ACTION QUALITY ASSESSMENT WITH MULTIVARIATE TIME SERIES

ÇOK DEĞİŞKENLİ ZAMAN SERİLERİ İLE EYLEM KALİTE DEĞERLENDİRMESİ

BURÇİN BUKET OĞUL

PROF. DR SUAT ÖZDEMİR

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Doctor of Philosophy

in Computer Engineering

November 2022

ABSTRACT

ACTION QUALITY ASSESSMENT WITH MULTIVARIATE TIME SERIES

Burçin Buket OĞUL

Doctor of Philosophy, Department of Computer Engineering

Supervisor: Prof. Dr. Suat ÖZDEMİR

November 2022, 80 pages

Action quality assessment using computerized methods is considered to be a promising direction in objective evaluation of actions in several domains including health, sport and education. In a typical architecture for quality assessment, a classification or regression system is asked to assign a query action to a predefined category or a continuous label that determines its quality level. Such systems are still trained manually, and they may have inconsistent annotations. Hence, an attempt to categorize or quantify the quality level can be biased due to potentially scarce or skewed training data.

In this thesis, we approach the quality assessment problem as a pairwise ranking task where we relatively assess two input actions to identify better performance instead of assessing their absolute levels. To this end, we propose a novel computational model that takes two action data in the form of multi-variate time-series acquired from motion sensors and reports the probability of a query sample having a better quality than a reference one. The ranking model is built upon an attention-enhanced Siamese Long Short-Term Memory (LSTM) Network fed by piecewise aggregate approximation of time-series data. A probabilistic ranking layer is proposed to make the final relative assessment. The pairwise model is further extended to create an empirical feature representation in a regression setup.

The model is adopted in three different applications, namely, gait assessment in Parkinson’s Disease (PD) patients using foot sensors, surgery skill assessment using kinematics sensors and diving quality assessment using estimated pose from video recordings. According to experimental results, the proposed model achieves higher assessment accuracy than the existing models for pairwise ranking in all common datasets. The new regression model with new ranking-based empirical feature representation also outperforms the existing models when applied in their experimental setup. The proposed model is further shown to be accurate in individual progress monitoring.

The model that is developed in this thesis can be considered as a generic model for several pairwise ranking tasks provided that the inputs are in the form of multi-variate 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. This relative assessment approach may overcome the limitations of having consistent annotations to define quality levels and provide a more interpretable means for objective skill assessment. Moreover, the model allows monitoring the skill development of individuals by comparing two activities at different time points. We expect that this model will find a wide range of applications in several domains, but more particularly in sports and healthcare.

Keywords: Pairwise ranking; relative assessment, multi-variate time-series analysis; gait assessment, surgical skill assessment, sport skill assessment

ÖZET

ÇOK DEĞİŞKENLİ ZAMAN SERİLERİYLE EYLEM KALİTE DEĞERLENDİRMESİ

Burçin Buket OĞUL

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Suat ÖZDEMİR

Kasım 2022, 80 sayfa

Eylem kalite değerlendirmesinde bilgisayarlı yöntemlerin kullanılması, sağlık, spor ve eğitim gibi çeşitli alanlardaki eylemlerin objektif olarak değerlendirilmesinde umut verici bir yön olarak kabul edilmektedir. Tipik bir eylem kalite değerlendirme mimarisinde amaç, herhangi bir eylemin önceden tanımlanmış bir kategoriye veya o eylemin kalite seviyesini belirleyen sürekli bir etikete atanması amacıyla bir sınıflandırma veya regresyon sistemi geliştirmektir. Bu tür sistemler manuel olarak, üstelik işaretleme yapan kişinin olası tutarsız etiketlemeleri ile eğitildiğinden, kalite düzeyini kategorize etme veya eylemin kalite başarı seviyesini tam olarak tahmin etme girişimi, potansiyel olarak az sayıda ve dengesiz eğitim verileriyle hatalı sonuçlara sebebiyet verebilir.

Bu tezde, eylem kalitesi değerlendirme problemini, eylemlerin mutlak seviyelerini doğrudan değerlendirmek ve tahmin etmek yerine, daha iyi performansı belirlemek için iki girdi eylemini göreceli olarak değerlendirdiğimiz ikili bir sıralama görevi olarak ele alıyoruz. Bu amaçla, hareket sensörlerinden elde edilen çok değişkenli zaman serisi türünde iki eylem verisini girdi olarak alan ve bir sorgu örneğinin referans olandan daha iyi kalitede olma olasılığını rapor eden yeni bir model öneriyoruz. Bu ikili sıralama modeli, zaman serisi verilerinin parçalı toplam yaklaşımıyla (*piecewise aggregate approximation*) eğitilen bir dikkat mekanizması (*attention-enhanced*) tabanlı Siyam Uzun Kısa-Süreli Bellek (*Siamese Long Short-Term Memory*) Ağı üzerine kurulmuştur. Mimarinin final katmanında, nihai göreceli değerlendirmeyi yapmak için de yenilikçi bir olasılıksal sıralama katmanı önerilmiştir. Ayrıca geliştirilen bu ikili model, eylem kalite değerlendirmesi problemini bir regresyon modeli olarak ele almak istediğimizde, o modelin eğitimindeki öznitelik kümesini oluşturmak için daha da genişletilmiştir.

Model sırasıyla, Parkinson hastalarında ayak sensörleri kullanılarak hastaların yürüyüşlerinin değerlendirmesi, kinematik sensörler kullanılarak cerrahi beceri değerlendirmesi ve video kayıtlarından elde edilmiş pozlar kullanılarak olimpik dalış kalitesi değerlendirmesi olmak üzere üç farklı uygulamada test edilmiştir. Deneysel sonuçlara göre, önerilen modelin, bu veri setlerini kullanan mevcut modellerden daha yüksek doğruluk sonuçlarına eriştiği görülmüştür. Ayrıca geliştirdiğimiz sıralama tabanlı deneysel öznitelik temsiline sahip yeni regresyon modelinin, aynı deneysel düzenekte uygulandığında mevcut modellerden daha iyi performans değerlerine ulaştığı da gösterilmiştir. Modelin ayrıca bireysel gelişim takibinde de anlamlı sonuçlar verdiği izlenmiştir.

Model, girdilerin çok değişkenli zaman serisi sinyalleri biçiminde olması sebebiyle, ikili sıralama görevi için genel bir model olarak düşünülebilir. Uzun Kısa-Süreli Bellek katmanı, modeli tüm sıralı sinyaller için uygulanabilir hale getirirken, dikkat mekanizması, farklı ölçüm türlerinden elde edilen yeni sinyalleri benimseme yeteneğini genişletir. Olasılıksal kayıp fonksiyonuna sahip önerilen sıra katmanı, Siyam modelinin, girdi eylemlerin benzerliklerini hesaplamak için doğrudan değerlendirmeleri yerine, bu eylemlerin birbirlerine göre göreceli olarak karşılaştırılmasına imkân sağlar. Bu göreceli değerlendirme yaklaşımı, kalite seviyelerini tanımlamak için yeterli etiketlemeye sahip olamamanın dezavantajlarının üstesinden gelebilir ve nesnel beceri değerlendirmesi için daha yorumlanabilir bir araç sağlayabilir. Ayrıca model, farklı zaman noktalarında iki aktiviteyi karşılaştırarak bireylerin beceri gelişiminin izlenmesine olanak tanır. Bu modelin çeşitli alanlarda, ancak özellikle spor ve sağlıkta geniş bir uygulama yelpazesi bulmasını bekliyoruz.

Anahtar Kelimeler: İkili sıralama; göreceli değerlendirme, çok değişkenli zaman serisi analizi; yürüyüş değerlendirmesi, cerrahi beceri değerlendirmesi, spor beceri değerlendirmesi.

ACKNOWLEDGEMENTS

Firstly, I am deeply grateful to my supervisor, Prof. Dr. Suat Özdemir, for his coaching, guidance, valuable edits and supports throughout this research. His comments and suggestions helped me a lot to improve this thesis. I would also like to thank to Matthias Felix Gilgien for his extensive knowledge and support at the Norwegian School of Sport Sciences where I was a researcher during my PhD.

I would like to express my deepest gratitude to my dear children, Yalın and Derin, for the endless understanding they have shown to me despite being without me days and nights; my mother and sister for their efforts to protect my mental and physical health and not make my children feel my absence; Semih Abi for believing in me during my doctoral journey.

Finally, without your endless effort, huge encouragement, brilliant ideas, and immense patience nothing would be possible my dear mentor and life mate!

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

CONTENTS

Page

TABLES

FIGURES

1. INTRODUCTION

1.1 Motivation

Ranking two or more entities, which can be either a person, an object, or an event, based on any criteria is a common practice in our daily or business life. When we are in the position of selecting something, we usually compare available options and decide after a rank-based evaluation. For example, the winner of a championship is determined by the rank of participants, but not based on their individual performance. Final grades of students in a class are often assigned using a so-called "grading by curve" system, where all students are sorted by their individual scores first and then letter grades are given in their categorical order. A recruitment process usually involves a step of ranking existing candidates in a certain stage. A collaborative recommendation system shows potential entries to a user based on their rank of likes by other similar users. All these ranking assessment results are usually inferred from pairwise ranking of given entities.

Although it is very common in daily practices, 'pairwise ranking' has not received as much attention as classification, clustering, and regression tasks in computer science, or particularly machine learning community. In general, pairwise models have been studied extensively in computer science literature [1-4]. Some of these studies can be grouped into multi-stream learning models, such as Siamese or triplet networks, for 'classification' of objects [1]. 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. The 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 [2]. The objective is to provide users most relevant entities in terms of their similarity to given query. Here again, a pairwise model aims to learn how similar two inputs are based on a training set of similarity scores, but not to rank them. In this context, ranking refers to relevance-based sorting of retrieved objects.

In ‘pairwise ranking’, particularly, the aim is 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 [3]. We have also recently seen some applications of pairwise ranking in video data for action quality assessment in sport activities as well [4].

In this thesis, we address the problem of ranking two human actions, where the activity information is acquired from multiple sensors over a time period. To this end, we offer a two-stream machine learning model that takes two input signals to be compared and returns the probability of first input ranked higher than second input without any prior information about the criteria of ranking but with the availability of previous human-annotated samples of ranked pairs. The model is an adoption of Siamese recurrent neural network [5] 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 a Long Short-Term Memory (LSTM) enhanced by an attention mechanism to capture remote dependencies in input signals relevant to gait skills. To cope with the sparsity of observed data, a pre-processing step based on piecewise aggregate approximation is adopted to the model. For convenience, the model will be referred to as *Ranking by Siamese Recurrent Network with Attention (RSRNA)* in the rest of the thesis.

The model encourages the use of a relative evaluation approach instead of an absolute scoring in action quality assessment. However, the model proposed can also be extended to enable an absolute assessment without any further feature extraction from raw data. To this end, we offer an empirical feature representation scheme to feed a regression framework built upon any mathematical model. In this scheme, each feature represents the pairwise rank between query sample and any other sample in the training set.

Although the models are applicable to any types of multi-variate time-series data, we adopt them in the scope of action quality assessment and demonstrate their ability in three different real-world scenarios: (1) Gait quality assessment for Parkinson's Disease (PD) patients using foot-worn sensors, (2) Surgical quality assessment using kinematic sensors of surgery robots, and (3) Sport action quality assessment by estimated pose information from activity videos. The literature review and motivation for pairwise ranking for each of these tasks are elaborated in relevant chapters.

1.2 Contribution

The contribution of the thesis is as follows:

(1) 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 existing models used in image data cannot be directly inherited for time-series signals. Here, we address this challenge using a *novel* pairwise deep learning model tailored for multi-variate time-series.

(2) The thesis introduces a *novel* empirical feature representation scheme for time-series signals based on pairwise rank of query signal against each of the signals present in the training set. The results show that this representation can achieve a higher accuracy in some cases when tested in a regression setup. Moreover, the representation scheme enables users to create feature vectors without any prior knowledge about the application domain or the problem in question.

(3) This is the *first study* which considers gait analysis problem as a pairwise ranking task. This approach allows neurologists to compare two different patients or to monitor the same patient in different stages of treatment. The proposed solution is not directly comparable existing methods, which attempt to classify a patient into known clinical categories. However, the new solution is benchmarked against available methods by re-implementing them in pairwise ranking setup. The results have shown that proposed method can achieve the highest accuracy these experiments.

(4) Similarly, this study is the *first* in which surgical skill analysis problem is considered as a relative assessment task instead of absolute scoring of skills. This approach enables users both to assess the skill development of new surgeons and to assign correct surgeon for specific operations. The experiments with this problem also shown that proposed method can achieve the highest accuracy in pairwise ranking when they were run in the same experimental setup.

1.3 Organization

The remaining part of the thesis is organized as follows:

Chapter 2 gives a formal definition of the problem and the details of computational solution offered in this thesis. The method is built upon the introduction of general architecture first, and the description of each sub models in the given architecture.

Chapter 3 introduces the new model for regression based on pairwise ranking results, including the description of the general framework for regression and the vectorization with pairwise ranks.

Chapter 4, 5 and 6 present the details of three different applications of the model, i.e. (1) gait skill assessment, (2) surgical skill assessment, and (3) sport skill assessment. Each chapter starts with relevant literature for given task, follows with experimental design and implementation details and concludes with experimental results. For each task, the methods are evaluated to discern the ability of sub models in the proposed architecture as well as benchmarking results with the existing approaches.

Chapter 7 concludes the thesis with a general overview of the results, evaluation of model components, particular discussions with each application scenario and potential future directions due to the current limitations of the proposed approaches.

2. PAIRWISE RANKING

2.1. Problem Definition

Given a human action represented by a multivariate time-series sensory signal with a length of *K*, denoted by $x = x_1x_2 \ldots x_K$, where x_i refers to a set of sensory measurements at time *i*, and an output variable $q(x)$ denoting the quality of this action, general action quality assessment problem is defined as predicting $q(x)$ from a model learned from number of annotated samples.

In this study, we propose a relative (rank-based) action quality assessment, which can be defined as determining which action is performed with better skill. Therefore, we consider the problem as a pairwise ranking task. In this case, given two actions, say *m* and *n*, with length of *K* and *L*, are denoted by $x^m = x_1^m x_2^m \ldots x_K^m$ and $x^n = x_1^n x_2^n \ldots x_L^n$ respectively, where x_i^m and x_i^n refers to a set of sensory measurements at time *i* for *m* and *n* respectively, and two output variables $q(x^m)$ and $q(x^n)$ denoting the quality of referred actions, pairwise ranking of *m* and *n* is defined as identifying whether $q(x^m)$ is higher than $q(x^n)$.

For the sake of generalizability, the output of the model is referred by p_{mn}, which is interpreted as the probability of the query action (*m*) being performed better than the reference action (*n*);

$$p_{mn} = \begin{cases} 1 & m \text{ performs better than } n \\ 0.5 & m \text{ and } n \text{ show equal performance} \\ 0 & n \text{ performs better than } m \end{cases}$$

2.2 A Siamese Model for Ranking

The goal is to train a model that minimizes the probabilistic loss in a set of samples annotated by experts. The model assumes that the annotations of the exact skill levels are not provided but all pairs are labelled by their pairwise rank for their skills by experts.

The general framework that we introduce is based on a Siamese network of attention-enhanced LSTM integrated with a probabilistic ranking layer (Figure 1). Siamese neural network is one of the artificial neural network architectures which contains two or more identical sub-networks [5]. Identical means these networks share the same weights as shown in Figure 1 and the setup is used to find similarities between inputs by comparing its feature vectors. The model is a novel adoption of Siamese Recurrent Neural Networks [5] for the task of pairwise ranking instead of pairwise similarity inference.

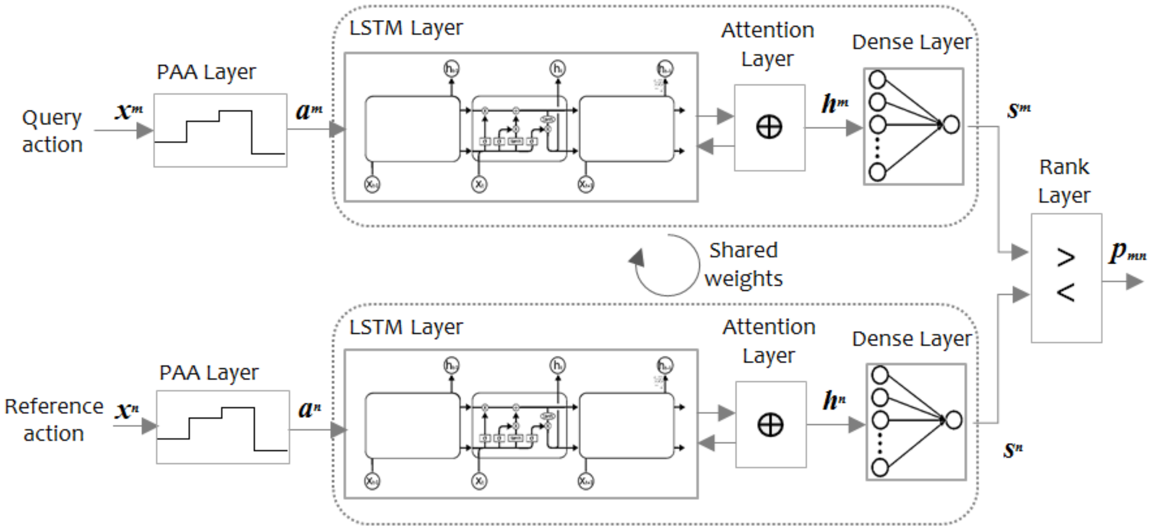


Figure 1. General framework for pairwise ranking of actions.

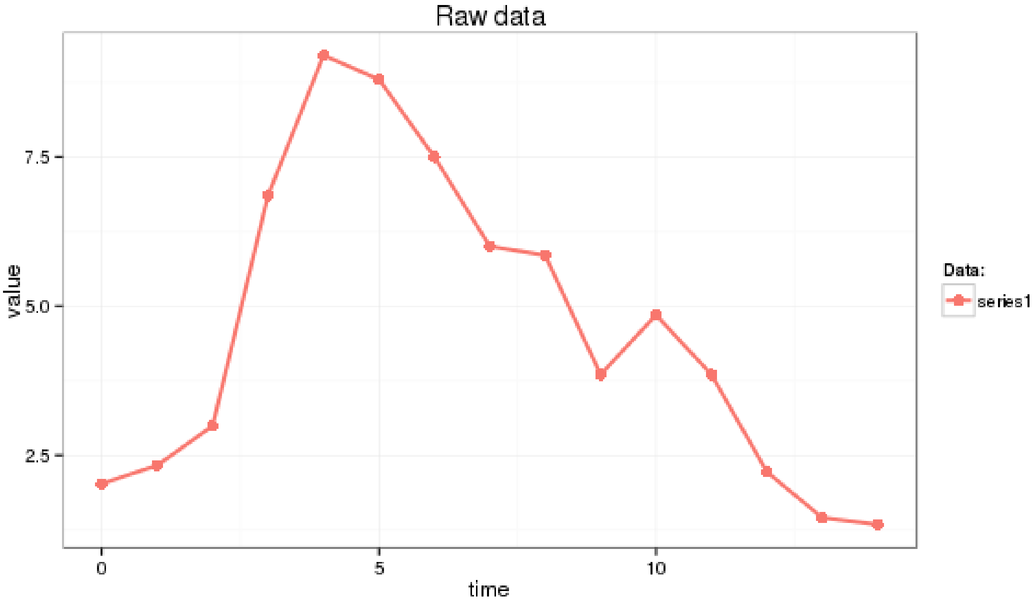
The framework involves an essential pre-processing step for input based on Piecewise Aggregate Approximation (PAA) to reduce the problem of sparsity in observed sequences. Variable-length time-series nature of the input signals is addressed by the LSTM sub-model. The balance of the contribution of relevant and irrelevant observations in the sequence is handled by an attention mechanism attached to LSTM. A dense layer transforms the multi-variate output of the LSTM into a single comparable value. The latent variables obtained from dense layer output are used to feed a loss layer based on pairwise rank. The overall model reports p_{mn} for actions m and n.

2.3 Piecewise Aggregation of Data

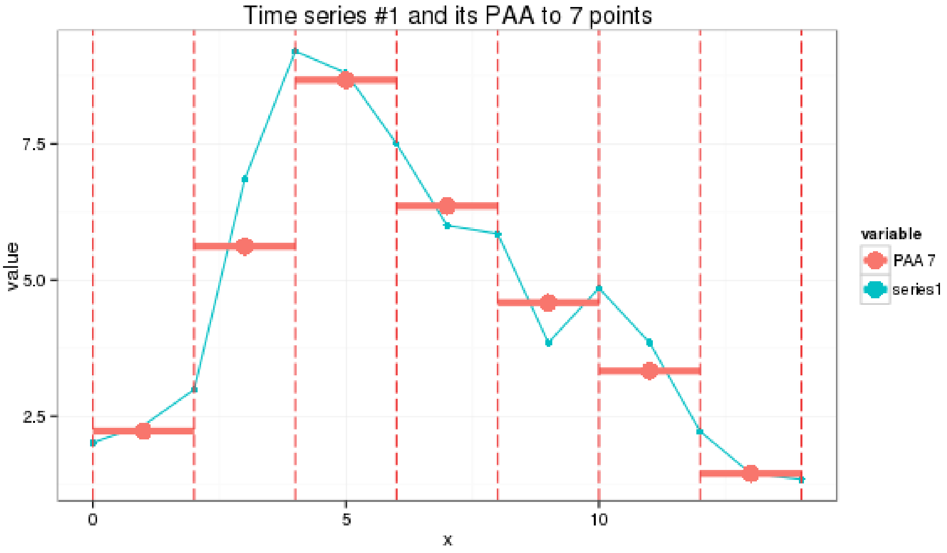
The action model based on attention-enhanced LSTM has an excessive number of parameters to be optimized in training phase (See 2.4). On the other hand, the kinematic data in our problem has a high dimensionality as opposed to the small number of samples in available datasets. This leads to a slow and insufficient learning of the model parameters in the proposed framework. To overcome this issue, we offer a pre-processing step based on Piecewise Aggregate Approximation (PAA) to reduce the dimensionality of the input signal while preserving the content that is representative for the skill level. PAA approximates a one-dimensional time-series signal x of length p into a of arbitrary length $q<p$, where each a_i is calculated by;

$$\boldsymbol{a_i = \frac{q}{p} \sum_{j=\frac{p}{q}(i-1)+1}^{\left(\frac{p}{q}\right)i} x_j}$$

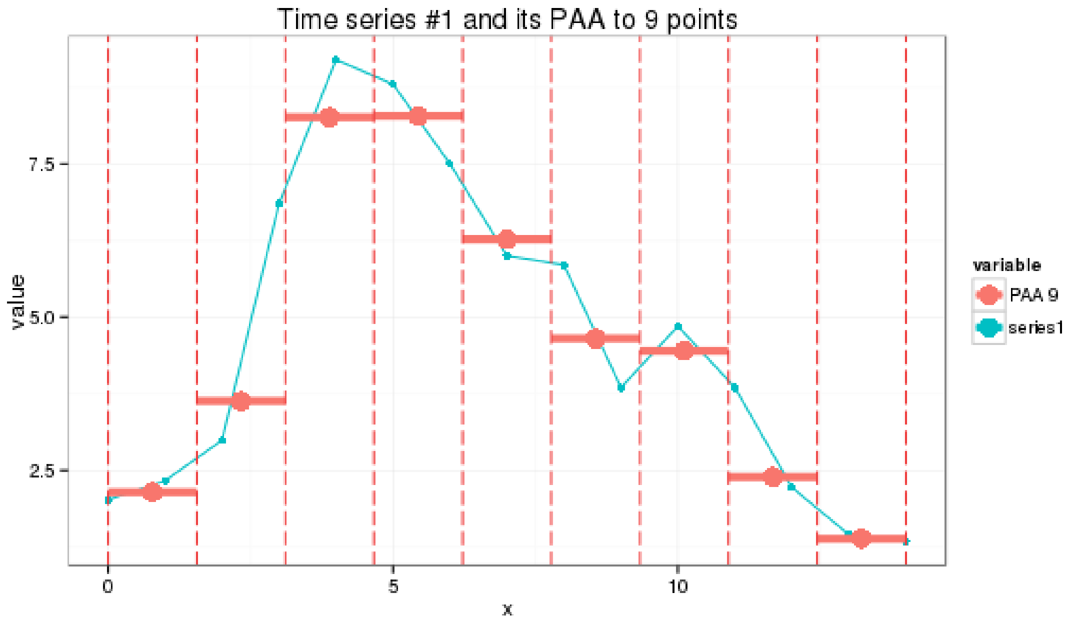
This approximation results with the reduction of the dimensionality of the signal by splitting it into equal-sized segments which are calculated by taking the average values in each segment. The equation above provides the mean of the elements in the equi-sized frame which makes up the vector of the reduced dimensional series. The most interesting aspect of the algorithm is how it creates these equi-sized frames. It is important to note here that before the actual mean approximates of the windows are computed, the input vector is Z-normed. Z-normalization is the process of normalizing the data to *zero mean and zero unit of energy* which is to say that the mean is 0 and the standard deviation is approximately 1. Once that is performed, the piecewise approximates can be computed. In Figure 2, a PAA transformation for different discretization levels is shown. Figure 2a is the raw values of the time series while 2b shows the PAA transformation for level 7 and 2c is for level 9.


Raw data
value
7.5
5.0
2.5
0
5
10
time
Data:
series1

(a)


Time series #1 and its PAA to 7 points
value
7.5
5.0
2.5
0
5
10
x
variable
PAA 7
series1

(b)



(c)

Figure 2. An example PAA transformation for different discretization levels: (a) raw time-series, (b) PAA transformation for level 7, (c) PAA transformation for level 9.

We apply PAA for each motion variable independently to get a smoother multi-variate signal at the input of the Siamese network.

2.4 Modeling Action: Attention-Enhanced LSTM

Both query and reference actions are pre-processed using PAA and given into different inputs of Siamese network. The pre-processed data, given in the form of multi-variate time-series, is used to feed an LSTM network at each stream:

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

where a_t and h_t are the input vectors at time t, where the superscript defining the stream is ignored. The LSTM model is parameterized by output, input and forget gates, controlling the information flow within the recursive operation. The following equations formally describe the LSTM function:

$$i_t \;=\; \sigma(W_i\,x_t +\; U_i\,h_{t-1} +\; b_i\,)$$

$$f_t \;=\; \sigma\big(W_f\,x_t +\; U_f\,h_{t-1} +\; b_f\,\big)$$

$$o_t \;=\; \sigma(W_o\,x_t +\; U_o\,h_{t-1} +\; b_o\,)$$

$$\tilde{c}_t = \;\tanh(W_c\,x_t +\; U_c\,h_{t-1} +\; b_c\,)$$

$$c_t = \;\; \sigma\,(i_t \circ \tilde{c}_t +\; f_t \circ c_{t-1}\,)$$

$$h_t = o_t \circ \tanh(C_t)$$

At every time step t, LSTM outputs a hidden vector h_t that reflects the skill representation of the kinematic motion at time point t. In our application, we used a bidirectional version of LSTM [6] to allow the modelling of two-way temporal dependencies in actions.

The LSTM layer is enhanced by an attention mechanism, which helps maximizing the contribution of the relevant encoding context vectors and minimize those of irrelevant vectors while building the decoding context [7]. The attention layer that we implement uses an attention function to assign weight to each hidden state produced by the LSTM layer. The weighted distribution of hidden states is used as a new representation of input signals. We calculate an attention function for each hidden state h_t, t=1,…,T, as follows;

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

where *Ws* is an attention hidden weight matrix and b is a bias parameter. From this function, softmax weights are calculated by;

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

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

$$c = \sum_{t=1}^{T} h_t \alpha_t$$

The attention-enhanced LSTM layer is followed by a fully-connected layer fed by the vector of skill representation, c^m for any of the input m. This layer transforms skill representations of query and reference actions into scalars, s^m and s^n, to make them explicitly comparable.

2.5 Ranking Loss

We adapt a probabilistic loss function for model learning, which was originally introduced to learn how to rank text objects using a gradient descent approach [8]. A probabilistic rank layer is built such that quality equivalence is taken into account. The pairwise rank between two inputs is desired to be represented by p_{mn}, which is interpreted as the probability of m having better quality than n. We denote the posterior probability distribution $P_{ij}=P(i›j)$, where › refers to the skill superiority of i to j and let $\bar{P}_{ij}$ be the desired target values for those posteriors, such that $\bar{P}_{ij}$ ∊ {1,0.5,0}. The goal is then 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;

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

Letting o_{mn} is the difference between rank orders of m and n, the probabilities are modelled by;

$$P_{ij} \equiv \frac{e^{o_{ij}}}{1+e^{o_{ij}}}$$

Then, the final cost function becomes;

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

where o_{ij}=(s^i – s^j), i.e. is the difference between rank orders of *i* and *j*,

The model parameters, including Siamese network, LSTM, attention functions and fully-connected layer weights, are then inferred by minimizing this loss for all (*i*, *j*) trial pairs in the training data.

3. RANKING-BASED REGRESSION

3.1. Regression Framework

Regression is an obvious way to model the original action quality assessment task. Although we propose to use a relative approach for quality assessment, here we want to show that the pairwise ranking model introduced in this thesis can also be useful for regression analysis to predict the exact value of action quality as well. To this end, we introduce a novel feature representation scheme to feed a conventional regression model.

If we refer back to the formal definition of the problem, we want to predict an output variable $q(x)$ indicating the quality of an action denoted by $x = x_1x_2...x_K$, where x_i refers to a set of sensory measurements at time i. Regression is then defined as:

$$q(x)=f(w,s)$$

where s is a vector representing the signal s. In its simplest case, where all samples have the same length of K, s can be same as x and w refers to set of weight in a linear function:

$$q(x)=w_1x_1+ w_2x_2+\ldots+ w_Kx_K.$$

In more generalized formulation, w will refer to a set of parameters for the model $f(.)$ and s will be a fixed number of features to be extracted from x.

Figure 3 gives an overview of conventional regression setup for any regression solution with time-series input. While enormous attempts have been found in the literature, there is no feature set that fit all problem [9]. Since many of the proposed features are “black-box”, it is not usually possible to select among possible features manually. Although an experimental grid search can make it possible to select a subset of available features, this is often time-consuming and requiring a huge amount of training and validation data.

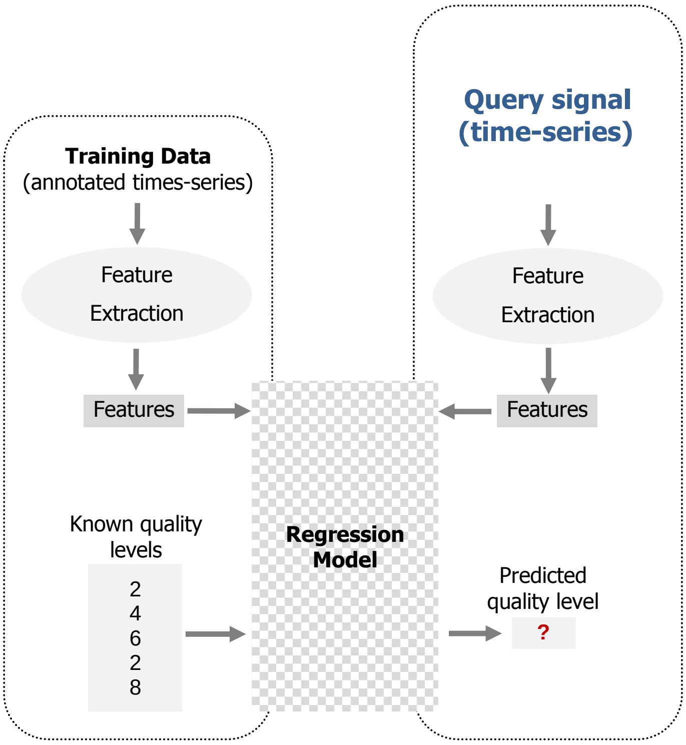
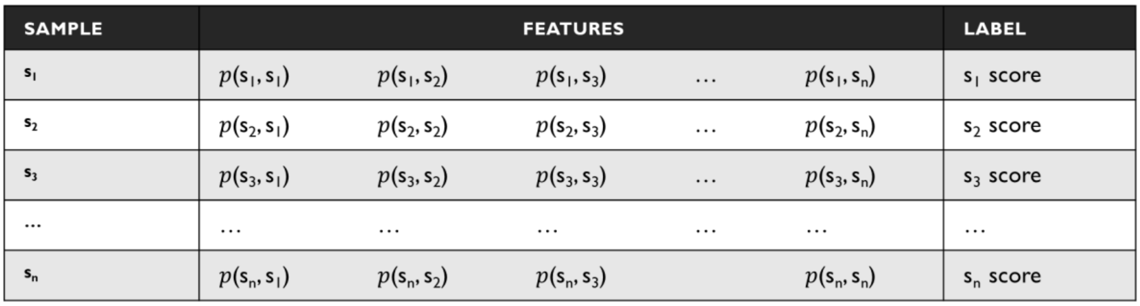


Figure 3. Generalized regression framework for absolute action quality assessment.

3.2. Using Pairwise Ranks as Features

An alternative way to conventional feature extraction approach is to use the training samples themselves as a part of features. One common approach that is used in several classification problems is to use the pairwise similarity/distance between query sample and each of the training sample as a single feature in the input vector [10] [11]. This empirical enables users to focus directly on own data regardless of the domain or problem being considered.

Since we try to model the superiority of an action to the others in regression setup, direct use of pairwise similarity/distance scores do not be convenient to feed the regression model. Instead, we propose a novel empirical feature representation scheme based on pairwise ranks in replace of pairwise similarities. Informally, we have used ranking values as features for regression as shown in Figure 4.



SAMPLE	FEATURES					LABEL
s_1	$p(s_1, s_1)$	$p(s_1, s_2)$	$p(s_1, s_3)$	…	$p(s_1, s_n)$	s_1 score
s_2	$p(s_2, s_1)$	$p(s_2, s_2)$	$p(s_2, s_3)$	…	$p(s_2, s_n)$	s_2 score
s_3	$p(s_3, s_1)$	$p(s_3, s_2)$	$p(s_3, s_3)$	…	$p(s_3, s_n)$	s_3 score
…	…	…	…	…	…	…
s_n	$p(s_n, s_1)$	$p(s_n, s_2)$	$p(s_n, s_3)$		$p(s_n, s_n)$	s_n score

Figure 4. Using pairwise ranks as features for regression

More formally, we need a mapping $\partial(x^m)$ → f^m, to transform a signal x for action m to a feature vector. Here, we define a n dimensional feature vector, where n is the number of samples in the training set and define fi as $p(s_m, s_i)$, which is the probability of the action m having better quality than the ith action in the training set. Then f^m becomes;

$$f^m=[\ p(x^m, x^1)\ p(x^m, x^2)\ \dots\ p(x^m, x^n)]$$

For the ease of implementation and to eliminate any bias from self-similarity $p(x^m, x^m)$is set to 0.

4. RANKING PD GAIT SKILLS

4.1. Background

Parkinson's disease (PD) is a neurodegenerative disorder of aging that affects dopamine-producing neurons in the substantia nigra area of the brain [12]. 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 [13] [14] [15] [16]. 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 [17]. Based on interviews by an examiner or a patient's self-assessment, scales such as the Unified Parkinson Disease Rating Scale (UPDRS) [18] 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 categories using conventional machine learning algorithms fed by a set of extracted features from sensory signals [19]. Lee et al. [20] 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. [21] 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. [22] 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 [21]. [23] 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. [24] 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 [21] and [22]. [25] 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. [16] 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 [24] [25], several researchers applied deep learning techniques, such as convolutional neural networks (CNN) and recurrent neural network (RNN), instead of using hand-crafted features. [26] used a two-channel model that combines long short-term memory (LSTM) and CNN to learn the spatio-temporal information behind the data. [27] 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 [28]. 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 [29] [19] [24]. [28] adapted their random forest model in a regression setup, instead of classification in [25], 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 [30]. 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 [17]. 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 [17].

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 [17] 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.

The thesis contributes the studies on remote PD monitoring in two ways: 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 a real-world PD gait dataset. 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.

4.2. Materials and Methods

4.2.1 Model Adoption

We adopt RSRNA model for PD monitoring as follows. Given two PD patients, *m* and *n*, with their gait data of $\boldsymbol{x^m}$ and $\boldsymbol{x^n}$, which are in the form of multi-variate time-series GRF signals, 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;

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

We interpret this as the probability of first patient having more severe symptoms than the second. Our goal is then to learn a model that minimizes the probabilistic loss in human-annotated samples for PD severity. To this end, we apply RSNA model which can take the case of the equivalence of severity into consideration (Figure 5).

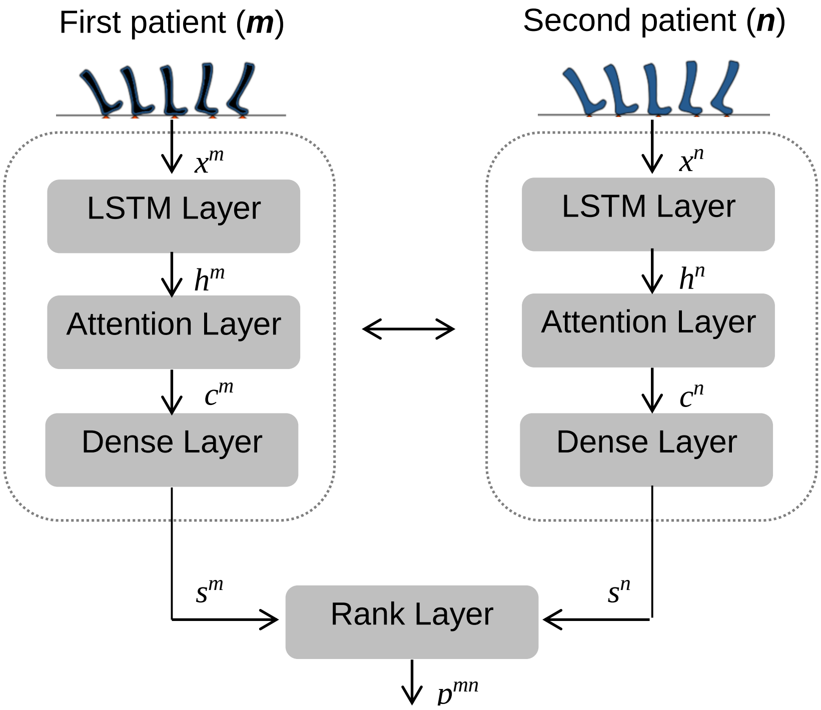


Figure 5. Adoption of RSRNA model for pairwise ranking of PD patients from gait signals.

4.2.2. Data

A public PhysioNet dataset [31] was used in this study. 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 sampling rate was 100 Hz. In our study, we use the digitized outputs of these 16 sensors to analyze the dynamics and characteristics of these multivariate time series.

Table 1. 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

The dataset also includes demographics information, measures of disease severity in terms of different metrics such as UPDRS and other related measures. As [21] 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.

4.2.3. Experimental Setup

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 sample has been used in a pair for training and the pairs in which the other sample 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 Ɛ = 0.01 in our evaluations.

Table 2. Evaluations of correct predictions and associated ground truth for different pairwise ranking schemes

Ranking scheme	p_{mn}	Ground truth
Ternary	≥ 0.5+ Ɛ	*m›n*
	≥ 0.5- Ɛ and < 0.5+ Ɛ	*m≡n*
	< 0.5- Ɛ	*m‹n*
Binary	≥ 0.5	*m›n*
	< 0.5	*m‹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

4.2.4 Implementation

We used an LSTM network to capture temporal representations in PD symptoms. For the attention layer, we followed the previous implementation by [32] 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 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.

4.3. Results

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 6 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 7.

In Figure 6, 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.

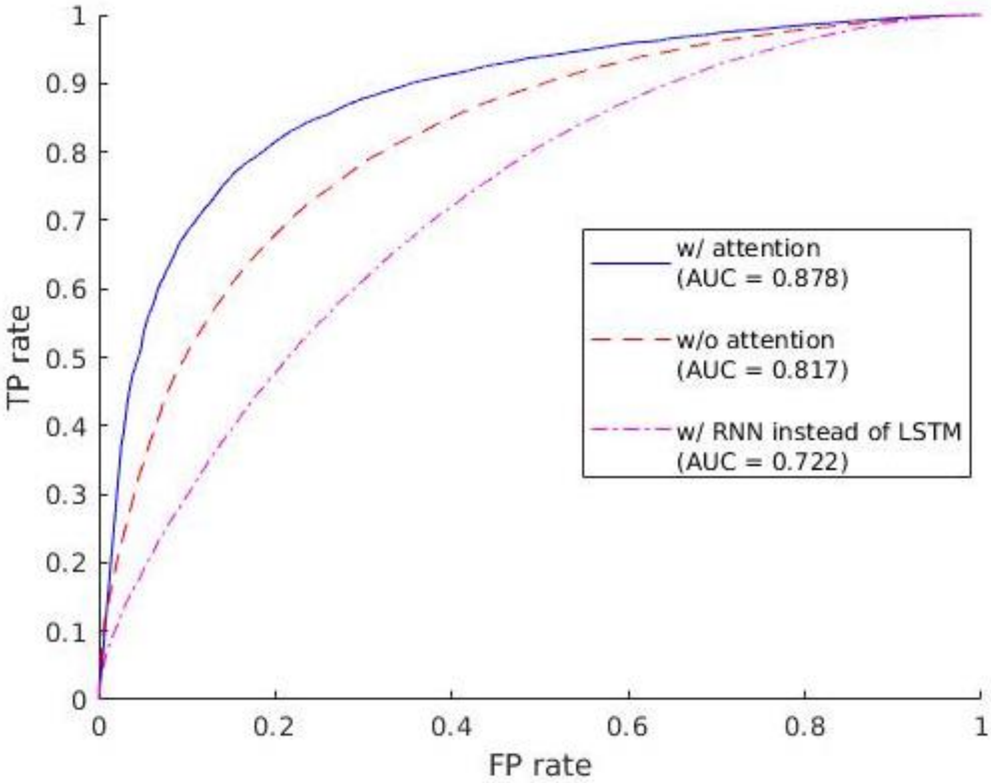
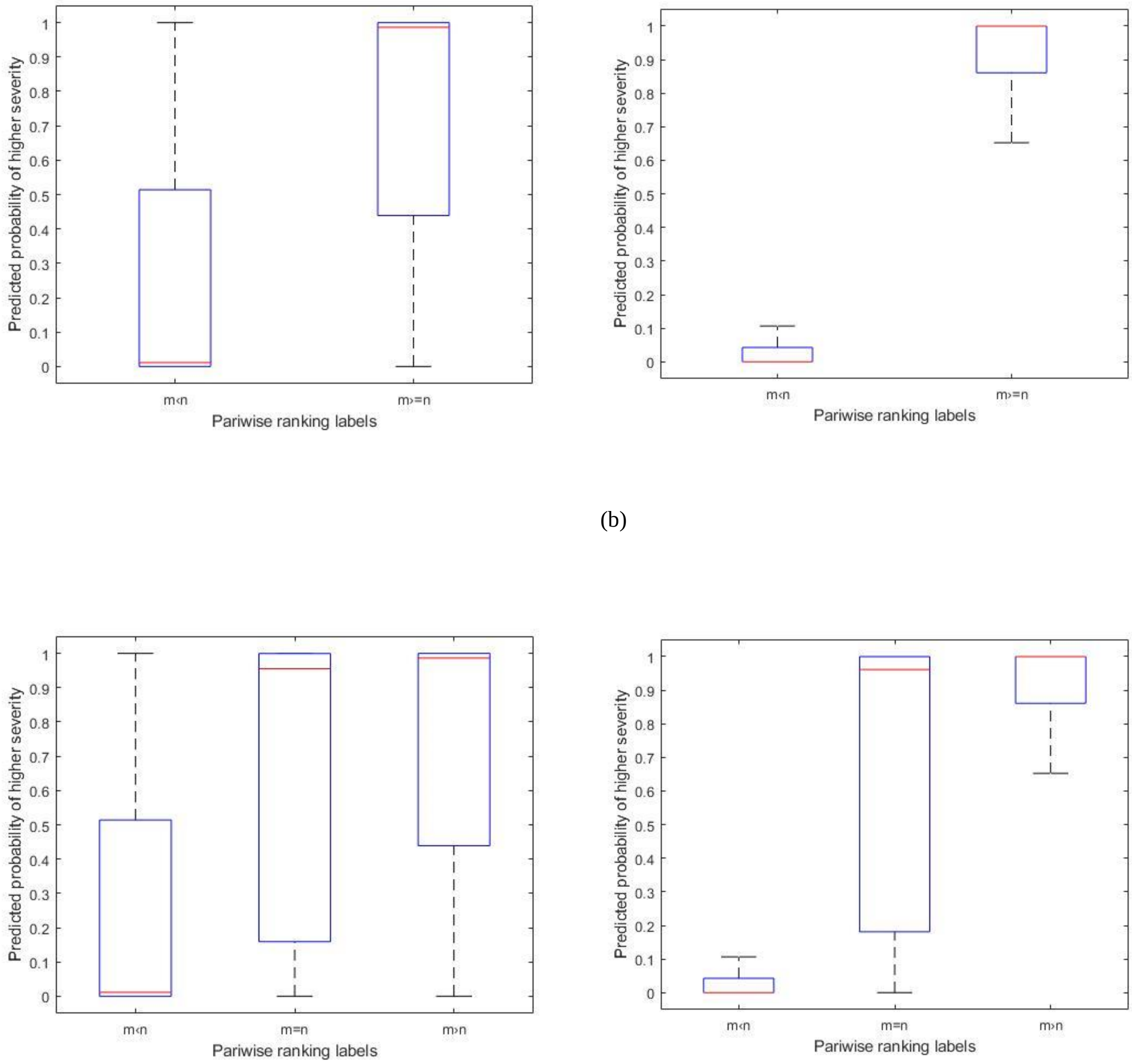


Figure 6. ROC curves for binary pairwise ranking by RSRNA model using alternative sub-models; (1) with attention, (2) without attention and (3) using RNN instead of LSTM.

The boxplots shown in Figure 7 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 7.a-b) and ternary (Figure 7.c-d) ranking schemes.



(a)

(b)

(c)

(d)

Figure 7. 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 3. Justification of the proposed model by comparison of relative assessment (pairwise ranking) performances of different architectures with alternative sub-models.

	Binary ranking		Ternary ranking	
Method	**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.

Previous Method 1: Daliri [21] classified patients as PD or not using an SVM with frequency domain features. Similarly, we reconfigured Daliri’s [21] 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. [28] 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. [27] 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. Comparison of the proposed model with previous studies in terms of their relative assessment (pairwise ranking) performances.

	Binary ranking		**Ternary ranking**	
Method	**Acc**	**AUC**	**Acc**	**AUC (avg)**
RSRNA (proposed model)	**81%**	**0.878**	**78%**	**0.862**
Previous Method 1* [21]	64%	0.623	58%	0.617
Previous Method 2* [28]	63%	0.744	59%	0.737
Previous Method 3* [27]	64%	0.698	61%	0.685
Previous Method 4* [27]	57%	0.579	55%	0.567

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

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.

Figure 8 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 [28]) 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.

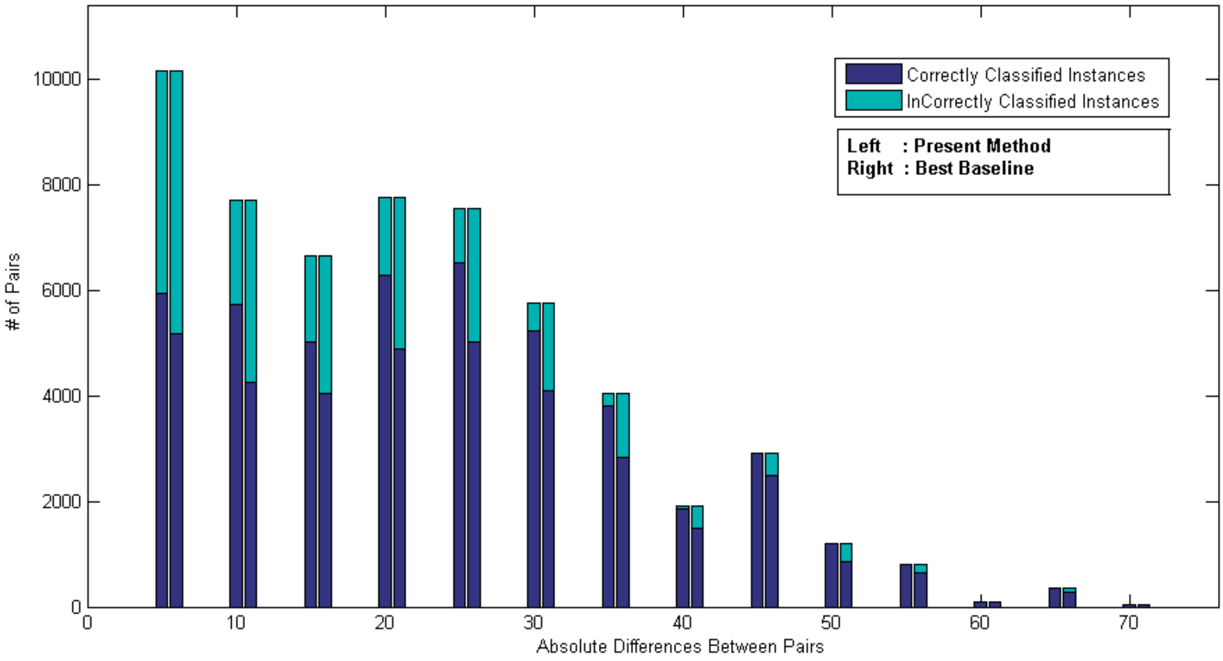


Figure 8. Bars of correctly classified pairs versus incorrectly classified ones based on patients' UPDRS scores. The left and right bars show the results of the present method and the best baseline, respectively.

5. RANKING SURGERY SKILLS

5.1. Background

Assessment of surgical skills may have three main objectives: (1) choosing appropriate surgeons for a specific operation, (2) examining current performance of candidate surgeons before credentialing, and (3) monitoring the progress of surgeon's skills during training activities. These assessment activities are usually performed manually in an operation room under supervision and feedback of expert surgeons. Manual assessment of surgical skills by individuals may lead to misinterpretations of the skill performance and hence lead to suboptimal training and organization of the surgical activities. Some structured methods such as Objective Structured Assessment of Technical Skills (OSATS [33]) have been employed to minimize the effect of the subjective nature of expert intervention. However, the process needs improvements to increase its efficiency since the application of these techniques still require significant effort of multiple experts over a long time period [34]. Considering the fact that evaluation of the candidates by senior surgeons has certain cost, there is an increasing need for alternative or complementary computerized assessment systems.

We have recently witnessed a significant attempt to computerize surgery skill assessment using machine learning algorithms [35]. Robot-assisted surgery helps this effort by providing data in different forms, such as kinematic sensor measurements derived from robot arms and video recording of a surgical action performed by an operator. An overview of recent methods for computerized skill assessment using machine learning is given in Table 5.

In one of the earliest studies, kinematic data collected during robot-assisted surgery were used to predict the expertise level of the surgeon [36]. A set of hand-crafted features were extracted from surgery action and fed into three different supervised classifiers (k-Nearest Neighbour, Support Vector Machine (SVM) and Linear Regression) for classification of surgeons into either “expert”, “intermediate” or “novice” levels. The authors employed several kinematic features including task completion time, path length, depth perception, speed, motion smoothness, curvature, turning angle and tortuosity to build the model. In a similar work [37], the authors used different time and frequency domain features of kinematic data, which were obtained through sequential motion texture, discrete Fourier transform, discrete cosine transform and approximate entropy analysis to train a linear SVM model. In addition to classification, i.e. assigning objects into predefined skill labels, they also considered to predict the level of skills by running the SVM in a regression setup. [38] proposed a deep learning architecture based on Convolutional Neural Networks (CNN) that can automatically extract relevant features and classify the expertise level using a fully-connected layer at the end. Similar architectures were used by [39] and [40] with slight modifications in layer organizations. [41] used video recordings of surgery actions instead of motion kinematics to feed a 3D CNN with the same objective (ternary classification). CNN was combined with Long Short-Term Memory (LSTM) model to analyze kinetic data for classification [42]. These studies reported very high classification accuracy, up to 100% for some surgery actions, in a public benchmark dataset for human gesture and skill assessment from surgical activity, called JIGSAWS [43]. The performance of conventional machine learning methods with hand-crafted features was recently re-evaluated in a larger in-house dataset [44], where they determined that an average accuracy of 91.5% can be achieved in binary classification of skill. The LSTM model was shown to be accurate in binary skill classification ("expert" or "novice") from kinematic signals in a private dataset [45]. The ability of CNN applied on video recordings was further assessed in another study with an in-house dataset [46]. However, they reported that the accuracy diminished from 86% to 70% when they increased the number of skill categories from two to five.

Table 5. Methods for computerized assessment of surgery skills

Reference	Data Type	Task	Dataset
Fard et al. 2018 [36]	Kinematic	Classification	JIGSAWS
Fawas et al., 2018 [39]	Kinematic	Classification	JIGSAWS
Wang and Fey, 2018 [38]	Kinematic	Classification	JIGSAWS
Zia and Essa, 2018 [47]	Kinematic	Regression	JIGSAWS

Dougthy et al., 2018 [48]	Video	Ranking	JIGSAWS
Fawas et al., 2019 [39]	Kinematic	Regression	JIGSAWS
Funke et al., 2019 [41]	Video	Classification	JIGSAWS
Nguyen et al., 2019 [42]	Kinematic	Classification	JIGSAWS
Li et al., 2019 [49]	Video	Ranking	JIGSAWS
Ogul et al., 2019 [50]	Kinematic	Ranking	JIGSAWS
Zhang et al., 2020 [40]	Kinematic	Classification	JIGSAWS
Kelly et al., 2020 [45]	Kinematic	Classification	In-house
Lavanchy et al., 2021 [46]	Video	Classification	In-house
Perez-Escamirosa et al., 2021 [44]	Video	Classification	In-house
This study	Kinematic	Ranking Regression	JIGSAWS, ROSMA

The major problem with these performance assessment systems is their limited ability to predict a fixed number of predefined, possibly inconsistent, categories for skill levels. As reported by [46], they are unable to model skill levels between these pre-defined categories. Recalling the three main objectives for surgical skill assessment, discussed at the beginning of the text, i.e. (1) choosing appropriate surgeon, (2) examining current performance of surgeons, and (3) monitoring the progress of a surgeon, the classification approach may support partially the second objective. However, it fails to provide an accurate solution for first and third tasks since the number of categories representing skill levels is not sufficient to model precise comparison of actions. Regression can be considered as a possible solution in general. However, in small dataset scenarios, where continuous labels representing skill levels are too sparse, it is not easy to provide generalizable models for exact value predictions. Two previous approaches for this [47] [39] indeed reported very low correlations between predicted and actual skill levels.

The skill assessment problem was recently considered as a task of learning to rank video recordings [48] [49] instead of assigning them into predefined labels. These studies aimed to build generic models with wide applicability of skill determination in any domain, but algorithms were also tested for surgical skill assessment with the JIGSAW dataset. First, the study introduced a two-stream Temporal Segment Network to capture both the type and quality of actions [48]. Second, the study integrated an attention pooling and temporal aggregation mechanism to a two-stream CNN model [49]. Skill assessments through video recordings have two main limitations. First, video data processing is time and resource inefficient, which makes it difficult to run the algorithms in conventional personal computers. Second, video can record the actions in two dimensions, if only one camera is used. This is unfortunate since tracking of trajectories and velocities can only be measured in two dimensions and important information of surgery skills is lost, if the third dimension is lacking.

Here, the surgical skill assessment problem is considered as a pairwise comparison task. The RSRNA model proposed in this thesis is adopted for the problem. The model was first tested on the JIGSAWS dataset to compare it with previous methods. According to the results, our model can significantly improve the state-of-the-art in both ranking and regression tasks for computerized surgical skill assessments. Further, the model was evaluated for monitoring tasks in a larger and more recent dataset, called ROSMA [51]. The results show that our model can achieve reasonably good accuracy.

5.2. Materials and Methods

5.2.1. Model Adoption

We adopt RSRNA model for the surgical skill assessment problem as follows. We compare a query surgical action (*m*) with a reference action (*n*) in order to infer if the query is performed better than the reference. Semantically, the reference may refer to a previous action of the same surgeon to monitor the skill improvement, or to an action performed by another surgeon to make a skill comparison for better assignment to a surgery. While the model is formally the same, it can be used in any semantic model based on how the model parameters are trained from available data.

The kinematic data of two actions with length K and L are denoted by $\boldsymbol{x^m}$ = $x_1^m x_2^m \dots x_K^m$ and $\boldsymbol{x^n}$= $x_1^n x_2^n \dots x_L^n$ respectively, x_i^m refers to a set of kinematics measurements at time i. A kinematic measurement can be position, angular velocity, gripper angle or any other motion-specific identifier of a particular hand at a given time point.

The model is similarly enhanced by adapting an attention mechanism to the LSTM; and a processing step, which calculates the Piecewise Aggregate Approximation (PAA) of input kinematic data to ease parameter optimization of the whole Siamese network. We show that these enhancements significantly improve the prediction accuracy. Then, we apply our new regression approach that uses pairwise ranks of a query action against a set of reference actions as features to train a regression model. This allows the pairwise ranking model to be turned into an exact skill prediction model when needed. Finally, we demonstrate that our model can serve as solution for the third objective of skill assessment, i.e. monitoring of surgeon's own progress. To the best of our knowledge, this is the first study that reports an empirical result in that respect.

5.2.1. Data

The performance of the entire model was evaluated in two different publicly available surgery data sets obtained from the da Vinci robot systems. They can provide both three-dimensional kinematic data and stereo video of surgery tasks. The kinematic data contain variables of both master and slave's left and right manipulators. The kinematic data for each sample is considered as a multi-variate time series, in which each variable corresponds to a different motion-specific parameter.

JIGSAW [43], is a common benchmark dataset in the field. It has surgical data collected from eight subjects with different skill levels performing three different surgical tasks. Self-defined skill levels were based on participant self-classifications based on hours of experience as novice (< 10 h), intermediate (10–100 h) or expert (> 100 h) operators. The tasks are 'throw suturing', 'needle passing', and 'knot tying' performed on benchtop training phantoms. The data consist of 76 motion variables collected at 30 Hz, including tooltip positions and orientation, linear and rotational velocities, and gripper angle. A trial is a part of the data set that corresponds to one subject performing one instance of a specific task. Each subject is categorized by a fixed expertise level, but each trial may have a different skill score. This score is annotated using OSATS as a grading system.

JIGSAWS dataset consists of three different surgical tasks which have been performed by study subjects (surgeons) on bench-top models. All these three exercises (suturing, knot-tying and needle passing) are typically part of surgical skills training curricula. Kinematic data were collected directly from the da Vinci API. The details about these exercises are given below:

Suturing (SU): The surgeon picks up needle, proceeds to the incision (designated as a vertical line on the bench-top model), and passes the needle through the "tissue", entering at the dot marked on one side of the incision and exiting at the corresponding dot marked on the other side of the incision. The surgeon extracts the needle out of the tissue after the first needle pass. Then s/he passes it to the right hand and repeats this process three more times.

Knot-Tying (KT): The surgeon picks up one end of a suture tied to a flexible tube attached at its ends to the surface of the bench-top model and ties a single loop knot.

Needle-Passing (NP): The surgeon picks up the needle (in some cases not captured in the video) and passes it through four small metal hoops from right to left. The hoops are connected to the bench-top type at a small height above the surface. It was forbidden for the surgeons to move the camera. Moreover, they were not allowed to apply the clutch while performing the surgical operations. Figure 9 shows snapshots of the three surgical tasks

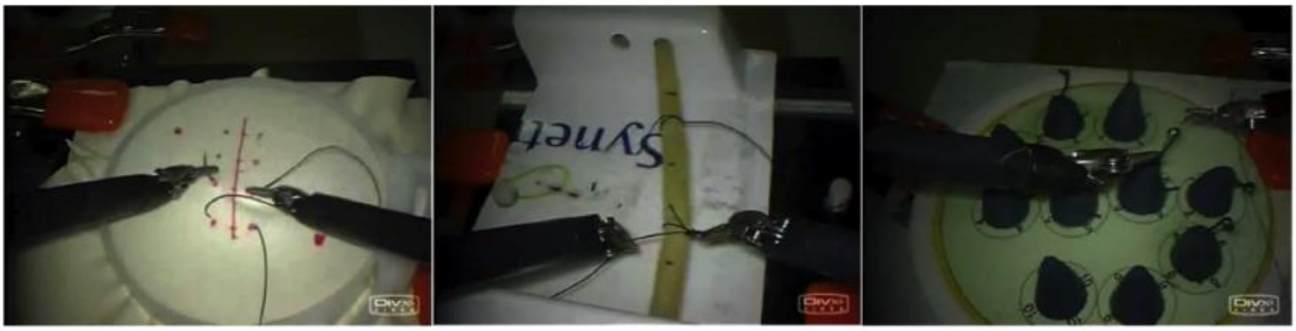
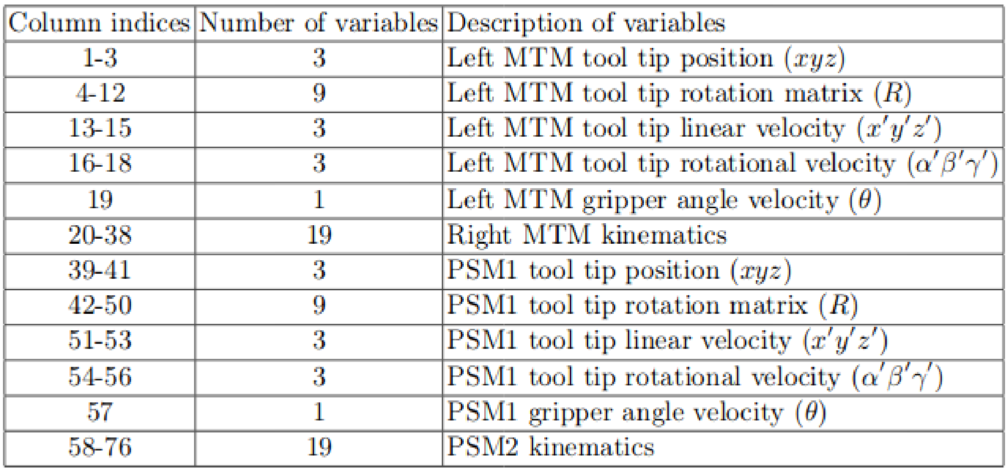


Figure 9. Snapshots of three different surgical exercises in JISGAWS dataset (from left to right): Suturing, knot-tying, and needle-passing

The dataset includes i) kinematic data, ii) video data, and iii) manual annotations. The kinematic dataset is collected from the da Vinci robot systems using its API at 30 Hz. The left and right MTMs, and the first and second PSMs (PSM1 and PSM2, also referred as the right and left PSMs in this dataset), are included in the dataset. The motion of each manipulator was described by a local frame attached at the far end of the manipulator using 19 kinematic variables, which brings us a 76-dimensional data in order to describe the kinematics information for all manipulators listed below. The 19 kinematic variables foreach manipulator include Cartesian positions, a rotation matrix, linear velocities, angular velocities, and a gripper angle. Cartesian positions are denoted by x, y, z variables, the rotation matrix is defined by 9 variables, denoted by R, linear velocities are denoted by x', y', z', angular velocities are denoted by α', β', γ', where α, β, γ are Euler angles. Finally, a gripper angle is denoted by θ. A common coordinate system is used to represent the kinematic variables. The details of the variables in the kinematic dataset are given in the Table 6. The sampling rates were the same which have been used to synchronize the kinematic data for the MTMs, PSMs, and the video data.

Table 6. Variables in kinematic data



Column indices	Number of variables	Description of variables
1-3	3	Left MTM tool tip position (xyz)
4-12	9	Left MTM tool tip rotation matrix (R)
13-15	3	Left MTM tool tip linear velocity ($x'y'z'$)
16-18	3	Left MTM tool tip rotational velocity ($\alpha'\beta'\gamma'$)
19	1	Left MTM gripper angle velocity (θ)
20-38	19	Right MTM kinematics
39-41	3	PSM1 tool tip position (xyz)
42-50	9	PSM1 tool tip rotation matrix (R)
51-53	3	PSM1 tool tip linear velocity ($x'y'z'$)
54-56	3	PSM1 tool tip rotational velocity ($\alpha'\beta'\gamma'$)
57	1	PSM1 gripper angle velocity (θ)
58-76	19	PSM2 kinematics

ROSMA [51] was recently released to facilitate the research in the field. It contains more samples and longer actions compared with JIGSAWS. Twelve subjects operated the da Vinci Research Kit to perform three different surgery tasks: post and sleeve, pea on a peg and wire chaser (Figure 10).

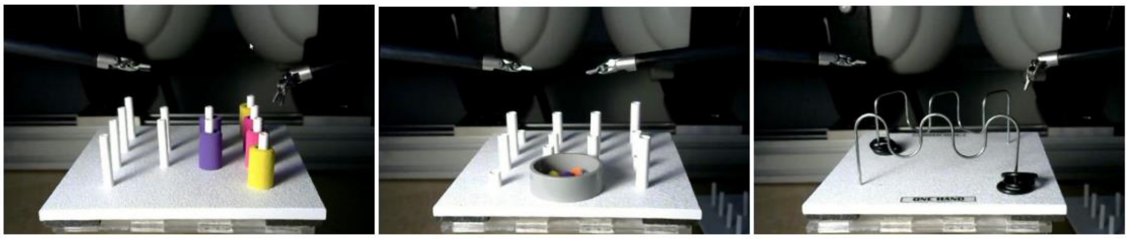
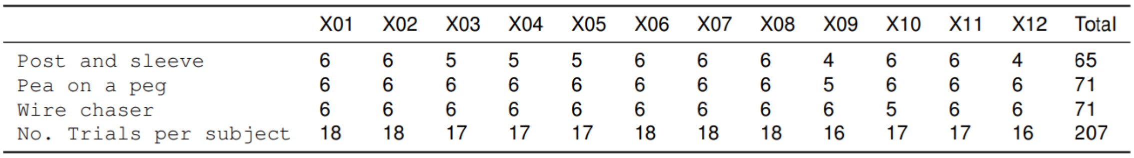


Figure 10. Snapshot of the three tasks in the ROSMA dataset at the starting position: (from left to right) post and sleeve, pea on a peg and wire chaser.

The twelve subjects (X01-X12) attempted each of the surgical task 4-6 different times to a total of 207 trials (Table 7). The obtained dataset includes all the kinematic and dynamic information provided by the da Vinci robot (both master and slave side). A board of human experts defined an objective performance scale by introducing penalty points for each surgery task. Then, each trial (subject + task) was given a score based on penalty points and completion time in seconds.

Table 7. Number of trials of each subject and exercise in the ROSMA dataset



	X01	X02	X03	X04	X05	X06	X07	X08	X09	X10	X11	X12	Total
Post and sleeve	6	6	5	5	5	6	6	6	4	6	6	4	65
Pea on a peg	6	6	6	6	6	6	6	6	5	6	6	6	71
Wire chaser	6	6	6	6	6	6	6	6	6	5	6	6	71
No. Trials per subject	18	18	17	17	17	18	18	18	16	17	17	16	207

Using JIGSAW and ROSMA data, we performed experiments in three different evaluation setups for (1) pairwise ranking of different surgeons, (2) regression to predict the exact skill level, and (3) monitoring of individual skill.

5.2.2. Experimental Setup

5.2.2.1 Ranking

We aim first to evaluate our framework in a common setup to justify our own model parameters and to benchmark against current state-of-the-art for pairwise ranking. To this end, we built an experimental setup that performed a four-fold cross validation to evaluate the prediction performance. In this setup, the pairs between ¾ of the surgery actions were used for training and the remaining pairs were used for testing. The fold were organized such, that the test samples included both the pairs where neither video has been used in a pair for training and the pairs where the other video was used for training in a different pairing. The model performance is discerned using pairwise ranking accuracy, which is the percentage of correctly ordered pairs, produced by each testing fold. This scheme reports two different accuracy results for the cases where the skill equivalence is considered and where it is not. When skill equivalence is considered, the accuracy gives the evaluation of ternary ranking performance. Otherwise, it evaluates the binary ranking. Table 8 lists the conditions of correct ordering of a pair (*m*,*n*) in binary and ternary cases. We used ε=0.01 in our evaluations.

Table 8. Conditions for correct predictions of pairwise ranking

Ranking type	p_{mn}	Ground truth
Ternary	≥ 0.5+ Ɛ	*m›n*
	≥ 0.5- Ɛ and < 0.5+ Ɛ	*m≡n*
	< 0.5- Ɛ	*m‹n*
Binary	≥ 0.5	*m›n*
	< 0.5	*m‹n*

We applied our model for each surgery task separately to rank surgery actions by their skills.

5.2.2.1 Regression

We argue that the results of pairwise rankings can be used for prediction of the exact score of surgery skill. To do this, we offer a method which could translate a list of pairwise ranks into an exact score of skill level. The conventional way of regression involves extracting a number of features from input signals to represent the sample in a machine learning model. Instead, we use an empirical representation where each feature refers to the pairwise rank between the query sample and another sample from a reference list. A pairwise rank here refers to the probability of the query action being performed better than the corresponding sample in a reference list.

In the regression setup, the performance of predictions was evaluated using Spearman's Correlation Coefficient (SCC) between actual and predicted values of skill levels, as suggested by [47] [39], two previous studies that adopted the idea of using regression for surgery skill assessment. We followed the same procedure to benchmark our method against these methods in the same dataset. SCC is a nonparametric metric that evaluates how well the relationship between two distributions can be described by a monotonic function. Ten-fold cross-validation was performed to measure the performance.

5.2.2.1 Monitoring

Our last objective is to demonstrate that the pairwise ranking model can be used for measuring the progress of a candidate surgeon during training activities. This demonstration is done using the ROSMA dataset, in which different trials are available from the same surgeon on the same surgery task. Instead of a typical k-fold cross-validation, we performed a leave-user-out (LUO) procedure for testing. In this procedure, the trials of one user (surgeon) are left out for prediction, while all other pairs of the remaining trials on the same surgery task are used for training. This was repeated 12 times for each surgeon independently. Final, accuracy was determined by averaging the pairwise ranking accuracy of these folds.

5.2.3 Implementation

PAA is implanted 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 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.

5.3. Results

5.3.1 Ranking

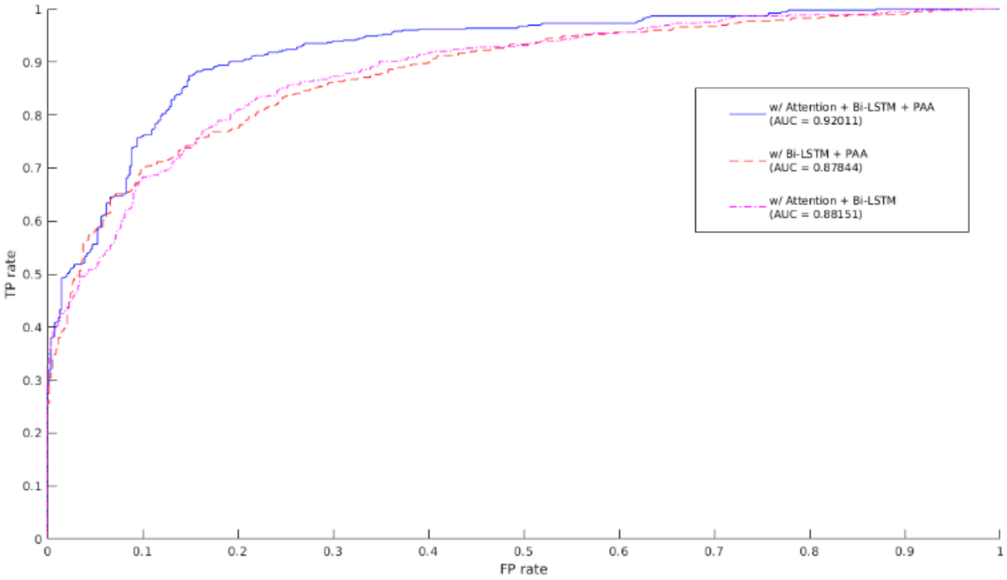
Table 9 discerns the accuracy for each task for ternary and binary ranking.

Table 9. Results of pairwise ranking with the present framework.

Surgery type	Ternary ranking (Including skill equivalence)	Binary ranking (Excluding skill equivalence)
	Acc	Acc
Knot tying	79.2	83.65
Needle passing	78.87	82.48
Suturing	69.29	72.89
AVG	**75.8**	**79.67**

Figure 10 shows Receiver Operating Characteristic (ROC) curve for the proposed model when applied for binary pairwise ranking. The ROC curve depicts the performance of the model is also discerned when the attention layer is removed. The figure shows that the attention enhancement has a significant contribution for the prediction performance. Reported ranking accuracy decreased to 74.64% when attention mechanism is eliminated. The contribution PAA step is also shown in the figure. The PAA can boost the prediction accuracy around 74%.

a.



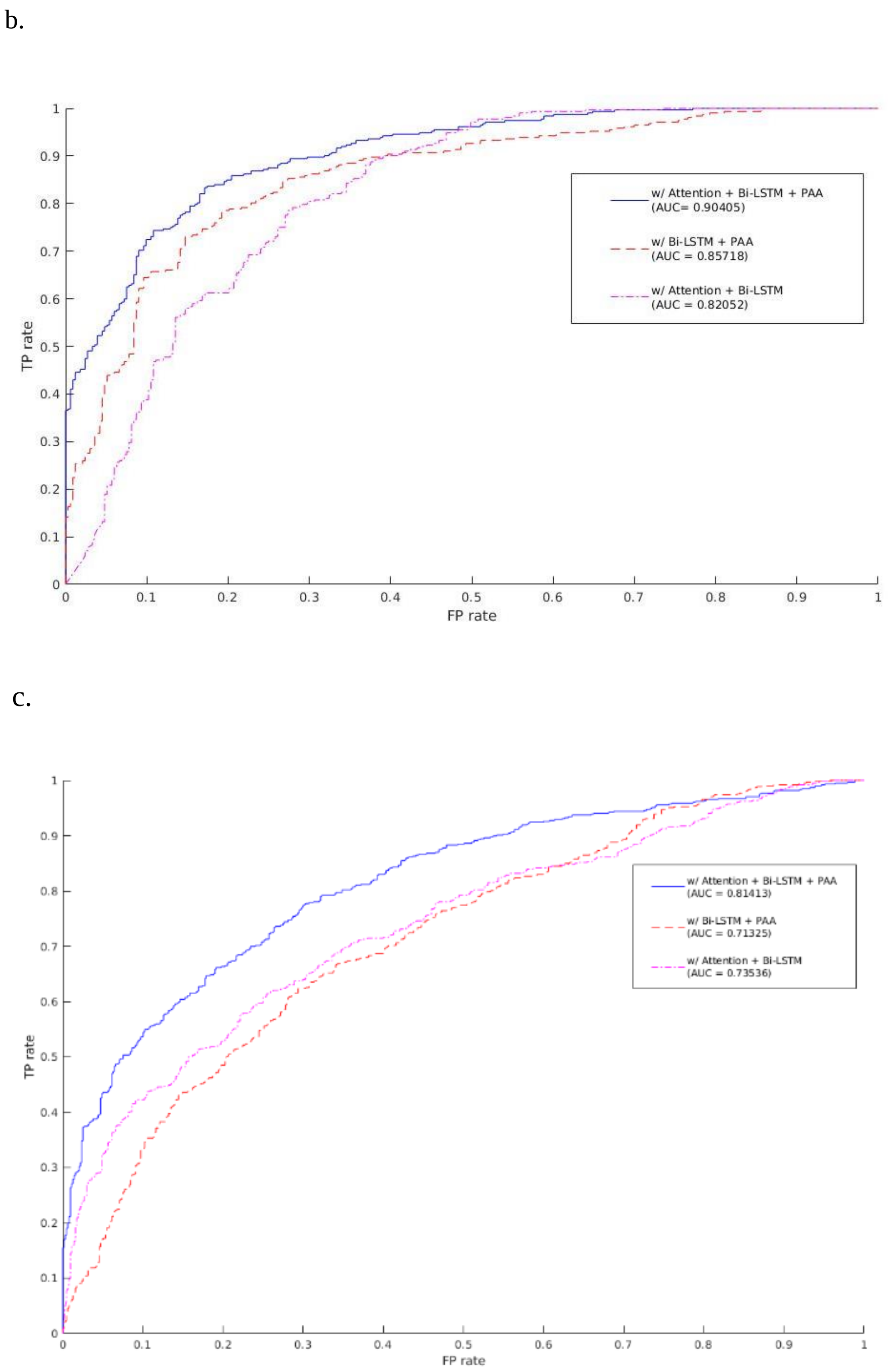


Figure 11. ROC curves for binary ranking for surgery skill assessment for

(a) knot tying,

(b) needle passing,

(c) suturing.

Although kinematic data is a multivariate signal with so many sensory measurements, it involves two main characteristic channels. One represents the changes in the position of the arms and the other refers to varying velocity over time. To understand the contribution of these two characteristics, we run binary ranking experiments with positional features and velocity features separately. The experiments revealed that the binary ranking accuracies with positional characteristics are 77.33%, 74.99% and 67.83% for knot tying, needle passing and suturing respectively. With velocity characteristics only, the model can achieve the accuracies of 71.95%, 67.88% and 71.1% for the same tasks. According to the results, positional features contribute more on ranking performance, however, the integration of velocity features improves the final accuracy.

The present model was compared with three most relevant studies in the literature. Two of them used video data for skill ranking and tested their methods in the same dataset. The third study is our own preliminary model on kinematic data presented in [50]. Video-based methods work for only binary ranking cases since their loss function didn't support the evaluation of equivalence in skills. They did not give accuracies separately for each task, but rather reported overall performance in surgery dataset. To make a comparison with these methods we ran our model with a subset of the original data in which the equally-rated pairs were removed. We calculated the average of accuracies achieved with three surgery types.

The results are shown in Table 10. Our model can significantly outperform both video-based methods and the kinematic-based method in terms of pairwise ranking accuracy. Moreover, the present model built upon kinematic data reduces the computational resource requirements compared to approaches which use video recordings. [48] reported that average running time to train a single fold is 18 hours with NVIDIA TITANX GPU, whereas learning a fold in our model is conducted in less than an hour with a conventional CPU.

Table 10. Results of pairwise ranking excluding skill equivalence

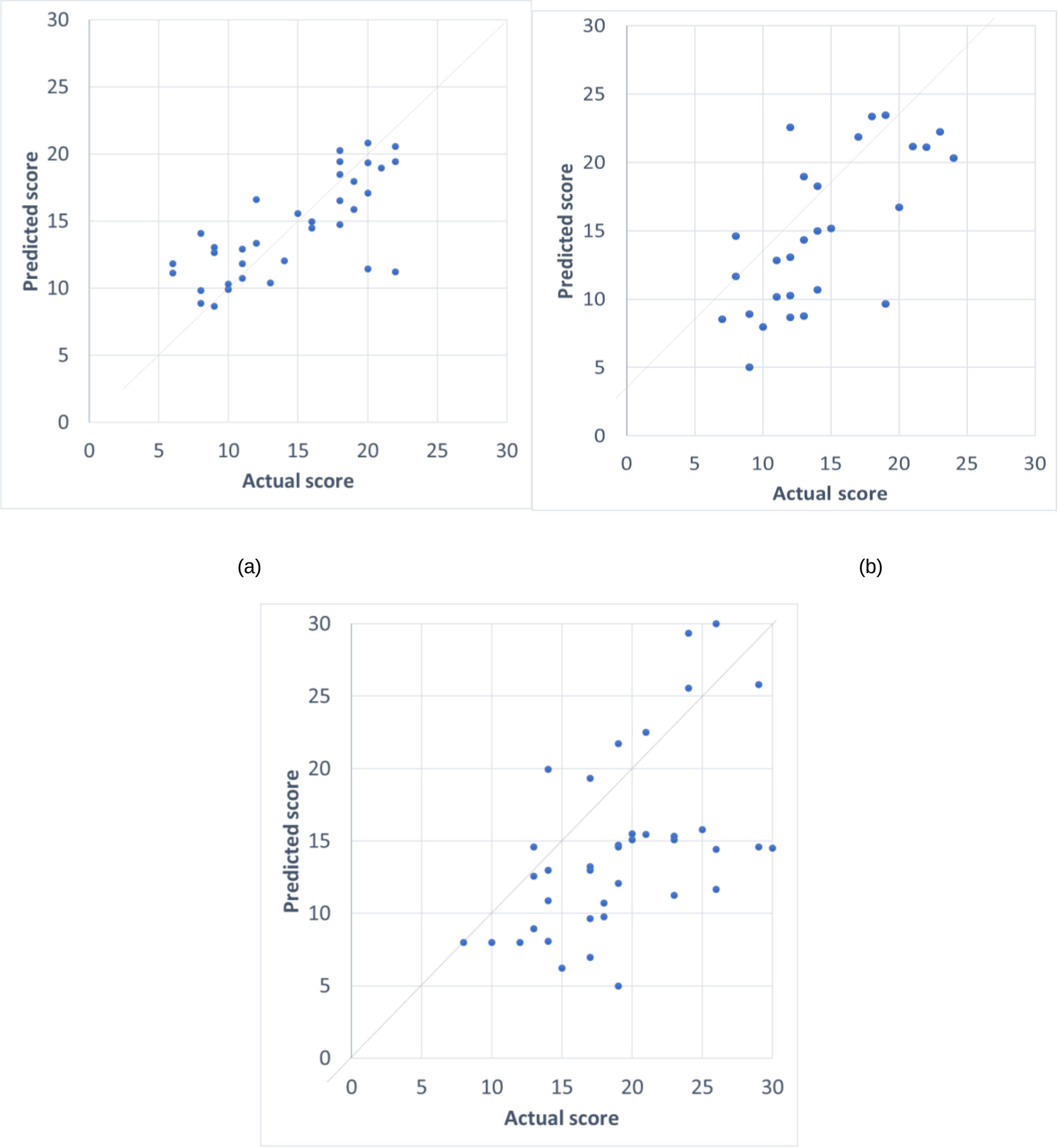
Method	Action data	Surgery type	Accuracy (%)
Doughty et al. 2018 [48]	Video	-	74.4
Li et al. 2019 [49]	Video	-	73.1
Ogul et al. 2019 [50]	Kinematic	Knot tying	79.6
		Needle passing	77.5
		Suturing	63.5
		Average	73.5
Present study	Kinematic	Knot tying	83.7
		Needle passing	82.5
		Suturing	72.9
		Average	**79.7**

Table 11 shows the results of the same architecture on ROSMA dataset. This performance is also consistent with the results of pairwise rankings that we obtained in first dataset, which therefore constitutes a validation of our model in an independent dataset.

Table 11. Results of pairwise ranking with present framework on ROSMA.

Action	Acc
Wire chaser	75.6
Post and sleeve	75.1
Pee on a peg	74.9
AVG	**75.2**

Figure 12 shows the results of exact value predictions as the comparison of predicted scores against actual scores for each task.



(a)

(b)

(c)

Figure 12. Scatter plots for predicted skill scores vs actual scores for the tasks of (a) knot tying, (b) needle passing, and (c) suturing

5.3.2 Regression

The results are given in Table 12. This experiment validates that the pairwise ranking model could be turned into a regression model with increased performance.

Table 12. Comparison of regression models for surgery skill assessment.

Method	SCC		
	Knot tying	Needle passing	Suturing
Zia and Essa, 2018	0.66	0.45	0.59
Fawas et al., 2019	0.65	0.57	0.60
Present method (with actual ranks)	0.99	0.99	0.99
Present method **(with predicted ranks)**	**0.71**	**0.65**	**0.59**

5.3.3 Monitoring

According to Table 13, our model achieved 70% pairwise ranking accuracy.

Table 13. Performance of our method in individual progress monitoring.

Action	Ranking accuracy (%)
	Present method
Wire chaser	73.9
Post and sleeve	66.7
Pee on a peg	69.4
Average	**70.0**

6. RANKING DIVING SKILLS

6.1. Background

Modelling and analyzing human motion have been subject to extensive research in computer vision [52], in terms of feature extraction [53], action representation [54] [55], action recognition [56] [57], and abnormality detection [58]. These efforts mostly address the challenging tasks of motion and action detection and recognition for many applications in various domains like sports and healthcare [55].

Sports analysis is a useful application of technology, providing value to athletes, coaches, and sports fans by producing quantitative evaluations of performance [59]. During recent years, the focus on vision-based analysis of sports has increased significantly in both research and commercial systems [60]. According to the survey by [61], some of the best-known current application areas are in sports analysis for broadcast, for example showing the position of players [62] or the ball as 3D models to allow the locations or trajectories to be explored in detail by a TV presenter. The ability to track a ball [63] [64] [65] [66] in low-latency real-time is important for both analysis in broadcast TV and helping the referee or umpire. Similarly, once players have been located and player modeling has been done, analyzing the motion of players [67] [68] at key moments in sport can give useful insights for both trainers and broadcasters. Although today's commercial systems apply many fundamental techniques such as tracking players and ball and analyzing the motion of both individual players and teams, to fully automate the video analysis of sports events many issues are still open for research [61]. For sport as a human centered activity, the initial step in many of the automatic analysis tools or software is to extract the athlete. This is done by locating and segmenting each person of interest in the video. Moreover, in some cases it becomes essential to follow the athlete through the entire video. However, there are several challenges that effects these tasks. The first aspect is the body posture of an athlete. It can vary greatly during sports exercises, decreasing the performance of any standard human/pedestrian detector. Occlusion is another significant challenge. Athletes can be partly or fully occluded by, e.g., other athletes, people, equipment, or obstacles. In any contact sport or team sport occlusion between people is a frequent problem and includes cases of collisions and interactions between several players simultaneously.

As mentioned above, human-centered video analysis has witnessed a significant increase of research activities during the last years [69]. Among various analysis techniques, two main questions arise: “What action” (recognition problem, without any localization) [70] [71] [72] [73] and “Where in the video?” (localization problem) [74] [75] [76] [77] [78]. Among these considerable amounts of work that study the recognition of human actions in video, the problem of “How Well?” people perform actions is a relatively new area that needs the attention of researchers in this field. The term “action quality assessment” refers to how well a person performed an action [79]. However, the problem of action quality assessment is directly addressed by only a handful of works. [80] and [81] use human pose features to assess action quality of Olympic sporting events. In the first study, a pose estimator is run on every frame. The results are concatenated in order to form a large action descriptor. This approach learns a regression model from spatio-temporal features. The authors use two types of features. The first set of features are low-level features which capture gradients and velocities directly from pixels. The second type are high-level features which are related with the trajectory of human pose. The descriptor is post-processed (DCT, DFT) into features which are used for estimating the parameters of a support vector regression (SVR) model to predict the event scores (quality). [81], uses approximate entropy-based feature representation to model the dynamics in human movement to achieve temporal segmentation in untrimmed motion capture data and fine-grained quality assessment of diving actions in videos. Different from traditional dynamical modeling approaches [82] which do not use any information about the interdependencies between joints, [81] develops a dynamical model by extending conventional ideas to quantify the interdependencies between body joints. Using the estimated pose for each frame, they propose a new approach – approximate entropy-based feature representation to model the dynamics in human movement by quantifying dynamical regularity. All of the APEN features are concatenated to get a high-dimensional feature vector. The dynamical information in these videos is better encoded by their feature vector than by DCT. They indicate that their superior performance is because APEN feature encodes the dynamical information in the time series of poses while DCT does not. Additionally, the proposed framework incorporates the interdependency between the joints, while traditional approaches consider each joint independently. One of the disadvantages of employing human poses as a feature is that incorrectly estimated poses have a negative impact on the final output. Due to atypical body positions, pose estimation has been demonstrated to be difficult on diving and figure skating datasets [80]. Pose only descriptors neglect important cues used in sport "execution" scoring such as splash size in diving. In addition, relative pose does not reflect absolute positional information, which could be crucial for scoring (e.g., entry position of a dive). Therefore, visual features such as 3D convolutional neural networks (C3D) are expected to perform better. Three different frameworks are proposed by [79] to evaluate Olympic sports which utilize spatiotemporal features learned using C3D. Score regression is performed with: SVR, LSTM, and LSTM followed by SVR. Rather than extracting human pose explicitly, the authors proposed a system which leverages visual activity information to assess quality of actions. The first step of their sports assessment system is extracting spatiotemporal features from video. To do that, they use C3D network which has been shown to be effective at preserving temporal information in video and outperform 2D ConvNets [83]. The proposed system, which is implemented using Caffe [84], shows significant improvement over existing quality assessment approaches on the task of predicting scores of Olympic events. They found that C3D-SVR gave best results. In this work, they also introduce new datasets for sports score assessment: The existing MIT diving dataset is doubled from 159 samples to 370 examples. Also, they have introduced a new gymnastic vault dataset which contains 176 samples.

Learning a measure of similarity between pairs of objects is an important generic problem in computer vision. Rather than using popular machine learning techniques to learn a pairwise similarity function [85], many deep learning models have also been used in learning to rank [86] [87]. Deriving from the idea of exploring relative relationship through ranking in previous works, [87] uses a pairwise deep ranking model to learn the spatial and temporal DCNN architectures for performing highlight detection in egocentric videos by using pairs of highlight and non-highlight segments. Using this pairwise ranking idea, Doughty *et al.* (2017) is the first paper which determines skills from 4 different datasets by the pairwise deep ranking model. The authors use Temporal Segment Networks (TSN) architecture [88], which achieves state of the art performances for action recognition on UCF101 [89] and HMDB51 [90], to model long range temporal structure. They build a Siamese version of the two-stream TSN composed of a spatial and temporal stream. [48] is an extension of the previous study. Here, the authors test their same method on both stationary and egocentric recordings. They conclude that in Dough-Rolling, the kitchen-based CMU-MMAC dataset [91] and Chopstick-Using datasets, the egocentric allows for better performance, due to the camera position's closeness to the action as well as information in the head motion. Similar to [48], [92] aims to evaluate skill in a more general sense for a variety of tasks. In this work, the authors propose a novel deep network for evaluating user's video compared with the instructional video in terms of semantic similarity. Action units were encoded from dense trajectories and FV aggregation with LSTM network. The variable-length action unit features were then evaluated by the Siamese LSTM network on breakfast dataset [93].

In sport healthcare systems, the correct execution of well-defined movements also plays a crucial role in physical rehabilitation [94]. Human proprioception may not be sufficient to spot movement mistakes. Thus, expert trainers observing the movement can give the trainee proficient feedback for timely improving the quality of the performance and avoiding persistent inaccuracies. However, it is not the case that a personal trainer is always available to assess the quality of movements during their execution. Therefore, there is a strong motivation to develop automatic systems able to detect mistakes during the performance of well-defined routines for providing feedback in real time. [94] presented a learning-based method that provides visual assistance to the person performing an exercise by displaying real-time feedback, thus enabling the user to correct inaccurate body motion. They use a novel recursive neural network, the MGWR, that uses growing self-organization for the efficient learning of input sequences and evaluate their approach with a data set containing 3 powerlifting exercises performed by 17 athletes. With respect to their previous model [95], the current approach accounts also for learning motion intensity to better predict and assess the dynamics of actions. [79], cast the quality assessment in physical therapy as a classification problem. In this work, 3D pose information from a Kinect is used to determine if an exercise repetition was "good" or "bad" using popular ML techniques for classification. Among support vector machines (SVM), single and double layered neural networks (NN), boosted decision trees, and dynamic time warping (DTW) features, the study shows AdaBoosted tree performed best. The authors also introduce the pilot LAM (large amplitude movement) Exercise Quality Dataset which is designed to treat CP (cerebral palsy) in an automated fashion. HTKS [96] is a game-like cognitive assessment method, designed for children between four and eight years of age. [97] have introduced the CogniLearn system, which is used for automated video capture and performance assessment during the HTKS assessment. For this, they first perform human body pose estimation using the pose estimator called DeeperCut [98], then using the body pose estimates from DeeperCut, they act a classification module that determines whether the subject touched his or her head, shoulders, knees, or toes. The CogniLearn system compares the part that was touched with the part that should have been touched based on the rules of the game and assesses the overall accuracy score of the person playing the game. Different from these physical exercises in healthcare, [69] proposes a framework that assesses how well people practice Sun Salutation, which is a simple, and effective series of Yoga Postures that invigorates the whole body, using Hidden Markov models with STIP features. The performers who either perform an action too fast (small time interval) or are not consistent across multiple cycles or take unequal rests after attaining a key pose while performing slow-paced Sun Salutation is detected as bad performers by the proposed system.

Third application of the RSRNA model introduced here is the relative assessment of diving skills of athletes in Olympic championships. In this application, the motion signals are extracted from video sequences of diving sessions using pose estimation techniques. When the body positions from several locations are obtained over time, this sequence is fed into RNSA model in the same way as the other two applications. Here, it is important to notice that pose estimation may not work in %100 accuracy, therefore, an information is loss is expected in the input layer, which will affect the final accuracy of ranking. The adopted model is tested in a public video dataset and benchmarked against conventional machine learning models based of absolute assessment of skills. Again, the existing methods were run in the same experimental setup with ours to discern their abilities in pairwise ranking.

6.2. Materials and Methods

6.2.1. Model Adoption

The RSRNA model is designed to work for comparing two multi-variate time-series motion signals in terms of one of their attributes which represents the quality of the action.

Main problem in the adoption of diving skill comparison in Olympic events is the fact that the athletes are not allowed to wear any kind of sensors, but their motions can be captured by video recordings. Since the video recording contains many other components such the environment of the action, changing background and potentially other people around, a preprocessing method is needed at the input layer of the RSRNA model. For this, we used a pose estimation approach where video recording can be transformed into a signal of multi-position coordinates of the body on temporal dimensions (See Figure 13 for an example human pose estimation). The next steps are followed in the same way as how PD monitoring is implemented.

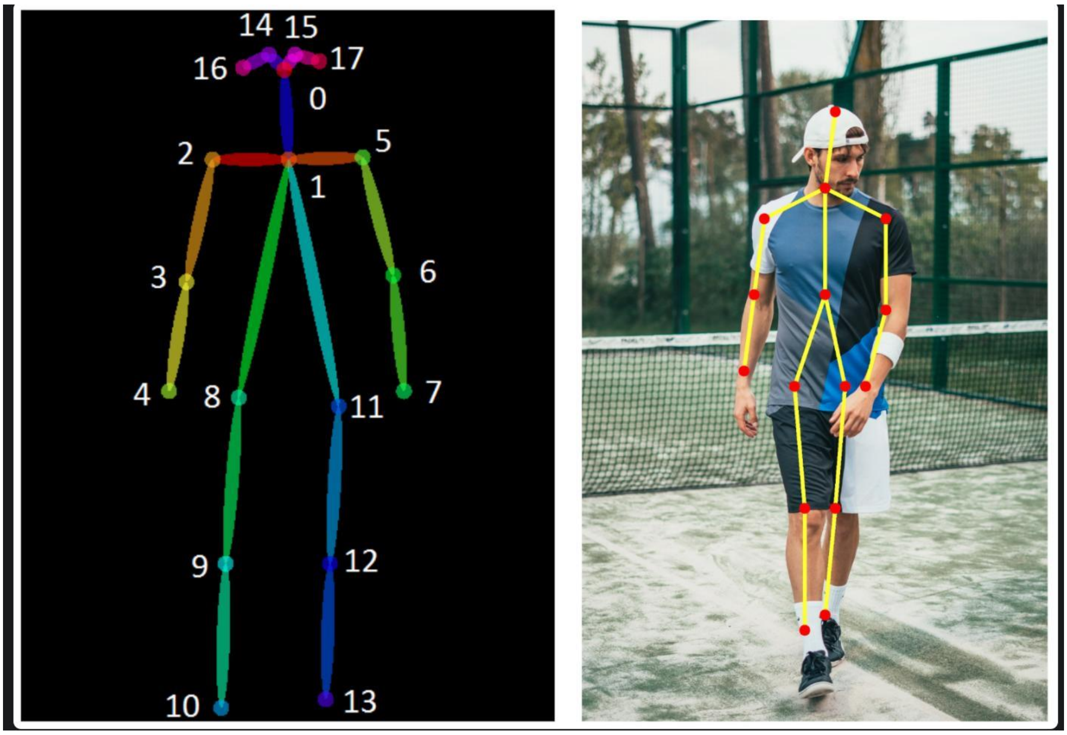
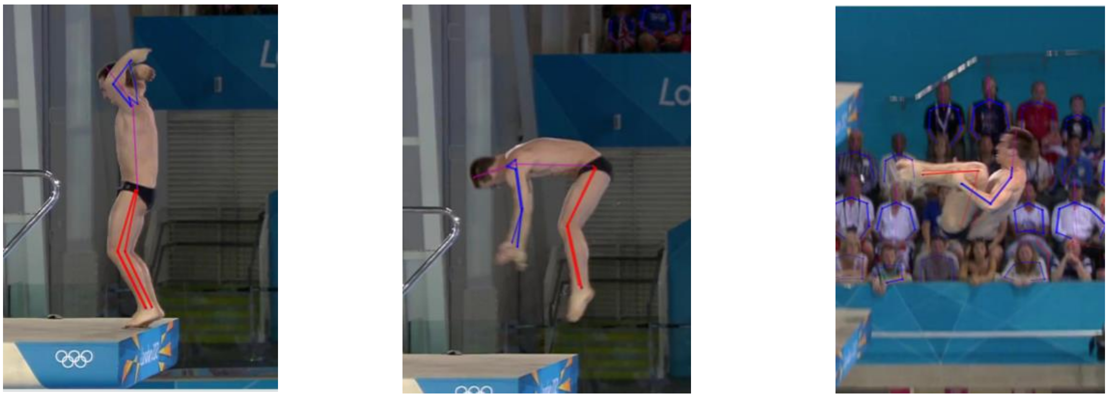


Figure 13. An example of how a video recording shot can be transformed into a pose signal.

In a video, our main assumption was we know the pose information of the athlete in each frame which is gathered from either through ground truth or automatic pose estimation. Let's say *q(j)(t)* be the *x* component of the j^{th} joint in the t^{th} frame of the video. All the joint positions that are relative to the head position are normalized because we want our position features to be translation-invariant using this formula: $f(j)(t) = q(j)(t) - q(0)(t)$ where *q(0)(t)* refers to the head of the performer based on our assumptions.

To extract the position of the body joints of an athlete from the entire video $q(j)(t)$, we run a pose estimation algorithm for every frame. We considered several pose estimation methods in this stage. The first one, alpha pose, [99] is a top-down approach which first detects individual people and then estimate each person's pose. This kind of approaches interpret the process of detecting key points as a two-stage pipeline, that is, firstly locate and crop all persons from image, and then solve the single person pose estimation problem in the cropped person patches. The second method we considered is a bottom-up approach [100]. The bottom-up approaches first detect body parts and then group these parts to human instances. [100] formulate the problem of multi-person pose estimation as part grouping and labeling via a Linear Program.

The results of these methods on diving dataset are shown in Figure 14 and 15. Although Alpha Pose is able to capture the poses which cannot be estimated by 2^{nd} method, it cannot be considered as successful. There are many frames (Figure 14) where the athlete pose cannot be found. Such cases force us to do missing data imputation to feed the pairwise rank model. Consequently, we choose to use sensory data as inputs to our deep learning framework for quality assessment.



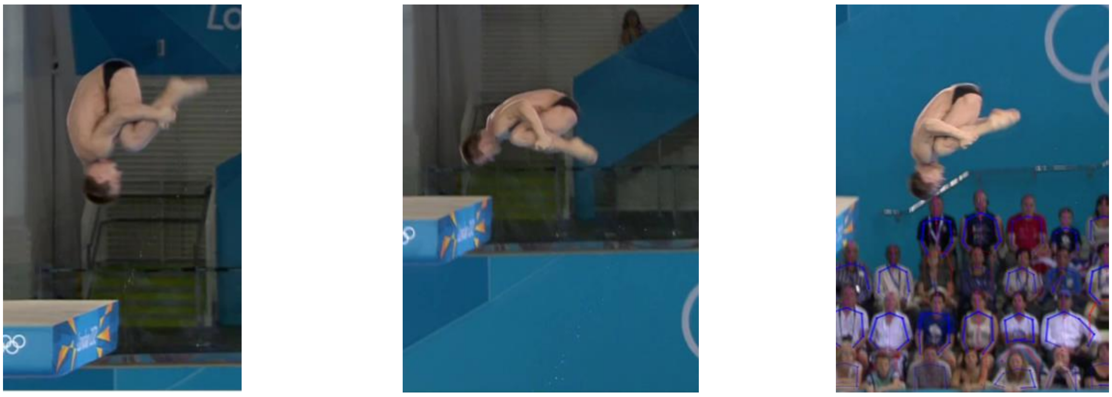
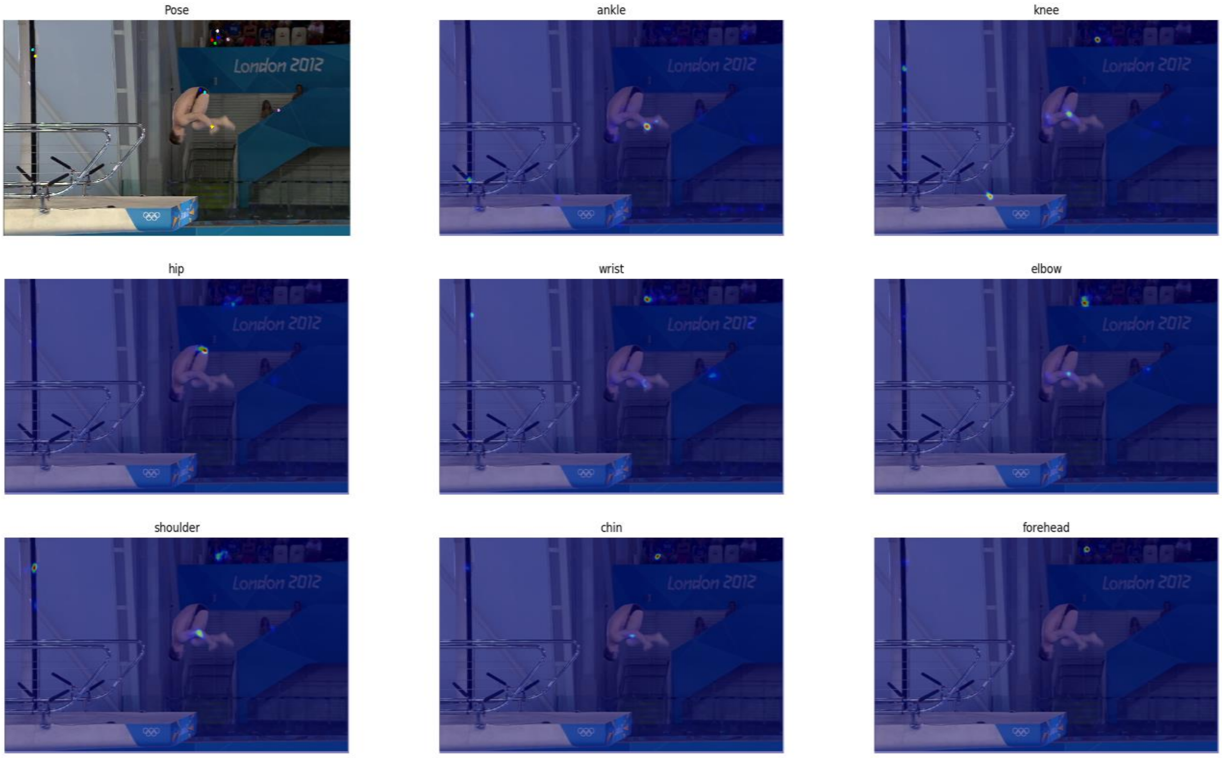


Figure 14. Alpha Pose Results on diving dataset (Top row: Successful results. Bottom row: Unsuccessful results)



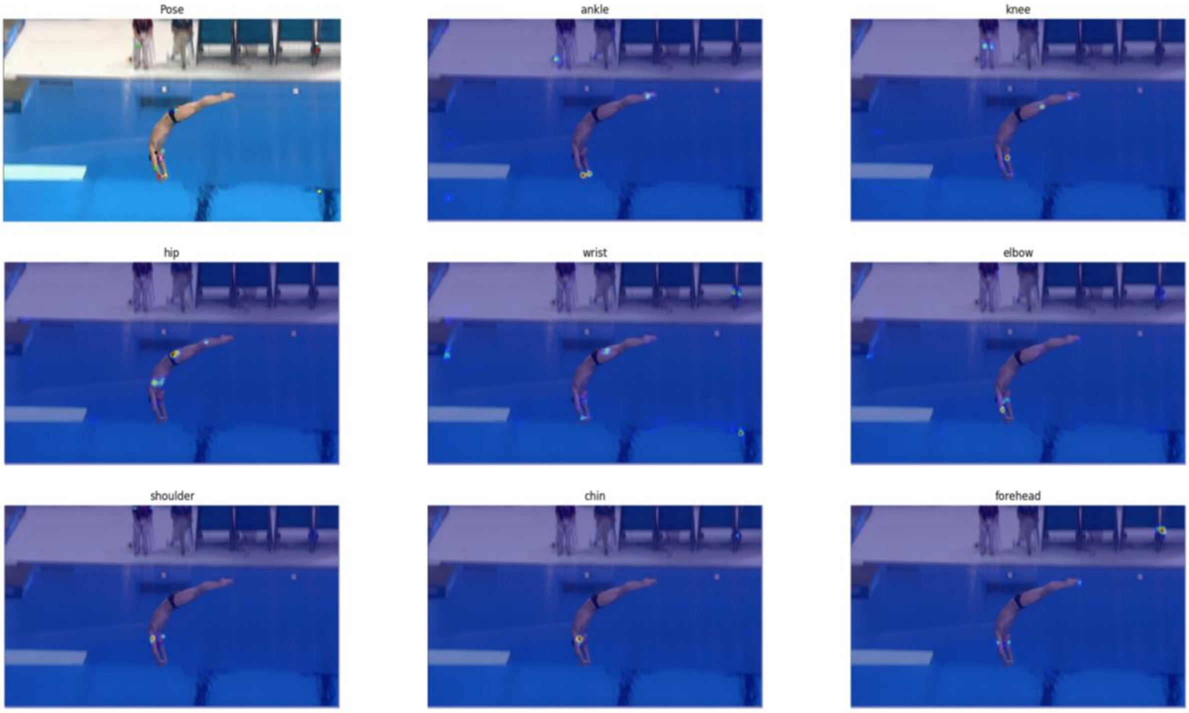


Figure 15. Pose Tensorflow results on diving dataset.

(Top row: Successful results. Bottom row: Unsuccessful results)

Finally, we considered the model suggested by [80] for action quality assessment. They used this pose estimations to extract features based on Discrete Cosine Transform (DCT) before feeding an SVM regression model. We extract the pose information using Flexible Parts Model [101] for each frame independently, as done by [80]. Yang and Ramanan [101] used dynamic programming in order to find the best pose in a single frame. Since our aim is to find the best pose through the entire video, we extract the N-best pose from each frame using [102]. Then, using a dynamic programming method, we correlate the poses to discover the best track in the entire video. The association looks for the single best smooth track covering the whole temporal span of the video.

6.2.2. Data

In [80] an Olympics video dataset is introduced for the action quality assessment problem. Sports footage has the advantage that it can be obtained freely, and the expert judge's scores are frequently released publicly. The dataset consists of YouTube videos from recent Olympics and other worldwide championships for two categories of sports: diving, and figure skating. The videos are long with several instances of multiple people performing different Olympic actions. Annotation is done with the start and end frame for each Olympic video. The dataset is publicly available. (http://people.csail.mit.edu/hpirsiav/quality.html).

There are 159 videos in the diving dataset. Because the videos are slow-motion broadcasts from television channels, the effective frame rate is 60 frames per second. Each video is approximately 150 frames, and there are 25,000 frames in the entire dataset. The ground truth judge scores vary between 20 and 100 where 20 is the worst performance. In the paper, authors use 100 instances for training. The remaining instances are used for testing procedures. Each experiment was repeated 200 times with different random splits, and the authors find the average of the results. The study does not only calculate the score of the event but also gives the feedback to the athlete so that s/he can improve his/her dive. Some examples diving actions in the dataset are shown in Figure 16.

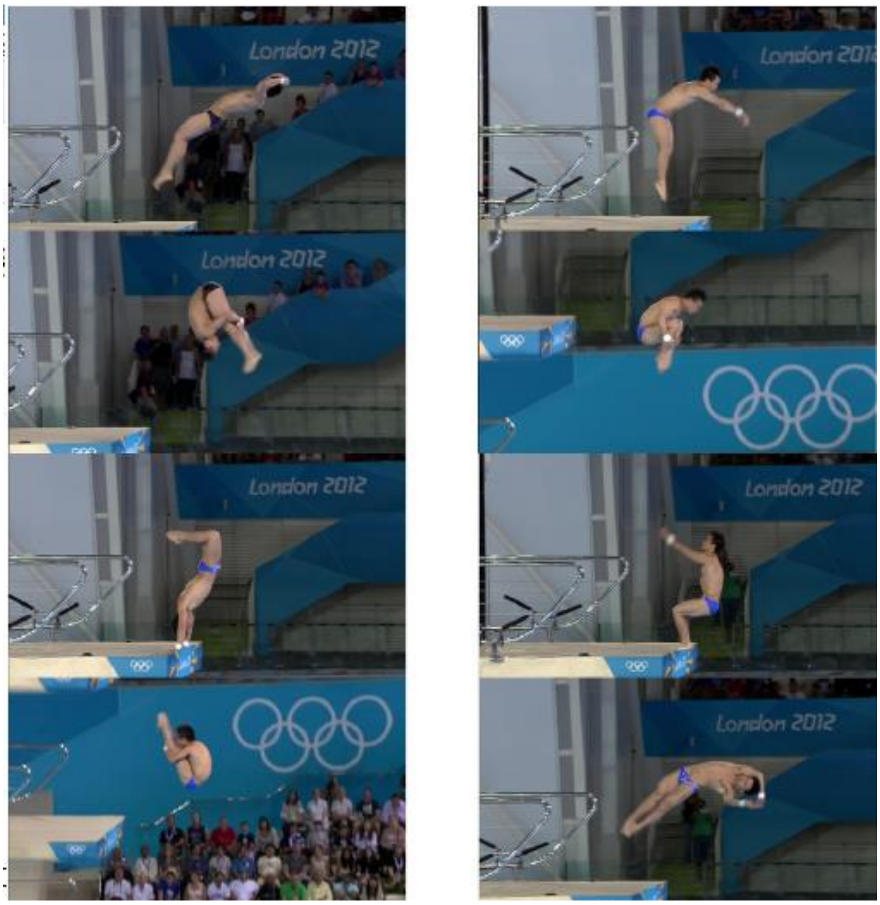


Figure 16. Example diving video shots from Olympics video dataset for action quality assessment.

6.2.3. Experimental Setup

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, pairwise ranking accuracy (Acc) is used. This is the percentage of correctly ordered pairs generated by each testing fold. Depending on whether the rank layer models the equivalence of diving skills of two athletes, 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.

6.2.4 Implementation

For pose estimations, we used a Matlab implementation of the human pose estimation algorithm described in [101] [103]. It includes pre-trained full-body and upper-body models. Much of the detection code is built on top of deformable part-based model implementation [104]. The training code implements a quadratic program (QP) solver described in [105]. The code is trained and tested using positive images from the PARSE dataset [106] the BUFFY dataset [107] and negative images from the INRIA Person Background dataset [108].

The other components of RSRNA model are implemented in the same way as done for PD monitoring application (See 4.2.4). A sigmoid activation layer was used to model the probabilistic rank layer, which is followed by a binary cross-entropy loss function in the 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.

6.3. Results

We have used pose information of the athletes to rank their performances. Table 14 shows the pairwise ranking accuracies for binary and ternary cases with different configurations of the RSRNA model

Table 14. Results of pairwise ranking on Diving Dataset using body joints

Method	**Binary ranking**	**Ternary ranking**
Siamese bi-LSTM	70.02	68.68
Siamese bi-LSTM w/ PAA	70.16	69.09
Siamese bi-LSTM w/ attention	**73.18**	**72.07**

7. DISCUSSION AND CONCLUSION

Although pairwise ranking of entities is an old task in computing, relative assessment of actions from motion signals has not been studied well in the literature. We have seen some examples of pairwise ranking for image and video data with different applications. This thesis however addresses a new problem; relative assessment of actions based on their quality using a pairwise ranking model of sensory motion signals in the form of a multi-variate time-series. To this end, the study proposes a deep learning model built upon a two-stream Siamese architecture, called RSRNA. Each stream of the framework represents the other motion where the output of the model refers to the probability of first action having a better quality than the second one. To the best of author's knowledge, pairwise ranking of time-series has been studied for the first time in this thesis.

The framework is adopted for three real-time applications in different domains, namely, remote healthcare, surgery training and sports in Olympic events. The first application aims to monitor PD patients through gait analysis using a relative assessment scheme for gait skills. The second application is the assessment of surgery skills via kinematic signals obtained via a surgery robot used by surgeons during training. The objective of the last application is to rank the diving athletes during Olympic events based on their performance quality.

The RSRNA framework adapts an effective pre-processing step for smoothing the motion signal to overcome the problem of sparsity in observed position information over the action. A piecewise aggregate approximation scheme, which attempts to provide a balanced probabilistic distribution over the vertical dimension of the signal, enable the model to run in reduced space and converge faster in parameter learning. The experiments showed that this scheme provides better accuracy in surgical skill and diving quality assessments while it makes no difference in the PD monitoring application. This can be contributed to the fact that the motions in first two applications span in a larger scale with a smaller number of training samples compared with PD gait dataset. On the other hand, PD gait signals acquired from foot motions fluctuates less in all dimensions and attains a lower sparsity in general with the addition of more training samples.

The framework models temporal information using an LSTM at each stream of the framework. Although all RNNs have feedback loops in the recurrent layer which let them maintain information in memory over time it can be difficult to train standard RNNs to solve problems that require learning long-term temporal dependencies. This is because the gradient of the loss function decays exponentially with time, which is called as vanishing gradient problem. LSTM uses special units in addition to standard units which include a 'memory cell' that can maintain information in memory for long periods of time. A set of gates is used to control when information enters the memory, when it's output, and when it's forgotten. This architecture lets them learn longer-term dependencies. Our experimental results in all three applications justify that LSTM performs better than standard RNNs for skill assessment. This is probably due to the fact that the quality of an action usually dependent on several temporal movements over time but not only the nearest changes. On the other hand, this dependency may not be structured enough to be modeled by special gates of LSTM. That’s why an attention enhancement is integrated in the proposed framework. Again, the results have justified the use of attention mechanism for pairwise ranking of skills in all experimental setups.

As with most AI applications, the main limitation of current model is small data. Especially in the medical field, there is considerably less data than in video and text-based systems, which brings us unreliable results while training regression or classification models. More specifically, in our applications, we have had 275 subjects in gait and 159 athletes in diving datasets. For this reason, building a regression or a classification model on these small data may overfit or may not represent the data well. However, using a pairwise approach, we have addressed the small data problem indirectly. Since we study on all the pairs and compare the performances of each pair, we have more than 35000 data in gait dataset, 12000 data in diving dataset.

The thesis brings a new modality for computational quality/performance/skill assessment. We have considered the skill assessment problem as a relative assessment task. Relative assessment approach provides a more interpretable and reliable view of the progress while it overcomes the limitations caused by inconsistencies in subjective grading scales.

Moreover, we have showed that the model is not data type dependent. The same model is useful on different data type: Sensory, Kinect and Video (Body Joints).

We have also showed that we can consider pairwise rankings as an intermediate step for a regression task. Using pairwise scores as features, we can train a regression-based models which brings us a succeeding correlation result. This approach has an additional advantage that the empirical representation scheme is independent from the problem under consideration but can adapt to the case directly based on the pairwise rank of training samples.

Each application of RSRNA model comes together with particular arguments in additional to general conclusion inferred from the proposed model and its applicability in several domains. We first introduce 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 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 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 [109] or monitoring patients in intensive care units via physiological vital signs [110]. 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.

A novel framework for objective skill assessment for robot-assisted surgery using kinematic data is also introduced, that shall be used for choosing, credentialing and monitoring of surgeons. The model provides a more interpretable and reliable view of skill assessment for surgical operations. The experimental results justify that this model can achieve better accuracy than the state-of-the-art methods in both ranking and regression setups for surgery skill assessment. This model may help to overcome the limitations caused by inconsistencies in subjective skill grading scales, that are used to train such systems. Compared to video-based solutions, the use of kinematic data reduces the demands on computational power and is therefore a more applicable alternative for the practical implementation in a hospital setting.

To our knowledge, this is the first study that has considered and experimented the task of individual progress monitoring for surgery skills from a computational perspective. We describe how our model can be used in this context and validate it empirically in a recent dataset. The empirical results are promising; these results will serve as a strong baseline for future studies in monitoring task.

One of the limitations of the current study is the fact that reported pairwise rankings may violate triangular consistency, which will result in an unidentifiable full ranking of all actions. Although this information is not always requested in real-life surgeon trainings, considering the consistency in full ranking in the loss function may improve the prediction accuracy of the model. This is left for future work. The need for further validation of the ROSMA dataset with deeper statistical analysis challenges another future study. Another limitation is related to the kinematic data. Although kinematic data has an advantage over video data in capturing three-dimensional motion information, kinematic data does not contain contextual and semantic information such as the smoothness and strength of the movement, and the interaction between tools and tissue. Therefore, it may be a future direction to integrate video and kinematic data for more accurate ranking predictions with the expense of increasing computational costs. As a result, the assessment of surgical skill needs further investigation to perform in an objective way. Current progress in kinematic sensor data analysis is considered as a powerful complementary tool to manual assessment. It is reasonable to suggest that assessing surgical skill requires multiple simultaneous assessments, including machine-learning-based decision support systems as offered in the present study.

Computerized skill assessment of athletes in Olympic event has physical limitations due to unacceptability of wearable sensors during the activity. Therefore, video-based solutions have received more attention recently by computer vision researchers. The greater ability of deep learning systems supported by transfer learning in image and video data has pushed more the research in this direction. Still, we have applied RSRNA model for this problem assuming that human pose information is available as a multi-variate time-series to represent the action of the athletes. However, since we have to estimate the pose information from video stream, the input is prone to information loss due to false predictions in this stage. Although this approach is to competitive with more resource-demanding three dimensional convolutional deep learning models, we have shown that our approach can perform better that other conventional methods based on hand-crafted features.

In general, 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 will find wide range of applications in several domains, but more particularly in the health domain, to compare patients based on their physiological recordings or motion signals.

8. REFERENCES

[1] E. Hoffer and N. Ailon, "Deep Metric Learning Using Triplet Network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015.

[2] F. Cakir, K. He, X. Xia, B. Kulis and S. Sclaroff, "Deep metric learning to rank," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Y. Shi, Y. Niu, W. Guo, Y. Huang and J. Zhan, "Pairwise learning to rank for image quality assessment," *IEEE Access*, vol. 8, pp. 192352-192367, 2020.

[4] H. Doughty, W. Mayol-Cuevas and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Conf. Comput. Vis. Pattern Recognit., CVPR*, 7862-7871.

[5] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *AAAI Conference on Artificial Intelligence*, 2016.

[6] A. Graves, S. Fernández and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition," in *International Conference on Artificial Neural Networks*, 2005.

[7] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations*, 2015.

[8] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton and G. Hullender, "Learning to rank using gradient descent," in *International Conference on Machine Learning*, 2005.

[9] B. Fulcher. and N. Jones, "Highly comparative feature-based time-series classification," *IEEE Trans Knowl Data Eng*, vol. 26, p. 3026–3037, 2014.

[10] R. J. Kate, "Using dynamic time warping distances as features for improved time series classification," *Data Mining and Knowledge Discovery volume,* vol. 30, p. 283–312, 2016.

[11] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of Computational Biology,* vol. 10, p. 857–868, 2003.

[12] L. V. Kalia and A. E. Lang, "`Parkinson's disease," *Lancet,* vol. 14, pp. 896-912, 2015.

[13] R. Bhidayasiri and D. Tarsy, "Parkinson's disease: hoehn and yahr scale," in *Movement Disorders: A Video Atlas,* Berlin, Germany, Springer, 2012, p. 4–5.

[14] V. Dibilio, A. Nicoletti, G. Mostile, G. Portar, A. Luca, F. Patti and M. Zappia, "Computer-assisted cognitive rehabilitation on freezing of gait in Parkinson's disease: A pilot study," *Neuroscience Letters,* vol. 654, pp. 38-41, 2017.

[15] B. Kostek, K. Kaszuba, P. Zwan, P. Robowski and J. Slawek, "Automatic assessment of the motor state of the Parkinson's disease patient--a case study," *Diagnostic Pathology* , vol. 7, pp. 1-8, 2012.

[16] S. Patel, K. Lorincz, R. Hughes and P. Bonato, "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors," *IEEE Trans. Inf. Technol.,* vol. 13, pp. 864-873, 2009.

[17] J. S. Perlmutter, "Assessment of Parkinson disease manifestations," in *Current Protocols in Neuroscience,* 2009.

[18] C. Ramaker, J. Marinus, A. M. Stiggelbout and B. J. v. Hilten, "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease," *Movement Disorders,* vol. 17, pp. 867-876, 2002.

[19] P. Srivastava, A. Shukla, P. Vepakomma, N. Bhansali and K. Verma, "A survey of nature-inspired algorithms for feature selection to identify Parkinson's disease," *Computer Methods and Programs in Biomedicine,* vol. 139, pp. 171-179, 2017.

[20] S. H. Lee and J. S. Lim, "Parkinson's disease classification using gait characteristics and wavelet-based feature extraction," *Expert Systems with Applications,* vol. 39, pp. 7338-7344, 2012.

[21] M. R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson’s disease," *Biomedical Signal Processing and Control,* vol. 8, pp. 66-70, 2013.

[22] Y. N. Jane, H. K. 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.

[23] O. Ertugrul, Y. Kaya, R. Tekin and M. Almali, "Detection of Parkinson's disease by shifted one dimensional local binary patterns from gait," *Expert Systems with Applications,* vol. 56 , pp. 156-163, 2016.

[24] D. Joshi, A. Khajuria and P. Joshi, "An automatic non-invasive method for Parkinson's disease classification," *Computer Methods and Programs in Biomedicine,* vol. 145, pp. 135-145, 2017.

[25] K. Acici, C. Erdas, T. Asuroglu, M. Toprak, H. Erdem and H. Ogul, "A random forest method to detect Parkinson’s disease via gait analysis," in *International Conference on Engineering Applications of Neural Networks*, 2017.

[26] 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. 315, pp. 1-8, 2018.

[27] Y. Xia, Z. Yao, Q. Ye and N. Cheng, "A dual-modal attention enhanced deep learning network for quantification of Parkinson's disease characteristics," *IEEE Trans. Neural Syst. Rehabil. Eng.,* vol. 28, pp. 42-51, 2020.

[28] T. Asuroglu, K. Acici, C. B. Erdas, M. K. Toprak, H. Erdem and H. Ogul, "Parkinson's disease monitoring from gait analysis via foot-worn sensors," *Biocybern. Biomed. Eng.,* vol. 38, pp. 760-772, 2018.

[29] S. Aghanavesi, J. Westin, F. Bergquist, D. Nyholm, H. Askmark, S. Aquilonius, R. Constantinescu, A. Medvedev, J. Spira and F. Ohlsson, "A multiple motion sensors index for motor state quantification in Parkinson's disease," *Computer Methods and Programs in Biomedicine,* vol. 189, p. 105309, 2020.

[30] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artif. Intell. Med.*, vol. 75, pp. 51-63, 2017.

[31] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, P. 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, pp. 215-220, 2002.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[33] J. A. Martin, G. Regehr, R. Reznick, H. Macrae and J. Murnaghan, "Objective structured assessment of technical skill (osats) for surgical residents," *British Journal of Surgery*, vol. 84, pp. 273-278, 1997.

[34] P. v. Hove, G. Tuijthof, E. Verdaasdonk and et al., "Objective assessment of technical surgical skill," *Br J Surg*, vol. 97, pp. 972-987, 2010.

[35] I. Rivas-Blanco, C. Perez-Del-Pulgar, I. Garcia-Morales and V. Munoz, "A Review on Deep Learning in Minimally Invasive Surgery," *IEEE Access*, vol. 9, pp. 48658-48678, 2021.

[36] M. Fard, S. Ameri, R. D. Ellis, R. Chinnam, A. Pandya and M. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, 2018.

[37] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin and I. Essa, "Video and Accelerometer-Based Motion Analysis for Automated Surgical Skills Assessment," *International journal of computer assisted radiology and surgery*, p. 443–455, 2018.

[38] Z. Wang and A. M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International Journal of Computer Assisted Radiology and Surgery,* vol. 13, p. 1959–1970, 2018.

[39] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar and P.-A., "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery,* vol. 14, p. 1611–1617, 2019 .

[40] D. Zhang, Z. Wu, J. Chen, A. Gao, X. Chen, P. Li, Z. Wang and G. Yang, "Automatic Microsurgical Skill Assessment Based on Cross-Domain Transfer Learning," *IEEE Robot. Autom. Lett,* vol. 5, p. 4148–4155, 2020.

[41] I. Funke, S. T. Mees, J. Weitz and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery,* vol. 14, p. 1217–1225, 2019.

[42] X. Nguyen, D. Ljuhar, R. N. M. Pacilli and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer Methods and Programs in Biomedicine,* vol. 177, pp. 1-8, 2019.

[43] Y. Gao, S. Vedula, C. Reiley, N. Ahmidi, B. Varadarajan, H. Lin, L. Tao, L. Zappella, B. Béjar, D. Yuh and et al., "JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling.," in *MICCAI Workshop: Modeling and Monitoring of Computer Assisted Interventions (M2CAI),* 2014.

[44] F. Pérez-Escamirosa, A. Alarcón-Paredes, G. A. Alonso-Silverio and et al., "Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches," *International Journal of Computer Assisted Radiology and Surgery,* vol. 15, p. 27–40, 2020.

[45] J. D. Kelly, A. Petersen, T. S. Lendvay and T. M. Kowalewski, "Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks," *International Journal of Computer Assisted Radiology and Surgery,* vol. 15, p. 2079–2088, 2020.

[46] J. L. Lavanchy, J. Zindel, K. Kirtac, I. Twick, E. Hosgor, D. Candinas and G. Beldi, "Automation of surgical skill assessment using a three-stage machine learning algorithm," *Scientific Reports,* vol. 11, p. 1–9, 2021.

[47] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," *International Journal of Computer Assisted Radiology and Surgery* , vol. 13, p. 731–739, 2018.

[48] H. Doughty, D. Damen and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Computer Vision and Pattern Recognition*, 2018.

[49] Z. Li, Y. Huang, M. Cai and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[50] B. B. Oğul, M. F. Gilgien and P. D. Şahin, "Ranking Robot-Assisted Surgery Skills Using Kinematic Sensors," in *European Conference on Ambient Intelligence*, 2019.

[51] I. Rivas-Blanco and et al, "A surgical dataset from the da Vinci Research Kit for task automation and recognition," arXiv:2102.03643, 2021.

[52] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt and I. Craddock, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Computer Vision and Image Understanding,* vol. 148, pp. 136-152, 2016.

[53] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision,* vol. 3, p. 177–280, 2008.

[54] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang and J. Gall, "A Survey on Human Motion Analysis from Depth Data," in *CVPR*, 2013.

[55] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," in *ACM Computing Surveys*, 2011.

[56] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing,* vol. 28, p. 976–990, 2010.

[57] S. R. Ke, H. L. U. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo and K. H. Choi, "A review on video- based human activity recognition," *Computers,* vol. 2, pp. 88-131, 2013.

[58] O. Popoola and K. Wang, "Video-based abnormal human behavior recognition a review,," *IEEE Trans. Syst., Man, Cybern, Part C,* vol. 6, p. 865–878, 2012.

[59] B. Reily, H. Zhang and W. Hoff, "Real-time Gymnast Detection and Performance Analysis with a Portable 3D Camera," *Computer Vision and Image Understanding,* 2016.

[60] R. Gade, R. G. Larsen and T. B. Moeslund, "Measuring energy expenditure in sports by thermal video analysis," in *Computer Vision and Pattern Recognition,* 2017.

[61] G. Thomas, R. Gade, T. B. Moeslund, P. Carr and A. Hilton, "Computer vision for sports: current applications and research," in *Computer Vision and Image Understanding,* 2017.

[62] C. Bialik, "The people tracking every touch, pass and tackle in the world cup," 2014. [Online]. Available: https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/. [Accessed 15 06 2022].

[63] M. Fast, "What the Heck is PITCHf/x? - The Physics of Baseball," The Hardball Times Baseball Annual, 2010.

[64] S. Anderson, "Forsgren helps revolutionize golf on TV with Protrace," 2013.

[65] N. Owens, C. Harris and C. Stennett, "Hawk-eye tennis system," in *International Conference on Visual Information Engineering,* 2003.

[66] Innovations, "http://www.hawkeyeinnovations.co.uk./sports/cricket .," Hawk-Eye in Cricket, 22 11 2017. [Online].

[67] P. Prandoni, E. Reusens, M. Vetterli, L. Sbaiz and S. Ayer, "Automated Stroboscoping of Video Sequences". Patent US7042493B2, 2006.

[68] M. Reusens, M. Vetterli, S. Ayer and V. Bergnozoli, "Coordination and Combination of Video Sequences with Spatial and Temporal Normalization.," *European Patent Specification EP 4,,* 2007.

[69] H. Jain and G. Harit, "A framework to assess sun salutation videos.," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 2016.

[70] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *Pattern Analysis and Machine Intelligence,* p. 1644–1657, 2014.

[71] T. Lan, T. C. Chen and S. Savarese, "A hierarchical representation for future action prediction.," in *ECCV*, 2014.

[72] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos.," in *ICCV*, 2011.

[73] Y. Kong, D. Kit and Y. Fu, "A discriminative model with multiple temporal scales for action prediction.," in *ECCV,* 2014.

[74] G. Yu, N. A. Goussies, J. Yuan and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search.," in *IEEE Transactions on Multimedia*, 2011.

[75] Y. Tian, R. Sukthankar and M. Shah, "Spatiotemporal deformable part models for action detection.," in *CVPR,* 2013.

[76] M. Jain, J. Gemert, H. Jegou, P. Bouthemy and C. Snoek, "Action localization with tubelets from motion.," in *CVPR,* 2014.

[77] D. Oneata, J. Revaud, J. Verbeek and C. Schmid, “Spatiotemporal object detection proposals.,” in *ECCV,* 2014.

[78] L. Wang, Y. Qiao and X. Tang, "Video action detection with relational dynamic-poselets," in *ECCV*, 2014.

[79] P. Parmar and B. T. Morris, "Learning To Score Olympic Events," in *arXiv preprint arXiv:1611.05125*, 2016.

[80] H. Pirsiavash, C. Vondrick and A. Torralba, "Assessing the quality of actions.," in *European Conference on Computer Vision* , 2014.

[81] V. Venkataraman, I. Vlachos and P. K. Turaga, "Dynamical regularity for action analysis," in *Proceedings of the British Machine Vision Conference*, 2015.

[82] S. Ali, A. Basharat and M. Shah., "Chaotic invariants for human action recognition.," in *International Conference on Computer Vision*, 2007.

[83] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks.," in *ICCV*, 2015.

[84] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding.," in *arXiv preprint arXiv:1408.5093*, 2014.

[85] G. Chechik, V. Sharma, U. Shalit and S. Bengio, "Large scale online learning of image similarity through ranking.," *Journal of Machine Learning Research*, p. 1109–1135, 2010.

[86] Y. S. J. Wang, «Learning fine-grained image similarity with deep ranking.,» %1 içinde *CVPR*, 2014.

[87] T. Yao, T. Mei and C. W. Ngo, "Learning query and image similarities with ranking canonical correlation analysis.," in *ICCV*, 2015.

[88] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. V. Gool, "Temporal segment networks: towards good practices for deep action recognition.," in *ECCV*, 2016.

[89] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild.," preprint arXiv:1212.0402, 2012.

[90] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition.," in *ICCV*, 2011.

[91] F. D. l. Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database.," in *Robotics Institute*, 2008.

[92] S. T. Kim and Y. M. Ro, "Evaluationnet: Can human skill be evaluated by deep networks?," arXiv preprint arXiv:1705.11077, 2017.

[93] H. Kuehne, A. B. Arslan and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Computer Vision and Pattern Recognition,* 2014.

[94] G. I. Parisi, S. Magg and S. Wermter, "Human motion assessment in real time using recurrent self-organization.," in *Robot and Human Interactive Communication (RO-MAN),* 2016.

[95] G. I. Parisi, F. v. Stosch, S. Magg and S. Wermter, "Learning human motion feedback with neural self-organization," in *International Joint Conference on Neural Networks,* 2015.

[96] M. M. McClelland, C. E. Cameron, R. Duncan, R. P. Bowles, A. C. Acock, A. Miao and M. E. Pratt., "Predictors of early growth in academic achievement: The head-toes-knees-shoulders task.," in *Frontiers in psychology*, 2014.

[97] S. Gattupalli, D. Ebert, M. Papakostas and V. A. F. Makedon, "CogniLearn: A Deep Learning-based Interface for Cognitive Behavior Assessment," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017.

[98] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model.," in *CoRR abs/1605.03170,* 2016.

[99] H.-S. Fang, S. Xie, Y.-W. Tai and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[100] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres and B. Schiele, "Articulated multi-person tracking in the wild," 2016.

[101] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Conference on Computer Vision and Pattern Recognition,* 2011.

[102] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *International Conference on Computer Vision,* 2011.

[103] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," in *IEEE Pattern Analysis and Machine Intelligence*, 2013.

[104] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, p. 1627–1645, 2010.

[105] D. Ramanan, "Dual Coordinate Solvers for Large-Scale Structural SVMs.," 2014. [Online]. Available: http://arxiv.org/abs/1312.1743. [Accessed 20 06 2022].

[106] D. Ramanan, "Learning to parse images of articulated bodies," *Advances in Neural Information Processing Systems (NIPS),* vol. 19, p. 1129–1136, 2006.

[107] V. Ferrari, M. Marin-Jimenez and A. Zisserman, "Pose Search: Retrieving People using Their Pose," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[108] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005.

[109] K. C. Siontis, P. A. Noseworthy, Z. I. Attia and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology,* vol. 18, pp. 465-478, 2021.

[110] A. Davoudi, K. R. Malhotra, B. Shickel and et al, "Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning," *Scientific Reports,* vol. 9, 2019.

Publications

1. BB Oğul, M Gilgien, S Özdemir (2022)
   Ranking surgical skills using an attention-enhanced Siamese network with piecewise aggregated kinematic data
   Int. J. of Computer Assisted Radiology and Surgery 17, 1039–1048 ***(SCI-Exp)***.
   https://doi.org/10.1007/s11548-022-02581-8

2. BB Oğul, S Özdemir (2021)
   A Pairwise Deep Ranking Model for Relative Assessment of Parkinson's Disease Patients From Gait Signals
   IEEE Access 10, 6676-6683 ***(SCI-Exp)***.
   https://doi.org/10.1109/ACCESS.2021.3136724.

3. BB Oğul, MF Gilgien, PD Şahin (2019)
   Ranking robot-assisted surgery skills using kinematic sensors
   European Conference on Ambient Intelligence 2019,
   Lecture Notes in Computer Science 11912, 330-336 ***(SCI-CPI)***.
   https://doi.org/10.1007/978-3-030-34255-5_24