.
[32] C. Quan, K. Ren, Z. Luo, Z. Chen, and Y. Ling, "End-to-end deep learning approach for Parkinson's disease detection from speech signals," *Biocybernetics Biomed. Eng.*, vol. 42, no. 2, pp. 556–574, Apr. 2022.
[33] T. Kurita, "Principal component analysis (PCA)," in *Computer Vision: A Reference Guide*. Cham, Switzerland: Springer, 2020, pp. 1013–1016.
[34] P. Steiner, A. Jalalvand, and P. Birkholz, "Cluster-based input weight initialization for echo state networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7648–7659, Oct. 2023.
[35] H. Liu, E. Q. Wang, L. V. Metman, and C. R. Larson, "Vocal responses to perturbations in voice auditory feedback in individuals with Parkinson's disease," *PLoS ONE*, vol. 7, no. 3, Mar. 2012, Art. no. e33629.
[36] A. R. Bradshaw and C. McGettigan, "Convergence in voice fundamental frequency during synchronous speech," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0258747.
[37] H. Pan, Y. Wang, Z. Li, X. Chu, B. Teng, and H. Gao, "A complete scheme for multi-character classification using EEG signals from speech imagery," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 8, pp. 2454–2462, Aug. 2024.
[38] X. Xiao, Y. Li, Q. Wu, X. Liu, X. Cao, M. Li, J. Liu, L. Gong, and X.-J. Dai, "Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: A large and longitudinal machine learning cohort study," *Alzheimer's Res. Therapy*, vol. 17, no. 1, p. 103, May 2025.
[39] N. Li, J. Ou, H. He, J. He, L. Zhang, Z. Peng, J. Zhong, and N. Jiang, "Exploration of a machine learning approach for diagnosing sarcopenia among Chinese community-dwelling older adults using sEMG-based data," *J. NeuroEng. Rehabil.*, vol. 21, no. 1, p. 69, May 2024.
[40] H. Khachnaoui, R. Mabrouk, and N. Khlifa, "Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: A review," *IET Image Process.*, vol. 14, no. 16, pp. 4013–4026, Dec. 2020.
[41] C. Zhu, "Computational intelligence-based classification system for the diagnosis of memory impairment in psychoactive substance users," *J. Cloud Comput.*, vol. 13, no. 1, p. 119, Jun. 2024.
[42] R. Khaskhoussy and Y. B. Ayed, "Speech processing for early Parkinson's disease diagnosis: Machine learning and deep learning-based approach," *Social Netw. Anal. Mining*, vol. 12, no. 1, p. 73, Dec. 2022.
[43] K. Singh and S. Dash, "Early detection of neurological diseases using machine learning and deep learning techniques: A review," in *Artificial Intelligence for Neurological Disorders*, 2022, pp. 1–24.
[44] S. Chandrasekaran, V. Dutt, N. Vyas, and A. Anand, "Fuzzy KNN implementation for early Parkinson's disease prediction," in *Proc. 7th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Feb. 2023, pp. 896–901.
[45] C. Quan, Z. Chen, K. Ren, and Z. Luo, "FedOcw: Optimized federated learning for cross-lingual speech-based Parkinson's disease detection," *npj Digit. Med.*, vol. 8, no. 1, p. 357, Jun. 2025.

WEIKANG HOU received the B.S. degree in computer science from Southwest Petroleum University, Chengdu, China, in 2020, and the M.S. degree from the Graduate School of Computer Software, Southwest Petroleum University, in October 2024. He is currently pursuing the Ph.D. degree in system informatics with Kobe University, Japan.

CHANGQIN QUAN received the Ph.D. degree from the University of Tokushima, Tokushima, Japan, in 2011, the M.E. degree from Central China Normal University, Wuhan, China, in 2005. She is currently an Associate Professor with Kobe University. Her research interests include machine learning algorithms, natural language processing, human–computer interface, and medical.



ZHONGLUE CHEN received the M.S. degree in software engineering from the Huazhong University of Science and Technology. He was the Manager with the Research Department of the GYENNO Science, since 2013. His research interests include the application of motion sensors, machine learning, human kinematics, and the application of these technologies in the field of central nervous system diseases.



SHENG CAO (Member, IEEE) received the Ph.D. degree from the Graduate School of System Informatics, Kobe University, in 2017. He is currently an Assistant Professor with Kobe University. His research interests include data-driven control, robotic rehabilitation, human–robot interaction, biomechanics analysis, and robot's safe control.

KANG REN received the Ph.D. degree in computational science from Kobe University, Japan, and the M.S. and B.E. degrees in control science and engineering from Huazhong University of Science and Technology (HUST), China. He is the Founder of GYENNO Science, where he leads research efforts focused on artificial intelligence and its applications in the diagnosis, assessment, treatment, rehabilitation, and management of neurological disorders. He has authored more than 30 peer-reviewed publications in related fields.

WEN SU received the M.M. degree from Peking Union Medical College, Beijing, China, in 2002. He was a Postdoctoral Fellow with the Department of Molecular Biology, Laval University, Canada, in 2003. She was a Visiting Scholar with Cedars-Sinai Hospital, USA, in 2018. She is currently a Professor with the Peking University and the Director of the Department of Neurology in Beijing Hospital. Her research interests include Parkinson's disease, and related movement disorders.

ZHIWEI LUO received the M.E. and Ph.D. degrees from Nagoya University, Japan, in 1991 and 1992, respectively. He was an Assistant Professor with Toyohashi University of Technology, in 1992, a Frontier Researcher with RIKEN, in 1994, and an Associate Professor with Yamagata University, in 1999. He was the Team Leader with RIKEN, in 2001, where he leaded the development of the world first human care robot RI-MAN. Since 2006, he has been a Professor with Kobe University. His research interests include system control, robotics, human–computer interface, and health engineering. He honored Fellow of SICE, in 2016. He is currently a Board Member of SCI, Japan.