Parkinson's Detection Based On Combined CNN And LSTM Using Enhanced Speech Signals With Variational Mode Decomposition

Mehmet Bilal ER
Harran University
Esme ISIK
Malatya Turgut Özal University
Ibrahim ISIK (✉ ibrahim.isik@inonu.edu.tr)
İnönü University

Research Article

Keywords: Parkinson's disease, Long short-term memory, Variational Mode Decomposition

DOI: https://doi.org/10.21203/rs.3.rs-305818/v1

Parkinson's Detection Based on Combined CNN and LSTM Using Enhanced Speech Signals with Variational Mode Decomposition

Mehmet Bilal ER[1], Esme ISIK[2], Ibrahim ISIK[3*]

[1]Department of Computer Engineering, Harran University, 63050 Şanlıurfa, Turkey
[2]Department of Optician, Malatya Turgut Özal University, 44280, Malatya, Turkey
[3]Department of Electrical-Electronics Engineering, İnönü University, 44280, Malatya, Turkey

Abstract

The dysfunction of the cells in the brain that contain the substance known as dopamine, which enables brain cells to interact with each other, results in Parkinson's disease (PD). PD can cause many non-motor and motor symptoms such as speech and smell. One of the difficulties that Parkinson's patients can experience is a change in speech or speaking difficulties. Therefore, the right diagnosis in the early period is important in reducing the possible effects of speech disorders caused by the disease. Speech signal of Parkinson patients shows major differences compared to normal people. In this study, a new approach based on pre-trained deep networks and Long short-term memory (LSTM) by using mel-spectrograms obtained from denoised speech signals with Variational Mode Decomposition (VMD) for detecting PD from speech sounds is proposed. The proposed model consists of four stages. In the first step, the noise is removed by applying VMD to the signals. In the second stage, Mel-spectrograms are extracted from the enhanced sound signals with VMD. In the third stage, pre-trained deep networks are preferred to extract deep features from the Mel-spectrograms. For this purpose, ResNet-18, ResNet-50 and ResNet-101 models are used as pre-trained deep network architecture. In the last step, the classification process is occured by giving these features as input to the LSTM model, which is designed to define sequential information from the extracted features. Experiments are performed with the PC-GITA dataset, which consists of two classes and is widely used in the literature. The results obtained from the proposed method are compared with the latest methods in the literature, it is seen that it has a better performance in terms of classification performance.

Keywords: Parkinson's disease, Long short-term memory, Variational Mode Decomposition

*Corresponding authors: ibrahim.isik@inonu.edu.tr

1. Introduction

A progressive neurodegenerative condition caused by the early death of dopaminergic neurons in Parkinson's disease [1]. Parkinson's disease is the second most common neurological disorder after Alzheimer's [2], it is estimated to affect about 1% of the population over 60 years old [3]. The causes of falls in parkinson's disease are not fully known. The course of the disease is unstable and progresses at different rates. Parkinson's disease symptoms can be treated with a variety of medications [4]. Parkinson's disease symptoms are classified as non-motor symptoms and motor. Motion is associated with motor symptoms and they are more noticeable relative to non-motor symptoms. In motor symptoms, the parkinson's patient complains of slowness movement. Non-motor symptoms are apparent within a certain region; symptoms such as sleep disruptions, difficulty swallowing, chewing, and speaking apparent. The effect on the speech of Parkinson's disease is described as phonation, articulation and prosody [5]. Speech signals are also used as one of the main techniques for diagnosing Parkinson's disease [6]. In Parkinson's patients, physicians and speech pathologists have embraced subjective approaches dependent on auditory signs to identify various diseases [2]. In [6], in order to identify speech and pronunciation concerns, the authors did some experiments on speech signal recordings in three languages. It is noted that speech pronunciation is impaired by this disease, such as vowels, phrases, terms. Therefore, speech is an important method of diagnosis Parkinson's disease. In most traditional methods that support the automatic detection of Parkinson's-related vocal characteristics, subjects are asked to produce a fixed vowel in terms of both amplitude and frequency [7]. Various machine learning algorithms are implemented in common Parkinson's detection studies to learn the relationship between the extracted features and class labels. In most recent studies, disease detection has been carried out by using handwriting and speech datasets [8,9]. Support Vector Machines

(SVM) and acoustic properties were mostly considered for Diagnosis Parkinson's disease. In addition, it is observed that Gaussian models and Convolutional Neural Network (CNN) are used in PD classification in the literature.

The main contributions of this research are as follows:

- Speech signals denoised by applying VMD.
- A new system based on deep learning is proposed for the detection of PD by using speech signals.
- Mel-spectrograms that have very effective results in sound processing are used instead of spectrograms in the time frequency domain.
- The effectiveness of combined pre-trained deep networks and LSTM model in PD classification has been revealed.

The rest of this paper is organized as follows: In Section 2, the literature studies are analyzed and the differences between them are shown. The information about the material is given in Section 3 and the methodology is summarized in Section 4. In Section 5, the dataset used in the study and experimental applications for defining Parkinson's patients is given. In Section 6, the findings of the study are explained.

2. Related Works

In this section of the study, some important studies about the classification of Parkinson’s disease will be given. Many approaches have been proposed with the development of pattern recognition and artificial intelligence. In [10], a new technique is also proposed for stratification of subjects with Parkinson's disease. Data were collected from 31 people, consisting of 195 continuous vowel phonations. 23 of these people are Parkinson's patients and 8 are healthy people. Their methodology consists of three stages. These stages are pre-processing, feature extraction, and classification and feature selection. The linear kernel SVM is used for classification purposes. The accuracy value obtained from the proposed model is 91.4%. In [11], a data mining method known as weka was used to extract healthy subjects from Parkinson's patients. SVM which is preferred for classification purposes, is known as a supervised learning algorithm. Data pre-processing was performed on the dataset before the classification process. To achieve the best possible accuracy, different kernels were evaluated by using LibSVM. The linear kernel SVM produces the best accuracy of 65.21%, while the Radial Based Function (RBF) kernel achieves 60.86% accuracy. In [7], The authors proposed a new model. In the study, data were collected from 40 subjects (20 healthy, 20 Parkinson's). 26 speech samples were recorded from each subject, including short sentences, words, numbers, and continuous vowels. SVM and K-Nearest Neighbor (K-NN) were used for the classification process. The values 1, 3, 5 and 7 are preferred for the number of neighbors for K-NN, while linear and RBF kernel are constituted for SVM. The accuracy value of 82.50% with K-NN and the accuracy value of 85% when using SVM classifier was achieved. In [12], for Parkinson's disease diagnosis, different speech signal processing algorithms are compared. In the study, a new tool known as Tunable Q-factor wavelet transform (TQWT) was introduced. The classifiers were trained by using different classifiers in different feature subsets. It was observed that the Mel Frequency Kepsturm Coefficients (MFCC) and TQWT have reached the highest accuracy and therefore they are considered the important features in the classification problem of Parkinson's disease. In the study, the average accuracy of 86% was obtained with SVM. In [13], the authors proposed two frameworks based on CNN architectures for the classification of Parkinson's disease using vocal features. In both frameworks, various feature datasets were evaluated both separately and by combining both of them. In the first step, the feature datasets are set to 9-layer CNN as input, and in the second step, deep features are extracted simultaneously by passing the feature datasets through parallel layers that are directly connected to the convolution layers. According to experimental results, second framework appears to be very promising. In [14], a new artificial intelligence-based method has been developed to help early diagnosis of Parkinson's disease. Dysphonic measurements and clinical scores in 68 subjects were obtained by using the UCI Machine Learning database. Weights which are derived from the Multi-Layer Perceptron (MLP), were used for feature selection. This set of reduced features is then provided as a Lagrange Support Vector Machine (LSVM) input for classification process. This hybrid algorithm has been compared with other classifiers used in the study. Speech recordings were used for the detection of Parkinson's diseases in [15]. For sample and function collection, the two-dimensional technique of data selection is suggested. By using the chi-square statistical model, the proposed approach ranks the features and searches for the best subset of the specified features and selects samples recursively. The proposed method shows promissing results in terms of accuracy. In [16], a new method is proposed for the classification of vocal disorders with Hilbert-Huang Transform (HHT) and K-NN. The 12 features of each recorded voices were calculated with HHT. Also, a sample has been characterized by 9 different features based on the Linear Prediction Coefficients (LPCC) algorithm. Then, after each sample was expressed by 21 features, the classifier is established based on KNN. Additionally, the same experiments were performed by using the Random Forest classifier and Decision Tree to evaluate the performance of the K-NN classifier. The experimental results reveal that the classifier based on K-NN seems to be better than the other two classifiers with an accuracy rate of 93.3%. In [17], MR images of healthy and Parkinson's patients were classified by using deep neural networks. CNN architecture AlexNet has been used to improve diagnosis of Parkinson's disease. It was retrained with a pre-trained deep network with MR images and the classification process was performed. The accuracy value of 88.9% was achieved with the proposed system.

3. Background

3.1. Convolutional Neural Network

Technically, the CNN model includes the pooling, convolution, fully connected layers and classification layer [18]. The convolution is the first layer from which an input image derives features. The convolution is a filter which applied to the input image to extract a feature map from the input image. The height and weights of the filters are smaller than the input volume. The formula for the convolution process is assumed in Equation 1. The input image and kernel are denoted by f and h, respectively. The row and column indexes of the result matrix are represented by m and n, respectively.

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k] \tag{1}$$

The pooling layer is another architecture of pre-trained deep networks. The pooling layer is used to reduce the the dimensions of the input image after convolutional layer and to accelerate the calculations. Fully connected layers are an important constituent of deep networks that have proven to be very successful in identifing and classifying images. The completely linked input layer takes and flattens the output of the previous layers after turning them into a single vector that can be an input for the next level. The last layer is the output layer and probabilities are estimated for each label. Softmax is generally selected in this layer. Softmax formula is computed in Equation 2.

$$Softmax(x)_j = \frac{e^{xi}}{\sum_{n=1}^{N} e^{x_n}} \ , j = 1 \ldots . N \tag{2}$$

3.2. Long Short-Term Memory

Long Short-Term Memory networks are often referred to as "LSTM" and they are a special type of Recurrent Neural Networks (RNN). In its hidden layer, the LSTM neural network has a complex structure called LSTM cell. There are three gates in the LSTM cell shown in Figure 1, namely the input gate, the forgotten gate and the output gate that govern the information flow through the cell and neural network.

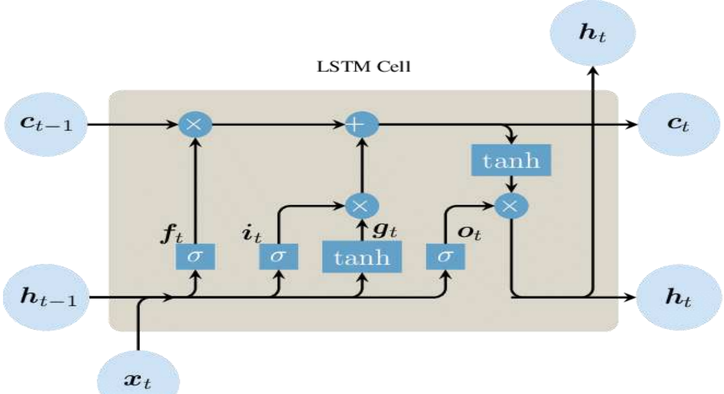


Figure 1. The architecture of LSTM

The LSTM architecture is designed in a chain like structure [19]. The first step of composing an LSTM network is to determine the information to be extracted from the cell as given in Equation 3. The sigmoid function decided to the data definition and exclusion process. Furthermore, the sigmoid feature decides which part of the output is to be extracted. The forget gate (or f_t) is a gate in which h_{t-1} is a vector ranging from 0 to 1 and C_{t-1} corresponds to each number in the cell state.

$$f_t = \sigma(W_F[h_{t-1}, X_t] + b_f) \tag{3}$$

Here, σ is denoted as the sigmoid function and (W_f) and (b_f) are refered the weight matrices and the bias vector of forget gate. The equations for the forward pass of an LSTM unit in Equations 4-6. In equation 4, it decides, stores the information (X_t) from the new input and also updates the cell state. This stage consists of two parts which are the sigmoid layer and the tanh layer. The sigmoid layer must first determine whether or not to update the new data (0 or 1). It was put the cell state via tanht in Equation 5 to force the values to be between -1 and 1. And it was multiplied to update the cell state by the output of the sigmoid gate. The new memory was added to the old one, C_{t-1}, into the new cell state C_t.

$$i_t = \sigma(W_i[h_{t-1}, X_t + b_i]), \tag{4}$$

$$N_t = tanh(W_n[h_{t-1}, X_t] + b_n), \tag{5}$$

$$C_t = C_{t-1}f_t + N_t i_t \tag{6}$$

Here, in the time between t-1 and t, C_{t-1} and C_t are denoted as cell states. In Equation 7, the sigmoid layer defines the portions of the cell state go to the output. (O_t) In the next step, the output of the sigmoid gate (O_t) is multiplied by the new values (C_t) produced by the tanh layer in Equation 8.

$$O_t = \sigma(W_0[h_{t-1}, X_t] + b_0]), \tag{7}$$

$$h_t = O_t tanh(C_t) \tag{8}$$

Here w_0 and b_0 are weight matrices and bias vector of the output gate, respectively.

4. Proposed Method

In this section of the study, the proposed methodology for Parkinson's disease and its main components are discussed in detail. The proposed method consists of four stage. In the first stage, the raw speech signals are preprocessed to obtaine the input of the proposed model and then VMD is applied. Mel-spectrograms are secondly extracted from the enhanced signals with VMD. Thirdly, the fc-1000 layer of ResNet models is used to extract outstanding and distinctive features from mel-spectrograms. Finally, CNN features are input to the deep bi-directional LSTM to attribute the temporal clues and recognize the sequential information in a set. Detailed explanation of each stage of the architecture is discussed in the next sections. The pseudo code of the proposed model is given in Algorithm 1. The schematic representation of the proposed architecture is shown in Figure 2.

Algorithm 1. Algorithm of the proposed method

```
Input: Parkinson and Normal speech signals, wav data 44.1 KHz
Output: y = [P ($y_{parkinson}$, P ($y_{Normal}$ ),] where P(α)>> Probability of each class
1:  Generate Variational mode decomposition paremeters k=5, α=120 ve tol=10^(-7)
2:  Apply preprocessing to speech data
3   $\min_{\{s_k\}\{w_k\}} \left\{\sum_{k=1}^{K} \| \partial_t \left[\left(\partial(t) + \frac{j}{\pi t}\right) * s_k(t)\right] e^{-jw_k} \|_2^2 \right\}$, Apply 5 mod VMD to all raw speech signals
4:  $L(\{u_k\}, \{w_k\}, \lambda) = \alpha \sum_k \| \partial_t \left[\left(\partial(t) + \frac{j}{\pi t}\right) * M_k(t)\right] e^{-jw_k} \|_2^2 + \|f(t) - \sum_k u_k(t)\| + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle$, Get denoised Speech Signal Using VMD
5:  Mel-spectrogram extraction from denoised speech signals
7:  Separate data for testing and training with the 10-fold verification method
8:  Use the fc-1000 layer of ResNet-18, ResNet-50, and ResNet-101 models to extract features from mel-spectrrograms.
9:  Get deep features vector from fc-1000 layers
10: Give the deep features to the LSTM model
```

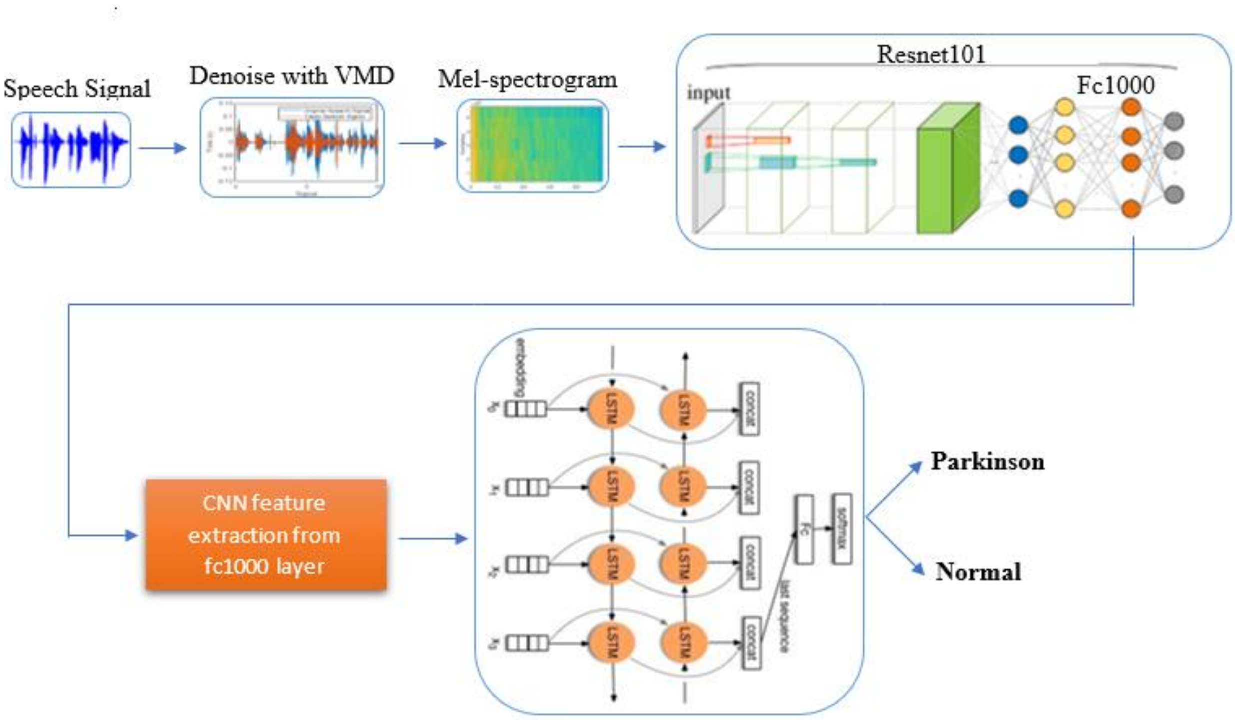


Figure 2. Proposed Model

4.1. Preprocessing

The speech signals reach a steady state in a certain period of time as are not static. Therefore, the speech signal is firstly framed and feature extraction process is performed in a short frame in time. The length of each frame is chosen as 25ms. The widely used Hamming window is used after frame process. 50% overbinning is set between the consecutive frames to achieve a smoothing transition between frames.

4.2. Denoising with Variational Mode Decomposition

The real signal is decomposed into finite number of sub-signals and modes by using the VMD method [20]. It is a adaptive signal decomposition technique, where each sub-signal compact around their respective center frequencies. The bandwidth of each modal is evaluated in three stage. In the first stage, Hilbert transform is used to obtain the frequency spectrum of each modal. In the second stage, it is shifted to the frequency spectrum of mode to the corresponding baseband. In the third stage, the bandwidth of the modal is estimated by the Gaussian smoothness of the demodulated signal [20]. The result can be formulated as a constrained variational problem.

$$\underset{\{s_k\}\{w_k\}}{min}\left\{\sum_{k=1}^{K}\|\,\partial_t\left[\left(\partial(t)+\frac{j}{\pi t}\right)*s_k(t)\right]e^{-jw_k}\|_2^2\right\} \tag{9}$$

$$\text{so that } \sum_k u_k = f \tag{10}$$

where, u_k denotes as kth decomposed mode, w_k indicates the center frequency of the kth mode signal. f (t) is the input signal and $\left[\left(\partial(t)+\frac{j}{\pi t}\right)*M_k(t)\right]$ is the Hilbert transform of $u_k(t)$. The exponential term e^{-jw_k} shifts the frequency spectrum of each mode to the center frequency. The constrained optimization problem in Equation 9 can be solved using augmented Lagrangian multipliers as follows;

$$L(\{u_k\},\ \{w_k\},\lambda)=\alpha\sum_k\|\partial_t\left[\left(\partial(t)+\frac{j}{\pi t}\right)*M_k(t)\right]e^{-jw_k}\ \|_2^2$$
$$+\|f(t)-\sum_k u_k(t)\|+\langle\lambda(t),f(t)-\sum_k u_k(t)\rangle \tag{11}$$

It is indispensable to define the initial parameters for the decomposition. These parameters are the decomposition number w_k, the balancing parameter of the data compliance constriction is α and the convergence criteria tolerance is tol. It is picked as $w_k = 5$, $\alpha = 120$ and $tol = 10^{-7}$. The above parameters are the constant used in many studies for efficient decomposition of the speech signal [21,22]. The pseudo code of VMD is given in algorithm 2. 5 modes decomposed speech signal is given in Figure 3. In addition, the original speech signal and the denoised signal by applying VMD are given in Figure 4.

Algorithm 2. Algorithm of VMD

```
Input: raw_signal=Raw Parkinson and Normal Speech Signals, wav data
Output:  denoised_signal= wav data
[x, residual] = vmd(raw_signal,'NumIMF',6);
for n = 1:5
    ax(n) = nexttile(t);
    plot(tm,imf(:,n)')
    xlim([tm(1) tm(end)])
end
denoised_signal = sum(imf(:,2:4),2);
plot(denoised_signal, raw_signal)
```

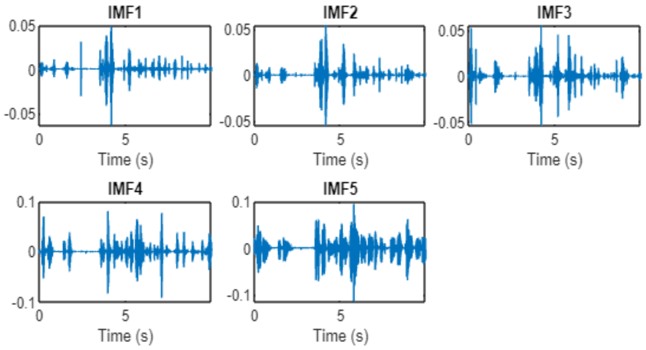


Figure 3. 5 mode VMD of Speech Signal

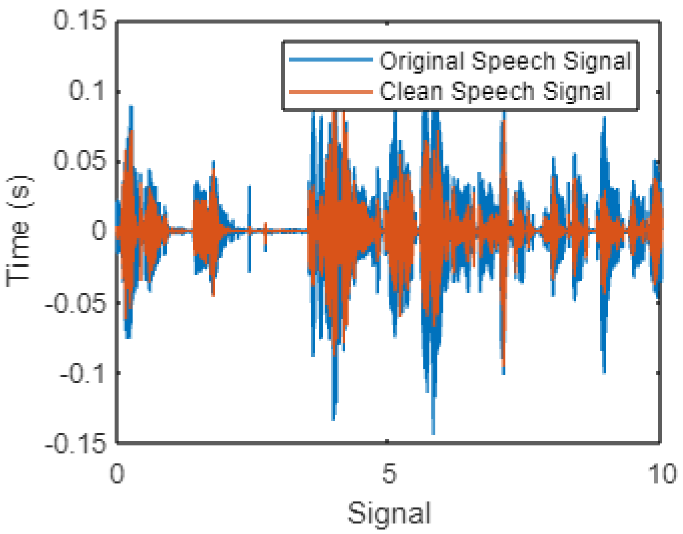


Figure 4. Original and denoised signal

4.3. Mel-Spectrogram Extraction

MFCC is widely used for voice recognition. Mel is the scale of the frequency of a sound tone picked by the human ear. As a result of the studies, it has been observed that the scales are linear up to 1 kHz, and a logarithmic increment in higher values. The relationship between Mel spectrum (M) and frequency (Hz) is shown in Equations 12-13.

$$f_{mel} = 2595log_{10}(1+\frac{f}{700}) \quad (12)$$

$$f = 700(10^{\frac{m}{2595}} - 1) \quad (13)$$

Discrete Fourier transform (DFT) has been applied to transform the speech signal from time-domain to frequency domain. The formula for the transformation of the speech sound signal x(n) to the frequency domain by using DFT is given in Equation 14.

$$X_k = \sum_{k=0}^{N-1} x_n \, e^{-2\pi jkn\,/\,N}, \qquad \mathrm{n} = 0, 1, 2, \dots\dots, \mathrm{N} - 1 \tag{14}$$

The next stage is the calculation of the power of spectrum. The power of spectrum P(k) is obtained as in Equation 15.

$$P(k) = \frac{1}{N} |X(k)|^2 \tag{15}$$

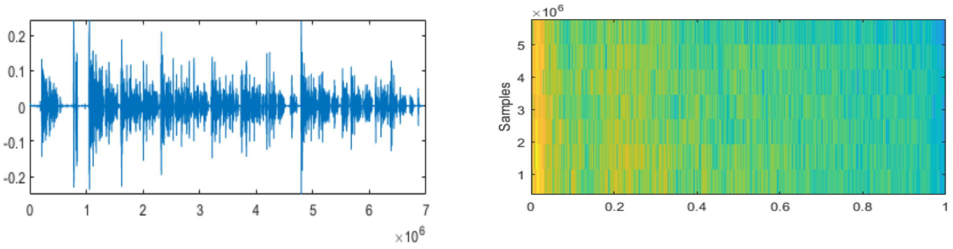
The power of spectrum P(k) is passed through a series of Mel scale triangular filter windows to obtain the Mel spectrum. The frequency of the triangle filter, $H_m(k)$ is calculated as follows:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{16}$$

f(m) is the center frequency of the Mel triangle filter. The logarithm energy spectrum in each frame is S(m) which is obtained using a logarithmic process.

$$S(m) = ln[\textstyle\sum_{k=0}^{N-1} P(k) H_m\,(k)], 0 \le m \le M \tag{17}$$

P_m is the power spectrum, H_m(k) is filter window and M is the number of the filter windows. The samples of the speech sound signal and the mel spectrogram are given in Figure 5.



(a) (b)

Figure 5. Ilustration of Speech Sound Signal and Mel spectrogram **(a)** Speech Sound Signal, **(b)** Mel spectrogram

4.4. Combined ResNet-LSTM network

In this research, a method based on CNN and LSTM model has been developed for the detection of Parkinson's disease using mel spectrogram images of speech signals. In the structure of this architecture, ResNet and LSTM combined models are used together. ResNet models are used to extract feature from the mel spectrogram images of VMD denoised speech signals in this study. ResNet models are preferred to observe the effect of network depth on performance. There are many variants of ResNet and the widely used ResNets18, ResNet50 and ResNet101 models were prefered in this study. ResNet ranked first in the ILSVRC competition held in 2015 with the lowest error rate with 3.37% [23]. ResNet18 has 73 sub-layers, ResNet50 177 sub-layers, and ResNet101 347 sub-layers. Also, all of these networks have fc-1000 (1000 neurons) and this layer is used as feature extraction layer. Detailed features of the layers in ResNet-18, ResNet-50 and ResNet-101 models are given in Table 1. The features obtained from the fc-1000 layers of ResNet-18, ResNet-50 and ResNet-101 models were used as input to the designed LSTM. Softmax in the output layer of the designed two-layer LSTM model with 32 and 64 outputs were used for Parkinson diagnosis. The designed hybrid ResNet-LSTM network is shown in Figure 6.

Table 1. The general architecture of ResNet Models

Layer name	Output Size	ResNet-18	ResNet-50	ResNet-101
Conv1	112×112	7 × 7,64, stride 2	7 × 7, 64, stride 2	7 × 7, 64, stride 2
Conv2_x	56×56	3 × 3 max pool, stride 2	3 × 3 max pool, stride 2	3 × 3 max pool, stride 2
		$\begin{bmatrix} 3x3, 64 \\ 3x3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1x1, 64 \\ 3x3, 64 \\ 1x1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1x1, 64 \\ 3x3, 64 \\ 1x1, 256 \end{bmatrix} \times 3$
Conv3_x	28×28	$\begin{bmatrix} 3x3, 128 \\ 3x3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1x1, 128 \\ 3x3, 128 \\ 1x1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1x1, 128 \\ 3x3, 128 \\ 1x1, 512 \end{bmatrix} \times 4$
Conv4_x	14×14	$\begin{bmatrix} 3x3, 256 \\ 3x3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1x1, 256 \\ 3x3, 256 \\ 1x1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1x1, 256 \\ 3x3, 256 \\ 1x1, 1024 \end{bmatrix} \times 23$
Conv5_x	7×7	$\begin{bmatrix} 3x3, 512 \\ 3x3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1x1, 512 \\ 3x3, 512 \\ 1x1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1x1, 512 \\ 3x3, 512 \\ 1x1, 2048 \end{bmatrix} \times 3$
Output	1×1	average pool, fc1000, softmax	average pool, fc1000, softmax	average pool, fc1000, softmax

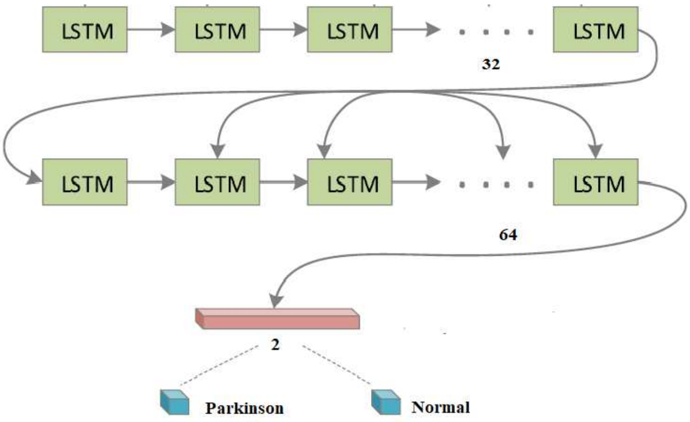


Figure 6. Combined ResNet-LSTM Model

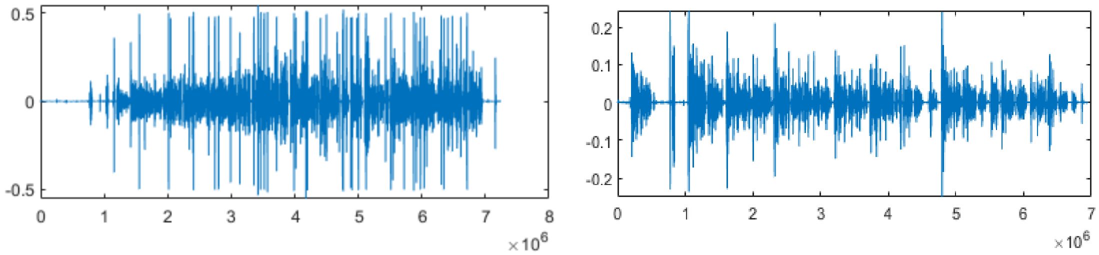
5. Experimental Applications

5.1. Dataset

PC-GITA Spanish dataset is used in this research [24]. The dataset contains speech records of 50 Parkinson's disease patients and 50 healthy people speaking Spanish. In addition, the records were taken from 25 men and 25 women. The dataset details are shown in Table 2. The dataset consists of recording monologues, vowels and reading the text. Each record consists of various sound characteristics discussed below. The sound signals of healthy and Parkinson's patients are given in Figure 7.

Table 2. The Age and gender dispersion of Parkinson's disease and healthy people in the dataset.

Data		Parkinson	Normal	Gender	Age (Male/Female)
Vowels	A	150	150		Parkinson
	E	150	150	75 M	62.2±11.2 M
					60.1±7.8 F
	I	150	150	75 F	
	O	150	150		
	U	150	150		Normal
Words	Apto	50	50	25 M	61.2±11.3 M
				25 F	60.7±7.7 F
	Atelta	50	50	25 M	
				25 F	
Monologues		50	50	25 M	
Words		50	50	25 F	



(a) (b)

Figure 7. Ilustration of normal and parkinson speech signals **(a)** Normal speech **(b)** Parkinson speech

5.2. Evaluation Metrics

The results obtained from the proposed method were calculated using the evaluation criteria explained below. Accuracy is measured by the number of predictions that are accurate, divided by the total number of predictions. The accuracy evaluation is given in Equation 18.

$$Accuracy = \frac{|TP|+|TN|}{|YN|+|YP|+|TN|+|DP|} \tag{18}$$

Precision (P) is a evaluation that predicts the probability that a positive forecast is correct. The precision evaluation is given in Equation 19.

$$Precision(P) = \frac{|TP|}{|TP|+|FP|} \tag{19}$$

The F-score is a harmonious mean of positive prediction ratio and sensitivity criteria and is calculated as shown in Equation 20.

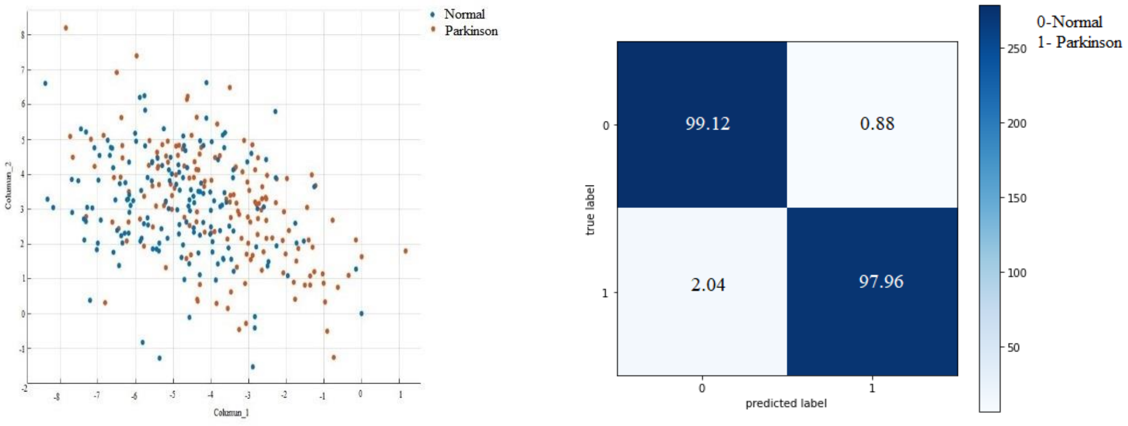
$$F - Score = \frac{2*|TP|}{2*|TP|+|FP|+|FN|} \tag{20}$$

5.3. Experimental Results

In this section, the results obtained from the proposed approach and the performance comparison explains in detail. After extracting the mel-spectrogram images using the VMD sound recordings, these images were given as input to the Resnet models. The high-level detected features of ResNet models are extracted by using the fc 1000 layer. In order to define hidden gates from deep learning networks and to detect Parkinson's disease, the features were given as input to LSTM. Finally, the classification was occured using Softmax. Data were decomposed for training and testing using a 10-fold cross validation technique. The different parameters were used for the training process in order to ensure that the model becomes sufficient and optimal. The experiments were realized with different batch sizes and different learning rates for the most ideal solution. The batch sizes are set to 32, 64, and 128, respectively. The learning rate was set as 0.01, 0.001 and 0.0001, respectively. The number of epochs is defined as 20. The results obtained with various parameters are given in Table 3-5. The classification results obtained from ResNet18 + LSTM is illustrated in Table 3. With a batch size of 32 and learning rate of 0.1, the highest accuracy rate is obtained as 95.31%, as the batch size choosen to 64 and learning rate to 0.001, the highest accuracy is measured 94.91%, the batch size is set to 128 and learning rate to 0.001, the highest accuracy rate is obtained 95.47% without appling VMD. After VMD is applied, the highest accuracy rate is measured 98.30% with the batch sizes of 32 and the learning rate of 0.01, the batch size is set to 64 and learning rate to 0.01, the highest accuracy rate is acquired 98.54%, the highest accuracy rate is obtained as 98.37% it was set the batch sizes 128 and learning rate 0.01. The features scatter obtained from the ResNet18 + LSTM model is given in Figure 8 (a), and the confusion matrix belonging to the best classification result obtained from this model is given in Figure 8 (b).

Table 3. The classification results obtained from ResNet-18 + LSTM.

Model	Batch-Size	Learning rate	Accuracy %	Precision %	Sensitivity%
ResNet-18&LSTM (without VMD)	32	0.1	95.31	94.89	95.62
		0.01	94.08	94.19	95.82
		0.001	94.88	94.80	94.96
	64	0.1	94.18	95.28	95.52
		0.01	94.06	94.35	94.65
		0.001	94.91	95.12	94.39
	128	0.1	95.02	95.36	95.28
		0.01	95.06	95.53	94.28
		0.001	95.47	94.51	95.24
ResNet-18&LSTM (with VMD)	32	0.1	96.12	98.14	96.74
		0.01	98.30	96.97	96.86
		0.001	97.97	97.81	97.81
	64	0.1	98.45	97.30	95.53
		0.01	98.54	96.84	95.91
		0.001	98.24	96.42	97.32
	128	0.1	95.85	95.68	98.42
		0.01	98.37	96.88	96.71
		0.001	96.06	98.40	95.65



(a)

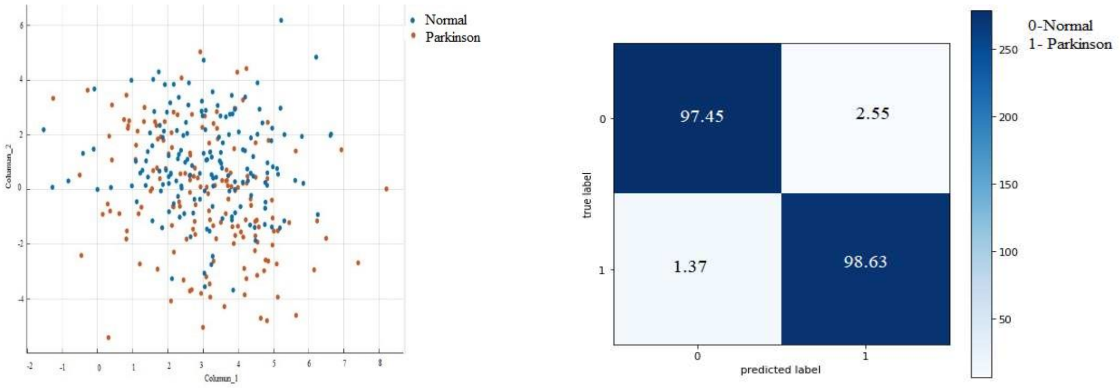
(b)

Figure 8. Ilustration of feature scatter and confusion matrix for Resnet18+LSTM **(a)** Feature scatter plot, (b) Confusion matrix

The classification results from ResNet-50+LSTM is given in Table 4. The batch sizes, and the learning rate is set to 32 and 0.001, respectively, the highest accuracy rate is measured as 95.68. The batch size is set to 64 and learning rate to 0.01, the highest the accuracy rate is obtained as 95.81, the highest accuracy rate was acquired as 95.17 with the learning of 0.01 and batch sizes of 128, without appling VMD. After applying VMD, the highest accuracy rate was 98.04% when the Batch sizes were set to 32 and learning rate to 0.01, the highest accuracy rate was 97.84% when the batch sizes were set to 64 and learning rate 0.001, and the highest accuracy rate was measured as 97.68% with batch sizes of 128 and learning rate of 0.01. The features obtained from the ResNet-50+LSTM model is given in Figure 9 (a), and the confusion matrix of the best classification result obtained from this model is given in Figure 9 (b).

Table 3. The classification results obtained from ResNet50 + LSTM.

Model	Batch-Size	Learning rate	Accuracy %	Precision %	Sensitivity%
ResNet-50&LSTM (without VMD)	32	0.1	94.44	94.64	95.26
		0.01	94.26	95.96	95.77
		0.001	95.68	95.98	94.30
	64	0.1	95.78	95.00	94.34
		0.01	95.81	95.72	95.88
		0.001	94.13	95.70	94.28
	128	0.1	94.15	94.42	95.39
		0.01	95.17	94.76	94.77
		0.001	94.34	95.45	94.99
ResNet-50&LSTM (with VMD)	32	0.1	95.96	96.70	96.21
		0.01	98.04	96.05	96.56
		0.001	95.34	97.94	98.32
	64	0.1	96.10	95.79	97.18
		0.01	95.64	95.92	97.92
		0.001	97.84	97.27	97.03
	128	0.1	97.59	97.80	96.19
		0.01	97.69	97.33	97.28
		0.001	97.25	96.51	95.75



(a)

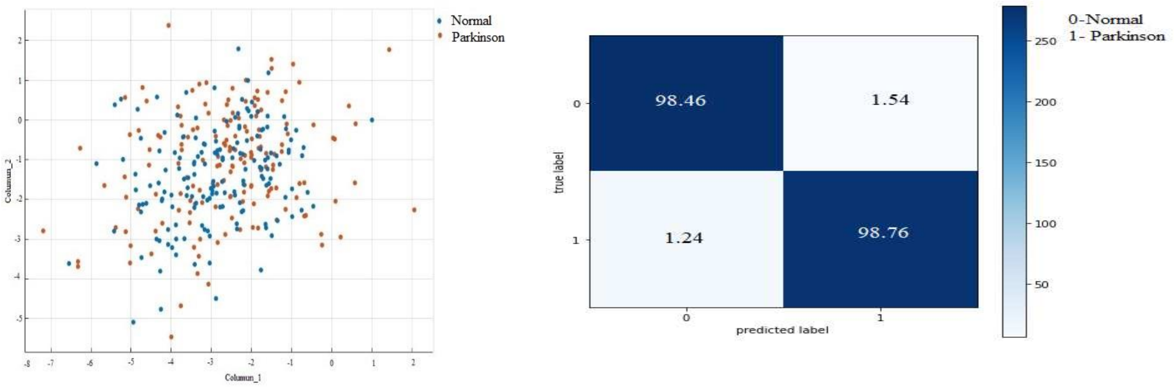
(b)

Figure 9. Ilustration of feature scatter and confusion matrix for Resnet-50+LSTM **(a)** Feature scatter plot, (b) Confusion matrix

The classification results obtained from ResNet101 + LSTM is given in Table 5. Without applying VMD, when the batch size of 32 and learning rate of 0.01 the highest accuracy rate is 95.94%, when the batch size of 64 and the learning rate of 0.001 the highest accuracy rate is measured as 94.75%, when the batch sizes 128 and learning rate of 0.01 the highest accuracy rate is obtained as 95.62%. After VMD is applied, when the batch size is set to 32 and learning rate to 0.001, the highest accuracy rate is measured as 97.21%, when the batch size of 64 and the learning rate of 0.001 the highest accuracy rate is acquired as 97.48%, when the batch sizes of 128 and the learning rate of 0.001 the highest accuracy rate is obtained as 98.61%. The distribution of the features obtained from the ResNet-101 + LSTM model is given in Figure 10 (a), and Figure 10 (b) gives the uncertainty matrix with the best classification outcome obtained from this model.

Table 5. The classification results obtained from ResNet101 + LSTM.

Model	Batch-Size	Learning rate	Accuracy %	Precision %	Sensitivity%
ResNet-101&LSTM (without VMD)	32	0.1	95.48	94.61	95.90
		0.01	95.94	95.28	94.30
		0.001	95.12	94.52	95.92
	64	0.1	94.39	94.88	94.74
		0.01	94.60	95.66	94.52
		0.001	94.75	95.54	94.95
	128	0.1	94.29	94.53	94.94
		0.01	95.62	95.93	94.98
		0.001	94.13	95.42	95.60
ResNet-101&LSTM (with VMD)	32	0.1	97.20	97.80	95.78
		0.01	96.41	96.77	97.88
		0.001	97.21	97.67	97.65
	64	0.1	96.33	96.97	95.81
		0.01	96.19	98.34	96.03
		0.001	97.48	95.51	97.42
	128	0.1	98.52	96.45	96.85
		0.01	97.08	96.54	96.52
		0.001	98.61	96.11	95.93



(a) (b)

Figure 10. Ilustration of feature scatter and confusion matrix for Resnet-101+LSTM **(a)** Feature scatter plot, (b) Confusion matrix

In this study, the results are compared with the other methods used in the literature in order to determine the best performance of the proposed process. Some important studies that focused on the detection Parkinson's from speech signals have been summarized in Table 6.

Table 6. Performance comparison with other approaches

Author	Technique	Data Sample	Accuracy%
Vásquez et al. [25]	Bayesian ridge regression (BRR)	Pc-Gita (Wovel, words, sentences)	94.90
Arias et al. [26]	CNN, SVM	Pc-Gita (monologues)	84
Rueda et al. [27]	Random Forest, SVM	Pc-Gita (Wovel, words)	70
Zahid et al. [28]	Deep Features with Random Forest	Pc-Gita (Wovel, words, sentences, read text, monologues)	98.30
Our Approach	ResNet-101&LSTM (With VMD)	Pc-Gita (Wovel, words, sentences, read text, monologues)	98.61

Vásquez *et al.* [25] developed a methodology based on voice processing and machine learning methods to measure the dysarthria level of Parkinson's patients. The features extracted from various dimensions of speech such as phonation, articulation, and prosody was modeled. Dysarthria level was measured using linear and nonlinear regression models with the accuracy 94.90%. Arias *et al.* [26] evaluated the phonation, articulation, and prosody data. SVM and CNN are used for classification and extraction of optimal features. The findings of the study indicated that prosody features were more effective than others and an accuracy of 84% was obtained. Rueda *et al.* [27] focused on electing features that best represent the pathophysiology of dysarthria caused by Parkinson's disease. The number of features was reduced to 15 by applying two-stage feature selection to the features extracted from sound recordings. The accuracy of 70% for classification has been achieved by using SVM and Random Forest. Zahid et al. [28] proposed three methods for the detection of Parkinson's disease. In the first method, an approach based on transfer learning using spectrograms of speech recordings was preferred. Using machine learning classifiers, deep features obtained from speech spectrograms were evaluated in the second technique. Simple acoustic properties of the recordings have also been evaluated using machine learning classifiers in the third technique. It was achieved 98.30% accuracy with deep features and Random Forest classifier.

6. Conclusion

People all over the world suffer from Parkinson's disease, which is one of the most common diseases. Early disease diagnosis is an open research topic, and considerable work has been done by several researchers to achieve the highest precision in detection and diagnosis. Computational models are effective in diagnosing medical disease and data are the primary factor. In this experimental study, a method based on deep networks is proposed for the detection of Parkinson's disease from sound recordings. In order to extract features from the mel-spectrogram images of the VMD applied voice signals, the fc-100 layer of the ResNet models is preferred. The designed LSTM network is used to recognize sequential information from the features obtained from ResNet models. Softmax is placed in the last layer of the designed architecture for classification. In the proposed method, it is tested on the PC-GITA Spanish dataset consisting of two classes. Data are decomposed for testing and training by using the 10-fold cross validation technique. The highest classification performance is obtained from the ResNet-101 + LSTM model with VMD as 98.61%. The comparison of the proposed model with the state of the art of models demonstrates our model's efficacy in PD detection.

Declarations

Conflict of interest The authors declare that they have conflict of interest.

Funding The authors did not receive support from any organization for the submitted work.

Availability of data and material The dataset is not publicly available.

Human and animal rights statement No involvement of humans or animals.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

Authorship contributions

Bilal Er: Conceptualization, Software Esme Isik.: Data curation, Writing- Original draft preparation. Ibrahim Isik: Methodology, Visualization.

References

[1] Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, et al. Parkinson disease. Nat Rev Dis Prim 2017;3. https://doi.org/10.1038/nrdp.2017.13.

[2] Benba A, Jilbab A, Hammouch A. Detecting Patients with Parkinson's disease using Mel Frequency Cepstral Coefficients and Support Vector Machines. Int J Electr Eng Informatics 2015;Volume 7:297–307. https://doi.org/10.15676/ijeei.2015.7.2.10.

[3] Reeve A, Simcox E, Turnbull D. Ageing and Parkinson's disease: why is advancing age the biggest risk factor? Ageing Res Rev 2014;14:19–30. https://doi.org/10.1016/j.arr.2014.01.004.

[4] Arena JE, Stoessl AJ. Optimizing diagnosis in Parkinson's disease: Radionuclide imaging. Parkinsonism Relat Disord 2016;22:S47–51. https://doi.org/10.1016/j.parkreldis.2015.09.029.

[5] Parra-Gallego LF, Arias-Vergara T, Vásquez-Correa JC, Garcia-Ospina N, Orozco-Arroyave JR, Nöth E. Automatic Intelligibility Assessment of Parkinson's Disease with Diadochokinetic Exercises. Commun Comput Inf Sci 2018:223–30. https://doi.org/10.1007/978-3-030-00353-1_20.

[6] Hosseini-Kivanani N, Vásquez-Correa JC, Stede M, Nöth E. Automated Cross-language Intelligibility Analysis of Parkinson's Disease Patients Using Speech Recognition Technologies. Proc 57th Annu Meet Assoc Comput Linguist Student Res Work 2019. https://doi.org/10.18653/v1/p19-2010.

[7] Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. IEEE J Biomed Heal Informatics 2013;17:828–34. https://doi.org/10.1109/jbhi.2013.2245674.

[8] Rios-Urrego CD, Vásquez-Correa JC, Vargas-Bonilla JF, Nöth E, Lopera F, Orozco-Arroyave JR. Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and non-linear features. Comput Methods Programs Biomed 2019;173:43–52. https://doi.org/10.1016/j.cmpb.2019.03.005.

[9] Trinh NH, O'Brien D. Pathological Speech Classification Using a Convolutional Neural Network. Proc. IMVIP, Irel., 2019.

[10] Little M, McSharry P, Hunter E, Spielman J, Ramig L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nat Preced 2008. https://doi.org/10.1038/npre.2008.2298.1.

[11] Bhattacharya I, Bhatia MPS. SVM classification to distinguish Parkinson disease patients. Proc 1st Amrita ACM-W Celebr Women Comput India - A2CWiC '10 2010. https://doi.org/10.1145/1858378.1858392.

[12] Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, et al. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput 2019;74:255–63. https://doi.org/10.1016/j.asoc.2018.10.022.

[13] Gunduz H. Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. IEEE Access 2019;7:115540–51. https://doi.org/10.1109/access.2019.2936564.

[14] Parisi L, RaviChandran N, Manaog ML. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst Appl 2018;110:182–90. https://doi.org/10.1016/j.eswa.2018.06.003.

[15] Ali L, Zhu C, Zhou M, Liu Y. Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. Expert Syst Appl 2019;137:22–8. https://doi.org/10.1016/j.eswa.2019.06.052.

[16] Chen L, Wang C, Chen J, Xiang Z, Hu X. Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbor (KNN). J Voice 2020. https://doi.org/10.1016/j.jvoice.2020.03.009.

[17] Sivaranjini S, Sujatha CM. Deep learning based diagnosis of Parkinson's disease using convolutional neural network. Multimed Tools Appl 2019;79:15467–79. https://doi.org/10.1007/s11042-019-7469-8.

[18] Cireundefinedan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J. Flexible, High Performance Convolutional Neural Networks for Image Classification. Proc. Twenty-Second Int. Jt. Conf. Artif. Intell. - Vol. Vol. Two, AAAI Press; 2011, p. 1237–1242.

[19] Olah C. Understanding LSTM Networks n.d. http://colah.github.io/posts/2015-08-Understanding-

LSTMs/ (accessed August 21, 2020).

[20] Dragomiretskiy K, Zosso D. Variational Mode Decomposition. IEEE Trans Signal Process 2014;62:531–44. https://doi.org/10.1109/TSP.2013.2288675.

[21] Karan B, Mahto K, Sahu SS. Detection of Parkinson Disease Using Variational Mode Decomposition of Speech Signal. 2018 Int. Conf. Commun. Signal Process., 2018, p. 508–12. https://doi.org/10.1109/ICCSP.2018.8524445.

[22] Deb S, Dandapat S, Krajewski J. Analysis and Classification of Cold Speech Using Variational Mode Decomposition. IEEE Trans Affect Comput 2020;11:296–307. https://doi.org/10.1109/TAFFC.2017.2761750.

[23] Mittal S, Agarwal S, Nigam MJ. Real Time Multiple Face Recognition: A Deep Learning Approach. Proc. 2018 Int. Conf. Digit. Med. Image Process., New York, NY, USA: Association for Computing Machinery; 2018, p. 70–76. https://doi.org/10.1145/3299852.3299853.

[24] Orozco JR, Arias-Londoño JD, Vargas-Bonilla J, González-Rátiva M, Noeth E. New Spanish speech corpus database for the analysis of people suffering from Parkinsons disease. Proc 9th Lang Resour Eval Conf 2014:342–7.

[25] Vásquez-Correa JC, Orozco-Arroyave JR, Bocklet T, Nöth E. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. J Commun Disord 2018;76:21–36. https://doi.org/https://doi.org/10.1016/j.jcomdis.2018.08.002.

[26] Arias-Vergara T, Vasquez-Correa JC, Orozco-Arroyave JR, Klumpp P, Nöth E. Unobtrusive Monitoring of Speech Impairments of Parkinson'S Disease Patients Through Mobile Devices. 2018 IEEE Int. Conf. Acoust. Speech Signal Process., 2018, p. 6004–8. https://doi.org/10.1109/ICASSP.2018.8462332.

[27] Rueda A, Vásquez-Correa JC, Rios-Urrego CD, Orozco-Arroyave JR, Krishnan S, Nöth E. Feature Representation of Pathophysiology of Parkinsonian Dysarthria. Interspeech 2019 2019. https://doi.org/10.21437/interspeech.2019-2490.

[28] Zahid L, Maqsood M, Durrani MY, Bakhtyar M, Baber J, Jamal H, et al. A Spectrogram-Based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson's Disease. IEEE Access 2020;8:35482–95. https://doi.org/10.1109/ACCESS.2020.2974008.

Figures

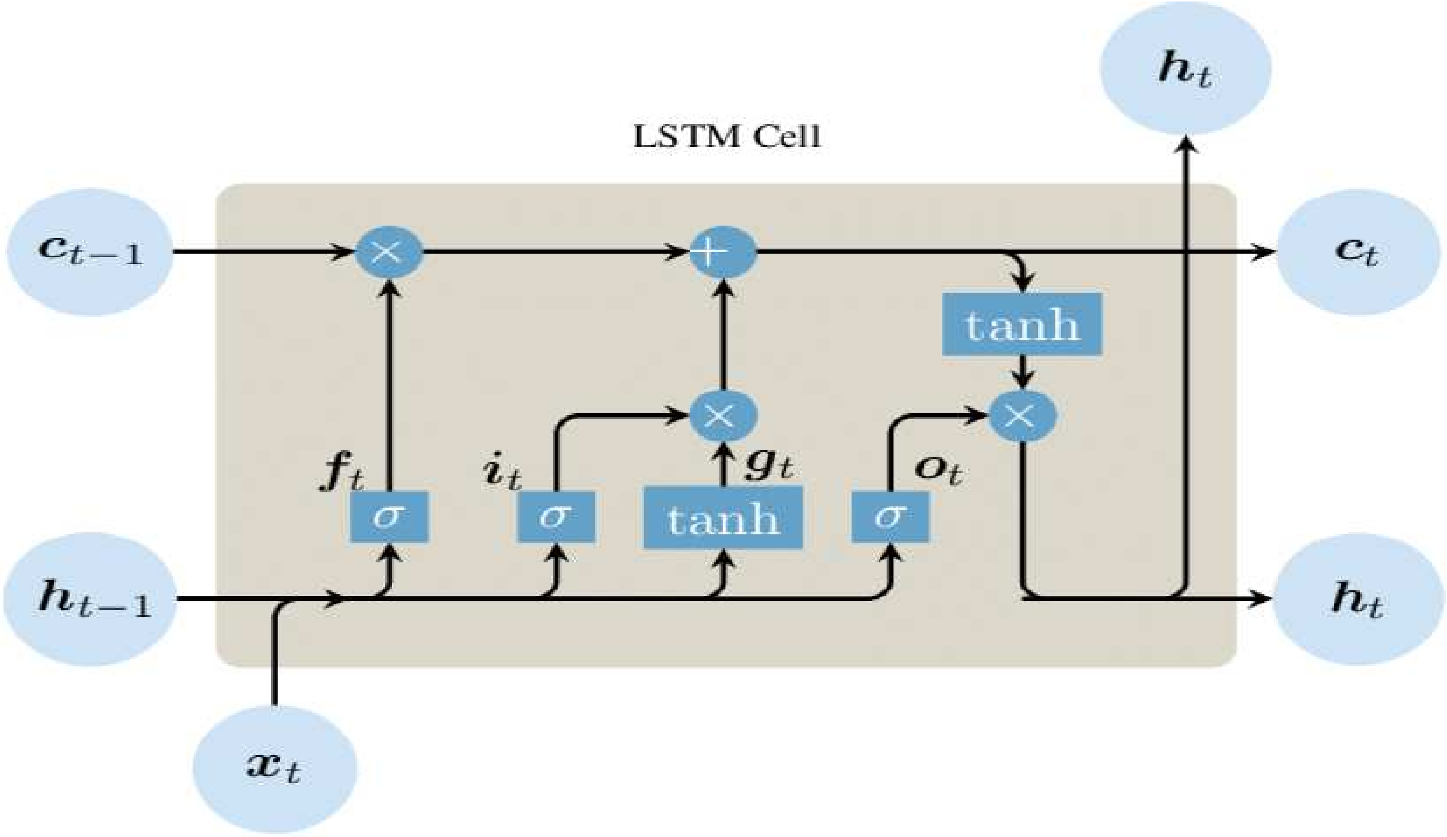


Figure 1

The architecture of LSTM

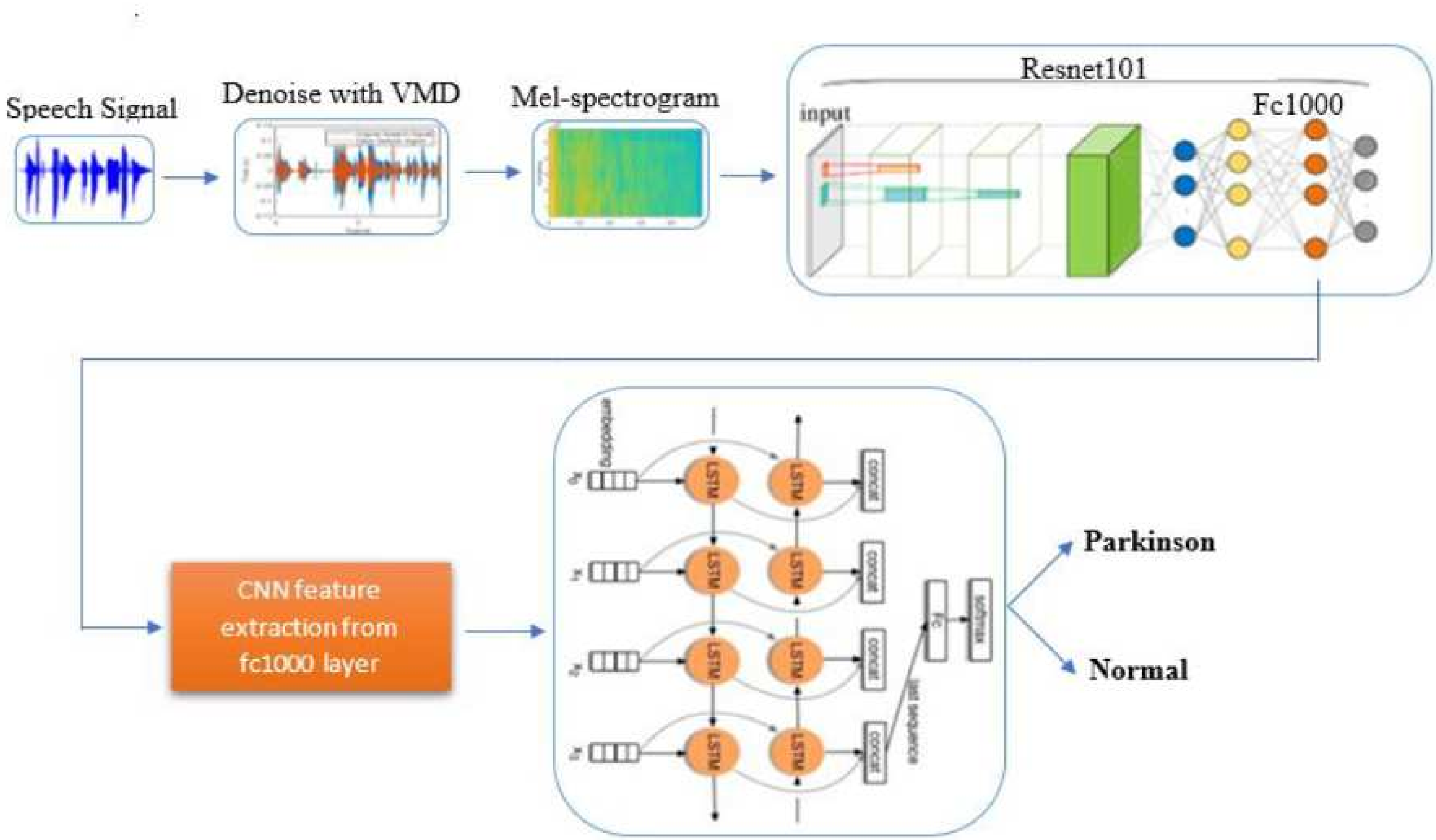


Figure 2

Proposed Model

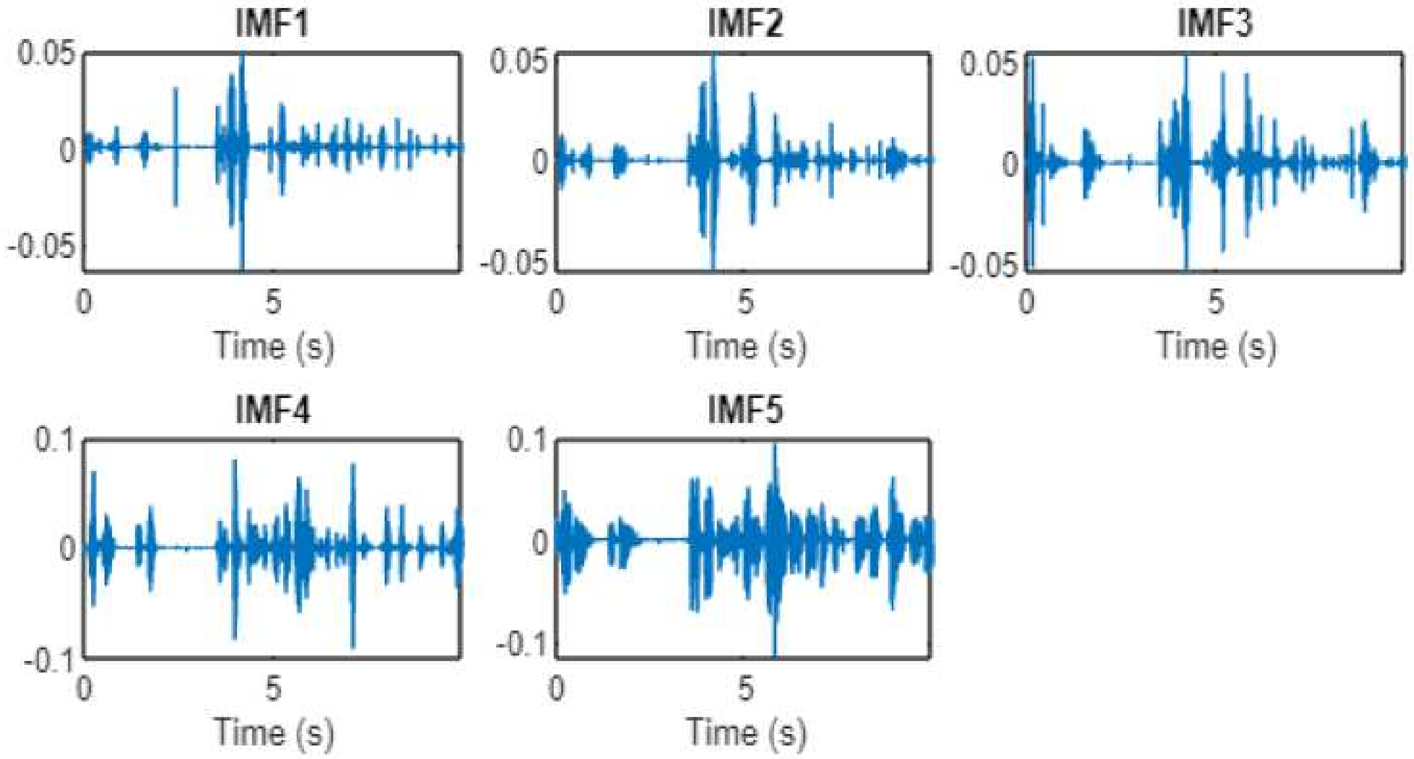


Figure 3

5 mode VMD of Speech Signal

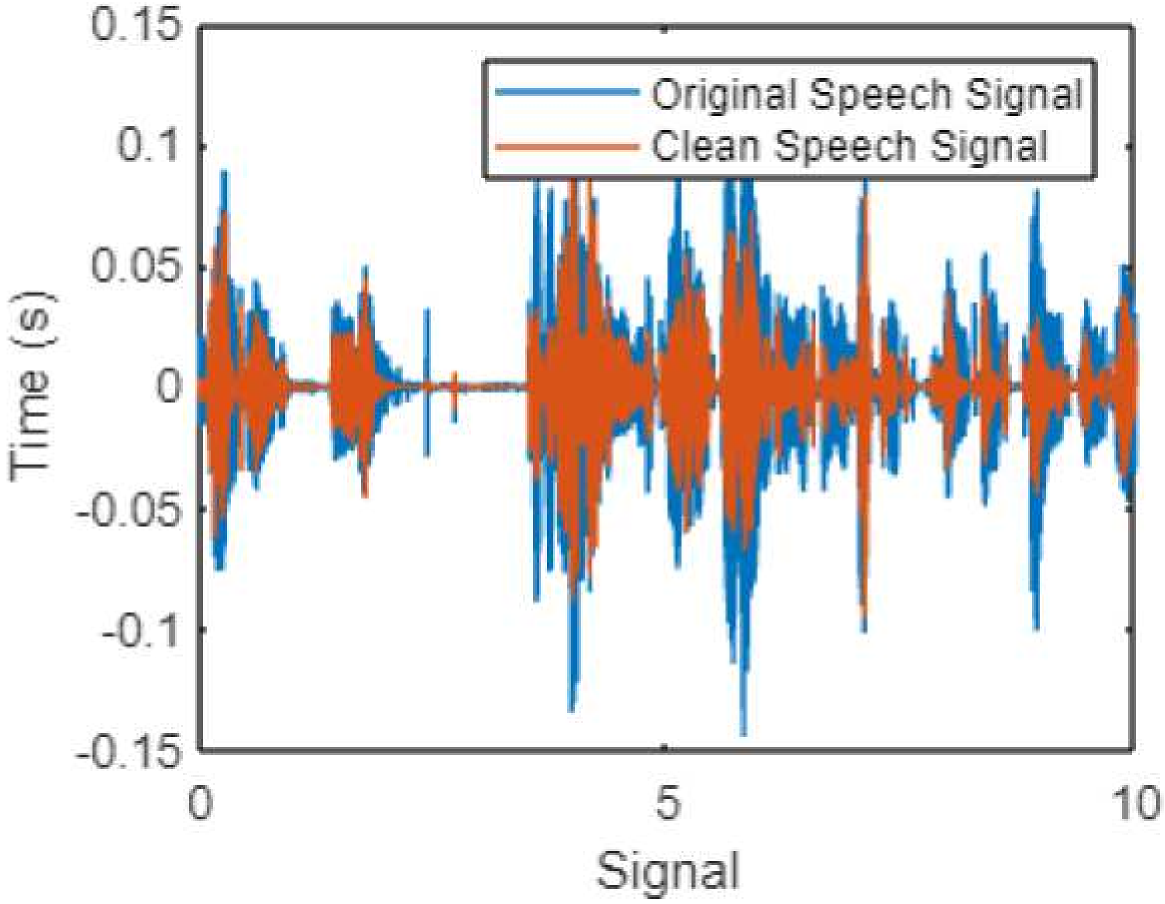


Figure 4

Original and denoised signal

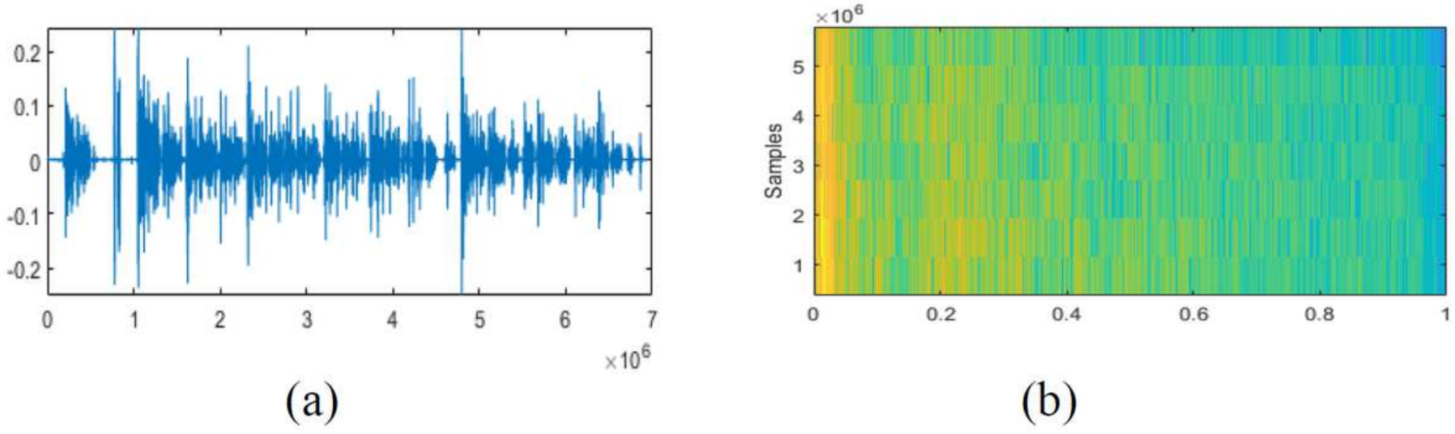


Figure 5

Ilustration of Speech Sound Signal and Mel spectrogram (a) Speech Sound Signal, (b) Mel spectrogram

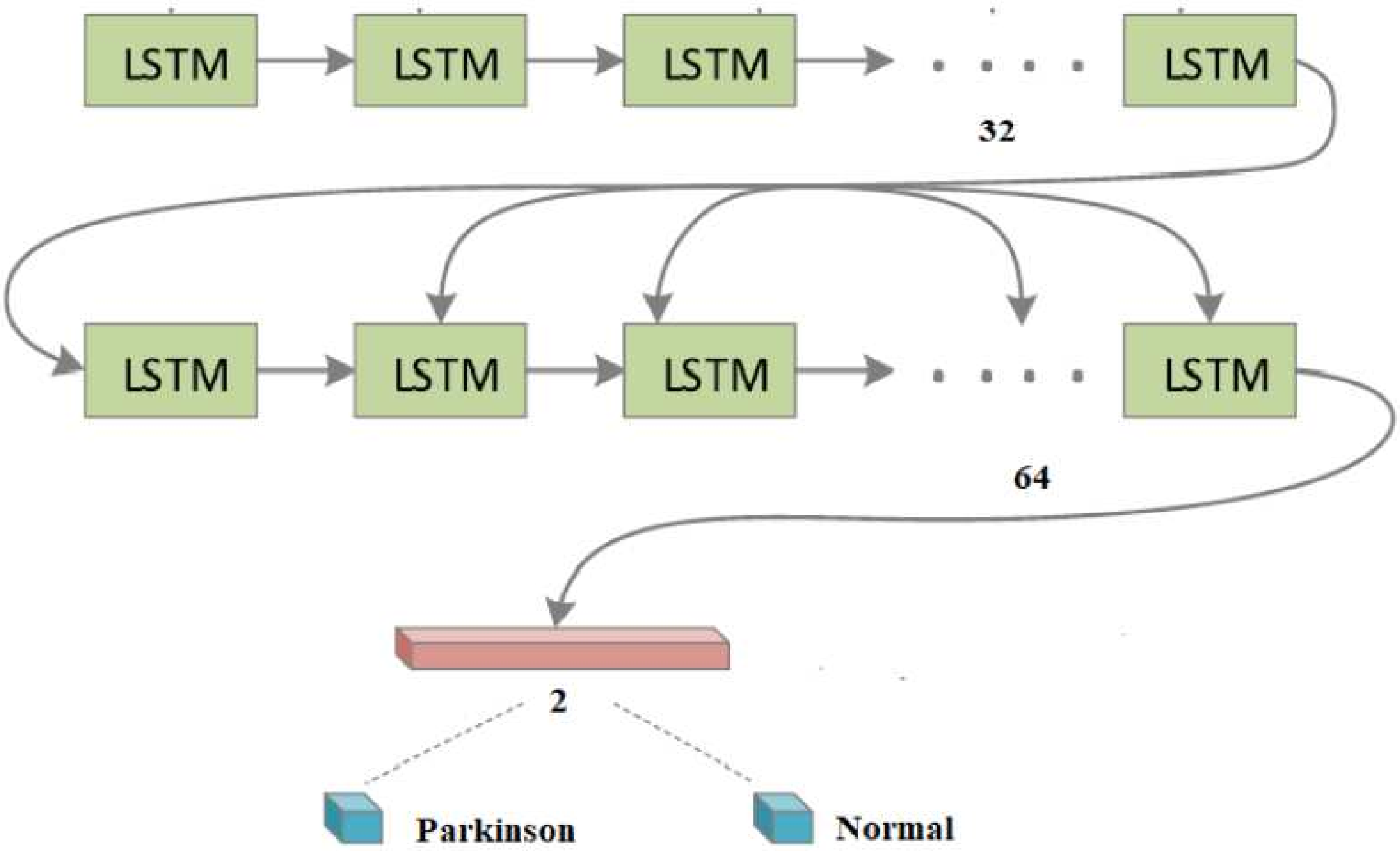


Figure 6

Combined ResNet-LSTM Model

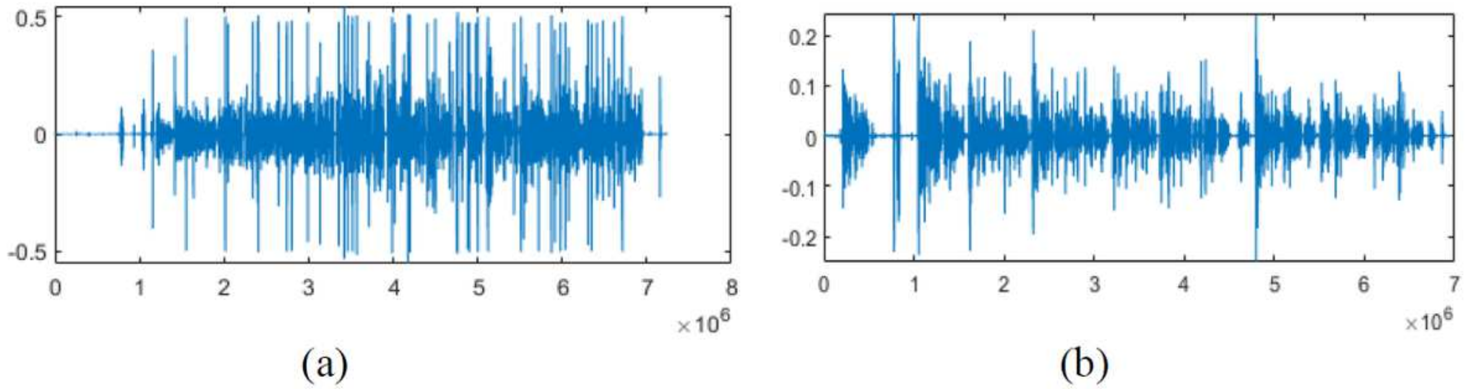
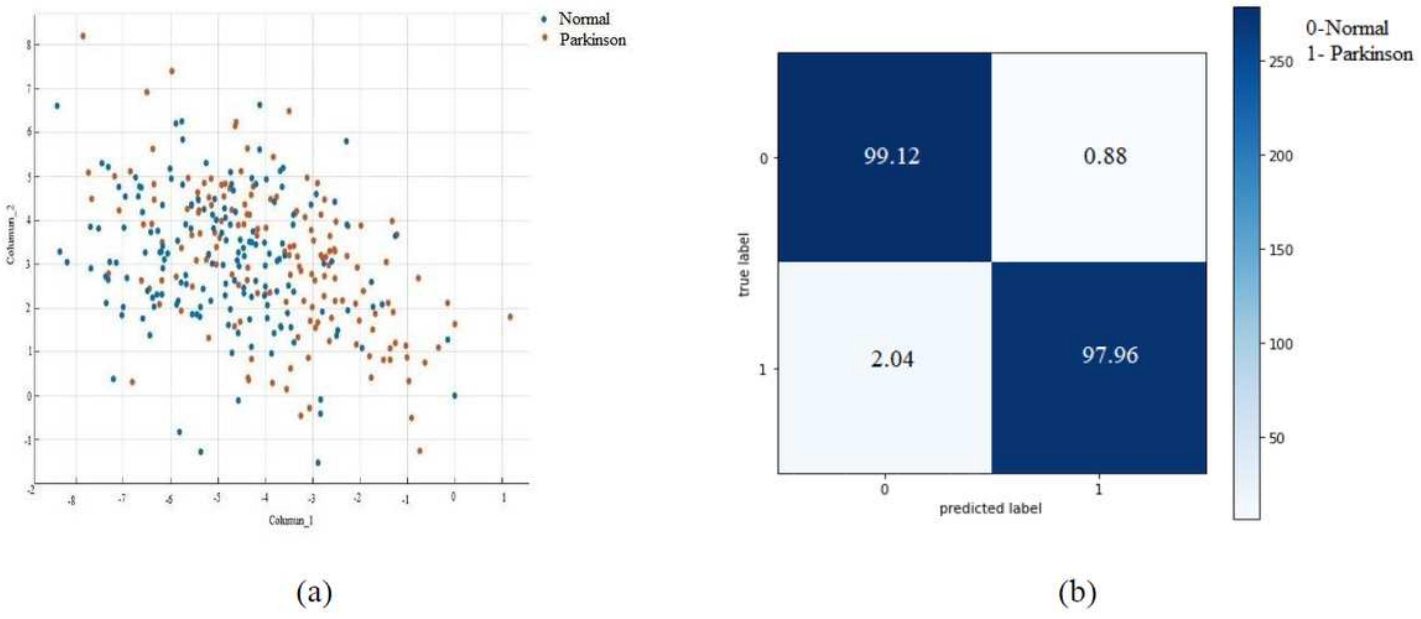


Figure 7

Ilustration of normal and parkinson speech signals (a) Normal speech (b) Parkinson speech

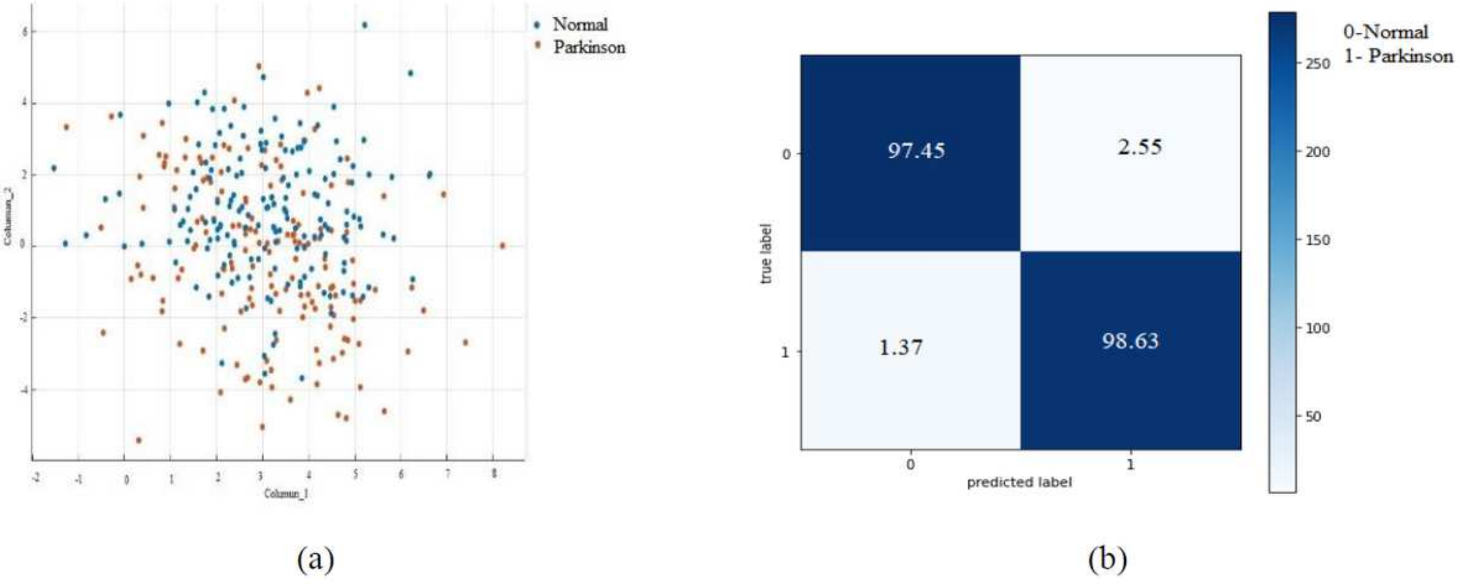


(a)

(b)

Figure 8

Ilustration of feature scatter and confusion matrix for Resnet18+LSTM (a) Feature scatter plot, (b) Confusion matrix

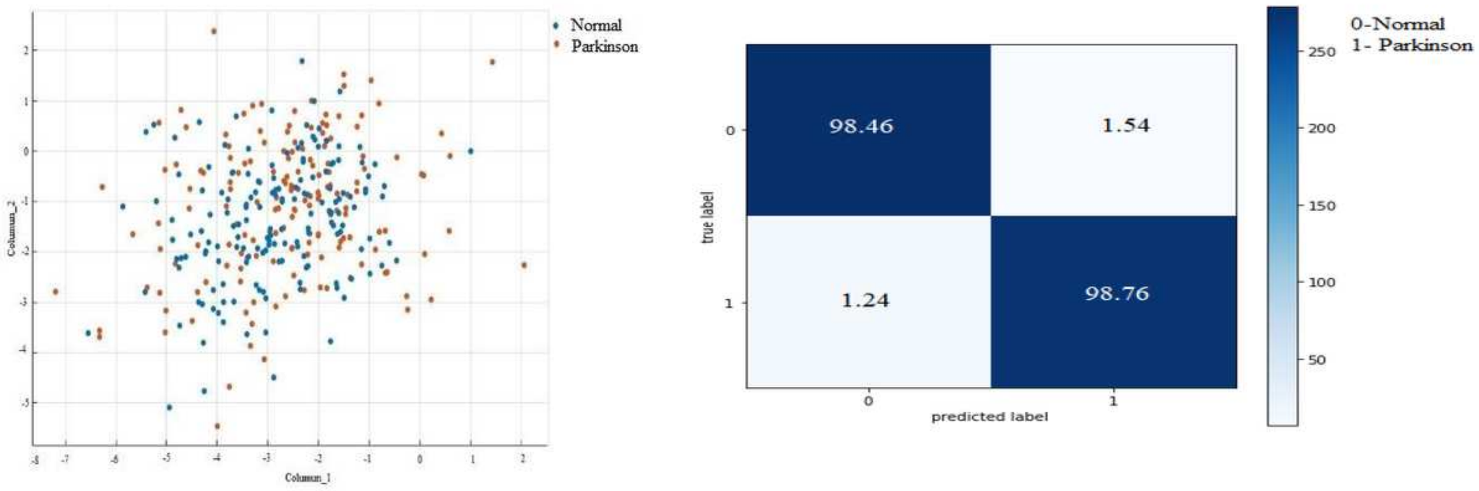


(a)

(b)

Figure 9

Ilustration of feature scatter and confusion matrix for Resnet-50+LSTM (a) Feature scatter plot, (b) Confusion matrix



(a)

(b)

Figure 10

Ilustration of feature scatter and confusion matrix for Resnet-101+LSTM (a) Feature scatter plot, (b) Confusion matrix