Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 8890592, 20 pages
https://doi.org/10.1155/2024/8890592

Research Article

A Triplet Multimodel Transfer Learning Network for Speech Disorder Screening of Parkinson's Disease

Aite Zhao, Nana Wang, Xuesen Niu, Ming Chen, and Huimin Wu

College of Computer Science and Technology, Qingdao University, Qingdao, China

Correspondence should be addressed to Aite Zhao; zhaoaite@qdu.edu.cn

Received 14 August 2023; Revised 29 January 2024; Accepted 27 February 2024; Published 20 March 2024

Academic Editor: Paolo Gastaldo

Deterioration in the quality of a person's voice and speech is an early sign of Parkinson's disease (PD). Although a number of computer-based methods have been invested to use patients' speech for early diagnosis of Parkinson's disease, they only focus on a fixed pronunciation test, such as the subjects' monosyllabic pronunciation is analyzed to determine whether they have potential possibility of PD. Moreover, only using traditional speech analysis methods to extract single-view speech features cannot provide a comprehensive feature representation. This paper is dedicated to the study of various pronunciation tests for patients with PD, including the pronunciation of five monosyllabic vowels and a spontaneous dialogue. A triplet multimodel transfer learning network is designed and proposed for identifying subjects with PD in these two groups of tests. First, multisource data extract mel frequency cepstrum coefficient (MFCC) features of speech for preprocessing. Subsequently, a pretrained triplet model represents features from three dimensions as the upstream task of the transfer learning framework. Finally, the pretrained model is reconstructed as a novel model that integrates the triplet model, temporal model, and auxiliary layer as the downstream task, and weights are updated through fine-tuning to identify abnormal speech. Experimental results show that the highest PD detection rates in the two groups of tests are 99% and 90% , respectively, which outperform a large number of internationally popular pattern recognition algorithms and serve as a baseline for other academic researchers in this field.

1. Introduction

Parkinson's disease (PD) is a degenerative disease commonly seen in the elderly, mainly manifested as motor retardation, static tremor, and muscle rigidity. Therefore, the examination of Parkinson's disease mainly depends on the medical history and physical examination, which should be completed by neurologists in hospitals. If the patient has obvious movement slowing, static tremor, 4–6 Hz per second, tremor is weakened or disappeared during movement, and there is a mask face, walking forward, small gait, muscle stiffness, and increased muscle tension, the patient is more likely to have Parkinson's disease. However, the motor symptoms of Parkinson's disease often occur late. In contrast, nonmotor symptoms, such as language and cognitive disorders, are manifested decades before the onset of motor symptoms, which is of great significance for the early diagnosis of the potential disease possibility of Parkinson's disease.

There is a lot of literature proving that early Parkinson's disease also has a small amount of speech impairment [1–5]. An assessment of vocal impairment was presented for separating healthy people from persons with early untreated Parkinson's disease (PD) [1]. The purpose of the study [2] is to determine if subjects in the early stages of untreated Parkinson's disease (PD) or PD treated with deprenyl alone suffer from motor speech abnormalities. Speech defects are common in advanced PD, including disturbances of respiration, phonation, and articulation. We studied 12 subjects with early PD (Hoehn and Yahr stage ≤ 2; mean duration disease 3.2 years) who were not taking symptomatic therapy and tested them under two conditions: on and off deprenyl. The study of [3] provides an evaluation of speech disorders in early Parkinson's disease. Moreover, evidence shows that speech difficulties were associated with greater autonomic dysfunction, sleep disturbances, and striatal dopaminergic deficit and can serve as a predictor of faster cognitive decline in early Parkinson's disease [4]. Detecting speech disorders in early Parkinson's disease by acoustic analysis is another study in 2018 [5].

Language dysfunction can show the cognitive ability of the brain and the speed of response to external stimuli. It is mainly manifested as throat voice and tongue movement disorders of different degrees, and the first manifestation is the weakening of voice. In addition, there are also situations of single pitch, slow speech speed, abnormal language pauses, continuous dysphonia, abnormal stress, vague and hoarse voice, decreased fluency of oral expression, and simplified syntactic expression. Furthermore, PD can also hinder voice production, making the voice of patients with Parkinson's disease soft and monotonous. Research shows that these symptoms often appear in the early stages of disease development, sometimes decades earlier than exercise-related symptoms. This year, a new study by neuroscientists at the University of Arizona showed that a specific gene usually associated with Parkinson's disease may be the reason behind these problems related to phonation. This discovery may help to early diagnosis and treatment of Parkinson's disease [6]. These representations are obtained from the patient's speech data. Therefore, using computer methods to analyze and process these speech data is the primary task.

The speech recognition system can be roughly divided into four parts (Figure 1):

(1) PD pronunciation test, including the reading vowel test and continuous conversation test

(2) Speech data collection: collect the test content using mobile phones or recording pens and other equipment equipped with microphones

(3) Speech recognition and PD detection: use deep learning or machine learning algorithm for feature extraction and recognition of speech data

(4) The prediction results of the model are fed back to doctors to help them make treatment plans

In order to detect Parkinson's disease, computer scientists try to capture the unique disease symptoms of PD patients and build models to compare with healthy people. Specifically, a large proportion of these methods are artificial intelligence technologies. For example, supervised traditional machine learning (ML) algorithms, such as random forest [7–9], decision trees [10, 11], and K-nearest neighbor (KNN) [12], have been highly effective in motor symptoms analysis of Parkinson's disease; support vector machines [13, 14] and XGBoost [15] have competitive performance in PD speech analysis and recognition. The deep network has achieved far more accuracy than ML methods, including speech, natural language processing, vision, and many other fields. In the analysis of speech, a classical ML algorithm usually requires complex feature engineering, while deep networks can usually achieve good performance by simply passing data directly to the network. Moreover, deep models can more easily adapt to different fields and applications. For instance, transfer learning makes it effective to apply the pretrained deep network to different applications in the same field. Benefiting from these strengths, deep models have also shown incomparable advantages in speech recognition [16] of Parkinson's disease, such as time series models (LSTM [17, 18] and GRU [19]), convolution-based neural networks (CNN [20–22] and ResNet [18, 23]), and hybrid or complex networks (transformer [24, 25], ensemble learning [26, 27], and few-shot learning [28]).

Despite that these methods have a place in a number of fields, they are also limited to concentrating on a single perspective, that is, using a single perspective to view speech data. For example, CNN is for spatial feature extraction, LSTM is for temporal feature extraction [29], and MFCC is for spectral feature extraction. However, the expression forms of each feature are diverse, and the degree of aggregation of the same feature of multiple samples to different spaces or different perspectives is greatly different. Therefore, we choose the method of multimodel fusion to express different levels of features in different dimensions and spaces and fuse them until the best effect is achieved. The multimodel fusion method makes up for the defect of one-sided feature representation of a single model and makes the fusion model more suitable for the input data.

In addition, we conduct the transfer learning framework to process speech data. Due to the long training time and large amount of data of the deep learning model, it is difficult for the complex model we designed to achieve excellent classification effects in a short time, and it is hard to change the details once the model is trained. The transfer learning framework is divided into an upstream task and a downstream task, which can perfectly optimize the system performance and improve the training efficiency, so that the reconstructed model of the downstream task can cover the shortcomings of the upstream task, pay more attention to the detailed feature description, and achieve faster modeling speed by fine-tuning the training weight.

The inner unit of the triplet network proposed integrates the attention mechanism, convolution, feature splicing, scalable structure, and other technologies. New elements have also been added to the block to improve the recognition ability of the model. This improved strategy successfully presents the advantages of each model and obtains a more robust hybrid model.

The main contributions of our work are summarized as follows:

(i) A triplet multimodel transfer learning network (TmmNet) is proposed for speech disorders screening of Parkinson's disease, which can not only extract the multidimensional spatiotemporal features of the input speech but also selectively enhance the significance of the features. The two-layer task framework adopted solves the problem of a large data volume and a long training process.

(ii) The proposed triplet network integrates a variety of improved new expansion units, adds multiscale convolution, multihead, spatial, and channel attention mechanisms, uses parallel mode for training and serial mode for feature splicing, and performs hierarchical feature representation and fusion.

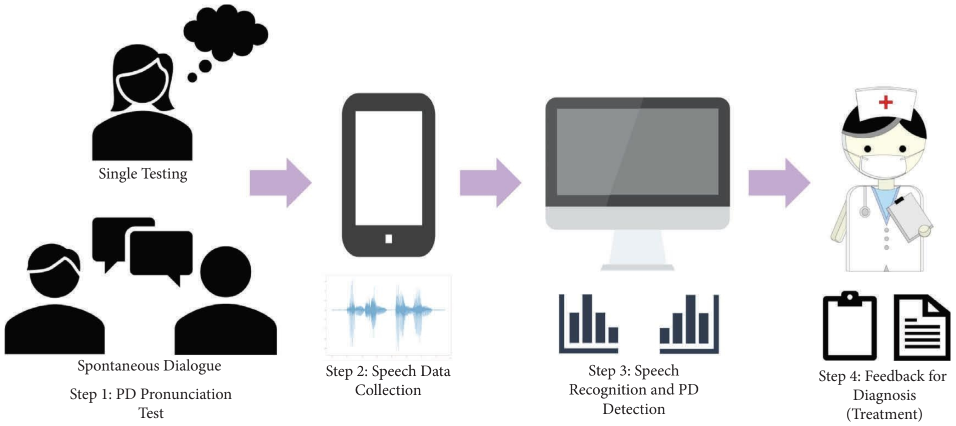

Single Testing
Spontaneous Dialogue
Step 1: PD Pronunciation Test
Step 2: Speech Data Collection
Step 3: Speech Recognition and PD Detection
Step 4: Feedback for Diagnosis (Treatment)

Figure 1: Speech recognition process. The speech recognition system can be divided into four parts: PD pronunciation test, speech data collection, speech recognition and PD detection, and feedback for diagnosis.

(iii) After multifeature fusion of the upstream pretraining model of the transfer learning structure is completed, the downstream model adds a bidirectional temporal recurrent memory network and two fully connected modules after the pretraining model for fine-tuning training. The significance of fusion features in the time series dimension is highlighted.

The rest of this paper is organized as follows: Section 2 illustrates the related work in recent years. Section 3 describes the framework and computing process of the proposed model. Section 4 provides the dataset introduction, experimental results, and settings. The conclusion discusses the strengthens and limitations in Section 5.

2. Related Work

As far as the algorithms mentioned above are concerned, we will introduce computer methods for speech recognition in the following three categories: manual feature extraction methods, machine learning methods, and deep learning methods. As illustrated in Table 1, we mark whether the investigated research involves these algorithm categories as "✓" and "−." "✓" means involved, and "−" means not involved.

Machine learning methods, such as SVM and XGBoost, have been widely applied in PD speech assessment [13, 15]. The study of [13] proposed to introduce the L1 regularization SVM for speech signal feature extraction and then trained an improved genetic algorithm and an SVM classifier for speech recognition. Wang et al. [15] compared XGBoost with support vector machines, random forests, and neural networks for the detection of the speech signal collected from Parkinson's patients, by identifying vocal fundamental frequency of speech. Although machine learning algorithms have made achievements in the field of voiceprint recognition, most of the machine learning algorithms used are still limited to feature classification, without more consideration on feature representation and description.

Table 1: The discussed studies.

Previous studies	MFCC feature	Deep learning	Machine learning
[13]			✓
[15]		✓	✓
[25]		✓	✓
[26]		✓	✓
[28]		✓	✓
[30]	✓		✓
[31]	✓		✓
[32]	✓	✓	✓
[33]	✓	✓	
TmmNet	✓	✓	✓

As popular deep learning methods, a large number of classical models [18, 19, 24, 34, 35] have achieved good performance in PD speech recognition. As the audio feature, MFCC was input into LSTM, GRU, CNN, ResNet, and other deep models for automatic speech recognition (ASR) in the study of [18]. GRU [19] was employed to assess speech impairments by computing static features from complete utterance. Hernandez et al. [24] explored the usefulness of using Wav2Vec self-supervised speech representation as the speech feature of dysarthria in training ASR systems and used a transformer-based context network for feature representation and classification. In addition, several hybrid fusion models [25, 26, 28, 36] have gradually emerged in the application of PD speech recognition. An audio spectrogram [25] transformer was proposed to analyze the multimodal PD speech and handwriting data. An ensemble model [26] was designed for the classification of PD speech data, which combined a deep sample learning algorithm with a deep network, realizing deep dual-side learning. A deep model based on iterative mean clustering [28] was established to obtain new high-level deep samples, which solved the problem of few-shot learning.

For MFCC feature extraction, some algorithms analyze and classify the MFCC features in speech data [30–33]. Qing et al. [32] designed a transfer learning network after extracting MFCC features from the raw speech data. In the study of [33], the MFCC parameters with the best performance in 12 dimensions were extracted to represent the acoustic characteristics of articulation disorders, which were utilized for automatic speech recognition based on the artificial neural network (ANN). Nivash et al. [31] carried out research in 2021 and used a series of machine learning algorithms to classify the MFCC features of speech, such as RF and naive Bayes, and naive Bayes was verified to be the best algorithm. MFCC was also utilized to detect patients with PD from healthy people. Literature [30] adopted an SVM classifier to distinguish the extracted voice and cepstrum features, and the results showed that MFCC is the best by comparison. These algorithms all involve MFCC feature extraction, which is sufficient to verify its availability.

Inspired by these approaches, we first extract the MFCC features of speech files in the preprocessing part and then develop a model of transfer learning structure that includes traditional machine learning and deep learning.

3. Triplet Multimodel Transfer Learning Network

For achieving successful speech analysis and recognition of PD patients and healthy controls in the real environment, we here propose a triplet multimodel transfer learning network for MFCC features, multilevel and scale feature extraction, and classification. First, we introduce the preprocessor for MFCC feature computation. Then, we describe the architecture of the pretrained model for multilevel and scale feature extraction, followed by detailed discussion on the individual components. Finally, we describe the reconstructed model for fine-tuning the upstream parameters and scoring the fused features before supplying the final diagnostic result.

3.1. The Overall Structure. Our proposed model is shown in Figure 2. In this framework, the speech data of healthy controls and PD are fed into the preprocessor first, which includes the progress of pre-emphasis, discrete Fourier transform (DFT), and inverse discrete Fourier transform (IDFT) to handle the input for MFCC feature production. Afterward, we express MFCC features in the form of sequences and reshape every ten extracted frames into one frame, forming a 40 $*$ 10 matrix as a training sample. Then, the training samples are input into the pretraining triple network in batch size, including two transformer blocks, multiscale convolution blocks, and a dense block. The output features of the triplet network are spliced together through the global maximum pooling layer to form multilevel fusion features, and then, the diagnostic results are obtained through two fully connected layers. Since there is no scalable structure in the pretraining model to represent the changes of speech data in the time dimension, we reconstruct the model in the downstream task of the transfer learning model as a hybrid model in series of a triplet network and a temporal network.

3.2. Data Preprocessor. As shown in Figure 2, first of all, we use a pre-emphasis method to compensate the high-frequency part of the voice. For the sampled value $x[n]$ of speech at time n, the output after pre-emphasis processing is

$$y[n] = x[n] - a * x[n-1]. \tag{1}$$

The pre-emphasis coefficient a is generally between 0.9 and 1. Then, the voice is divided into segments by windowing. The windowing function is nonzero only in some regions but zero in other regions. The next step of windowing and framing is a discrete Fourier transform (DFT). The function of the Fourier transform is to map the signal from the time domain to the frequency domain. Assuming that the number of sampling points after windowing is N, DFT for these N points includes

$$x[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi/Nkn}. \tag{2}$$

Then, the amplitude of each frequency component is obtained, and the frequency is mapped to mel frequency. The expression relationship between mel frequency and frequency (f) is as follows:

$$\mathrm{Mel}(f) = 1127 * \ln\left(1 + \frac{f}{700}\right). \tag{3}$$

Inverse discrete Fourier transform (IDFT): we take the logarithm of the mel feature in the previous step, which can be used as an acoustic feature, take the logarithmic frequency spectrum as a time-domain signal, and do a Fourier transform again, because the content of our voice is often determined by the path that the sound passes through from the sound location (similar to a series of filters) and is independent of the vibration frequency (fundamental frequency) of the sound location itself. The function of cepstrum is to separate the filter from the sound source to help identify the content of the sound. The calculation process of cepstrum is shown in the following formula, which only represents the calculation process of cepstrum, excluding the process of mel filtering. We can see that cepstrum is to take the frequency spectrum after the Fourier transform as the time-domain signal and do another Fourier transform on this frequency spectrum:

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n] e^{-j2\pi/Nkn}\right|\right) e^{-j2\pi/Nkn}. \tag{4}$$

The process of delta is as follows: for each frame, the first 12 dimension cepstrum coefficients passing through IDFT are selected, and then, the energy is used as the 13th dimension feature. The time-domain signal after adding a window can obtain energy characteristics through calculation. Assuming that the window length starts from t_1 and ends at t_2, then the energy of the frame is

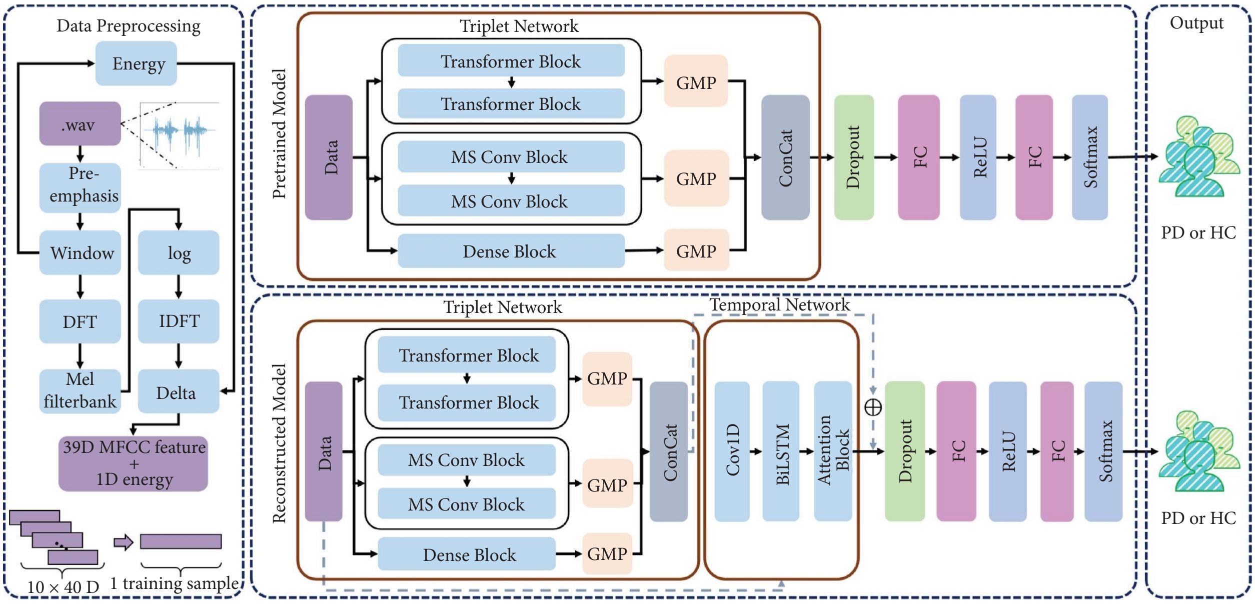


Figure 2: The structure of TmmNet. There are three modules: data preprocessor for MFCC feature extraction, pretrained model for multilevel and scale feature extraction, and reconstructed model for temporal information acquisition and feature classification.

$$\text{energy} = \sum_{t=t_1}^{t_2} x^2[t]. \tag{5}$$

The change of feature in time can represent the acoustic characteristics. Therefore, the change of feature in time is added to the original 13 dimensional features to obtain the delta feature, which represents the change of cepstrum coefficient and energy between frames.

First, we formulate the problem for speech recognition. The MFCC features are defined as $X = \{x_i \in \mathbb{R}^{40*10}, i = 1, 2, 3, \ldots, N\}$ with corresponding 2-class label sequences L, N being the sample numbers of the input data, and 10 and 40 being the width and height of each training sample after preprocessing.

3.3. Triplet Network. The triplet network is a model that integrates three functional blocks. The transformer block integrates a multihead attention mechanism and a multiple feed-forward neural network (Multi-FFN). The multiscale convolution module contains the feature output of different convolution kernels of the depth-wise, spatial, and channel attention components, a normalization module, two one-dimensional convolutions, and a fully connected layer. Finally, DenseNet with three dense blocks is adopted to reduce the possibility of information loss of the first two blocks.

3.3.1. Transformer Block. We redesigned the internal structure of the transformer block, which is composed of a multihead attention mechanism and a multiple feed-forward neural network (Multi-FFN). Since the input data are speech sequence data, a multihead attention can receive three sequences: query, key, and value. The output sequence length of the multihead attention is consistent with the input query sequence length. The length of the query is L_q, and the length of the key and value is L_k.

Multihead attention is composed of one or more parallel cell structures. We call each such cell structure a head. For convenience, we name this cell structure one head attention. Multihead attention consists of multiple one head attention. Remember that a multihead attention has n heads, and the weights of the i^{th} head are W^Q, respectively, W_i^Q, W_i^K, and W_i^V. Then,

$$\begin{aligned} \text{head}_i &= \text{Attention}\left(q \cdot W_i^Q, k \cdot W_i^K, v \cdot W_i^V\right), \\ \text{MultiHead}(q, k, v) &= \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_n\right) \cdot W^O. \end{aligned} \tag{6}$$

The input q, k, and v matrices are input into each one head attention, respectively. The output matrices of each head are spliced according to the characteristic dimensions to obtain a new matrix and then multiplied with the W^O matrix to obtain the output. The multihead attention process is illustrated in Figure 3. The multihead attention mechanism divides each attention operation into a single head and can extract feature information from multiple dimensions. The three transformation tensors perform linear transformation on Q, K, and V, respectively. Each head starts to segment the output tensor from the semantic level; each head wants to obtain a set of Q, K, and V for the calculation of the attention mechanism.

For the multiple feed-forward neural network (Multi-FFN), we embed three different feed-forward neural networks and fuse the three outputs obtained from these blocks. It includes an RBF block, an FC block, and a Conv block; the structure of the three blocks is demonstrated in Figure 4. This redesigned transformer block not only includes the multihead attention mechanism but also transforms a single feed-forward MLP network into a combination of three

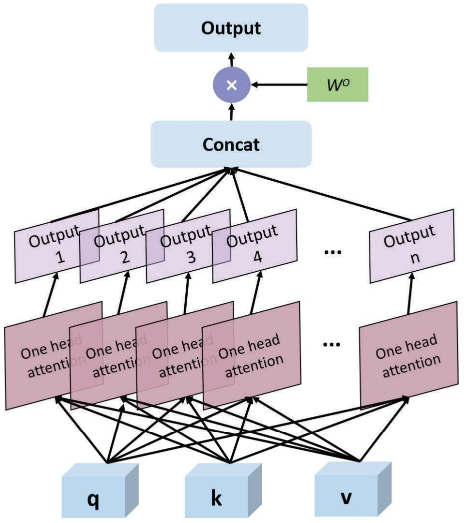


Figure 3: The process of the multihead attention mechanism.

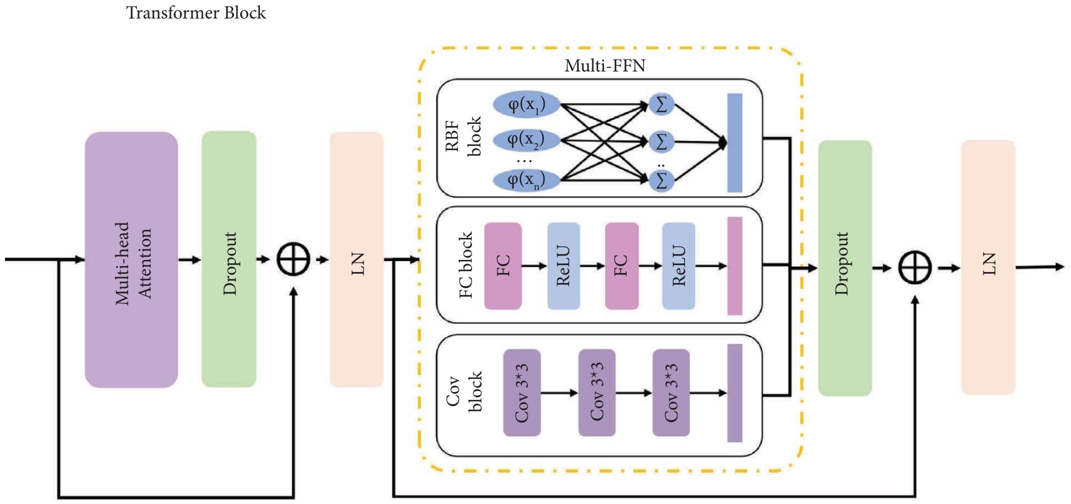


Figure 4: The structure of the transformer block.

feed-forward networks, aiming at extracting multimodel fusion features.

The RBF method is to select P basis functions, and each basis function corresponds to one training data. The interpolation function based on the radial basis function is as follows:

$$\phi(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right),$$
$$F(x) = \sum_{p=1}^{P} w_p \phi_p\left(\| X - X^P \|\right). \tag{7}$$

The input X is an m-dimensional vector, and the sample size is P, $P > m$. The input data point X_p is the center of the radial basis function ϕ. The function of the hidden layer is to map the vector from the low dimension m to the high dimension P. When the low dimension is linearly indivisible, it can be linearly separable from the high dimension. We select reflected sigmoidal function as a radial basis function ϕ. The Conv block contains three layers of convolution operations with a kernel of 3 ∗ 3. The down sampling layer is removed to avoid information loss.

3.3.2. *Multiscale Convolution Block.*

The multiscale convolution block follows the internal structure of the transformer. It uses group convolution to divide all channels into several groups, and convolution is performed in groups. The inverse bottleneck layer is used to perform the convolution operation in the order of dimension increasing (depth-wise convolution) to dimension reducing, and the order of depth-wise convolution is raised to the top. This is to facilitate the comparison of features after the 1 ∗ 1 convolution and prevent the parameter amount from rising. The structure of the multiscale conv block is shown in Figure 5. The speech data are first convolved through three different scales of depth-wise convolution kernels. The joint features of channel attention and spatial attention are extracted from the output of each layer, and then, the final multiconvolution

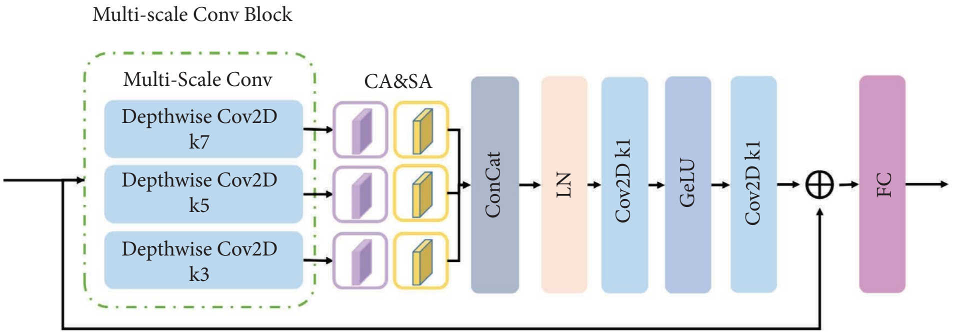


Figure 5: The structure of the multiscale Conv block.

fusion feature is obtained through two one-dimensional convolutions.

It is worth noting that this module uses multiscale convolution kernels, 7 ∗ 7, 5 ∗ 5, and 3 ∗ 3, in depth-wise convolution and processes the input data to obtain three features for fusion. After obtaining the feature map of these three blocks, the serial channel attention and spatial attention are added to highlight the landmark information and target location in the speech signal. Convolutional block attention (CBA) can improve the feature extraction ability of the network model without significantly increasing the amount of computation and parameters [37], which is shown in Figure 6. This module can serially generate attention feature map information in two dimensions of channel and space and then multiply the two feature map information with the original input feature map to generate the final feature map through adaptive feature correction. It includes two modules: channel attention and spatial attention, channel attention uses the relationship between feature channels to generate channel attention mapping. In order to effectively calculate channel attention, we compress the spatial dimension of input feature mapping. Average pooling is usually used to aggregate spatial information. Spatial attention uses the spatial relationship between features to generate spatial attention mapping. In order to calculate spatial attention, we first apply average pooling and max pooling operations along the channel axis and concatenate them to generate effective feature descriptors.

(a) *Channel Attention*. When compressing the spatial dimension of the input feature map, average pooling and max pooling methods are adopted to obtain a total of two one-dimensional vectors. Global average pooling has feedback for every pixel on the feature map, while global max pooling has gradient feedback only where the response is the largest in the feature map during gradient back propagation calculation. Meanwhile, average pooling and max pooling are employed to aggregate spatial dimension features to generate two spatial dimension descriptors: $F^c_{\max}$ and F^c_{avg}, and then, the weight is generated for each channel through an MLP network. Finally, the weight is multiplied by the original channel attention. The formula is as follows:

$$\begin{aligned} M_c(F) &= \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{\text{avg}}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right), \end{aligned} \tag{8}$$

where F represents the input feature map, F^c_{avg} and $F^c_{\max}$ are the features calculated by global average pooling and global max pooling, respectively, W_0 and W_1 denote two-layer parameters in a multilayer perceptron model, and the features between W_0 and W_1 in the multilayer perceptron model need to be processed with ReLU as the activation function.

(b) *Spatial Attention*. With the exception of generating the attention model on the channel, the author said that at the spatial level, the network also needs to understand which parts of the feature map should have higher response. First, average pooling and max pooling are utilized to compress the input feature map at channel levels, and the input features are subject to mean and max operations on the channel dimension, respectively. Finally, two 2D features are obtained and stitched together according to the channel dimension to obtain a feature map with two channels. It is then convolved with a hidden layer containing a single convolution kernel. It must be ensured that the final features are consistent with the input feature map in the spatial dimension:

Max and average pooling operations are also used, but they are executed in the channel dimension. In order to reduce the number of channels in the C dimension of the original feature to 1 dimension, so as to learn spatial attention. The formula is

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{\text{avg}}; F^s_{\max}\right]\right)\right), \end{aligned} \tag{9}$$

where the feature map after max pooling and average pooling is defined as $F^s_{\text{avg}} \in \mathbb{R}^{1*H*W}$ and $F^s_{\max} \in \mathbb{R}^{1*H*W}$ and σ represents sigmoid activation function. The convolution layer shown in this part uses 7 × 7 of the convolution kernel.

Channel attention and spatial attention can be expressed by the following formula:

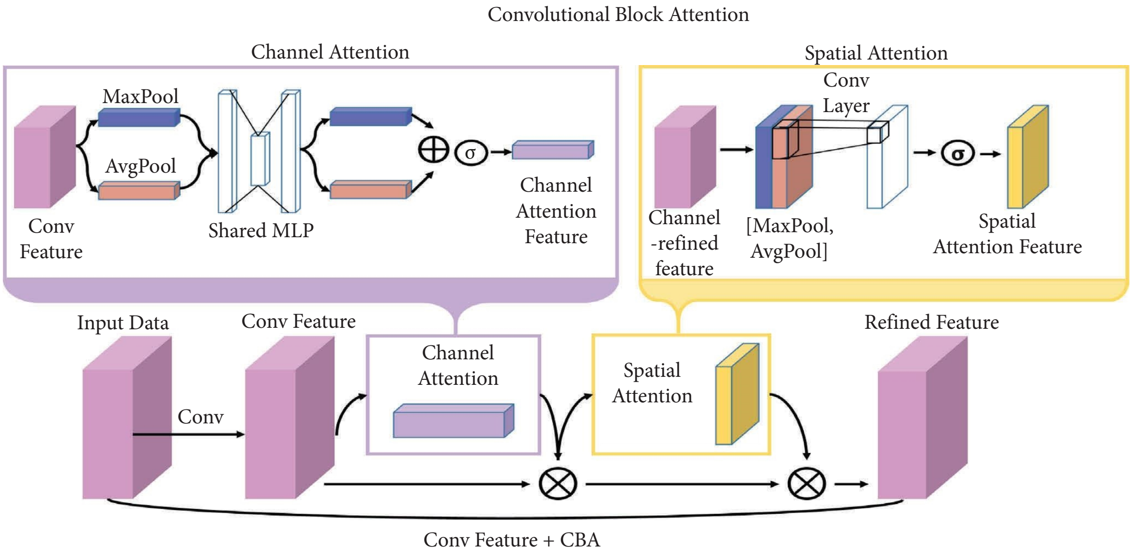


Figure 6: The internal structure of the convolutional block attention.

$$\begin{cases} F' = M_c(F) \otimes F, \\ F'' = M_s(F') \otimes F'. \end{cases} \tag{10}$$

where F is the feature map, $M_s(F)$ and $M_c(F)$ represent channel-based and space-based attention, ⊗ represents the element by element multiplication, and F' and F'' represent the output feature map after channel attention and spatial attention, respectively. Because the input and output sizes of the convolutional block attention module are the same, it can be inserted anywhere in the existing model.

Subsequently, after two layers of $1 * 1$ ordinary convolution layer, a layer of GeLU activation function is inserted in the middle to preserve the probability and dependence on input and to avoid the gradient disappearing.

3.3.3. DenseNet. DenseNet includes three dense blocks and uses a more aggressive dense connection mechanism. Each layer will accept all the layers in front of it as its additional input. The size of the feature map in each dense block is unified to facilitate the concatenation operation. The dense block + transition structure is used in the DenseNet network. A dense block is a module that contains many layers. The feature map size of each layer is the same. Dense connection is adopted between layers. The transition module connects two adjacent dense blocks and reduces the size of the feature map through pooling. As shown in Figure 7, the network structure of DenseNet is mainly composed of dense block and transition (convolution + pooling). The feature transfer method is to directly concatenate the features of all the previous layers to the next layer, instead of pointing to all the layers behind. The details can be illustrated in the literature of [38].

3.4. Reconstructed Model. As the downstream task of the transfer learning model, we reconstruct the network into a triplet network, a temporal network and a two-layer fully connected layer.

We keep the triplet network unchanged in the pretraining process. Since the input speech data are in a sequential state, we implant a temporal network composed of a 1-D convolution layer and a bidirectional LSTM (BiLSTM) with attention to conduct retraining and fine-tuning of the original network weights, followed by two fully connected layers. BiLSTM employs a two-layer internal extensible unit as its structure and adds an attention mechanism as the temporal feature extraction module of the fine-tuning downstream task.

We integrate the output features of the triplet network and the temporal network, preserve the bidirectional information transmission between the speech sequence data frames, make up for the shortcomings of the triplet network, and strengthen the attention to the value in the unique position of the output matrix through attention, and the working mechanism is illustrated in Figure 8.

4. Experiments

This section presents our experimental settings and the performance of the proposed model, compared against several state-of-the-art methods on two challenging speech datasets.

First, we provide a brief introduction to the dataset. Then, we briefly describe the experimental settings. Finally, we give the global evaluation of the experimental results on the two speech datasets.

4.1. Dataset Specifications. In this section, we give a brief description of two speech datasets, i.e., MDVR-KCL dataset [39] and IPVS dataset [40], including data collected from microphones, and the format is in ".wav." The details are introduced as follows.

MDVR-KCL dataset: The MDVR-KCL dataset is a voice file of early and late Parkinson's disease patients and healthy controls recorded with mobile devices. It was collected at

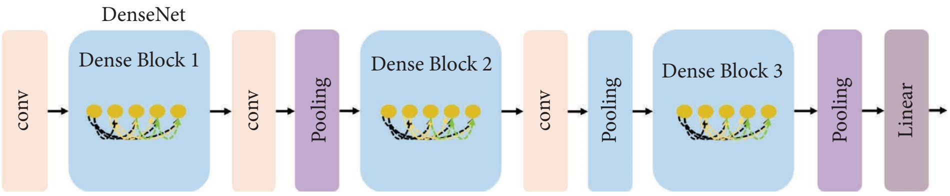


Figure 7: The structure of DenseNet.

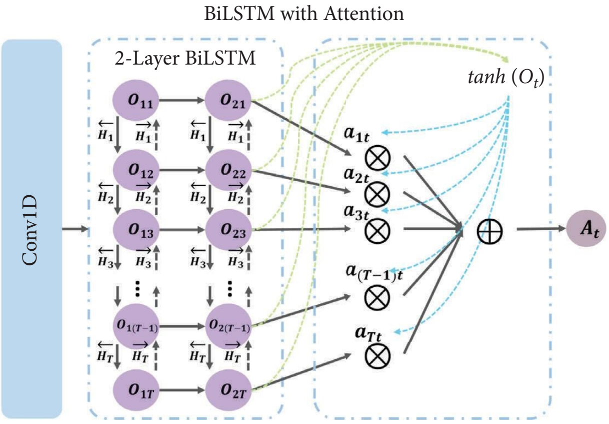


Figure 8: The temporal network. Vectors in the hidden state sequence O_t are fed into the learnable function $\tanh(O_t)$ to produce a probability vector a. The vector A_t is computed as a weighted average of O_T with weighting given by a.

King's College London (KCL) Hospital in Brixton, London, from September 26 to 29, 2017. A typical examination room with about ten square meters area and a typical reverberation tome of approx were utilized to perform the voice recording with 500 ms. The recording was carried out in the real situation of the call (that is, the participant puts the phone on the preferred ear and the microphone is directly close to the mouth). It can be assumed that all recordings are made within the reverberation radius, so it can be considered as "clean" [39].

A Motorola Moto G4 smart phone was used as a recording device. Through the developed application, high-quality recording with a sampling rate of 44.1 kHz and a bit depth of 16 bits (audio CD quality) was finally achieved. The format was ".wav," and the collecting process was as follows:

(i) Ask participants to relax a little and then call the test executor

(ii) Please read the article aloud

(iii) According to the constitution of the participants, they are required to read the text

(iv) Start a spontaneous conversation with the participants, and the test executor starts to randomly ask questions about the scenic spots, local transportation, or personal interests (if acceptable)

(v) The test executor ends the call by saying goodbye

The dataset included data from 16 PD patients and 21 healthy controls. For each HC and PD participant, the data regarding scores were labeled on the Hoehn and Yahr (H and Y) scale, as well as the UPDRS II part 5 and UPDRS III part 18 scales.

IPVS dataset: The IPVS dataset included voice recordings of 28 PD patients and 20 healthy controls, all of which were collected at a 16 kHz sampling rate in a quiet, echoless, warm room. The microphone was located 15 to 25 cm from the people. The participants performed the following tasks: two phonation of the vowels /*a*/, /*e*/, /*i*/, /*o*/, and /*u*/ and syllable execution of "ka" and "pa" for 5 s. In this study, the reading of phonetically balanced phrases and vowel recording were utilized, and a phonetically balanced text was read twice [40]. The reading rules are as follows:

(i) (a) 2 readings of a phonemically balanced text spaced by a pause (30 sec)

(ii) (b) execution of the syllable "pa" (5 sec), pause (20 sec), and execution of the syllable "ta" (5 sec)

(iii) (c) 2 phonation of the vocal "a"

(iv) (d) 2 phonation of the vocal "e"

(v) e) 2 phonation of the vocal "i"

(vi) (f) 2 phonation of the vocal "o"

(vii) (g) 2 phonation of the vocal "u"

(viii) (h) reading of some phonemically balanced words, pause (1 min), and reading of some phonemically balanced phrases

It should be emphasized that there is a one-minute break between the execution of (a) and (b) and between (g) and (h). Before the implementation mentioned in points (c), (d), (e), (f), and (g), it is necessary to inhale as much air as possible and continue to make sound until the lungs are empty. A 30-second pause is required between the execution of (c), (d), (e), (f), (g), and (h).

4.2. Experimental Settings. The experiment was implemented on two speech datasets, and appropriate settings were arranged according to the features of each dataset. The device had a graphics card of GeForce RTX 3080, the memory of an RAM of 32.0 GB, and a CPU of Intel(R) Core(TM) i7-11700. The settings were described in accordance with the dataset.

For the two speech datasets, we shuffled and randomly selected 80% of the data for training and 20% for testing, with a data capacity of 10000+ for the MDVR-KCL dataset and 20000+ for the IPVS dataset, respectively. The final testing time on each dataset was approximately 15 ms (MDVR-KCL dataset) and 27 ms (IPVS dataset). We utilized the spontaneous dialogue file in the MDVR-KCL dataset, as well as the monophonic pronunciation (*/a/*, */e/*, */i/*, */o/*, and */u/*) files collected by the microphone in the IPVS dataset, corresponding to points (c), (d), (e), (f), and (g) in the collection process.

For the MDVR-KCL dataset, we had ".wav" files of 16 PD patients and 21 healthy controls, each containing about two minutes of recording. First, we extracted MFCC features (40 dimensions) through a data preprocessing module: 13 dimensional static coefficients + 13 dimensional first-order difference coefficients + 13 dimensional second-order difference coefficients + 1 dimensional frame energy. The sampling rate was set to 8000, which meant taking 8000 points per second. This way, a segment of audio can output $N \times 40$ dimensional vectors, as these audios were continuous. We took 10 × 40-dimensional sequences as one training sample. Then, these 400 dimensional MFCC features were fed into the triplet network for pretraining and save the model parameters. The processed data were then input into the reconstructed model's triplet network and temporal network for retraining. The triplet network used pretrained parameters, the temporal network used initialization parameters, the time step was set to 100, and the batch size of the entire network was set to 128. For the IPVS dataset, we had ".wav" files of 28 PD patients and 20 healthy controls, the process parameters for MFCC feature extraction were consistent, and the differences were mainly reflected in the amount of data.

4.3. Speech Recognition. The experiment is implemented in four parts, ablation experiment, comparison experiment of machine learning models and deep models, and global evaluation.

4.3.1. Ablation Experiment. We split TmmNet into four constituting components, i.e., TmmNet without a fine-tuning process (TmmNet NoFT), TmmNet without an MS Conv block (TmmNet NoConv), TmmNet without an ST-Attn block (TmmNet NoAttn), and TmmNet without a temporal network and an ST-Attn block (TmmNet NoTN) for the ablation experiment. Due to the small difference in the effect of various models on the IPVS dataset, we only use the MDVR-KCL dataset to carry out the ablation experiment.

We used four evaluation indicators, precision, recall, F1 score, and accuracy, to evaluate the performance of the four constituting components and the overall model. As shown in Table 2, the precision represents the proportion of positive cases in the samples with positive predicted results. The performance of several split modules here varies greatly. It can be seen that TmmNet, TmmNet (NoConv), and TmmNet (NoAttn) perform best, while TmmNet (NoFT) and TmmNet (NoTN) perform poorly, because the temporal network in the fine-tuning process and downstream tasks has a greater impact on precision. For recall, the performance of the overall model and the split modules was not satisfactory, but the overall model reaches 75%, ranking first. For F1 score, TmmNet performs the best, followed by TmmNet (NoConv). TmmNet (NoTN) is the worst, which proves that the TN module of the fine-tuning part has the greatest impact on the F1 score value of the overall model. By comparing the accuracy of these components, TmmNet (NoFT) and TmmNet (NoTN) perform worse than other models, indicating the importance of fine-tuning and temporal network in the overall model.

The confusion matrix of the components is shown in Figure 9. We can see that TmmNet achieves 100% of the detection rate of PD, but the misclassification rate of HC is still high and also better than that of other component modules. The worst detection rate for PD is TmmNet (NoTN), and the highest error rate for HC is TmmNet (NoFT). It can be seen that the fine-tuning part, attention, and temporal information play a significant role in the TmmNet framework. 24.74% of healthy subjects are classified as PD patients, because the pronunciation in the training data of some mild PD patients is similar to that of healthy people.

We also utilize ROC (receiver operating characteristic) curves and AUC (area under curve) values to compare the performance of these constituting components (Figure 10). The closer the ROC curve is to the upper left corner, the better the performance is. AUC is defined as the area under the ROC curve enclosed by the coordinate axis. It is a machine learning performance metric used to evaluate the binary model. The degree of AUC greater than 0.5 measures the extent to which the algorithm is superior to the randomly selected algorithm. It can be seen that TmmNet, TmmNet (NoAttn), and TmmNet (NoConv) are more than 80% and that TmmNet (NoFT) and TmmNet (NoTN) are more than 70%.

4.3.2. Results of Machine Learning Models. In this section, we compare eight machine learning methods, i.e., DT, GBDT, LDA, KNN, LightGBM, LR, RF, and XGBoost, for the classification of speech signals in PD patients and healthy people.

MDVR-KCL dataset: As illustrated in Table 3, we still adopt four classification evaluation indicators to compare

Table 2: Results of the four constituting components on the MDVR-KCL dataset.

Components	Evaluation			
	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
TmmNet (NoFT)	86.39	56.87	68.59	80.46
TmmNet (NoConv)	97.16	69.02	80.71	86.82
TmmNet (NoTN)	71.73	61.92	66.46	75.44
TmmNet (NoAttn)	96.32	66.74	78.85	85.22
TmmNet	**100.00**	**75.26**	**85.88**	**90.23**

The bold values in Table 2 represents the highest results among the split models.

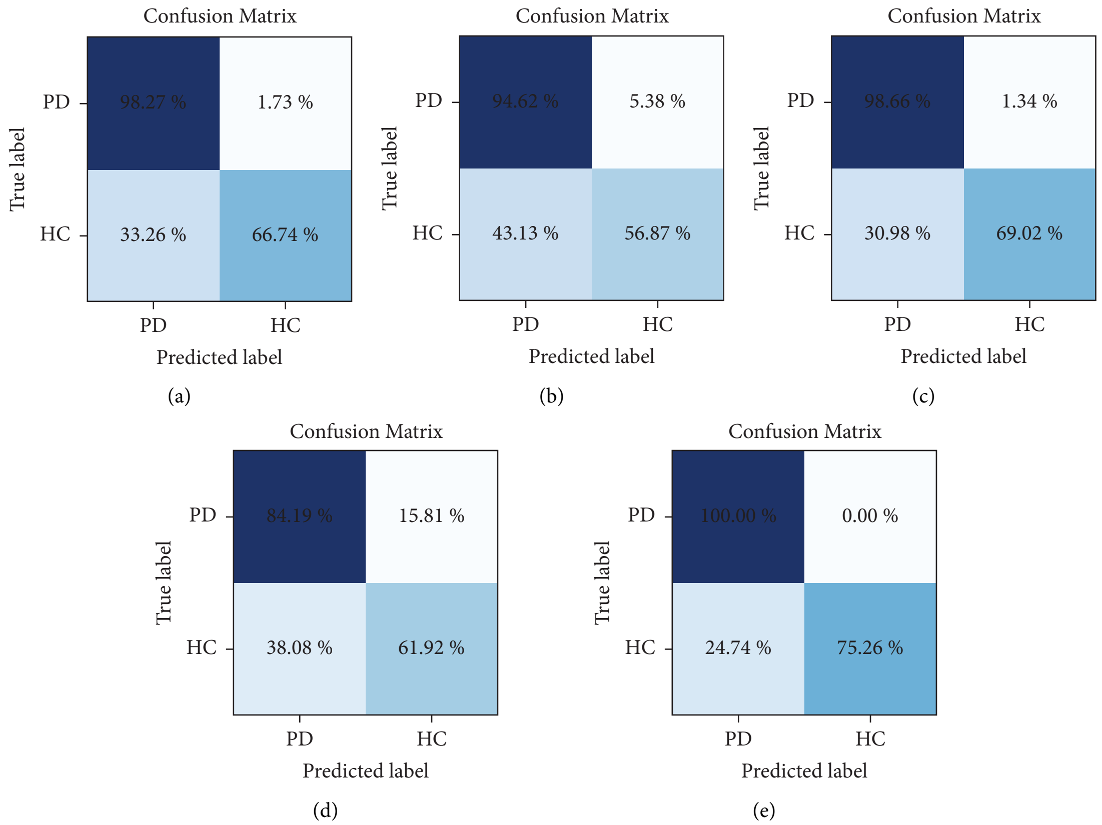


Figure 9: The confusion matrix of TmmNet and its components on the MDVR-KCL dataset. (a) TmmNet (NoAttn). (b) TmmNet (NoFT). (c) TmmNet (NoConv). (d) TmmNet (NoTN). (e) TmmNet.

the performance of different machine learning models. The best models for these four indicators are LR (precision 96.32%), LightGBM (recall 90.47%), LightGBM (F1 score 77.49%), and XGBoost (Acc 80.79%). The classification accuracy of KNN (80.73%) and XGBoost (80.79%) is similar, but there is still a big gap compared with TmmNet (90.23%).

In addition, the ROC curve is demonstrated in Figure 11. The AUC value of TmmNet is the highest, followed by LightGBM, which is a gradient boosting framework and utilizes a decision tree-based learning algorithm. It is distributed and suitable for samples with large datasets. DT and KNN have the lowest effect. When KNN treats the sample imbalance, the predicted accuracy of rare categories is low. DT is prone to overfitting, and it is easy to ignore the correlation of attributes in the dataset. For the input speech data, each frame is interrelated, and the sample data volume is large, which is also the reason why machine learning methods can be applied.

IPVS dataset: We also used the IPVS dataset to distinguish the characteristics of healthy subjects and patients with Parkinson's disease by following the pronunciation of five vowels /*a*/, /*e*/, /*i*/, /*o*/, and /*u*/.

The classification performance of machine learning methods is discussed in Table 4. The evidence shows that the results of our proposed TmmNet in five syllable pronunciation classification are significantly superior to the traditional machine learning algorithm, with an average accuracy of more than 99%. It also means that our model can be directly used for speech disorders prediagnosis. By comparing the pronunciation of five vowels, it is found that

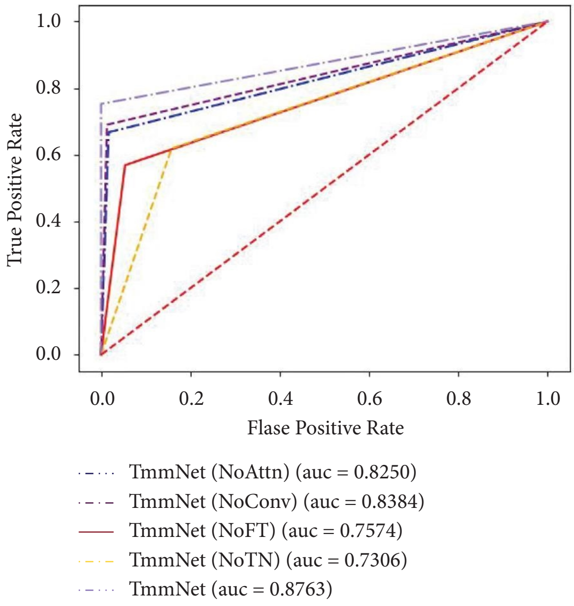


Figure 10: ROC curve of four constituting components and the overall model.

Table 3: Results of the eight machine learning methods on the MDVR-KCL dataset.

Methods	Evaluation			
	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
DT	65.81	46.95	54.80	68.52
GBDT	82.42	38.01	52.02	71.31
LDA	70.25	41.73	52.36	71.48
KNN	89.70	59.42	71.49	80.73
LightGBM	67.76	90.47	77.49	79.00
LR	96.32	71.68	43.81	54.38
RF	91.71	51.17	65.69	78.27
XGBoost	89.66	58.44	70.76	80.79
TmmNet	**100.00**	**75.26**	**85.88**	**90.23**

The bold values in Table 3 represents the highest results compared with the machine learning classifiers.

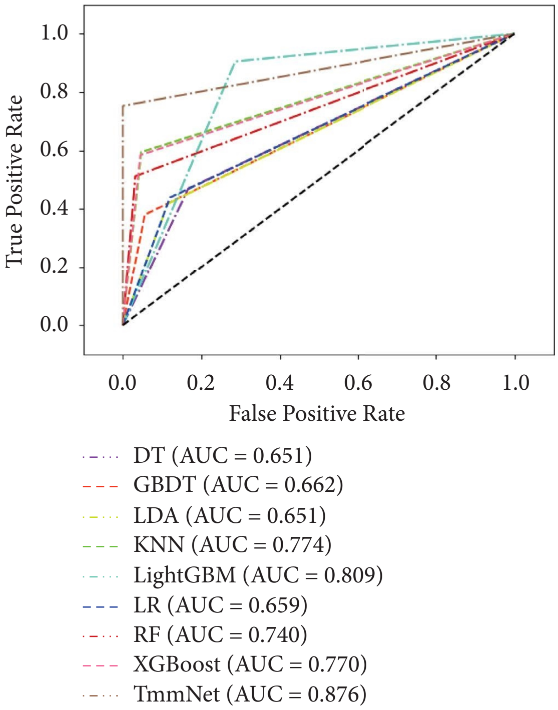


Figure 11: ROC curve of eight machine learning methods and TmmNet.

Table 4: Performance of machine learning methods on the IPVS dataset.

	Vowel /*a*/					Vowel /*e*/					Vowel /*i*/			
Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
DT	98.46	98.06	98.26	98.33	DT	98.17	98.81	98.49	97.95	DT	97.76	97.18	97.47	97.23
GBDT	97.92	98.47	98.19	98.26	GBDT	97.47	99.49	98.47	97.90	GBDT	97.68	98.08	97.88	97.67
KNN	97.07	97.07	97.07	97.21	KNN	97.50	98.48	97.99	97.23	KNN	95.04	97.87	96.43	95.97
LR	97.48	98.57	98.02	98.09	LR	97.92	99.11	98.51	97.97	LR	97.15	97.32	97.24	96.96
RF	99.94	99.69	99.82	99.82	RF	99.42	99.96	99.69	99.58	RF	99.75	99.40	99.58	99.54
LDA	97.07	97.07	97.07	97.21	LDA	97.50	98.48	97.99	97.23	LDA	95.04	97.87	96.43	95.97
LightGBM	99.69	100.00	99.84	99.85	LightGBM	99.89	99.73	99.81	99.74	LightGBM	99.93	99.52	99.72	99.69
	Vowel /*o*/					Vowel /*u*/					TmmNet			
Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
DT	98.71	98.46	98.58	98.34	DT	98.40	98.14	98.27	98.05	Vowel /*a*/	**99.89**	**99.89**	**99.89**	**99.90**
GBDT	97.58	99.28	98.42	98.13	GBDT	98.11	99.57	98.84	98.68	Vowel /*e*/	**99.86**	**99.83**	**99.84**	**99.79**
KNN	99.74	99.79	99.76	99.72	KNN	97.62	97.26	97.44	97.09	Vowel /*i*/	**99.72**	**99.61**	**99.66**	**99.63**
LR	96.41	97.79	97.09	96.57	LR	97.24	99.15	98.18	97.93	Vowel /*o*/	**99.95**	**99.85**	**99.90**	**99.87**
RF	99.79	99.79	99.79	99.75	RF	99.89	99.89	99.89	99.88	Vowel /*u*/	**99.89**	**99.94**	**99.92**	**99.91**
LDA	96.65	97.14	96.90	96.33	LDA	97.62	97.26	97.44	97.09					
LightGBM	99.89	99.59	99.74	99.69	LightGBM	99.89	99.68	99.78	99.76					

The bold values in Table 4 represents the highest results among all the compared methods.

the subjects are difficult to distinguish between /*e*/ and /*i*/ pronunciation patterns, and the average effect of various machine learning methods is the worst, but the effect of TmmNet is still more than 99%. It is worth mentioning that the effect of RF stands out among many methods and can be comparable with the proposed TmmNet.

In addition, the ROC curve can also clearly show which machine learning method is more suitable for speech datasets. The comparison of AUC values is shown in Figure 12. We only list the evaluation results of resolving vowel /*a*/. The AUC values of LightGBM and RF rank in Top 1 and Top 2, respectively, which shows their advantages in traditional classification. The performance gap of other classifiers is relatively small, which indicates that the data are highly separable and fully suitable for machine learning methods.

4.3.3. Results of Deep Models. In this section, we evaluate and compare 6 CNN-based models, i.e., CNN, DNN [41], DenseCNN [42], ResCNN [43], ResNet50 [44], and ThinResNet [45], and 5 temporal models, i.e., HMM, LSTM, LSTM (Attn), BiLSTM (Attn) [46], and BiGRU(Attn) [47], for the speech recognition in PD patients and healthy people.

MDVR-KCL dataset: In Table 5, by comparing the CNN-based model with the temporal model, we can see that the average classification performance of the temporal model is better, which is related to the sequential form of speech data. On the one hand, among the CNN-based models, DenseCNN obtains the highest accuracy. Because it proposes a more radical intensive connection mechanism, which connects all the layers and directly concatenates the feature maps from different layers, it can achieve feature reuse and improve efficiency, so that it can obtain superior performance. The results of ResNet50 and DenseCNN are relatively poor. Due to the sparsity of the data in the training process, it leads to the overfitting phenomenon, which is inferior to other models. On the other hand, the bidirectional memory model with an attention mechanism (BiGRU (Attn), BiLSTM (Attn)) in the temporal network performs satisfactorily due to its special gating mechanism and the construction of the expandable unit. Inspired by the advantages of these models, our proposed TmmNet has the attribute of integrating spatiotemporal features and has a transfer learning infrastructure. It goes beyond these mainstream models and becomes an effective tool that best fits the speech data being trained.

Furthermore, the ROC curve of 11 different deep models on the MDVR-KCL dataset is demonstrated in Figure 13. The results in Table 5 are consistent with the performance ranking of the ROC curve. The performance based on ResNet and HMM is relatively poor. The corresponding AUC can also see that the inflection point of TmmNet is closer to the upper left, while the results of ResNet50, HMM, ResCNN, and LDA have a large gap compared with other deep models, which is clearly reflected.

IPVS dataset: We also compared five deep learning methods in Table 6, and the performance is significantly better than that of machine learning methods, because the accuracy of pronunciation resolution for each syllable is more than 99%, of which the effect of HMM is obviously at a disadvantage and also shows its limitations. We will give priority to other networks as speech recognition algorithms.

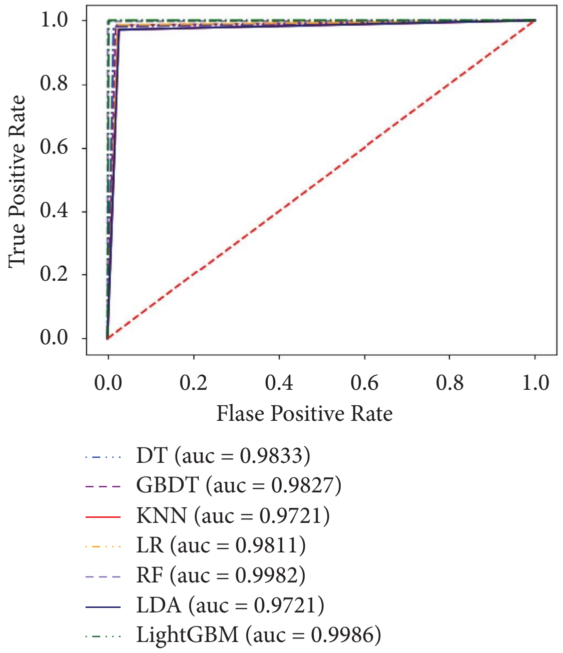


Figure 12: ROC curve of machine learning and TmmNet.

Here, we can see that all the machine learning methods and deep models compared in this paper have generally excellent classification effects on this dataset, indicating that the monosyllabic pronunciation of subjects is easier to distinguish than reading a long passage or finishing a dialogue. The research in this paper can serve as a reference for the prediagnosis and severity assessment of Parkinson's disease.

ROC and AUC are utilized to evaluate the classification performance of the above deep learning model in Figure 14. Similarly, the performance of deep models is not far from that of traditional machine learning algorithms. Due to the high sensitivity of the temporal network to speech sequences, the average performance is slightly higher than that of machine learning and other deep learning models.

4.3.4. Global Evaluation. In this section, we evaluate the overall effect of the TmmNet model and use the confusion matrix of TmmNet on two datasets to describe the classification accuracy. In addition, a perceptual experiment is conducted to evaluate the classification results of speech disorders.

MDVR-KCL dataset: We use the confusion matrix to show the classification results of TmmNet on the MDVR-KCL dataset in Figure 15. The accuracy of screening for PD reached 100%, although some healthy subjects were wrongly classified as PD patients. At the lower left corner of the confusion matrix, 24.74% of the samples of healthy subjects were wrongly divided into PD samples. We extracted a wrongly divided sample and found that its feature distribution was

Table 5: Performance of deep models on the MDVR-KCL dataset.

CNN-based methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Temporal network-based methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
CNN	89.38	57.50	69.99	80.01	HMM	61.74	50.48	55.54	67.73
DNN [41]	80.76	51.26	62.71	76.06	LSTM	84.28	59.39	69.68	79.37
ResCNN [43]	63.40	46.09	53.38	68.13	LSTM (Attn)	88.42	51.08	64.76	77.79
ThinResNet [45]	86.96	59.60	70.73	79.96	BiGRU (Attn) [47]	92.45	64.49	75.98	84.04
DenseCNN [42]	85.71	86.84	87.16	86.47	BiLSTM(Attn) [46]	92.77	67.27	77.99	85.07
ResNet50 [44]	65.00	78.00	70.91	59.75	TmmNet	**100.00**	**75.26**	**85.88**	**90.23**

The bold values in Table 5 represents the highest results compared with the deep learning models.

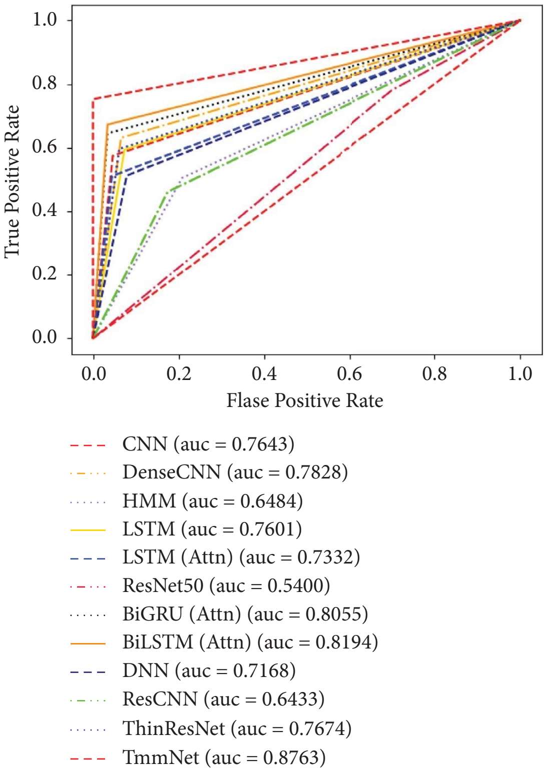


Figure 13: ROC curve of deep models and TmmNet on the MDVR-KCL dataset.

similar to the samples in the PD category. Therefore, the similarity of the data would also affect the screening and early diagnosis of Parkinson's disease.

IPVS dataset: We use the confusion matrix to show the classification results of TmmNet for pronunciation /*a*/ in Figure 16. It can be seen that the probability of correct classification of samples on the diagonal is more than 99%. Compared with the MDVR-KCL dataset, the accuracy of monosyllabic follow-up classification in the IPVS dataset is significantly higher, which is inevitably related to the high complexity of long texts. Therefore, when we conduct speech tests on subjects, we can follow the vowels first and then the long text, which can effectively detect Parkinson's disease and evaluate its severity.

Perceptual Experiment. There are a total of 20 nonmedical subjects conducting hearing experiments in a quiet room of 20 square meters. We randomly select an audio from a dataset of PD patients, and each person is limited to 10 seconds to listen to a recording before giving a judgment on whether it is an audio from a PD patient. After conducting a hearing test on 20 people for a segment of audio, it was found that 12 people correctly recognized the audio for PD patients, with a comprehensive accuracy rate of 60.00%. After inquiry, it is not ruled out that there is a possibility of speculation. This recognition rate is much lower than the model results proposed in this article. We also invited a PD expert to conduct hearing tests on samples from 50 datasets, including 25 PD samples. The test results showed that the audio recognition accuracy of PD patients was 84%, HC audio recognition accuracy was 92%, and the overall accuracy was 88.00%. Therefore, it can be seen that the probability of using the proposed scheme for accurate diagnosis of Parkinson's disease using audio exceeds 90%, providing a reference for automated diagnosis research.

Table 6: Performance of deep models on the IPVS dataset.

	Vowel /*a*/					Vowel /*e*/					Vowel /*i*/			
Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
BiLSTM	99.24	99.94	99.59	99.60	BiLSTM	99.39	99.83	99.61	99.47	BiLSTM	99.41	99.68	99.55	99.50
LSTM	99.36	99.94	99.65	99.68	LSTM	99.39	100.00	99.69	99.58	LSTM	99.39	99.76	99.57	99.52
BiLSTM (Attn)	99.28	99.89	99.59	99.60	BiLSTM (Attn)	99.63	99.93	99.78	99.70	BiLSTM (Attn)	99.24	99.58	99.41	99.35
HMM	96.34	96.30	96.32	96.50	HMM	97.63	95.56	96.58	95.37	HMM	96.94	93.42	95.14	94.70
CNN	98.82	99.64	99.23	99.26	CNN	99.01	99.76	99.38	99.17	CNN	98.68	99.59	99.13	99.02
	Vowel /*o*/					Vowel /*u*/					TmmNet			
Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)	Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
BiLSTM	99.15	99.89	99.52	99.42	BiLSTM	98.91	99.73	99.32	99.22	Vowel /*a*/	**99.89**	**99.89**	**99.89**	**99.90**
LSTM	99.23	99.89	99.56	99.48	LSTM	99.21	99.73	99.47	99.40	Vowel /*e*/	**99.86**	**99.83**	**99.84**	**99.79**
BiLSTM (Attn)	98.69	99.74	99.21	99.06	BiLSTM (Attn)	99.11	99.89	99.50	99.43	Vowel /*i*/	**99.72**	**99.61**	**99.66**	**99.63**
CNN	99.11	99.50	99.30	99.15	CNN	98.70	99.68	99.19	99.07	Vowel /*o*/	**99.95**	**99.85**	**99.90**	**99.87**
HMM	97.77	96.23	96.99	96.48	HMM	97.63	97.11	97.37	97.00	Vowel /*u*/	**99.89**	**99.94**	**99.92**	**99.91**

The bold values in Table 6 represents the highest results among all the compared methods.

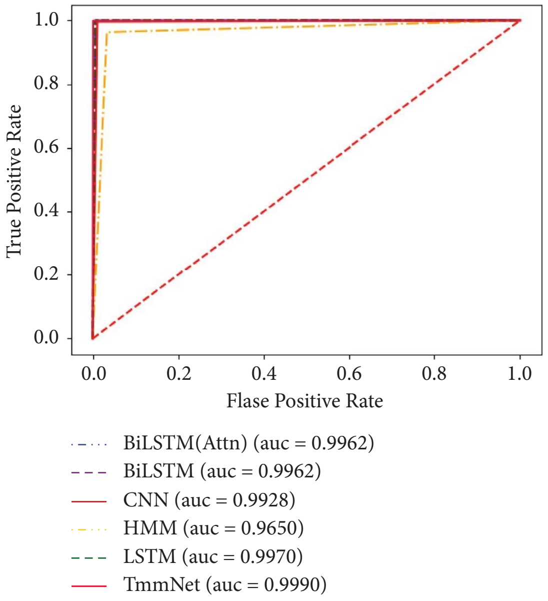


Figure 14: ROC curve of deep models and TmmNet on the IPVS dataset.

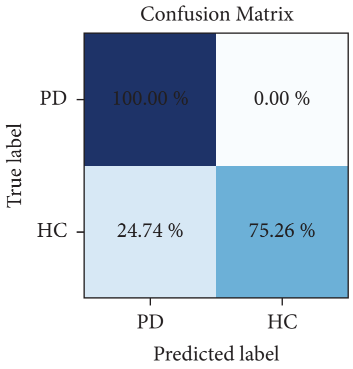


Figure 15: The confusion matrix of TmmNet on the MDVR-KCL dataset.

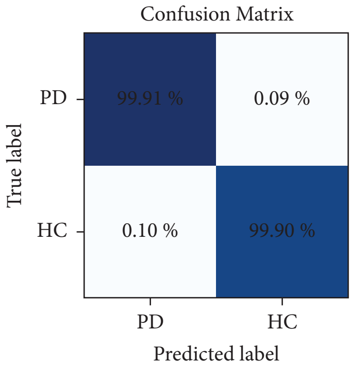


Figure 16: The confusion matrix of TmmNet on vowel /*a*/.

Cohen's Kappa. Cohen's kappa is an indicator used for consistency testing and can also be used to measure the effectiveness of classification. For classification problems, consistency refers to whether the predicted results of the model are consistent with the actual classification results. The calculation of Cohen's Kappa is based on the confusion matrix, with values ranging from −1 to 1, usually greater than 0.

The formula for calculating the kappa coefficient based on the confusion matrix is as follows:

$$\text{kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{11}$$

where p_o is the sum of the number of correctly classified samples for each class divided by the total number of samples, which is the overall classification accuracy, and p_e represents the sum of the product of actual and predicted quantities for all classes, divided by the square of the total number of samples.

It can be divided into five groups to represent different levels of consistency: low consistency is [0.0, 0.20], fair consistency is [0.21, 0.40], moderate consistency is [0.41, 0.60], substantial consistency is [0.61, 0.80], and almost perfect consistency is [0.81, 1].

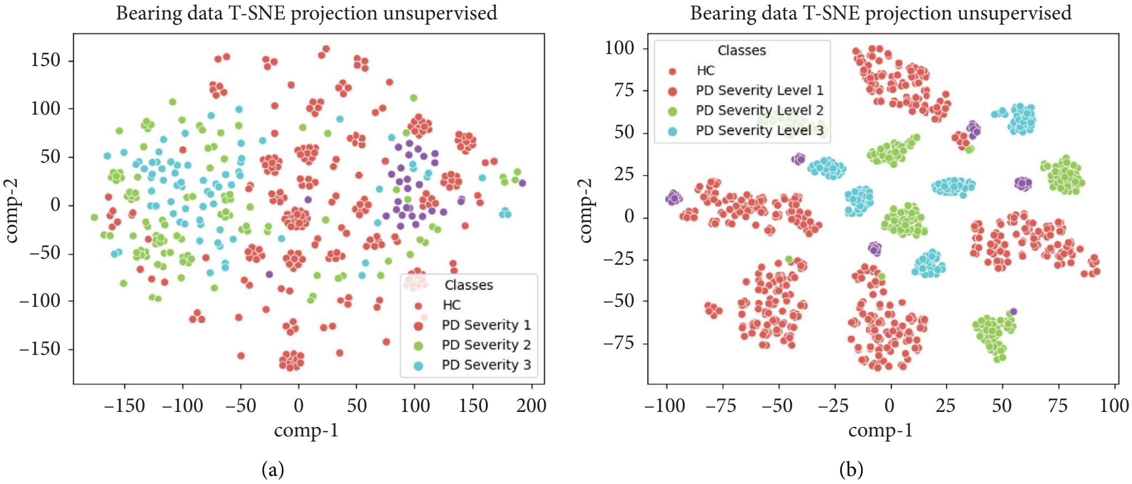
Through consistency testing, the Kappa values of TmmNet on two datasets are 0.9980 (/*a*/), 0.9952 (/*e*/), 0.9926 (/*i*/), 0.9974 (/*o*/), 0.9981 (/*u*/), and 0.7863 (MDVR-KCL dataset), respectively. It can be clearly seen that the TmmNet model performs much better on IPVS than on MDVR-KCL, achieving almost perfect consistency, while achieving high consistency on MDVR-KCL. The confusion matrix on the MDVR-KCL dataset is relatively imbalanced, as the detection rate of PD in the test set is 100%, there is a problem of data imbalance. Other models also have the same problem, and the data should be filtered or added later.

Severity Assessment. Furthermore, we also adopt a speech dataset from Parkinson's disease to validate the proposed model, and the results showed that TmmNet also has good performance in classifying the severity of PD speech. This study can first detect patients with potential Parkinson's disease based on speech data and then evaluate their severity. The experimental results are shown in Table 7. According to the Hoehn and Yahr scale, speech data in the MDVR-KCL dataset are classified into four categories: healthy individuals, PD1 level, PD2 level, and PD3 level, which is completely marked by expert evaluation scores. We compared five deep learning methods [48–51], including models based on convolutional neural networks, transformers, and transfer learning for speech emotion recognition, and achieved good results, which is sufficient to prove that these deep models can evaluate the severity of Parkinson's disease speech, and the effectiveness of the proposed TmmNet is also remarkable.

Furthermore, we also utilize the t-SNE visualization method to show the classification ability of our proposed method. The experimental results are shown in Figure 17.

Table 7: Performance of deep models on the MDVR-KCL dataset for severity rating.

Methods	Precision (%)	Recall (%)	F1 score (%)	Acc (%)
Resnet50	91.60	87.25	86.47	95.64
LIGHT-SERNET [48]	99.98	99.93	99.95	99.94
CTL-MTNET [49]	99.87	99.64	99.75	97.38
GM-TCN [50]	99.88	99.56	99.72	99.87
TIM-NET [51]	99.95	99.86	99.90	99.88
TmmNet	100.00	99.97	99.93	99.95



(a)

(b)

Figure 17: Comparison of t-SNE feature dimensionality reduction on the MDVR-KCL dataset. (a, b) Comparison between the original data and the feature extracted from our model after dimensionality reduction on the MDVR-KCL dataset.

Subfigures (a) and (b) represent the comparison between the features extracted from our model and the original data after dimensionality reduction. We can clearly see that the original data are more chaotic than the extracted features, and the data of the four classes have cross coverage, while the features extracted by the proposed model have a distance between different classes and a large degree of aggregation for the same class, showing better separability.

5. Conclusion

We have presented techniques for screening out PD patients or samples with potential PD from the speech data of subjects, including MFCC feature extraction, and a pre-trained triplet hybrid model and a reconstructed temporal model achieve transfer learning for high-level expression of the MFCC feature. Experiments have shown that our method can not only be applied to the detection of monosyllabic vowels in patients with Parkinson's disease but also have the function of analysis and recognition for a period of time of the spontaneous dialogue. Although the effect is not as good as the former, it can be used as a reference for the detection and classification of PD speech. By the abovementioned strong results, we hope to stimulate more research in this direction so that we can eventually improve the ability of transfer learning models to process speech sequence data and promote speech modeling.

Data Availability

Data are available on request from the authors.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Aite Zhao was responsible for supervision, writing of the original draft, and writing, reviewing, and editing of the manuscript. Nana Wang and Xuesen Niu were responsible for methodology, software, and writing of the original draft. Ming Chen and Huimin Wu were responsible for formal analysis, methodology, software, and visualization.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant no. 62106117, China Postdoctoral Science Foundation under Grant no. 2022M711741, Natural Science Foundation of Shandong Province under Grant no. ZR2021QF084, and Shandong Province Higher Education Institutions Youth Innovation and Technology Support Program (2023KJ365).

References

[1] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 02 2011.

[2] C. Stewart, L. Winfield, A. Hunt et al., "Speech dysfunction in early Parkinson's disease," *Movement Disorders*, vol. 10, no. 5, pp. 562–565, 1995.

[3] S. Polychronis, F. Niccolini, G. Pagano, T. Yousaf, and M. Politis, "Speech difficulties in early de novo patients with Parkinson's disease," *Parkinsonism & Related Disorders*, vol. 64, pp. 256–261, 2019.

[4] J. Rusz, R. Cmejla, H. Ružičková et al., "Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy," *Journal of Neural Transmission*, vol. 120, no. 2, pp. 319–329, 2013.

[5] J. Rusz, "Detecting speech disorders in early Parkinson's disease by acoustic analysis," Habilitation Thesis, Czech Technical University in Prague, Prague, Czechia, 2018.

[6] C. Medina, E. Vargas, S. Munger, and J. Miller, "Vocal changes in a zebra finch model of Parkinson's disease characterized by alpha-synuclein overexpression in the song-dedicated anterior forebrain pathway," *PLoS One*, vol. 17, no. 5, Article ID e0265604, 2022.

[7] M. I. A. Ferreira, F. A. Barbieri, V. C. Moreno, T. Penedo, and J. M. R. Tavares, "Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters," *Gait and Posture*, vol. 98, pp. 49–55, 2022.

[8] A. Zhao, J. Dong, J. Li, L. Qi, and H. Zhou, "Associated spatio-temporal capsule network for gait recognition," *IEEE Transactions on Multimedia*, vol. 24, p. 1, 2021.

[9] R. Liu, Z. Wang, S. Qiu et al., "A wearable gait analysis and recognition method for Parkinson's disease based on error state kalman filter," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4165–4175, 2022.

[10] G. Cesarelli, L. Donisi, A. Coccia et al., "Ataxia and Parkinson's disease patients classification using tree-based machine learning algorithms fed by spatiotemporal features: a pilot study," in *Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications*, pp. 1–6, MeMeA, Messina, Italy, June 2022.

[11] A. Zhao, J. Li, J. Dong et al., "Multimodal gait recognition for neurodegenerative diseases," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9439–9453, 2022.

[12] H. Zhao, R. Wang, Y. Lei, W.-H. Liao, H. Cao, and J. Cao, "Severity level diagnosis of Parkinson's disease by ensemble k-nearest neighbor under imbalanced data," *Expert Systems with Applications*, vol. 189, Article ID 116113, 2022.

[13] J. Qin, T. Liu, Z. Wang, Q. Zou, L. Chen, and C. Hong, "Speech recognition for Parkinson's disease based on improved genetic algorithm and data enhancement technology," in *Data Science*, Y. Wang, G. Zhu, Q. Han, H. Wang, X. Song, and Z. Lu, Eds., pp. 273–286, Springer Nature, Singapore, 2022.

[14] B. M. Pati, M. Mahanta, and A. Taparugssanagorn, "Assessment of spectrum sensing using support vector machine combined with principal component analysis," *International Journal of Sensor Networks*, vol. 39, no. 4, pp. 256–278, 2022.

[15] X. Wang, X. Chen, Q. Wang, and G. Chen, "Early diagnosis of Parkinson's disease with speech pronunciation features based on xgboost model," in *Proceedings of the 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pp. 209–213, Xiamen, China, June 2022.

[16] Z. Ma, Y. Liu, X. Liu, J. Ma, and F. Li, "Privacy-preserving outsourced speech recognition for smart iot devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8406–8420, 2019.

[17] B. Adam and S. Tenbohlen, "Classification of multiple pd sources by signal features and lstm networks," in *Proceedings of the 2018 IEEE International Conference on High Voltage Engineering and Application*, pp. 1–4, ICHVE, Athens, Greece, September 2018.

[18] M. Dhakal, A. Chhetri, A. K. Gupta, P. Lamichhane, S. Pandey, and S. Shakya, "Automatic speech recognition for the Nepali language using cnn, bidirectional lstm and resnet," in *Proceedings of the 2022 International Conference on Inventive Computation Technologies*, pp. 515–521, ICICT, Nepal, July 2022.

[19] J. C. Vásquez-Correa, N. Garcia-Ospina, J. R. Orozco-Arroyave, M. Cernak, and E. Nöth, "Phonological posteriors and gru recurrent units to assess speech impairments of patients with Parkinson's disease," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds., pp. 453–461, Springer International Publishing, Singapore, 2018.

[20] Q. Zhang, J. Lin, H. Song, and G. Sheng, "Fault identification based on pd ultrasonic signal using rnn, dnn and cnn," in *Proceedings of the 2018 Condition Monitoring and Diagnosis*, pp. 1–6, CMD, Perth, Australia, September 2018.

[21] H. V. Dang, H. Tran-Ngoc, T. V. Nguyen, T. Bui-Tien, G. De Roeck, and H. X. Nguyen, "Data-driven structural health monitoring using feature fusion and hybrid deep learning," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 4, pp. 2087–2103, 2021.

[22] A. Zhao, H. Wu, M. Chen, and N. Wang, "A multi-level feature attention network for covid-19 detection based on multi-source medical images," *Multimedia Tools and Applications*, vol. 1, no. 1, pp. 1–32, 2024.

[23] O. E. Ariss and K. Hu, "Resnet-based Parkinson's disease classification," *IEEE Transactions on Artificial Intelligence*, vol. 4, pp. 1–11, 2022.

[24] A. Hernandez, P. A. P'erez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. K. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," 2022, https://arxiv.org/abs/2204.01670.

[25] M. Mohaghegh and J. Gascon, "Identifying Parkinson's disease using multimodal approach and deep learning," in *Proceedings of the 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications*, pp. 1–6, CITISIA, Sydney, Australia, November 2021.

[26] J. Ma, Y. Zhang, Y. Li et al., "Deep dual-side learning ensemble model for Parkinson speech recognition," *Biomedical Signal Processing and Control*, vol. 69, Article ID 102849, 2021.

[27] Y. Liu, Y. Li, X. Tan, P. Wang, and Y. Zhang, "Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 63, Article ID 102165, 2021.

[28] Y. Li, L. Zhou, L. Qin et al., "Deep double-side learning ensemble model for few-shot Parkinson speech recognition," 2020, https://arxiv.org/abs/2006.11593.

[29] M. Xu, J. Du, Z. Xue, Z. Guan, F. Kou, and L. Shi, "A scientific research topic trend prediction model based on multi-lstm

and graph convolutional network," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 6331–6353, 2022.

[30] S. S. Upadhya, A. Cheeran, and J. H. Nirmal, "Discriminating Parkinson diseased and healthy people using modified mfcc filter bank approach," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 1021–1029, 2019.

[31] S. Nivash, E. N. Ganesh, K. Harisudha, and S. Sreeram, "Extensive analysis of global presidents' speeches using natural language," in *Sentimental Analysis and Deep Learning*, S. Shakya, V. E. Balas, S. Kamolphiwong, and K.-L. Du, Eds., pp. 829–850, Springer Singapore, Singapore, 2022.

[32] Y. U. Qing, M. A. Yi, and L. I. Yongfu, "Enhancing speech recognition for Parkinson's disease patient using transfer learning technique," *Journal of Shanghai Jiaotong University*, vol. 1, p. 27, 2022.

[33] S. R. Shahamiri and S. S. Binti Salim, "Artificial neural networks as speech recognisers for dysarthric speech: identifying the best-performing set of mfcc parameters and studying a speaker-independent approach," *Advanced Engineering Informatics*, vol. 28, no. 1, pp. 102–110, 2014.

[34] A. Zhao, H. Wu, M. Chen, and N. Wang, "A spatio-temporal siamese neural network for multimodal handwriting abnormality screening of Parkinson's disease," *International Journal of Intelligent Systems*, vol. 2023, Article ID 9921809, 18 pages, 2023.

[35] F. Ullah, M. R. Naeem, H. Naeem, X. Cheng, and M. Alazab, "Crolssim: cross-language software similarity detector using hybrid approach of lsa-based ast-mdrep features and cnn-lstm model," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5768–5795, 2022.

[36] A. Zhao, Y. Wang, and L. Jianbo, "Transferable self-supervised instance learning for sleep recognition," *IEEE Transactions on Multimedia*, vol. 25, p. 1, 05 2022.

[37] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *European Conference on Computer Vision*, Springer, Singapore, 2018.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, CVPR, Honolulu, Hawaii, July 2017.

[39] H. Jaeger, D. Trivedi, and M. Stadtschnitzer, "Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls," https://zenodo.org/records/2867216.

[40] G. Dimauro and F. Girardi, "Italian Parkinson's voice and speech," 2019, https://ieee-dataport.org/open-access/italian-parkinsons-voice-and-speech.

[41] Y. Yin, M. Wu, X. Wang, and X. Huang, "Speech recognition for power customer service based on dnn and cnn models," in *Digital TV and Wireless Multimedia Communications*, G. Zhai, J. Zhou, H. Yang, P. An, and X. Yang, Eds., pp. 453–468, Springer, Singapore, 2022.

[42] W. Si, C. Wan, and C. Zhang, "Towards an accurate radar waveform recognition algorithm based on dense CNN," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 1779–1792, 2021.

[43] S. Bose and A. Dey, "Rescnn: an alternative implementation of convolutional neural networks," in *Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering*, pp. 1–5, UPCON, Dehradun, India, November 2021.

[44] B. Mandal, A. Okeukwu, and Y. Theis, "Masked face recognition using resnet-50," 2021, https://arxiv.org/abs/2104.08997.

[45] A. Hajavi and A. Etemad, "Fine-grained early frequency attention for deep speaker recognition," in *Proceedings of the 2022 International Joint Conference on Neural Networks*, pp. 1–6, IJCNN, Padua, Italy, July 2022.

[46] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based bilstm models for personality recognition from user-generated content," *Information Sciences*, vol. 596, pp. 460–471, 2022.

[47] J. Deng, L. Cheng, and Z. Wang, "Self-attention-based bigru and capsule network for named entity recognition," 2020, https://arxiv.org/abs/2002.00735.

[48] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "Light-sernet: a lightweight fully convolutional neural network for speech emotion recognition," in *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6912–6916, IEEE, Singapore, May 2022.

[49] X.-C. Wen, J.-X. Ye, Y. Luo et al., "Ctl-mtnet: a novel capsnet and transfer learning-based mixed task net for the single-corpus and cross-corpus speech emotion recognition," 2022, https://arxiv.org/abs/2207.10644.

[50] J.-X. Ye, X.-C. Wen, X.-Z. Wang et al., "Gm-tcnet: gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition," *Speech Communication*, vol. 145, pp. 21–35, 2022.

[51] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition," in *Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, Rhodes Island, Greece, June 2023.