This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

A dual-modal attention-enhanced deep learning network for quantification of Parkinson's disease characteristics

Yi Xia, ZhiMing Yao, Qiang Ye, Nan Cheng

***Abstract*—It is well known that most patients with Parkinson's disease (PD) have different degree of movement disorders, such as shuffling, festination and akinetic episodes, which could degenerate the life quality of PD patients. Therefore, it is very useful to develop a computerized tool to provide an objective evaluation of PD patients' gait. In this study, we implemented a novel gait evaluating approach to provide not only a binary classification of PD gaits and normal walking, but also a quantification of the PD gaits to relate them to the PD severity level. The proposed system is a dual-modal deep-learning-based model, where left and right gait is modeled separately by a convolutional neural network (CNN) followed by an attention-enhanced long short-term memory (LSTM) network. The left and right samples for model training and testing were segmented sequentially from multiple 1D vertical ground reaction force (VGRF) signals according to the detected gait cycle. Experimental results indicate that our model can provide state-of-the-art performance in terms of classification accuracy. It is expected that the proposed model can be a useful gait assistance to provide a quantitative evaluation of PD gaits with high confidence and accuracy if trained suitably.**

***Index Terms*—Parkinson's disease (PD), Attention mechanism, Classification, Long Short-Term Memory (LSTM), Vertical Ground Reaction Force (VGRF)**

I. Introduction

PARKINSON'S disease (PD) is a representative degenerative disease that occurs due to a deficiency of nerve cells, called dopamine neurons, distributed in the substantia nigra of the brain [1]. Aging remains the biggest risk factor for developing idiopathic Parkinson's disease [2]. Hence, the increasingly aging population indicates that more and more elder persons could be affected by PD. Typical symptoms of the PD patients mainly consist of two parts: motor disabilities and non-motor disabilities. Characteristic motor features include tremor, bradykinesia (i.e., slowness of movement), rigidity (i.e., resistance to externally imposed movements), and impaired postural balance [3]. While the non-motor symptoms could be sleep disturbance, impaired heart rate variability, sialorrhea, sweating, and orthostatic hypotension [4]. PD is a chronic neurodegenerative disease, which means that the condition can become progressively worse as neuronal (brain) tissue dies and neurotransmitter levels fall [5]. Therefore, regular clinical assessment and close monitoring of the disease progression are very important to realize a dynamic management of the PD patients, and thus ensure their life quality can maintain a level as high as possible.

Current clinical assessment methods for PD patients still rely on patients' or caregivers' self-descriptions and clinician-mediated questionnaires, such as Unified Parkinson's Disease Rating Scale (UPDRS), freezing of gait questionnaire (FOG-Q) [6], Activities of Daily Living (ADL) part 14 [7], etc. In practice, skilled clinicians often rate the severity of a PD patient in certain aspect according to the performance of conducting the tasks designed in the questionnaires. This kind of subjective and time-consuming method tends to produce inconsistent results and thus plays a restricted role in diagnosing and monitoring. Hence, the diagnosis in current clinical practice requires regular review as certain 'red flag' features, such as rapid progression or early falls, point towards related diagnosis such as progressive supranuclear palsy or multiple system atrophy [5].

How to assess PD-related signs and symptoms objectively, efficiently and unobtrusively is very appealing to both the clinicians and the patients since it can boost the effectivity and efficiency of clinical visits and disease management. With the advent of new sensing technology and corresponding analysis methods, such a clinical wish appears to be coming within reach. For years, substantial efforts have been undertaken to develop various approaches to characterize almost all known PD symptoms (e.g., gait [8, 9], bradykinesia [10], and dyskinesia [11]) by using different type of signals (e.g., acceleration [9], plantar pressure [8], and surface electromyogram [12]). Comprehensive reviews about this topic can be found in [13].

Our main interest of this study is to provide accurate differentiating PD patients from normal controls (CO) as well as automatic rating of their severity levels by using the gait information. As we know, PD is caused by the degeneration of dopamine and other sub-cortical neurons in the basal ganglia of

This work is supported by National Natural Science Foundation of China (NSFC) for Youth (grant number 61701483); Anhui Provincial Natural Science Foundation (grant number 1608085MF136).

Yi Xia Author is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (Corresponding author)* (e-mail: xiayi@ahu.edu.cn)

ZhiMing Yao Author is with the Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China

Qiang Ye Author is with the Department of sport and health science, Nanjing Sport Institute, Nanjing 210014, China

Nan Cheng Author is with the Hospital Affiliated to Anhui Chinese Medical University, Hefei 230061, China

the brain. Dopamine functions as a neurotransmitter, which is responsible for relaying messages that plan and control movement and balance. Basal ganglia, and specifically the putamen in the neural network involves in timing control during movement [14]. Thus, the individuals with PD might be expected to have gait disorders with different forms as disease processes, such as a reduced gait speed, shorter stride length, increased stride-to-stride variability, shuffling gaits, and reduced angular excursion of the joints. According to the description of UPDRS rating scale, patterns of progression in PD can be generally classified into five stages. Stage 1 shows tremor and other movement symptoms on one side of the body without functional impairment. Stage 2 involves first signs of tremor and other movement symptoms on both sides of the body without impairment of balance. Stage 3 shows loss of balance and slowness of movement that fails to protect against falling. Stage 4 shows severe disability of standing and walking. Stage 5 causes stiffness in the legs and restricts the patient to bed or wheelchair [15]. Therefore, gait analysis, using kinetics [16, 17] and kinematics [8, 9] information, is a very useful diagnosing assistance in assessment of PD patients.

The ground reaction force (GRF) during walking can be measured either as a whole by using force plate or at multiple positions by using pressure-sensitive sensors underneath the foot. Due to its usefulness in understanding how the force is exerted to initiate and maintain a walking, GRF was always an important gait signal that has attracted continuous interest in gait analysis. Ren et al. [18] analyzed the Wiener-Akaike-Granger-Schweder influences between vertical ground reaction force (VGRF) signals at different plantar areas of both feet. By employing statistical test, they proposed Gait Influence Diagrams (GIDs) to display whether the directed influence from its corresponding plantar area i to j is significantly different between the PD patients and the healthy subjects. Daliri [17] utilized the difference signal along the time axis at each sensor location, and then several frequency features were calculated from the spectrum obtained by using short-time Fourier transform (STFT). The best classification results they reported in differentiating PD patients and normal subjects is an accuracy of 91.20% by utilizing support vector machines (SVMs) as the classifier. Abdulhay et al. [19] extracted the stride time, stance time, swing time and foot strike profile as the classification features to classify normal and PD subjects. In their experiments, the maximum accuracy of 94.8% was obtained with the medium Gaussian SVM. Aşuroğlu et al. [20] utilized Locally Weighted Random Forest (LWRF) to perform regression analysis to predict the severity of PD symptoms in terms of Universal Parkinson Disease Rating Scale (UPDRS) and Hoehn and Yahr (H&Y) scale, and the correlation coefficient between the actual and the predicted values is 0.895 (for UPDRS) and 0.960 (for H&Y) respectively.

Recently, deep learning (DL) technique, as a representation learning approach from large-scale data, has achieved state-of-art performance in many different tasks, such as object retrieval, classification, recognition, and clustering. The huge success of DL is largely due to the use of deep neural networks that can extract powerful features, which are often more discriminative than handcrafting features. Moreover, the feature learning process can be seamlessly integrated into the whole framework in an end-to-end manner. In contrast, feature extraction and classification are two separate processes in traditional shallow machine learning approaches, where the domain-dependent features must be designed and selected with enough expert knowledge and the classifier also need to be carefully tuned to obtain the best performance. Encouraged by the impressive successes of DL, different efforts have also devoted to learning gait features that are characteristic of different subjects. For example, convolutional neural network (CNN) was utilized to detect freezing of gait, a motor symptom that may be the most incapacitating to PD patients [21, 22]. Nancy Jane et al. [23] presents a Q-backpropagated time delay neural network (Q-BTDNN) classifier that builds a temporal classification model for the task of diagnosing the severity of gait disturbances in PD affected patients, and the best result they reported is an accuracy of 92.19% on the dataset gathered in [24]. In another work, a two-channel model combining CNN and Long-Short-Term Memory (LSTM) network was constructed for the purpose of quantitatively assessing the PD patients' gaits [25]. On the dataset gathered in [26], such a network provided an accuracy of 97.48% to differentiate patients with four different H&Y values, i.e. 0, 2, 2.5, and 3.

Though the DL-based approaches have achieved very promising performance in differentiating different gaits, these deep networks were designed and trained without consideration of the biomechanical characteristics of the gait, which has been found to be very important in characterizing the gaits [27-29]. From the perspective of biomechanics, walking is a series of rhythmical, alternating movements of the left and right lower limb resulting in the forward progressing of the whole body. Furthermore, the left and the right lower limb both have their own characteristic kinematics and kinetics due to the different neuro-controlling signal they received. For PD patients, due to the disruption of neuromuscular control of gait, they are characterized by asymmetric motor deficits in both the upper and lower extremities [27], which were successfully utilized in several studies [28, 29] for differentiating PD gaits and normal gaits. Inspired by these observations, we hypothesized that modeling left and right gait separately should provide better discriminating power than a single model where left and right gait are considered as a whole at characterizing the gaits of PD patients. For this purpose, we designed a dual-modal DL network that has two separate channels for modeling the left and right gait respectively. In each channel, the multi-points gait VGRF time series were firstly fed into a two-layer CNN to learn features that reflect the inter-dependent relationship at positions where the force sensors are installed. Then, an attention-enhanced LSTM network was exploited to learn features that can represent the temporal embedding in a gait cycle. Attention is a mechanism that mimics human's capability of focusing on the salient parts of the object to be observed instead of paying attention to the whole object. The attention in LSTM is a learnable vector where each element is in fact a weight being connected to an observation in a time series. It is expected that larger weights should be assigned to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

those more important observations. The left and right gait features learned in two parallel channels were then combined with a fully connected (FC) layer and its output was fed into the softmax layer for the final classification. To our best knowledge, the proposed dual-modal DL-based model has never been proposed for modelling the gaits of PD patients. The experimental results on three gait VGRF datasets demonstrated that this novel model could obtain state-of-art performance in either binary classification of the PD gaits and normal gaits or multi-classification of the gaits from PD patients with different severity level.

The rest of the paper is organized as follows. Section II first describes the three gait VGRF datasets to be analyzed in the present study, then some technical background is introduced, and lastly, the details of the proposed model is presented. Section III describes the experimental setups. Section IV reports results of our model on different datasets. Section V discusses findings from the experimental results. Finally, Section VI concludes the present research.

II. Materials and Methods

A. Materials

1) Description of dataset

In this study, the gait dataset is provided by Hausdorff et al. [30]. The dataset is comprised of three different contributions by three research groups (Ga [26], Ju [31] and Si [24]). All these three studies recruited PD patients and healthy COs to collect their gait information for the purpose of quantitative analysis. In total, there are 93 patients with the idiopathic PD (age: 66.3 ± 9.5 (SD) years, 59 men and 34 women) and 73 healthy COs (age: 63.7± 8.7 (SD) years, 40 men and 32 women). The PD Severity level was graded according to two scales, i.e., Hoehn and Yahr (H&Y) and Unified Parkinson's Disease Rating Scale (UPDRS). The H&Y scale is comprised of 5 stages ranging from 1 to 5 originally and is further extended with two more stages, i.e. 1.5 and 2.5 [32]. The UPDRS scale is more complex and consists of 5 sections with a value ranging from 0 to 199 [33].

TABLE 1
DESCRIPTION OF THE DATA FORMAT

Column	Description	Units
1	Time stamp	Seconds
2-9	VGRF from each of eight sensors (L1-L8) under the left foot	Newton
10-17	VGRF from each of eight sensors (R1-R8) under the right foot	Newton
18	Total force under the left foot	Newton
19	Total force under the right foot	Newton

According to the experimental protocol, every participant was asked to walk in their usual, self-selected pace on level ground for about two minutes. VGRF (in Newton) generated when the foot contact the ground was measured via 8 force-sensitive sensors that located under each foot. The coordinates of these sensors in an underfoot plane when a person is comfortably standing with both legs parallel to each other are shown in Fig.1. The outputs of the sensors were recorded by an A/D converter at a sampling rate 100Hz. Every record had 19 parameters that corresponded to each row in the dataset file. As shown in Table 1, it includes the time stamp, 16 VGRF data corresponding to 16 force sensors under both feet and 2 total VGRF data.

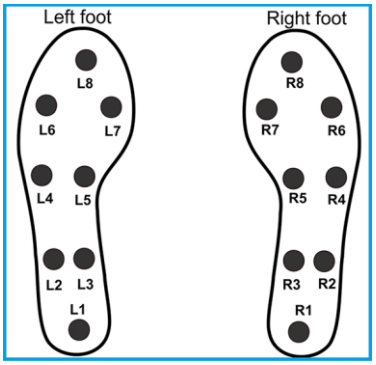

Fig. 1. The positions of the force-sensitive sensors underneath left and right foot

2) Preparation of data instances

As shown in Fig.2, the gait during normal walking for either left or right lower limb is a quasiperiodic process. Hence, for a more reasonable modelling and learning of the innate gait characteristics, it is better to segment gait data on the basis of walking cycles [34, 35]. For such purpose, this study utilized the total force to detect the walking cycles of the corresponding lower limb. A gait cycle for single lower limb is defined as a period from the beginning moment when the corresponding foot contacts the ground to the ending moment when the foot leaves the ground. In the total force vs. time curve, the force jumps from zero to non-zero value at the beginning moment, while it jumps from non-zero back to zero value at the ending moment. By simple thresholding technique, these key zero-crossing points can be detected and thus the gait cycles can be segmented. With the detected gait cycles, the multi-points VGRF data sequence can be divided into many data instances for training and testing in a DL-based model. As an illustration, Fig. 3 presents the sensor outputs on the basis of a segmented gait cycle for gaits of normal COs and PD patients with different H&Y scores.

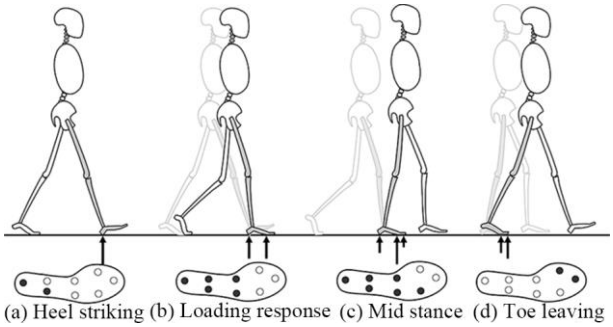

Fig. 2. A gait cycle for the right limb

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

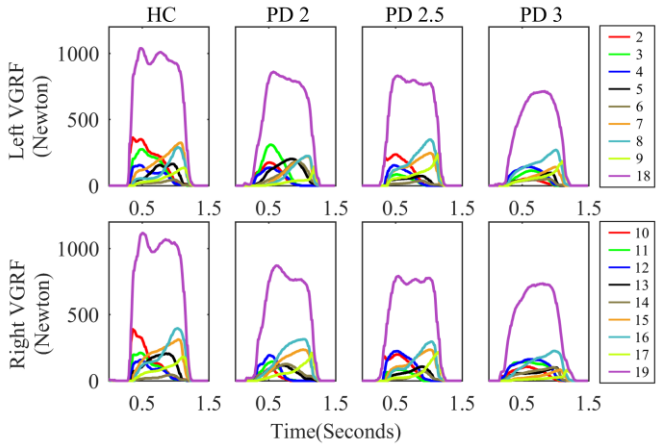

Fig. 3. VGRF of healthy COs and PD patients with different H&Y scores. The upper row is for the VGRF of the left foot, while the lower row is for the right foot. HC denotes healthy control, while PD 2, 2.5, and 3 represents the PD patient with different H&Y score. The number in the legend corresponds to the feature index in Table 1.

Due to the fluctuation of the walking speed and stride length from one cycle to another cycle, the durations of gait cycle are not always with the same length. Considering that feature extraction and pattern classification requires *N*-sized vectors for each cycle, where *N* is a fixed length, a zero-padding method is used in this study to make all the detected gait cycles have the same length of $N = 150$. This specific value of *N* is determined based on the sample rate (100 samples per second) and the slowest walking speed (about 1.5 seconds for the maximum duration of one gait cycle) that is observed in this study.

As a summary, The statistics of different classes in every investigated dataset as well as a merged dataset of them all is listed in Table 2.

TABLE 2
THE DIVISION OF DATA INSTANCES ACCORDING TO THE HEALTHY CONDITION

Dataset	H&Y score				
	Healthy COs	PD patients			
	0	2	2.5	3	Merged (2, 2.5, 3)
Ga [26]	4010	4498	2147	1344	7989
Ju [31]	1858	3475	4434	959	8868
Si [24]	3211	2938	722	0	3660
Merged	9079	10911	7303	2303	20517

B. Technical background

This paper mainly focuses on training a dual-modal attention-enhanced CNN-LSTM for discrimination analysis of gait patterns. Therefore, this section will briefly introduce some preliminary knowledge including CNN, LSTM and attention mechanism

1) CNN

By applying convolution operation with different kernel size at different layer with different scale, CNN can learn a hierarchy of progressively more abstract features that are hard to be designed manually. Consider a CNN with L convolutional layers, and its input is a 2D signal. The kernels for each layer have a size of $M_l \times N_l$, and they are parameterized by tensor $\mathbf{W}^{(l)} \in \mathbb{R}^{F_l \times m \times n \times F_{l-1}}$ and bias $\mathbf{b}^{(l)} \in \mathbb{R}^{F_l}$ where $l \in \{1, \cdots, L\}$ is the layer index, F_l is the feature map number of the *l*-th layer, and $m \times n$ is the kernel size. Then, for the *l*-th layer, the feature component $\mathbf{E}_{i,j}^{(l)} \in \mathbb{R}^{F_l}$ at the position of (i, j) in the 2D space is a function of the incoming activation tensor $\mathbf{E}^{(l-1)} \in \mathbb{R}^{F_{l-1} \times M_{l-1} \times N_{l-1}}$

$$\mathbf{E}_{i,j}^{(l)} = f\left(\mathbf{b}^{(l)} + \sum_{i'=1}^{m}\sum_{j'=1}^{n}\left\langle \mathbf{W}_{:,i',j',:}^{(l)}, \mathbf{E}_{:,i+m-i',j+n-j'}^{(l-1)} \right\rangle\right) \quad (1)$$

where $\langle \bullet \rangle$ denotes the inner product, and $f(\bullet)$ is an activation function (i.e. sigmoid(·), tanh(·) and rectifying linear unit (ReLU)).

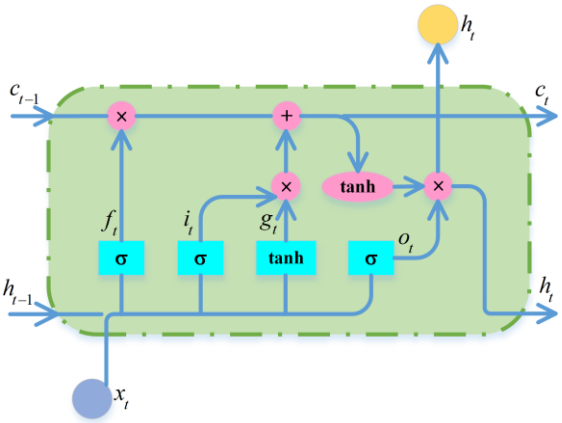

Fig. 4. Typical structure of LSTM, which consists of one memory state cell C and three gate functions (input i_t, forget f_t and output o_t).The update of each gate and state can be found in Eq. (2)-(8).

2) LSTM

Different from CNN, LSTM is a variant of recurrent neural network (RNN) that can better learn the dynamic timing behavior [36]. As shown in Fig 4, a basic LSTM unit explicitly introduces one memory state cell C and three gate functions (input i_t, forget f_t and output o_t) into their state dynamics. g_t maps the value of input to a range from -1 to 1. The input gate determines whether or not to let new input to alter the state of the memory cell, forget gate controls what to be forgotten and what to be remembered by the memory cell, and the output gate let the state of the memory cell impact the output at the current time step. Mathematically, each vector in the LSTM cell can be computed as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{bmatrix} \quad (2)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_\mathbf{f} \cdot \mathbf{X} + \mathbf{b}_\mathbf{f}) \quad (3)$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_\mathbf{i} \cdot \mathbf{X} + \mathbf{b}_\mathbf{i}) \quad (4)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot \mathbf{X} + \mathbf{b}_o) \quad (5)$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_c \cdot \mathbf{X} + \mathbf{b}_c) \quad (6)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (7)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (8)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

where $\mathbf{W}_i$, $\mathbf{W}_f$, $\mathbf{W}_o \in \mathbb{R}^{N\times 2N}$ are the weighted matrices and $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_o \in \mathbb{R}^N$ are bias vectors of LSTM to be learned during training, parameterizing the transformations of the input, forget and output gates respectively. $\mathbf{x}_t$ is the input of LSTM cell unit, σ is the sigmoid function, and $\odot$ stands for element-wise multiplication.

3) Attention mechanism

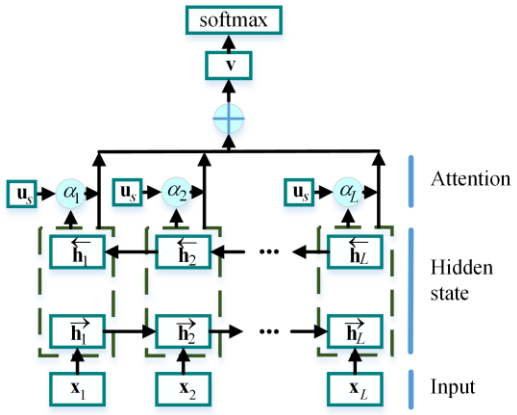

Fig. 5. The structure of a recurrent neural network with attention mechanism

For temporal sequential data, the introduction of attention is to assess each time step observation with an importance score or salience weight since not all observations contribute equally to the representation of a pattern, and thus construct a hidden representation by integrating these scores to obtain better classification performance. There are different ways [37, 38] for constructing and learning such importance score, but in general they all utilize the temporal context information. As an illustration, the attention mechanism proposed in [38] is shown in Fig. 5.

Given an input sequence $\mathbf{x}_{1,\cdots,N} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ of length N in which $\mathbf{x}_t \in \mathbb{R}^D$ denotes the observation at the t-th time step, the attention score a_t at the time step t is a scalar value, which indicates the importance of the current observation to the whole temporal sequence. Supposing that a bi-directional LSTM is used to model the input vector, then the LSTM's forward hidden state $\vec{\mathbf{h}}_t \in \mathbb{R}^M$ and the backward hidden state $\overleftarrow{\mathbf{h}}_t \in \mathbb{R}^M$ can be concatenated together, i.e., $\mathbf{h}_t = [\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{2M}$, to summarize the information embedded in the input vector $\mathbf{x}_t$. Thus, the attention score a_t can be calculated as the follows:

$$\mathbf{u}_t = \tanh(\mathbf{W}\mathbf{h}_t + \mathbf{b}) \tag{9}$$

$$\mathrm{a}_t = \frac{\exp(\mathbf{u}_t^T \mathbf{v})}{\sum_{t=1}^{N} \exp(\mathbf{u}_t^T \mathbf{v})} \tag{10}$$

where $\mathbf{W}$ is a linear transformation matrix and vector $\mathbf{b}$ is a bias term. Eq. (10) is in fact a softmax(·) function, which realizes the calculation of a normalized attention score. The attention score is measured as the similarity of $\mathbf{u}_t$ with a trainable context vector $\mathbf{v}$. After obtaining the attention score, the final output hidden state $\mathbf{h}_{wt} \in \mathbb{R}^{2M}$ can be calculated as a weighted sum of all the hidden states of LSTM:

$$\mathbf{h}_{wt} = \sum_{t=1}^{N} a_t \mathbf{h}_t \tag{11}$$

C. Dual-modal network for gait classification

To model the left and right gait explicitly in a single framework, we present a novel dual-modal framework for gait classification, which is comprised of two branches with same structure. In each branch, a neuro network was first applied to learn discriminative representations of the 2D input data, i.e., the multi-points VGRF time series. The attention-enhanced bi-directional LSTM was then utilized to model the temporal dynamics of the feature maps output by CNN. Finally, the two attention-weighted hidden state vectors output from the LSTM cells are concatenated and fed into a FC layer, whose output is the input to the softmax classifier. The proposed **D**ual-modal with each branch has a **C**onvolutional network followed by an **A**ttention-enhanced bi-directional **LSTM** is referred to as DCALSTM.

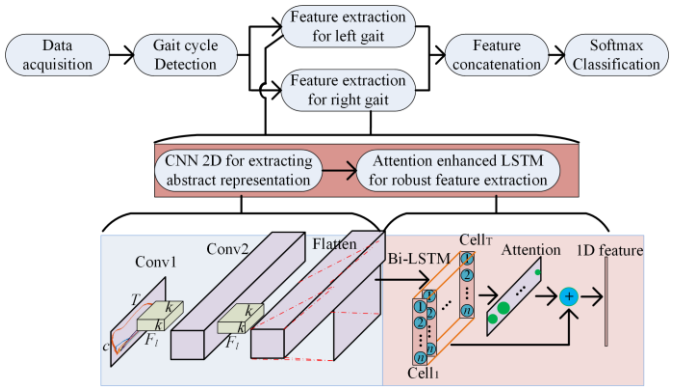

Fig. 6. The architecture of the proposed DCALSTM. Conv1, and Conv2 are two convolutional layers, Bi-LSTM represents bi-directional LSTM, and Flatten is a layer where all the other dimensions except the temporal dimension are flattened to a 1D vector. T is the length of data instance in the temporal dimension, c is the number of the force sensors underneath one foot, and n is the length of a hidden state vector (including both directions) in a bi-directional LSTM. F_l and $k\times k$ indicate the number of feature maps and the kernel size respectively.

1) Architecture of DCALSTM

The structure of the proposed DCALSTM is illustrated in Fig. 6. In each branch, the input has a dimension of $B\times 150\times 9\times 1$, where B is batch size of training or testing samples, 150 is the time span of a sample, and 9 is the number of the analyzed VGRF signals underneath a specific foot as mentioned in section II. The CNN has three layers, and in each layer a convolution operation is followed by a non-linear activation function (e.g., ReLU) to compute the feature map. For simplicity, in our network, the conventional kernel in all layers has the same size of $k\times k$, and the number of feature maps in the two layers is F_1 and F_2 respectively. Note that, for the purpose of learning the interaction between those different force sensing points during a gait cycle, a 2D instead of 1D kernel was utilized to perform both temporal (along the time axis) and spatial (between different sensing points) convolutional operation. Furthermore, in our DCALSTM, the convolutional layers do not include a pooling operation due to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

the limited size of the data samples corresponding to a gait cycle in both dimensions. Accordingly, the feature map output by the last convolutional layer has a size of $B\times150\times9\times C_3$. By dimension flatting, the above feature map can be transformed into a tensor with a size of $B\times150\times C_4$, where $C_4=9\times C_3$. Such a tensor will be fed into the followed attention-enhanced LSTM (AE-LSTM) whose structure is shown in Fig. 5.

The outputs of the AE-LSTMs in the two branches were integrated by a concatenation fusion, which was then passed into a FC layer to be transformed into a g-dimensional (g classes) vector. Finally, the softmax classifier realizes mapping the output of the FC layer to a probability distribution and discriminates whether a subject has PD disorders or rate the severity level of a PD patient.

2) Model Training

For ease of presentation, the model parameters to be trained can be divided into two groups. The first group includes the weight matrix $\mathbf{W}=\{\mathbf{w}_1,\cdots,\mathbf{w}_L\}$ involved in either CNN or AE-LSTM, and the second group includes the corresponding bias $\mathbf{B}=\{\mathbf{b}_1,\cdots,\mathbf{b}_L\}$. Here, the subscript 1 to *L* refers to the layer index, and the *L*-th layer is the softmax layer.

The loss function of our proposed model consists of two terms. The first loss term is the cross-entropy loss, which aims to minimize the classification error for the given training samples, and it is computed by

$$J_1(\mathbf{X},\mathbf{W},\mathbf{B})=-\frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{y}^{(i)},\log\hat{\mathbf{y}}^{(i)}\right\rangle \tag{12}$$

where $\hat{\mathbf{y}}^{(i)}$ is the predicted label, N is the number of training samples, $\mathbf{y}^{(i)}$ is a one-hot vector that represents the label of the *i*-th sample, $\mathbf{X}$ is the training dataset, and $\langle\bullet\rangle$ denotes the inner product.

The second item is the weight decay term, which is designed to decrease the magnitudes of the weights $\mathbf{w}_i$ and the bias $\mathbf{b}_i$, and hence it is beneficial to prevent over-fitting. It is calculated as

$$J_2(\mathbf{W},\mathbf{B})=\sum_{l=1}^{L}(\|\mathbf{w}_l\|_F^2+\|\mathbf{b}_l\|_2^2) \tag{13}$$

By incorporating the loss terms in Eq. (12) and (13) together, the training loss function of our proposed model can be formulated as

$$J=\min\left(J_1(\mathbf{X},\mathbf{W},\mathbf{B})+\lambda_1 J_2(\mathbf{W},\mathbf{B})\right) \tag{14}$$

$$J=\min\left(-\frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{y}^{(n)},\log\hat{\mathbf{y}}^{(n)}\right\rangle+\lambda_1\sum_{l=1}^{L}(\|\mathbf{w}_l\|_F^2+\|\mathbf{b}_l\|_2^2)\right) \tag{15}$$

where λ_1 is a tradeoff parameter that balances the relative importance of the two terms.

In our experiments, the AdaGrad algorithm [39] was utilized as our optimization method. As a self-adaptive learning mechanism, AdaGrad allows larger updates (large learning rate) for infrequent parameters, and smaller updates (small learning rate) for frequent parameters, resulting in better robustness and faster convergence. We didn't choose Adam (another popular self-learning algorithm) [40], because it has more parameters to be tuned and our experiments showed that AdaGrad can provide similar performance comparable to that by Adam.

III. *Experimental Settings*

This section introduces the experimental settings and evaluation criteria used to assess the effectiveness of the proposed DCALSTM for gait characterization. Two main experiments were conducted to evaluate the effectivity of our approach. The first experiment detects the PD patients from the healthy COs, which is a typical binary classification problem. In the second experiment, our proposed model was exploited to realize a multi-label classification problem for a quantitative rating of the PD patients' severity level.

For comparison, we also realized six baseline approaches, including four DL-based models (DCNN, DALSTM, DCLSTM, and CNN-LSTM) and two traditional classification schemes (TRAD1 and TRAD2). Some details above these models are given as follows:

DCNN: it has only a CNN instead of the attention-enhanced LSTM in each branch.

DALSTM: it has only an attention-enhanced LSTM in each branch.

DCLSTM: it removes the attention mechanism from the structure of DCALSTM.

CNN-LSTM: it is a model proposed in [41], whose input doesn't differentiate left or right gait. It also has two branches, where one branch was used to extract spatial features by using CNN, and the other branch exploited double-layer LSTM to learn temporal dynamics.

TRAD1: the feature set, as proposed by Abdulhay et al. [19], includes: stride time, stance time, swing time and foot strike profile for both lower limbs.

TRAD2: the feature set is designed by Aşuroğlu et al. [20]. For each force signal segment at different sensing position, total sixteen time-domain features (e.g., skewness, kurtosis, entropy, and energy) and seven frequency-domain features (e.g., mean, maximum, and minimum value of the FFT magnitude) are extracted.

After some performance tuning process, the classifiers used in TRAD1 and TRAD2 are chosen to be random forests (RF), and the number of decision trees in RF is set to be 50. The hyper-parameters (e.g., learning rate, dropout rate, and batch size) for DCNN, DALSTM, and DCLSTM were kept same as those for the proposed DCALSTM. The parameters for CNN-LSTM were determined according to the recommendations in [41]. In our objective function of Eq. (15), the weight decay parameter λ_1 was optimized by varying the parameter λ_1 from the set of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, and changing the iteration number from 1000 to 15000 with a stride of 1000 respectively. As a summary, the settings for important hyper-parameters are listed in Table 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

TABLE 3
THE HYPER-PARAMETERS FOR ALL THE MODELS REALIZED IN THIS STUDY

Learning parameter	Value or method
Batch size	128
Regularization	dropout (rate = 0.5)
Learning rate	0.01
Training iterations	15000
Kernel size	3×3
Feature map number F_1	32
Feature map number F_2	32
Hidden state vector length	256
Attention context vector length	128
λ_1	0.0001

For the purpose of evaluating all the classification models, we conducted 5-fold cross validation in a stratified manner. Firstly, the original data set is divided into 5 independent folds. Four of the five folds are combined and used as a training set; the remaining fold is used as a testing set. Each training set is resampled and resized by the SMOTE algorithm [42], so that the number of instances in every class is approximately balanced. The classification performance was measured by accuracy (Ac), sensitivity (Se), and specificity (Sp). Further, in order to determine whether the results obtained by the proposed model are statistically different from that by other methods, Wilcoxon's signed ranks tests [43] are conducted between the accuracy results achieved by these compared algorithms. In these tests, a *p-value* is computed after an experiment is performed twenty times. If the *p-value* is less than the significance level of 0.05, there are significant differences between these methods. All the models were implemented with Tensorflow library (version tensorflow-gpu-1.9.0), and our hardware platform was configured with Intel(R) Core(TM) i7-7800X CPU@3.50GHz, Nvidia GeForce Titan X 16GB GPU, and 64 GB RAM.

IV. RESULTS

The results of different loss regularization parameter λ_1 to the iteration number on four datasets are plotted in Fig.7. Considering the results on all these datasets, a value of 0.0001 is set for the loss regularization parameter λ_1 throughout our experiments. Fig. 8 shows the binary classification accuracies of training set and testing set on the merged dataset during the model learning process. As can be seen from Fig. 8, the proposed DCALSTM outperforms other models with an apparent margin on both training and testing set. Table 4 presents the results of all compared models on the binary classification problem, where the PD patients are discriminated from the healthy COs. For the two traditional classification approaches, TRAD2 outperforms TRAD1 with apparent margin in all experiments. While for the DL-based models, the DCNN model performs the worst with an accuracy of 96.17%, 96.22%, 96.94%, and 96.28% on Ga [26], Ju [31], Si [24], and the merged dataset respectively. The other models all have LSTM in its network structure, and they can obtain comparatively better results. The results for CNN-LSTM model are better than DALSTM and slightly inferior to that of DCLSTM in most cases. The difference between DCLSTM and DCALSTM is that DCLSTM has no attention mechanism, and it is found that DCALSTM can provide better results than DCLSTM in all cases.

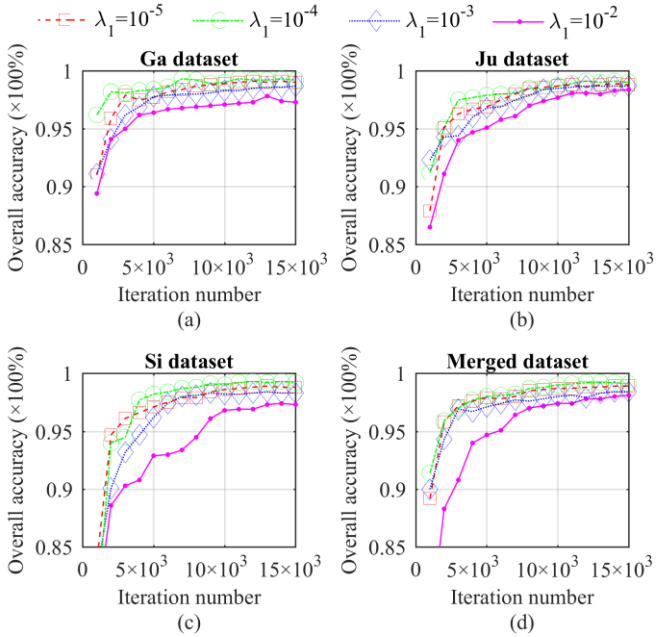

Fig. 7. The accuracies of our proposed DCALSTM model under different loss configurations. The binary classification performance was assessed on four different datasets including (a) Ga dataset [26] (b) Ju dataset [31] (c) Si dataset [24] (d) Merged dataset. The regularization parameter λ_1 varies from the set of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and the iteration number changes from 1000 to 15000 with a stride of 1000

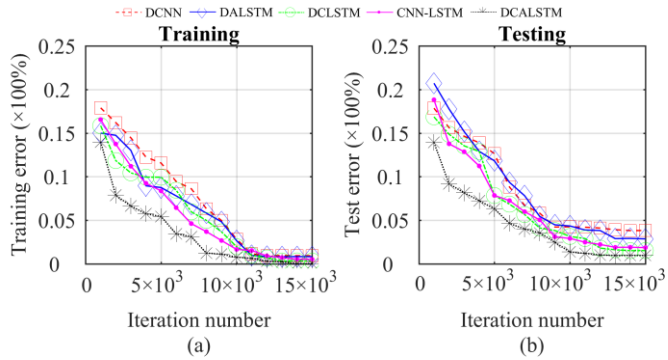

Fig 8. (a) Training error and (b) testing error on the merged dataset by all compared models. The iteration number changes from 1000 to 15000 with a stride of 1000.

Table 5 lists the results of PD H&Y scale rating problem obtained by different DL-based models. Similarly, the proposed DCALSTM provides the highest classification performance on all datasets in terms of accuracy. More specifically, the accuracy for the dataset gathered in Ga [26], Ju [31], Si [24], and the merged one is 98.11%, 98.36%, 99.01%, and 98.03%, respectively. Furthermore, for a visual display of the classification results between four different H&Y scores, the confusion matrices obtained by the proposed DCALSTM are shown in Fig. 9, where several observations can be made. First, for the healthy COs and the PD patients with the highest H&Y score 3, our DCALSTM can provide almost 100% detection accuracy. Second, for the Si dataset, though it has fewer severity levels, the accuracy of this dataset is similar,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

even slightly better than the other two datasets. This is because that the other datasets (Ga and Ju) both have more severity levels with approximate H&Y scores (i.e., 2, 2.5 and 3). More severity levels mean that there are more classes to be differentiated, and the PD patients with similar severity levels can be misclassified each other, which indicates that the gait difference between these levels of patients are not very apparent, and they could be wrongly labelled by the raters in preparing the dataset.

Table 6 tabulates the average *p-values* when the proposed model is against other classification schemes. Obviously, the proposed dual-modal DL network is statistically different from other counterparts with 5% significance level in terms of accuracy performance.

TABLE 4
THE RESULTS OF THE PROPOSED DCALSTM MODEL FOR DIFFERENTIATING GAITS OF PD PATIENTS FROM THAT OF NORMAL CONTROLS USING THE 5-FOLD CROSS-VALIDATION METHOD

Model	Ga [26]			Si [24]		
	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)	Ac (%)
TRAD1	94.99	93.99	94.66	94.84	94.99	94.91
TRAD2	98.20	97.63	98.01	98.09	98.16	98.12
DCNN	96.09	96.33	96.17	96.89	97.01	96.94
DALSTM	97.88	97.08	97.62	98.01	97.94	97.98
DCLSTM	98.61	98.98	98.73	98.66	99.10	98.86
CNN-LSTM	98.76	98.35	98.63	98.47	98.88	98.67
DCALSTM	99.35	99.23	99.31	99.23	99.35	99.29
Model	Ju [31]			Merged		
	Se (%)	Sp (%)	Ac (%)	Se (%)	Sp (%)	Ac (%)
TRAD1	94.96	93.60	94.72	92.21	91.79	92.08
TRAD2	97.94	97.58	97.87	97.77	97.02	97.54
DCNN	96.31	95.80	96.22	96.34	96.13	96.28
DALSTM	97.25	97.09	97.22	97.84	97.43	97.71
DCLSTM	98.40	97.90	98.31	98.90	98.46	98.76
CNN-LSTM	98.34	97.84	98.26	98.80	98.05	98.57
DCALSTM	99.20	98.98	99.16	99.10	99.01	99.07

TABLE 5
THE RESULTS OF THE PROPOSED DCALSTM MODEL FOR PD H&Y SCALE RATING (FOR HEALTHY CONTROLS, THE VALUE IS ZERO) USING THE 5-FOLD CROSS-VALIDATION METHOD.

Model	Ac(%)			
	Ga [26]	Ju [31]	Si [24]	Merged
TRAD1	92.13	93.31	93.36	89.57
TRAD2	96.46	97.58	97.67	96.74
DCNN	95.26	95.90	96.01	95.78
DALSTM	97.18	97.44	98.08	96.79
DCLSTM	97.89	98.09	98.88	97.74
CNN-LSTM	97.55	98.01	98.43	97.51
DCALSTM	98.11	98.36	99.01	98.03

TABLE 6
SUMMARY OF WILCOXON'S SIGNED RANKS TESTS. THE 5% LEVEL OF SIGNIFICANCE IS SELECTED

Model	The proposed DCALSTM	
	p-value	Significant?
TRAD1	< 1e-04	Yes
TRAD2	< 1e-04	Yes
DCNN	< 1e-04	Yes
DALSTM	< 1e-04	Yes
DCLSTM	1.75e-04	Yes
CNN-LSTM	1.03e-04	Yes

V. DISCUSSION

For the purpose of understanding the gait dynamics and the relationship of its variations to the pathological alterations in the locomotor control system, much research in recent years has focused on measuring gait-related signals and then extracting robust and discriminative features for an objective quantification. Measured by the force-sensitive sensors embedded either in shoes or a large-area mat, the VGRF reflects the force of the lower limb exerted on the ground to maintain a continuous and repetitive locomotion of the whole body. As shown in Fig.3, for persons with normal gait, the VGRF curves bear similar profiles, which have an approximately symmetric hump in the middle. However, the humps in the profiles tend to be flattened for PD patients who have abnormal gait as they offer friction to the ground for a longer time while walking. In addition, it can also be found from Fig.3 that for a normal person the heel force is greater than the toe force, but for a PD subject the toe force is slightly greater than the heel force because PD patients often exert more pressure on his toe than his heel [44]. Therefore, by analyzing the measured VGRF from both feet, discriminative gait characteristics can be derived for different applications, such as medical diagnosis of gait-related diseases [44] and person recognition [45]. Previous studies devoted much effort on manually designing discriminative features and then fine-tuning the classifier parameters to train a machine-learning tool for their applications. In contrast, our approach enjoys the benefits of DL technique where the feature learning and classification are performed seamlessly in an end-to-end pipeline. Furthermore, by recognizing that the VGRF time series measured from both feet represent motor functioning of brain areas in each hemisphere, a DL-based framework with two branches is constructed to model left and right gait separately. Based on this framework, several problem-specific considerations, such as the abstraction of spatial feature and attention mechanism, are also included to further enhance our model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2946194, IEEE Transactions on Neural Systems and Rehabilitation Engineering

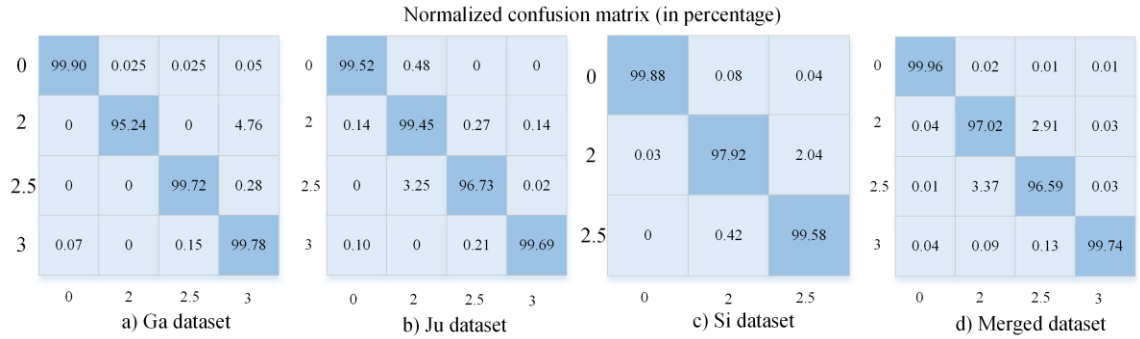

Fig. 9. Confusion matrix (values in percentage) for all the datasets (a: Ga dataset, b: Ju dataset, c: Si dataset, and d: Merged dataset) obtained by the proposed DCALSTM

The effectiveness of our approach was demonstrated by a series of experiments. As listed in Table 4 and 5, the importance of modeling left and right gait separately was demonstrated by the comparison between our model and CNN-LSTM, wherein the left and right gait signals were concatenated when they were fed into the model for training. Furthermore, the proposed DCALSTM model was also evaluated by being fed into only left or right gait data, and the obtained accuracy on the merged dataset for binary classification and H&Y scale rating problem was 98.08% and 97.45%, respectively. Such results were apparently lower than the corresponding results (Table 4 and 5, DCALSTM) obtained by using both left and right gait data. Therefore, left and right gait can be complementary in providing more discriminative feature vectors. As two traditional approaches, TRAD2 has much better performance than TRAD1, even better than some DL-based models in some cases. For TRAD2, the extracted temporal and spectral features focus on the force curve itself. While for TRAD1, the extracted features are mainly gait stride temporal parameters derived from force curve. This indicates that multiple force curves themselves could be more discriminative than several gait stride parameters in differentiating different group of subjects. Among DL-based models, DCNN provided relatively poorer discriminating performance than the models that have LSTM in their network structure. This indicated that the temporal structure of a VGRF time series learned by LSTM was more important than the spatial features between multiple sensing points learned by CNN. However, we must admit that 8 sensing points under each foot may be relatively few in number to provide very discriminative spatial features. We also noted that DCNN had only two layers, which was a relatively shallow network. To find whether a deeper CNN can improve the classification performance substantially, a five-layer CNN was also constructed in our experiments, and the obtained accuracy for binary classification problem and severity rating problem on the merged dataset was 97.32%, and 96.65% respectively. Though the performance has been improved, it was still slightly poorer than the performance obtained by those LSTM-based models. At the same time, such a result also indicated that it was important for CNN to learn abstract discriminative features from several concurrent VGRF time series measured at multiple sensing points. Therefore, a combination of CNN (learning the interconnections between different sensing points) and LSTM (learning the interdependence between adjacent context observations) should perform better than a separate CNN or LSTM for gait characterization. Such a hypothesis was supported by our experimental results since DCALSTM, DCLSTM, and CNN-LSTM obviously performed better than DCNN and DALSTM. In the end, the superiority of attention mechanism was demonstrated by the compassion between DCALSTM and DCLSTM, where DCALSTM performed consistently better than DCLSTM in all experiments.

In the future, more related investigations should be performed in the following aspects. First, the current dataset only provides VGRF data measured from 8 positions during walking. Though these positions are very important in capturing the gait kinetics, it is apparent that if more sensors are deployed underneath the foot, then the measured VGRF with higher spatial resolution (i.e., the plantar pressure distribution) can provide more discriminative information for characterizing different gaits [46, 47]. As for the proposed DCALSTM model, if the plantar pressure distribution is fed into the model, the spatial feature that learned will be more robust due to the increase of the force spatial resolution. Furthermore, the current dataset only provides vertical GRF. In fact, the medial-lateral and anterior-posterior GRF are also very useful in characterizing different gaits [48]. Therefore, it is expected that better performance can be obtained if they are included as the input to our model. Second, the proposed dual-modal model doesn't consider the concurrent coordination of left and right limb during walking. That is, by separating the left and right gait and then modelling respectively, our model doesn't consider the concurrent connection of left and right gait by viewing them as a whole, which should also be very important in characterizing gaits. Such importance can well explain why the LSTM model proposed in [25] can obtain very high classification accuracy by feeding the concatenated left and right gait data as a single vector into the model even without gait cycle detection. In fact, we can also add another channel into the proposed DCALSTM to model the left and right gait as a whole in a gait cycle. However, such a network design will complicate our model and may lead to the difficulty of training. Third, in the current study, only one walking cycle was utilized to determine whether a subject is a PD patient or his/her H&Y score. The score provided by the softmax classifier for specific class should obey certain probability density function (pdf).

Therefore, if the score for a specific data sample locates at the intersection part of the pdfs for two classes, then there is high uncertainty about the correct prediction of the class that the data sample belongs to [35]. To alleviate this kind of indetermination, it is better to jointly consider the scores from successive walking cycles. Hence, if we can estimate the pdf of the scores of specific class, the final decision can be made since it is reasonable to assume the output scores are independent and identically distributed.

VI. Conclusion

We have proposed a dual-modal attention-enhanced deep learning model for gait classification and H&Y scale rating of PD patients. The effectiveness of the proposed model was evaluated on three datasets via a 5-fold cross-validation method. In terms of accuracy, the best result for the problem of differentiating PD patients from normal COs was obtained on the Ga dataset [26], where a sensitivity, specificity, and accuracy was 99.35%, 99.23%, and 99.31% respectively. For the problem of classifying PD patients with different H&Y scores, the best result was obtained on the Si dataset [24] with an accuracy of 99.01%. These results indicate that the proposed model, if trained with more clinical gait data samples, has the potentials to provide better ratings of the PD patients, which is very helpful for the clinicians to make a better rehabilitation program.

Acknowledgment

This work is supported by National Natural Science Foundation of China (NSFC) for Youth (grant number 61701483); Anhui Provincial Natural Science Foundation (grant number 1608085MF136).

References

[1] W. C. Koller, S. Glatt, B. Vetere-Overfield, and R. Hassanein, “Falls and Parkinson's disease,” *Clinical neuropharmacology,* vol. 12, no. 2, pp. 98-105, 1989.

[2] A. Reeve, E. Simcox, and D. Turnbull, “Ageing and Parkinson's disease: why is advancing age the biggest risk factor?,” *Ageing research reviews,* vol. 14, pp. 19-30, 2014.

[3] S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. Growdon, D. Standaert, M. Akay, J. Dy, M. Welsh, and P. Bonato, “Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors,” *IEEE transactions on information technology in biomedicine,* vol. 13, no. 6, pp. 864-873, 2009.

[4] K. Seppi, D. Weintraub, M. Coelho, S. Perez‐Lloret, S. H. Fox, R. Katzenschlager, E. M. Hametner, W. Poewe, O. Rascol, and C. G. Goetz, “The Movement Disorder Society evidence‐based medicine review update: treatments for the non‐motor symptoms of Parkinson's disease,” *Movement Disorders,* vol. 26, no. S3, pp. S42-S80, 2011.

[5] L. M. Cunningham, C. D. Nugent, G. Moore, D. D. Finlay, and D. Craig, “Computer-based assessment of movement difficulties in Parkinson's disease,” *Computer Methods in Biomechanics and Biomedical Engineering,* vol. 15, no. 10, pp. 1081-1092, 2012.

[6] N. Giladi, J. Tal, T. Azulay, O. Rascol, D. J. Brooks, E. Melamed, W. Oertel, W. H. Poewe, F. Stocchi, and E. Tolosa, “Validation of the freezing of gait questionnaire in patients with Parkinson's disease,” *Movement Disorders,* vol. 24, no. 5, pp. 655-661, 2009.

[7] S. Fahn, *Recent developments in Parkinson's disease*: Raven Pr, 1986.

[8] J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, “Gait variability and basal ganglia disorders: Stride‐to‐stride variations of gait cycle timing in parkinson's disease and Huntington's disease,” *Movement Disorders,* vol. 13, no. 3, pp. 428-437, 1998.

[9] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Troster, “Wearable assistant for Parkinson’s disease patients with the freezing of gait symptom,” *Information Technology in Biomedicine, IEEE Transactions on,* vol. 14, no. 2, pp. 436-446, 2010.

[10] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. Vingerhoets, and K. Aminian, “Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system,” *Biomedical Engineering, IEEE Transactions on,* vol. 54, no. 2, pp. 313-322, 2007.

[11] M. G. Tsipouras, A. T. Tzallas, G. Rigas, S. Tsouli, D. I. Fotiadis, and S. Konitsiotis, “An automated methodology for levodopa-induced dyskinesia: assessment based on gyroscope and accelerometer signals,” *Artificial Intelligence in Medicine,* vol. 55, no. 2, pp. 127-135, 2012.

[12] J. L. Dideriksen, F. Gianfelici, L. Z. P. Maneski, and D. Farina, “EMG-based characterization of pathological tremor using the iterated Hilbert transform,” *Biomedical Engineering, IEEE Transactions on,* vol. 58, no. 10, pp. 2911-2921, 2011.

[13] G. Catarina, D. Josefa, C. Guilherme, A. T. Santos, R. M. Fernandes, A. Daisy, G. Nilza, M. Helen, I. Tom, and D. Joy, “A systematic review of the characteristics and validity of monitoring technologies to assess Parkinson’s disease,” *Journal of Neuroengineering & Rehabilitation,* vol. 13, no. 1, pp. 1-10, 2016.

[14] D. L. Harrington, K. Y. Haaland, and R. T. Knight, “Cortical networks underlying mechanisms of time perception,” *Journal of Neuroscience,* vol. 18, no. 3, pp. 1085-1095, 1998.

[15] T. D. Pham, “Pattern Analysis of Computer Keystroke Time Series in Healthy Control and Early-Stage Parkinson's Disease Subjects using Fuzzy Recurrence and Scalable Recurrence Network Features,” *Journal of Neuroscience Methods*, 2018.

[16] S. H. Lee, and J. S. Lim, “Parkinson’s disease classification using gait characteristics and wavelet-based feature extraction,” *Expert Systems*

with Applications, 2012.

[17] M. R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease," *Biomedical Signal Processing and Control*, 2012.

[18] R. Peng, E. Karahan, C. Chao, R. Luo, Y. Geng, J. F. B. Bayard, M. L. Bringas, D. Yao, K. M. Kendrick, and P. A. Valdes-Sosa, "Gait Influence Diagrams in Parkinson's Disease," *IEEE Transactions on Neural Systems & Rehabilitation Engineering,* vol. 25, no. 8, pp. 1257-1267, 2017.

[19] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Future Generation Computer Systems,* vol. 83, pp. 366-373, 2018.

[20] T. Aşuroğlu, K. Açıcı, C. B. Erdaş, M. K. Toprak, H. Erdem, and H. Oğul, "Parkinson's disease monitoring from gait analysis via foot-worn sensors," *Biocybernetics & Biomedical Engineering*, 2018.

[21] J. Camps, A. Samà, M. Martín, D. Rodríguez-Martín, C. Pérez-López, J. M. M. Arostegui, J. Cabestany, A. Català, S. Alcaine, and B. Mestre, "Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit," *Knowledge-Based Systems*, 2018.

[22] Y. Xia, J. Zhang, Q. Ye, N. Cheng, Y. Lu, and D. Zhang, "Evaluation of deep convolutional neural networks for detection of freezing of gait in Parkinson's disease patients," *Biomedical Signal Processing and Control,* vol. 46, pp. 221-230, 2018/09/01/, 2018.

[23] Y. Nancy Jane, H. Khanna Nehemiah, and K. Arputharaj, "A Q-backpropagated time delay neural network for diagnosing severity of gait disturbances in Parkinson's disease," *Journal of Biomedical Informatics,* vol. 60, pp. 169-176, 2016/04/01/, 2016.

[24] S. Frenkeltoledo, N. Giladi, C. Peretz, T. Herman, L. Gruendlinger, and J. M. Hausdorff, "Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson's disease," *Movement Disorders,* vol. 20, no. 9, pp. 1109-1114, 2010.

[25] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "A hybrid spatio-temporal model for detection and severity rating of Parkinson's disease from gait data," *Neurocomputing,* vol. 215, pp. 1-8, 2018.

[26] G. Yogev, N. Giladi, C. Peretz, S. Springer, E. S. Simon, and J. M. Hausdorff, "Dual tasking, gait rhythmicity, and Parkinson's disease: which aspects of gait are attention demanding?," *European Journal of Neuroscience,* vol. 22, no. 5, pp. 1248-1256, 2010.

[27] R. Djaldetti, I. Ziv, and E. Melamed, "The mystery of motor asymmetry in Parkinson's disease," *Lancet Neurology,* vol. 5, no. 9, pp. 796-802, 2006.

[28] B. Su, R. Song, L. Guo, and C. Yen, "Characterizing gait asymmetry via frequency sub-band components of the ground reaction force," *BIOMEDICAL SIGNAL PROCESSING AND CONTROL,* vol. 18, pp. 56-60, 2015.

[29] F. Liao, J. Wang, and P. He, "Multi-resolution entropy analysis of gait symmetry in neurological degenerative diseases and amyotrophic lateral sclerosis," *Medical Engineering & Physics,* vol. 30, no. 3, pp. 299-310, 2008.

[30] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation,* vol. 101, no. 23, pp. E215, 2000.

[31] J. M. Hausdorff, J. Lowenthal, T. Herman, L. Gruendlinger, C. Peretz, and N. Giladi, "Rhythmic auditory stimulation modulates gait variability in Parkinson's disease," *European Journal of Neuroscience,* vol. 26, no. 8, pp. 2369-2375, 2010.

[32] C. G. G. Chairperson, W. P. W. Committee, O. R. W. Committee, C. C. E. Consultant, and N. G. E. Consultant, "Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations The Movement Disorder Society Task Force on rating scales for Parkinson's disease," *Mov Disord,* vol. 19, no. 9, pp. 1020-1028, 2004.

[33] B. Post, M. P. Merkus, R. M. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?," *Movement disorders: official journal of the Movement Disorder Society,* vol. 20, no. 12, pp. 1577-1584, 2005.

[34] Y. Zhao, and S. Zhou, "Wearable Device-Based Gait Recognition Using Angle Embedded Gait Dynamic Images and a Convolutional Neural Network," *Sensors,* vol. 17, no. 3, pp. 478, 2017.

[35] M. Gadaleta, and M. Rossi, "IDNet: Smartphone-based Gait Recognition with Convolutional Neural Networks," *Pattern Recognition,* vol. 74, pp. 25-37, 2018.

[36] S. Hochreiter, and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[37] W. Pei, T. Baltrusaitis, D. M. J. Tax, and L. P. Morency, "Temporal Attention-Gated Model for Robust Sequence Classification," pp. 820-829, 2016.

[38] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification." pp. 1480-1489.

[39] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research,* vol. 12, no. 7, pp. 257-269, 2011.

[40] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "A Hybrid Spatio-temporal Model for Detection and Severity Rating of Parkinson's Disease from Gait Data," *Neurocomputing*, 2018.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P.

Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[43] F. Wilcoxon, "Individual comparisons by ranking methods," *Breakthroughs in statistics*, pp. 196-202: Springer, 1992.

[44] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Future Generation Computer Systems,* vol. 83, 2018.

[45] S. P. Moustakidis, J. B. Theocharis, and G. Giakas, "Subject recognition based on ground reaction force measurements of gait signals," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 38, no. 6, pp. 1476-1485, 2008.

[46] T. C. Pataky, T. Mu, K. Bosch, D. Rosenbaum, and J. Y. Goulermas, "Gait recognition: highly unique dynamic plantar pressure patterns among 104 individuals," *Journal of The Royal Society Interface,* vol. 9, no. 69, pp. 790-800, 2012.

[47] G. Xu, Z. Wang, H. Huang, W. Li, C. Liu, and S. Liu, "A Model for Medical Diagnosis Based on Plantar Pressure," *arXiv preprint arXiv:1802.10316*, 2018.

[48] A. Muniz, H. Liu, K. Lyons, R. Pahwa, W. Liu, F. Nobre, and J. Nadal, "Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait," *Journal of biomechanics,* vol. 43, no. 4, pp. 720-726, 2010.