IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Pairwise Deep Ranking Model for Relative Assessment of Parkinson's Disease Patients from Gait Signals

Burçin Buket Oğul[1,2], Suat Özdemir[1], Member, IEEE

[1]Department of Computer Engineering, Hacettepe University, Ankara, TR06810 Turkey
[2]Department of Computer Engineering, Çankaya University, Ankara, TR06810 Turkey

Corresponding author: B. Buket Oğul (e-mail: buket.ogul@cankaya.edu.tr).

B. B. Oğul was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the 2214-A program.

ABSTRACT Continuous monitoring of the symptoms is crucial to improve the quality of life for patients with Parkinson's Disease (PD). Thus, it is necessary to objectively assess the PD symptoms. Since manual assessment is subjective and prone to misinterpretation, computer-aided methods that use sensory measurements have recently been used to make objective PD assessment. Current methods follow an absolute assessment strategy, where the symptoms are classified into known categories or quantified with exact values. These methods are usually difficult to generalize and considered to be unreliable in practice. In this paper, we formulate the PD assessment problem as a relative assessment of one patient compared to another. For this assessment, we propose a new approach to the comparative analysis of gait signals obtained via foot-worn sensors. We introduce a novel pairwise deep-ranking model that is fed by data from a pair of patients, where the data is obtained from multiple ground reaction force sensors. The proposed model, called *Ranking by Siamese Recurrent Network with Attention*, takes two multivariate time-series as inputs and produces a probability of the first signal having a higher continuous attribute than the second one. Our detailed performance analysis shows that the accuracy of pairwise ranking predictions can reach up to 82% with an AUROC of 0.89 with ten-fold cross validation. The model outperforms the previous methods for PD monitoring when run in the same experimental setup. To the best of our knowledge, this is the first study that attempts to relatively assess PD patients using a pairwise ranking measure on sensory data. The model can serve as a complementary model to computer-aided prognosis tools by monitoring the progress of the patient during the applied treatment.

INDEX TERMS Siamese Network, Long Short-Term Memory, Parkinson's Disease, Gait Analysis, Pairwise Ranking.

I. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder of aging that affects dopamine-producing neurons in the substantia nigra area of the brain [1]. Although there is currently no known cure for the disease, patients are treated with medications to relieve symptoms such as tremor, bradykinesia, dyskinesia, and walking disorders to maintain and/or improve their quality of life [2-5]. To monitor PD patients, it is necessary to rate the degree of the severity of the disease. These measurements are based on the evaluation of motor manifestations, assessment of the difficulties experienced in daily living, and symptomatic response to medication [6]. Based on interviews by an examiner or a patient's self-assessment, scales such as the Unified Parkinson Disease Rating Scale (UPDRS) [7] provide estimations of the symptoms. UPDRS consists of four subscales each of which covers measurements related to "Mentation, Behavior, and Mood", "Activities of Daily Living", "Motor Examination," and "Complications of Therapy". However, the ratings in both the UPDRS and its subscales are not interval scales; that is, there are no quantitative distances between score values.

As an alternative to subjective assessments, measurements that are based on a set of sensors capturing the physical characteristics of human motion and/or physiological signals are also used to infer the state of the patient in terms of predefined criteria [8]. A common method for sensor-based evaluation is to automatically classify patients into one of the

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/ACCESS.2021.3136724, IEEE Access

categories using conventional machine learning algorithms fed by a set of extracted features from sensory signals [9]. Lee et al. [10] used gait characteristics to classify samples as PD or not. Wavelet features extracted using gait signals were then used to feed a neural network with weighted fuzzy membership functions so that they could distinguish PD patients from healthy control subjects. Daliri et al. [11] used support vector machines (SVM) applied to ground reaction force (GRF) signal features extracted by short-time Fourier transform (STFT) and reported 91.2% precision. Jane et al. [12] who used the Hoehn and Yahr (H&Y) scale to model a Q-backpropagated time-delay neural network for the data collected by GRF sensors achieved slightly better than the results obtained by Daliri et al. Ertugrul et al. [13] proposed a novel one-dimensional local binary pattern (LBP) approach, called shifted 1DLBP, to extract statistical features from histograms of gait signals. Joshi et al. [14] extracted wavelet-based features to be used in SVM-based classification. This hybrid method which combines the wavelet transform and SVM achieves similar accuracy results with [11] and [12]. Acici et al. [15] used a random forest (RF) algorithm for PD classification tasks based on the extracted set of features in the time and frequency domains. The RF algorithm resulted in 98.04% classification accuracy. Patel et al. [5] proposed estimating PD symptom severity with accelerometers. The authors classified the severity of different symptoms with an SVM using data gathered from an accelerometer. Their study presents promising results for the severity classification of symptoms such as tremor, bradykinesia, or dyskinesia. Although, this approach provides a categorical prediction, it is not sufficient for a quantitative assessment of PD symptoms. In recent studies [14,15], several researchers applied deep learning techniques, such as convolutional neural networks (CNN) and recurrent neural network (RNN), instead of using hand-crafted features. Zhao et al. [16] used a two-channel model that combines long short-term memory (LSTM) and CNN to learn the spatio-temporal information behind the data. Xia et al. [17] proposed a dual-modal attention enhanced deep learning model for quantification of Parkinson's disease features by modeling a CNN separately on the right and left gait, followed by an LSTM layer.

Classification-based evaluations provide limited understanding of the progress of the patient, since the categories are often binary, that is, in the form of presence/absence of defined symptoms [18]. A potential increase or decrease in the severity of symptoms cannot be inferred. One solution to this is to employ similar machine learning algorithms in a regression setup to directly quantify the severity, which serves as an absolution assessment of the symptoms [8, 9, 14]. Asuroglu et al. [18] adapted their random forest model in a regression setup, instead of classification in [15], to predict the exact value of the severity of PD symptoms from gait signals. Although this can provide a more precise evaluation of the current state of the patient, the generalization ability of such methods is limited due to the unavailability of a sufficient number of training samples with respect to the high granularity of grading scales used [19]. In fact, continuous labels that represent the severity is sparse to predict the model parameters accurately. Another limitation of the studies that use UPDRS values in a regression setup is that UPDRS and its subscales are not interval scales [6]. Since the distances between scores are not quantitative, regression-based approaches are not descriptive enough. Furthermore, severity assessment is usually considered to be subjective since they are not directly associated with a clinical test but the result of an expert evaluations. Therefore, predicted value of the severity is not found to be clinically reliable [6].

To overcome these limitations, we propose a novel model for the relative assessment of PD patients using gait signals acquired by foot-worn GRF sensors. We opt to use the scores of PD patients to be a ranking measure rather than a precise range change. This assessment is considered less prone to changes in different expert evaluations as Perlmutter et al. [6] suggested. Pairwise ranking labels were obtained by comparing the overall severity of PD symptoms in term of UPDRS. Given two patients' data as input, the model is asked to predict whether the first patient has more severe symptoms than the second.

In general, pairwise models have been studied extensively in computer science literature. Some of these studies can be grouped into multi-stream learning models, such as Siamese or triplet networks, for 'classification' of objects [20]. These models attempt to learn a number of parameters to keep the pairs in the same class together and the pairs in the opposite classes further. Final model can assign the query sample into a class based on the pairwise scores with training samples. Another group of studies, which is called 'learning to rank' deals with 'retrieval' of similar objects from a repository [21]. Here, a pairwise model aims to learn how similar two inputs are based on a training set of similarity ranks. In our study, we address a completely different problem; we aim to predict if the first sample is greater than the second sample in terms of an independent continuous label, which measures any quantity of the input signal. This problem has been tackled very recently for pairwise 'ranking' of image data in terms of their quality [22]. We have also recently seen some applications of pairwise ranking in video data for action quality assessment in sport activities as well [23].

To the best of our knowledge, present study is the first attempt for pairwise ranking of multi-variate time-series signals. Because of their non-spatial temporal characteristics, the models in image data cannot be directly inherited for time-series signals. Here, we address this challenge using a novel pairwise deep learning model. The model is an adoption of Siamese recurrent neural networks [24] for the task of pairwise ranking instead of pairwise similarity inference. This requires redefinition of the decision layer with a modified loss function. We offer a probabilistic loss layer for this purpose. The recurrent layer is implemented as

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/ACCESS.2021.3136724, IEEE Access

an LSTM enhanced by an attention mechanism to capture remote dependencies in input signals relevant to gait skills. For convenience, the model will be referred to as Ranking by Siamese Recurrent Network with Attention (RSRNA) in the rest of the paper.

The contribution of the study is hence twofold. From an application perspective, the present study introduces the idea of relative assessment of PD patients by analyzing motion signals. This approach promotes two applications: (1) prognosis by monitoring the progress of the same patient during applied treatments, (2) personalized medicine by referring to the success/failure stories of other relevant patients. The second contribution of the study is that we propose a novel pairwise ranking model, called RSRNA, for multi-variate time-series signals and evaluate it using real world datasets. The experimental results show that, compared to existing methods, the proposed RSRNA model provides better results for PD patient monitoring in terms of pairwise ranking accuracy.

II. APPROACH

A. RSRNA ARCHITECTURE FOR PAIRWISE RANKING

Given two PD patients, labeled by *m* and *n*, with their gait data of $\boldsymbol{x^m}$ and $\boldsymbol{x^n}$, which are multi-variate time-series of GRF signals measured during the experiment, the task is to determine which patient has more severe PD symptoms in terms of UPDRS scale. We denote this output by p_{mn} where;

$$p_{mn} = \begin{cases} 1 & m \text{ has more severe symptoms than } n \\ 0.5 & m \text{ and } n \text{ have same level of severity} \\ 0 & m \text{ has less severe symptoms than } n \end{cases} \quad (1)$$

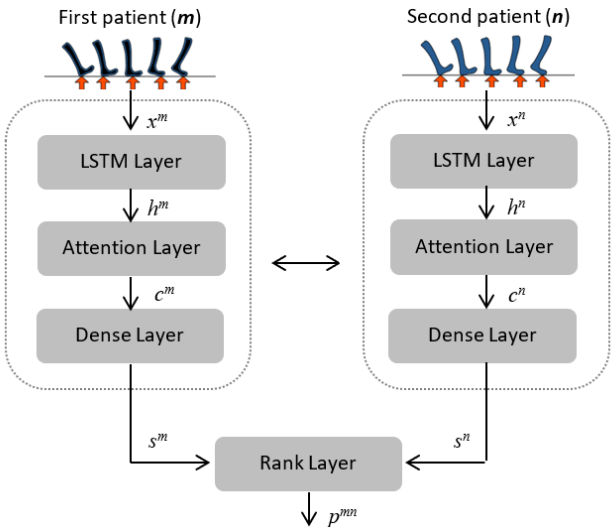


FIGURE 1. RSRNA model for pairwise ranking of PD patients from gait signals.

We interpret this as the probability of the first patient having more severe symptoms than the second. Our purpose is to learn a model that minimizes the probabilistic loss in human-annotated samples for PD severity. As a consequence, we propose a novel framework called RSRNA, which is based on a Siamese network of attention-enhanced LSTMs integrated with a probabilistic ranking layer in which the layer has the ability to consider the case of the equivalence of disease severity as well. The framework takes two gait signals, $\boldsymbol{x^m}$ and $\boldsymbol{x^n}$, of patients as input and reports a pairwise rank between them (Figure 1).

We feed an LSTM at one input of the Siamese network which is a powerful type of RNN used in deep learning [25] to model temporal data in the form of multi-variate time-series. This prevents the vanishing gradient problem which is the main limitation of RNN [26]. Since our data involves long term dependencies, we prefer to use LSTM to model single gait behavior of each patient. Ignoring the superscript, *m* or *n* above, defining the stream, i.e. the patient, LSTM can be considered a recurrent relation by Equation 2.

$$h_t = LSTM\ (h_t - 1, x_t) \quad (2)$$

Here, x_t refers to the vector of GRF measurements at time *t*. At every time step *t*, LSTM outputs a hidden vector h_t that reflects the disease representation by the gait signal at position *t*. The LSTM model is parameterized by output, input and forget gates, controlling the information flow within the recursive operation. It is implemented by following composite functions:

$$i_t = \sigma(W_i\, x_t + U_i\, h_{t-1} + b_i) \quad (3)$$
$$f_t = \sigma(W_f\, x_t + U_f\, h_{t-1} + b_f) \quad (4)$$
$$o_t = \sigma(W_o\, x_t + U_o\, h_{t-1} + b_o) \quad (5)$$
$$\tilde{c}_t = \tanh(W_c\, x_t + U_c\, h_{t-1} + b_c) \quad (6)$$
$$c_t = \sigma(i_t \circ \tilde{c}_t + f_t \circ c_{t-1}) \quad (7)$$
$$h_t = o_t \circ \tanh(C_t) \quad (8)$$

Here, σ is the logistic sigmoid function, i, f, o and c refer to input gate, forget gate, output gate, and cell input activation records, respectively.

B. LEARNING VIA ATTENTION

An attention mechanism has been recently introduced to improve conventional encoder-decoder structures, to maximize the contribution of the relevant encoding context vectors and minimize those of irrelevant vectors while building the decoding context [27]. The gait signals acquired from PD patients usually involve long series of measurements. Local variables in different temporal positions are expected to have varying effects on the PD symptoms [18]. Therefore, an attention layer is used in the framework to assign weight (importance) to each position associated with the PD symptoms to be predicted at the end. We implement an attention layer that uses an attention function to assign weight to each hidden state produced by LSTM layer. The weighted distribution of hidden states is

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/ACCESS.2021.3136724, IEEE Access

used as a new representation of input signals. We calculate an attention function, denoted by u_t, for each hidden state h_t, t=1,…,T, as follows:

$$u_t = \tanh(W_s h_i + b) \quad (9)$$

where W_s is an attention hidden weight matrix and b is a bias parameter. This function allows to calculation of a number of weight parameters, denoted by α_t, using the following equation;

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t'=1}^{T} \exp(u_{t'})} \quad (10)$$

These weights are used to produce a context vector c, which will be forwarded to the next layer:

$$c = \sum_{t=1}^{T} h_t \alpha_t \quad (11)$$

Before pairwise ranking, a fully connected layer takes the vector of skill representation at the output of the LSTM, c^m for any of the input m, and transforms it into a scalar, s^m, which is directly comparable with the output, s^m, at the other end of the Siamese network.

C. PAIRWISE RANKING FOR RELATIVE ASSESSMENT

A typical Siamese network models the loss function to infer the similarity between input signals [19]. Instead of similarity inference, we aim to rank these inputs. Hence, the framework allows the Siamese model to handle relative comparison of inputs instead of their direct evaluation for similarity. This is achieved by a rank layer adapted from a recent probabilistic loss function introduced in the *RankNet* approach [28]. In RankNet, the authors employ a probabilistic cost function that uses a pair of sample items to learn how to rank them. Their approach implements this cost function through a neural network optimized by gradient descent. In our case, we represent the pairwise rank between two patients having a PD disease by P_{mn} in which the probability of patient m having more severe symptoms than the patient n. We denote the posterior probability distribution $P_{ij} = P(i \rangle j)$, where $\rangle$ refers to the higher severity of i to j, and P_{ij} is assumed to be desired target values for those posteriors, such that $P_{ij} \in \{1, 0.5, 0\}$. Then, our aim is to minimize the distance between these two entities. We use a cross entropy cost function, C_{ij} to measure the closeness between two probability distributions, given by;

$$C_{ij} \equiv C_{(o_{ij})} = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij}) \quad (12)$$

Letting O_{mn} be the difference between rank orders of m and n, the probabilities are modelled by:

$$P_{ij} \equiv \frac{e^{o_{ij}}}{1 + e^{o_{ij}}} \quad (13)$$

Then, following the above definitions, the final cost function becomes:

$$C_{ij} = -\bar{P}_{ij} o_{ij} + \log(1 + e^{o_{ij}}) \quad (14)$$

III. EXPERIMENTS & RESULTS

A. DATASET

A public PhysioNet dataset (https://physionet.org/content/gaitpdb/1.0.0/) was used in this study [29]. The dataset contains the measurements of the gait signals of 93 PD patients and 73 healthy controls. Both groups have an average age of 66.3 years. Subjects wore eight sensors in each of their feet that measure force while performing their usual walking for approximately 2 minutes on level ground. The position of the sensors was as follows: assuming a person stands up with two legs parallel to each other, the point of origin is exactly in the middle of the legs and the person faces toward the positive side of the Y axis. X and Y coordinates of each sensor are displayed in Table 1. The sensors measured the force on the feet in Newtons as a function of time. The dataset includes demographics information, measures of disease severity in terms of different metrics such as Hoehn & Yahr staging, the UPDRS, and other related measures. As Daliri [9] stated, since the reaction force on the feet varies in time throughout a walking activity based on personal gait patterns, it could be leveraged as a convenient resource for individual gait analysis. In our study, we use the digitized outputs of these 16 sensors to analyze the dynamics and characteristics of these multivariate time series.

TABLE I
PLACEMENT OF INDIVIDUAL GRF SENSORS IN X AND Y COORDINATES UNDER THE FEET

Sensor	X	Y	Sensor	X	Y
Left 1	-500	-800	Right 1	500	-800
Left 2	-700	-400	Right 2	700	-400
Left 3	-300	-400	Right 3	300	-400
Left 4	-700	0	Right 4	700	0
Left 5	-300	0	Right 5	300	0
Left 6	-700	400	Right 6	700	400
Left 7	-300	400	Right 7	300	400
Left 8	-500	-800	Right 8	500	800

B. IMPLEMENTATION

We used an LSTM network to capture temporal representations in PD symptoms. For the attention layer, we followed the previous implementation by Yang et al. (30) with the suggested parameter set. A sigmoid activation layer was used to model the probabilistic rank layer, which is followed by a binary cross-entropy loss function in the

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/ACCESS.2021.3136724, IEEE Access

training model. We used the following hyper-parameters for learning by a stochastic gradient descent algorithm: a learning rate of 0.001, a unit size of 64 with a single hidden layer, and a batch size of 2. The framework was implemented in Keras using TensorFlow backend.

C. EVALUATION

The original dataset was reorganized to create new samples according to our relative assessment strategy. Each sample in the new dataset was composed of a pair of patients with their raw gait signals and a pairwise ranking label between them, which can be 1, 0.5 or 0. These ranking labels were obtained by comparing the overall severity of PD symptoms in term of UPDRS. The samples without UPDRS annotations were removed from the dataset. We assessed the accuracy of predictions using a ten-fold cross-validation setup. In this setup, the pairs between 1/10 of the patients were used for testing, and the remaining pairs were used for training. It should be noted that test samples included both pairs in which neither video has been used in a pair for training and the pairs in which the other video was used for training in a different pairing. To evaluate the performance, the following metrics were used.

Pairwise ranking accuracy (Acc): This is the percentage of correctly ordered pairs generated by each testing fold. Depending on whether the rank layer models the equivalence of PD severities of two patients, two different accuracy results may be reported. When the equivalence is considered, the accuracy gives the evaluation of ternary ranking performance. Otherwise, it evaluates binary ranking. Table 2 lists the conditions for the correct ordering of a pair (m, n) in binary and ternary cases. We used $\varepsilon = 0.01$ in our evaluations.

TABLE II
EVALUATIONS OF CORRECT PREDICTIONS AND ASSOCIATED GROUND TRUTH FOR DIFFERENT PAIRWISE RANKING SCHEMES

Ranking scheme	p_{mn}	Ground truth
Ternary	$\geq 0.5+\varepsilon$	$m \rangle n$
	$\geq 0.5-\varepsilon$ and $< 0.5+\varepsilon$	$m \equiv n$
	$< 0.5-\varepsilon$	$m \langle n$
Binary	≥ 0.5	$m \rangle n$
	< 0.5	$m \langle n$

Area under receiver operating characteristic (ROC) curve (AUC): An ROC curve plots true positive (TP) rate versus false positive (FP) rate at different classification thresholds. In our binary ranking case, a positive sample is a pair for which first patient have more severe symptoms than the second patient. This sample is referred as TP if it is correctly predicted, and as FP otherwise. AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). For perfect classification performance, the ROC curve is expected to be a full rectangle, and the AUC is expected to be 1. AUC is usually considered as an objective evaluation criterion for imbalanced datasets since it provides an aggregate measure of performance across all possible classification thresholds. Since the threshold change in classification phase may affect the performance of the model, we additionally use ROC curves to assess the robustness of our final model with some intermediate models using different sub-modules and model parameters. This enables us to choose the best model before comparing against other algorithms.

Boxplots: A boxplot is a graph that provides an indication of how the values in the data are spread out. It displays the distribution of data on a vertical bar with indicators for minimum, first quartile, median, third quartile and maximum. We used the boxplot to display the spread of predicted probabilities for higher severity of the first patient in different ranking labels. We expected that the probabilities would approach 1 when the first sample in the pair had a higher severity, and they would approach 0 when the first sample had lower severity. When equivalence is considered, the probabilities should accumulate around 0.5 for the pair samples with same severity. For each case, we expected small fluctuations around expected probabilities.

D. FINDINGS

In ten-fold cross-validation experiments, the RSRNA model achieved a binary pairwise ranking accuracy of 81% with an AUC of 0.878 and a ternary pairwise ranking accuracy of 78% with an average AUC of 0.862. Figure 2 shows the ROC curve for the proposed model when applied to binary pairwise ranking. Note that the ROC curve is not directly applicable for the ternary ranking scheme, but an AUC can be reported from the average of individual curves for all class labels. The boxplots of the predicted probabilities against pairwise ranking labels are shown in Figure 3.

In Figure 2, the performance of the model is also discerned when the attention layer was removed. The figure shows that attention enhancement has a significant contribution in the prediction performance. Reported ranking accuracy and AUC decreased to 74% and 0.817 when the attention mechanism was eliminated.

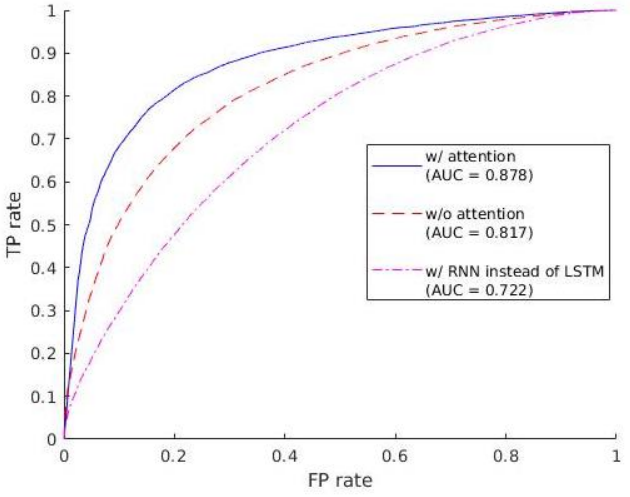
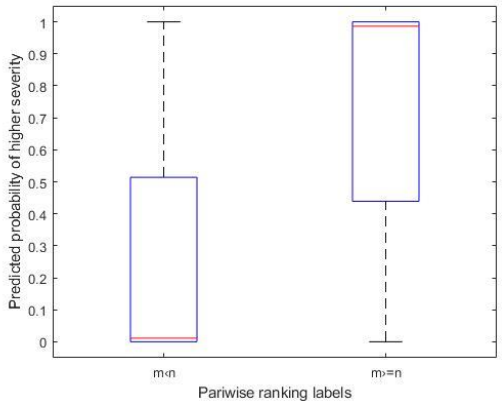


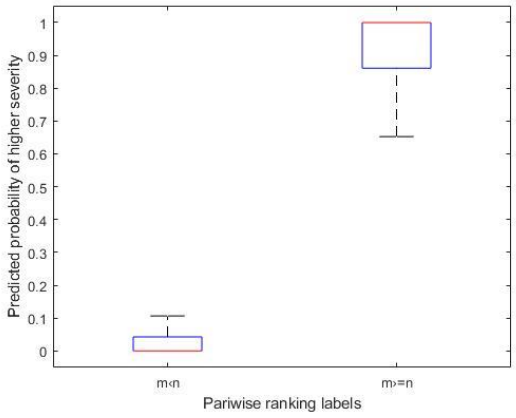
FIGURE 2. ROC curves for binary pairwise ranking by RSRNA model using alternative sub-models; (1) with attention, (2) without attention and (3) using RNN instead of LSTM.

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/ACCESS.2021.3136724, IEEE Access

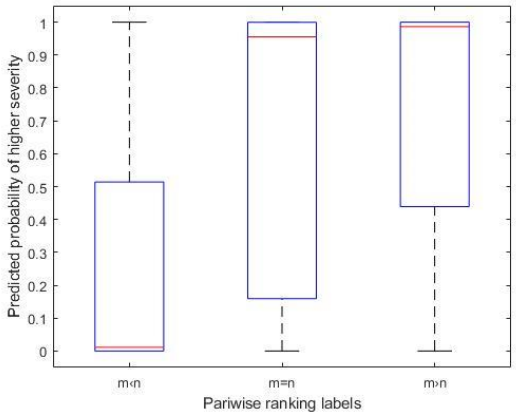
The boxplots shown in Figure 3 justify the argument that the attention mechanism is useful in detecting similarities between gait signals. As shown, using attention lowered fluctuations in the predictions in both binary (Figure 3.a-b) and ternary (Figure 3.c-d) ranking schemes.



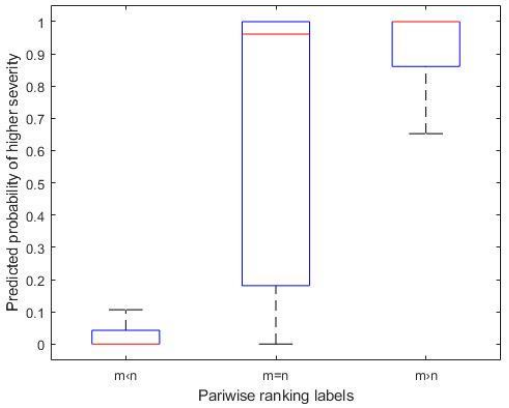
(a)



(b)



(c)



(d)

FIGURE 3. Boxplots of predicted probabilities against pairwise ranking labels for (a) binary ranking without attention, (b) binary ranking with attention, (c) ternary ranking without attention, and (4) ternary ranking with attention.

Selection of LSTM was evaluated by replacing the sub-model in this layer with a simpler RNN and evaluating the performance of the overall model in the same experimental setup. RNN was compiled with the following hyperparameters: hyperbolic tangent for activation, "orthogonal" initializer for recurrent initialization, "glorot_uniform" initializer for kernel initialization, and a unit size of 64. The model with RNN achieved a ranking accuracy of 63% with an AUC of 70.6 in the ternary scheme and a ranking accuracy of % with an AUC of 66 in binary scheme. In either of the cases, the performance of the model with RNN was lower than those with LSTM. This result justifies the fact that LSTM is a better choice in modeling temporal behavior of gait signals. The results with different configurations are summarized in Table 3.

TABLE III
JUSTIFICATION OF THE PROPOSED MODEL BY COMPARISON OF RELATIVE ASSESSMENT (PAIRWISE RANKING) PERFORMANCES OF DIFFERENT ARCHITECTURES WITH ALTERNATIVE SUB-MODELS.

Methods	Binary ranking		Ternary ranking	
	Acc	AUC	Acc	AUC (avg)
RSRNA	81%	0.878	78%	0.862
RSRNA – *without attention*	74%	0.817	71%	0.796
RSRNA – *with RNN instead of LSTM*	66%	0,722	63%	0.706

Since relative (pairwise) assessment of PD patients is proposed for the first time in this study, there is no existing work with which we can perform a direct comparison. However, we can refer to previous studies to create a number of baselines to benchmark our method.

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/ACCESS.2021.3136724, IEEE Access

Previous Method 1: Daliri [11] classified patients as PD or not using an SVM with frequency domain features. Similarly, we reconfigured Daliri's [11] model such that an SVM was fed by the fusion of frequency-domain features of two patients to be ranked. These features were extracted using fast Fourier transforms of gait signals.

Previous Method 2: Asuroglu et al. [18] attempted to quantify the exact value of symptoms in UPDRS scale. We reconfigured the model represented in this study so that it can report the pairwise rank when a pair of patients' data is presented in the input. To do this, we concatenated individual time-domain feature sets extracted from each patient sample to construct a new sample and feed a random forest model in the classification setup.

Previous Method 3: Xia et al. [17] used a model that combines a CNN followed by an LSTM layer. In this baseline, we used only the CNN section of the study to model the spatial features of the data. To adopt the spatial section of this model to our problem, we concatenated two input signals vertically and fed a CNN architecture, which included two convolutional layers, two max pooling layers, and a fully connected layer to classify if the first sample has a higher severity than the second. The convolution kernel in the two convolutional layers were both 3 × 3 and outputs 32 feature maps.

Previous Method 4: Using the same study as the third baseline, we modeled both spatial and temporal features of the dataset. We used the concatenation of two input signals to feed a CNN that had two convolution layers with the same parameters as Baseline 3, followed by an LSTM that had a length of 256 for hidden state vector to classify which of the two signals had a higher severity than the other.

The evaluation results of ten-fold cross-validation experiments with different baseline models are applied for the binary ranking prediction at the UPDRS scale are displayed in Table 4. As shown in Table 4., RSNA outperforms all benchmarked methods in both ranking schemes.

Table 4 also shows the results when the ternary ranking was applied. RSRNA model still outperformed benchmarked studies in terms of Acc and AUC when the case of severity equivalence was considered.

TABLE IV
COMPARISON OF THE PROPOSED MODEL WITH PREVIOUS STUDIES IN TERMS OF THEIR RELATIVE ASSESSMENT (PAIRWISE RANKING) PERFORMANCES.

Method	Binary ranking		Ternary ranking	
	Acc	AUC	Acc	AUC (avg)
RSRNA (proposed model)	**81%**	**0.878**	**78%**	**0.862**
Previous Method 1* [11]	64%	0.623	58%	0.617
Previous Method 2* [18]	63%	0.744	59%	0.737
Previous Method 3* [17]	64%	0.698	61%	0.685
Previous Method 4* [17]	57%	0.579	55%	0.567

**These methods was reconfigured for pairwise ranking and re-implemented by the authors.*

Figure 4 shows the superiority of the current pairwise ranking over other methods. Bars on the left-hand side were generated through the current method, RSRNA, while the bars on the right show the outputs of the best baseline in terms of AUC (Previous Method 2 [18]) provided in Table 4. The dark lower parts of the bars represent the number of correctly classified pairs. This result indicates that even if the absolute differences between pairs are as low as below 5, RSRNA is quite successful in modeling the differences.

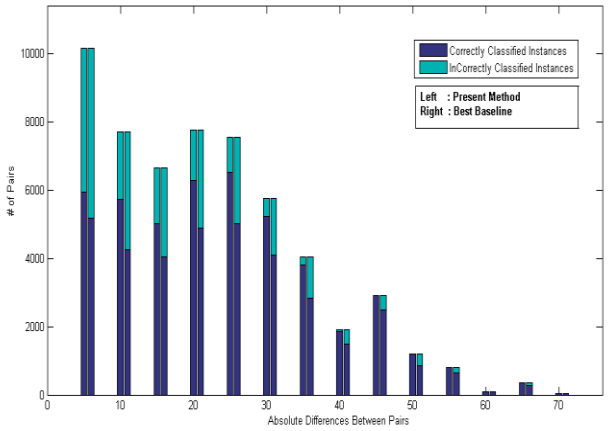


FIGURE 4. Bars of correctly classified pairs versus incorrectly classified ones based on patients' UPDRS scores. The left and right bars show the results of the present method and the best baseline, respectively.

IV. CONCLUSION

We introduced a novel approach for the relative assessment of the severity level of PD patients using gait sensors. To the best of our knowledge, this is the first attempt in the literature to assess PD patients by a pairwise comparison of gait signals. To this end, we proposed a novel deep learning architecture for pairwise ranking of multivariate time-series signals collected via GRF sensors worn under foot. According to the experimental results, the predictions were correlated with the clinical annotations. The accuracy of pairwise ranking predictions reached up to 81% with an AUC of 0.878 in ten-fold cross validation. The model outperformed the previous methods for PD monitoring when run in the same experimental setup.

The proposed RSRNA method can be considered as a generic model for several pairwise ranking tasks, as the inputs are multivariate time-series signals. While LSTM layer makes the model applicable for all sequential signals, attention enhancement extends its ability to adopt novel signals obtained from different measurement modalities. Proposed rank layer with probabilistic loss function allows the Siamese model to handle relative comparison of inputs instead of their direct evaluation for similarity. We expect that this model feasible for a wide range of applications, especially in the health domain, to compare patients based on their physiological recordings.

The relative assessment approach provides a more interpretable and reliable view of disease progress while overcoming the limitations caused by inconsistencies in subjective grading scales. This approach will promote two

applications. First, monitoring the progress of patients during applied treatments may support their prognosis and guide the organization of both preventive medicine and ongoing care practices [31]. As the present model allows comparison of patients' current data with their previous recordings, it can serve as a complementary model to new computer-aided prognosis tools. Second, this may support the *personalized medicine* effort by referring to the success/failure stories of the treatments of other relevant patients which can be obtained by retrieving similar cases using our RSRNA model. As the model is applicable to many other biomedical time-series signals, it may find applications in other health domains such as prognosing cardiovascular diseases using electrocardiograms [32] or monitoring patients in intensive care units via physiological vital signs [33].

Since PD patients usually suffer from the loss of basic motor abilities, remote monitoring is a recent challenge to provide satisfactory home care and clinical support. Our experiments showed that present model enables the relative assessment of current patient against others using wearable sensors, which can be easily used in home settings. Lack of multiple samples from individual patients prevented us to measure the performance of the system for assessing the progress of same person over time. This can be considered as a future clinical study. Providing multi-sensory data or video recordings used in remote monitoring of patients as inputs to the system may be another future aspect of the current study. Combining different modalities can be considered for developing an enhanced quality assessment system for PD patients.

REFERENCES

[1] Kalia L.V., Lang A.E. (2015) “Parkinson's disease”, The Lancet, 386;9996:896–912

[2] Bhidayasiri R., Martinez-Martin P. (2017), “Clinical Assessments in Parkinson's Disease: Scales and Monitoring”. International Review of Neurobiology 132:129–182

[3] Nicoletti A., Mostile G., Portar G., Luca A., Patti F., Zappia M., (2017) “Computer-assisted cognitive rehabilitation on freezing of gait in Parkinson’s disease: A pilot study”. Neuroscience Letters 654:38–41

[4] Kostek B., Kaszuba K., Zwan P., Robowski P., Slawek J. (2012) “Automatic assessment of the motor state of the Parkinson's disease patient--A case study”. Diagnostic Pathology 7(18)

[5] Patel S., Lorincz K., Hughes R., Huggins N., Growdon J., Standaert D., Akay M., Dy J., Welsh M., Bonato P. (2009) “Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors”. IEEE Transactions on Information Technology in Biomedicine 13;(6):864–873

[6] Perlmutter J.S. (2009) “Assessment of Parkinson disease manifestations”. Curr Protoc Neurosci Chapter 10:Unit 101

[7] Ramaker R., J. Marinus, A.M. Stiggelbout, B.J. van Hilten, (2002) “Systematic evaluation of rating scales for impairment and disability in Parkinson's Disease”, Mov Disord 17:867–876

[8] Aghanavesi S., Westin J., Bergquist F., Nyholm D., Askmark H, Sten Magnus A, Radu C (2020), “A multiple motion sensors index for motor state quantification in Parkinson’s disease”, Comput Methods Progr Biomed 189:105309

[9] Srivastava P., Shukla A., Vepakomma P., Bhansali N., Verma K. (2017), “A survey of nature-inspired algorithms for feature selection to identify Parkinson's disease”. Comput Methods Programs Biomed, 139:171–179

[10] Lee S.H., Lim J.S., (2012) “Parkinson’s disease classification using gait characteristics and wavelet-based feature extraction”. Expert Systems with Applications 39(8):7338–7344

[11] Daliri M.R., (2013) “Chi-square distance kernel of the gaits for the diagnosis of Parkinson’s disease”, Biomedical Signal Processing and Control 8(1):66–70

[12] Jane Y.N., Nehemiah H.K., Arputharaj K., (2016) “A Q-backpropagated time delay neural network for diagnosing severity of gait disturbances in Parkinson’s disease”. Journal of Biomedical Informatics 60:169–176

[13] Ertugrul O.F., Kaya Y., Tekin R., Almali M.N., (2016) “Detection of Parkinson’s disease by shifted one dimensional local binary patterns from gait”, Expert Systems With Applications 56:156–163

[14] Joshi D., Khajuria A., Joshi P., (2017) “An automatic non-invasive method for Parkinson's disease classification”, Comput Methods Programs Biomed, 145:35–145

[15] Acici K., Erdas C.B., Asuroglu T., Toprak M.K., Erdem H., Ogul H. (2017) “A random forest method to detect Parkinson's disease via gait analysis”. in: G Boracchi, L Iliadis, C Jayne, A Likas (Eds), Engineering Applications of Neural Networks EANN 2017 Communications in Computer and Information Science, 744:609–619

[16] Zhao A., Qi L., Li J., Dong J., Yu H. (2018) “A hybrid spatio-temporal model for detection and severity rating of Parkinson’s disease from gait data”. Neurocomputing, 315:1–8

[17] Xia Y., Yao Z., Ye Q., Cheng N. (2020) “A dual-modal attention-enhanced deep learning network for quantification of Parkinson’s disease characteristics”. IEEE Trans Neural Syst Rehabil Eng, 28:42–51,

[18] Asuroglu T., Acici K., Erdas C.B., Toprak M.K., Erdem H., Ogul H., (2018) “Parkinson’s disease monitoring from gait analysis via foot-worn sensors”. Biocybernetics and Biomedical Engineering, 38(3):760–772

[19] Shaikhina T., Khovanova, N.A. (2017) “Handling limited datasets with neural networks in medical applications: a small-data approach”. Artificial Intelligence in Medicine, 75. pp. 51-63.

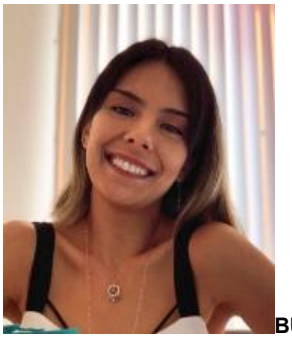
[20] Hoffer E., Ailon N. (2015) “Deep Metric Learning Using Triplet Network”. In: Feragen A., Pelillo M., Loog M. (eds) Similarity-Based Pattern Recognition. SIMBAD 2015. Lecture Notes in Computer Science, vol 9370. Springer, Cham.

[21] Cakir F., He K., Xia X., Kulis B., Sclaroff S. (2019) “Deep Metric Learning to Rank”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1861-1870

[22] Shi Y., Niu Y., Guo W., Huang Y. and Zhan J. (2020) "Pairwise Learning to Rank for Image Quality Assessment". in *IEEE Access* 8: 192352-192367

[23] Doughty H., Mayol-Cuevas W., and Damen D. (2019) “The pros and cons: Rank-aware temporal attention for skill determination in long videos”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7862–7871

[24] Mueller J., A. Thyagarajan (2016) "Siamese Recurrent Architectures for Learning Sentence Similarity". In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, Phoenix, Arizona, USA 2786–2792
[25] Hochreiter S., J. Schmidhuber (1997) "Long short-term memory". Neural Computation, 9(8):1735–1780
[26] Hochreiter S. (1998) "The vanishing gradient problem during learning recurrent neural nets and problem solutions". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6(02):107–116
[27] Bahdanau D., Cho K., Bengio Y. (2015) "Neural machine translation by jointly learning to align and translate". In Proc International Conference on Learning Representations. http://arxivorg/abs/14090473
[28] Burges C.J.C., Shaked T., Renshaw E., Lazier A., Deeds M., Hamilton N., and Hullender G. (2005) "Learning to rank using gradient descent". In Proceedings of the 22nd International Conference on Machine Learning (ICML), 89–96, Bonn, Germany, 2005
[29] Goldberger A., Amaral L., Glass L., Hausdorff J., Ivanov P.C., Mark P., Mietus J.E., Moody G.B., Peng C.K., Stanley H.E. (2002) "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals". Circulation [Online] 101;(23):e215–e220
[30] Yang Z., Yang D., Dyer C., He H., Smola A., Hovy H. (2016) "Hierarchical attention networks for document classification". 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
[31] Fok P., Farrell M., McMeeken J. (2010) "Prioritizing gait in dual-task conditions in people with Parkinson's". Human Movement Science 29;5:831–842,
[32] Siontis K.C., Noseworthy P.A., Attia Z.I. et al. (2021) "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management". Nature Review Cardiology 18, 465–478
[33] Davoudi A., Malhotra K.R., Shickel B. et al. (2019) "Intelligent ICU for Autonomous Patient Monitoring Using Pervasive Sensing and Deep Learning". Scientific Reports 9, 8020



BURÇİN BUKET OĞUL received her B.S degree from the Department of Computer Engineering in Başkent University and M.S degree from Information Systems Department at Middle East Technical University, Turkey. She is currently pursuing her PhD thesis at the Department of Computer Engineering in Hacettepe University, Turkey. Her research fields include machine learning and computer vision applications in health with particular interest in time-series data analysis.

SUAT ÖZDEMİR has been with the Department of Computer Engineering at Hacettepe University, Ankara, Turkey since May 2020. Before joining Hacettepe University, he worked at the Computer Engineering Department at Gazi University between 2007 and 2020. Prior to that, he received his MSc degree in Computer Science from Syracuse University (August 2001) and PhD degree in Computer Science from Arizona State University (December 2006). During his graduate study, he worked on wireless sensor network security. In Gazi University, He served as assistant department head between 2010 and 2015. He also worked for TUBITAK from 2010 to 2014 as senior researcher. He is a member of IEEE and currently serving as editor/TPC member/reviewer for various leading IEEE and ACM journals and conferences.