2021 24th International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2021.

Assessment of Rehabilitation Exercises from Depth Sensor Data

Shehzan Haider Chowdhury, Murshed Al Amin, A K M Mahbubur Rahman, M Ashraful Amin, Amin Ahsan Ali
Department of Computer Science and Engineering
Independent University, Bangladesh
{1720109, 1720069, akmmrahman, aminmdashraful, aminali}@iub.edu.bd

***Abstract*—Assessing the rehabilitation exercises are essential in the recovery and treatment of various musculoskeletal conditions following surgery. According to reports, over 90% of all rehabilitative exercise sessions are conducted in a home environment. As the number of patients grows, this method becomes prohibitively expensive. Providing technology support for home-based rehabilitation is an excellent approach to address this. The patient remains at home and does the exercises in front of the camera, with the footage or data being sent to the physician for comments on the exercises. In this paper, we propose two machine learning-based models to assess the quality of exercises where the data is captured by such kinect 3D sensors. The proposed models consist of a long short-term memory(LSTM) network which uses the time series skeletal data to predict the quality of the exercises. The first model uses the predefined features proposed by the physicians. For the second model, we extract features using graph convolutional network(GCN) on the skeletal data where each node represents a body part or joint in the body and the edges represent the connection between the body parts. We conclude that LSTM is more accurate at predicting the results when GCN features are used.**

***Index Terms*—Movement modeling, deep learning, health monitoring, Graph Convolutional Network (GCN), Long Short term memory (LSTM)**

I. Introduction

Rehabilitation exercises are a key part of a patient's post-operative recovery and treatment of many musculoskeletal conditions. Currently, physicians observe patients perform specific tasks or exercises ranging from walking and sitting-to-standing to deep squats, etc. to evaluate and set objectives for their physical mobility. Nevertheless in the long run it is neither feasible nor economical for a physician to be present for every rehabilitation exercise session [1]. Therefore, the initial stages of the exercise are performed under the direct supervision of a physician in a rehabilitation facility while the second stages consist of prescribed exercises that the patient performs at their home setting. Reports indicate that over 90% of all rehabilitation exercise sessions are being done in a home-based setting [2]. Even though in these circumstances the patients are required to record and report their progress and intermittently visit the physicians for an assessment, multiple medical sources have reported that patients are unable to perform the exercises correctly [3]. That leads to extensions of the recovery period. The use of an automated system to evaluate and provide feedback on how well the exercise is done would reduce the hassle of periodically visiting the physician. At the same time, the system would allow the patients to fix their own mistakes as the system would evaluate the movement and provide necessary feedback. By automating the task of patient exercise assessment, health service establishments can aim to reduce cost and improve home-based exercise to reduce the patient recovery period.

Fig. 1. Sample frames from KIMORE for 5 exercises [6].

The task of evaluating the quality of assessment of exercises falls under the general category of human action analysis. In recent years a large amount of research has been done to detect and classify human actions, for example, identifying standing-up, sitting down and walking motions from videos. However in exercise quality assessment we are particularly interested in analyzing the exercise.

Recent work on the quality of human movement assessment has gained attention resulting in various tools and devices to assist physical rehabilitation. For example, Sardari et al. [4] proposed a view-invariant method using a pre-trained convolutional neural network (CNN) to evaluate the quality of human movement. Their use of OpenPose, which is a real-time multiple-person detection library, fails to generate sufficient consistent heat maps resulting in lowered performance while also requiring heavy resources. On the other hand, as exercises are events that are related to time series, LSTM has been proven useful by Liao et al. [5]. However, their model is validated by measuring variations in movement data without any ground truth assessment. Therefore, there is still a lack of robust, lightweight systems for automatic monitoring and assessment of patient's performance.

The aim of the paper includes establishing and comparing models for the assessment of physical rehabilitation exercises using the skeleton dataset: KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical REhabilitation (KIMORE). Figure 1 shows sample frames from the KIMORE dataset for different exercises. The first

978-1-6654-9435-9/21/$31.00 © 2021 IEEE

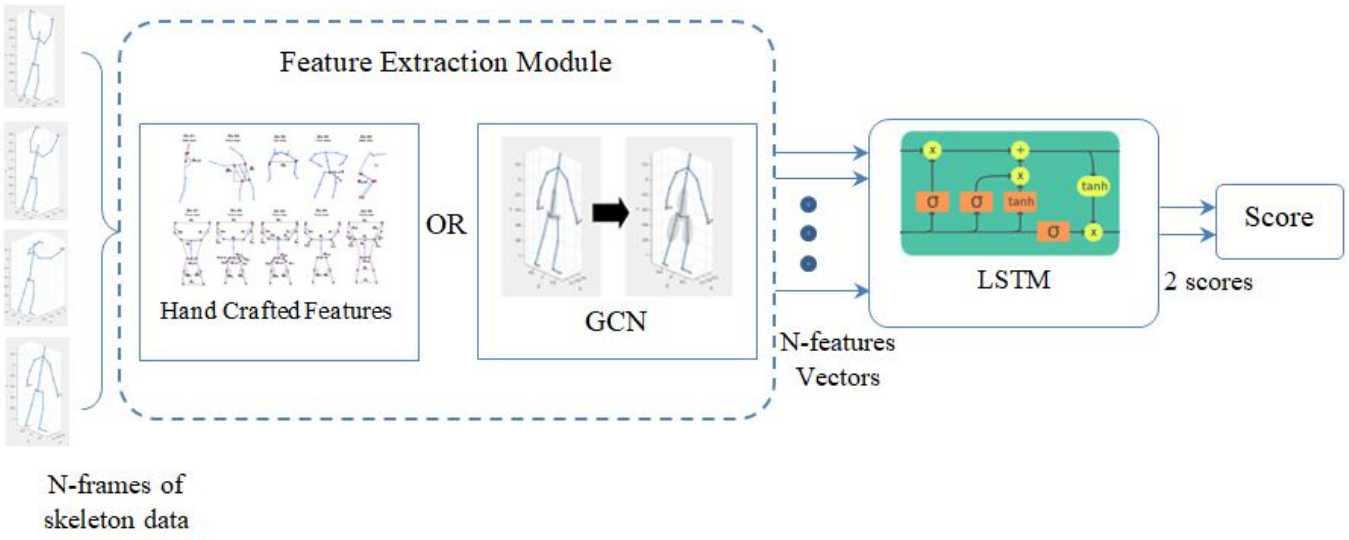

Fig. 2. Proposed model has a feature extraction module that feeds into an LSTM. The feature extraction module calculates the features from each frame and the LSTM uses these features to predict a score.

model uses handcrafted features(angles and distance between joints) (left on fig. 2) defined by the physicians in KIMORE while the second model generates features computationally using a GCN (right on fig. 2). With the proposal of two models, this paper is also able to contrast the feature extraction processes. Therefore this paper aims to answer the question whether physician's defined features are necessary for automated assessment of physical rehabilitation exercises. Intuitively, handcrafted features are expected to provide a good result as doctors use them to assess their patients, but using a GCN allows the model to encode the whole skeleton allowing the LSTM to provide better predictions. Thus, the main contributions of this paper are as follows:

- We propose a new model for the assessment of rehabilitation exercises using physician's prescribed features calculated from skeletal data.
- We provide a new model for the assessment of rehabilitation exercises using GCN features learned from skeletal data.
- We provide a contrast of automatically generated features and physician's features.

II. Related Work

Action analysis has recently picked up pace in the past few years, of which most research in the area is either based on physical rehabilitation, skill assessment or sports analysis. Within this sub-field of study, many researchers have worked on similar goals using non-skeleton-based methods, such as CNN for sports scoring, and then applied their work on physical rehabilitation [7]. Out of the many movement assessment studies, many of them are image or depth-image-based models. The image datasets traditionally consist of RGB data using a regular camera while depth-based datasets are often collected using sensors such as Kinect, a motion-sensing input device produced by Microsoft. Using image-based models, recent research has produced promising results using CNN. For example, Crabbe et al. [8] suggested using a CNN network to map depth images onto a high-level pose within a manifold space. Next, they used the high-level poses information onto a statistical model, to evaluate the quality of movement.

As motion analysis is dependent on the efficiency of the movement over a time period, some researchers however focus more on a temporal-based model. For example, Liao et al. [5], adapted a long short term memory (LSTM) based model, using 3D motion capture skeleton data to assess rehabilitation movement. Then they used a performance metric based on Gaussian mixture model log-likelihood to provide an estimation of the movement score. On the other hand, Elkholy et al. [9] used motion capture to calculate spatio-temporal descriptors to assess the quality of movement for walking on stairs, stand-up, sitting down and walking motions. They estimated a score by modeling the spatio-temporal descriptors of movements into a linear regression.

Similar to the above-mentioned studies, Sardari et al. [4] proposed a view-invariant method to assess the quality of human movement. They implemented an end-to-end CNN that is made up of two stages. A view-invariant trajectory descriptor is used to form a collection of trajectories for all joints. They use this in an adaptive, pre-trained 2D CNN to establish spatial relationships and predict a score for the movement quality. They applied their model on their own generated dataset, a multi-view, non-skeleton, non-motion capture, rehabilitation movement dataset (QMAR), and also in a Kinect based on the KIMORE dataset. As the KIMORE dataset is not a multi-viewed dataset, the model was limited to using a single view.

Nor Rashid et al. [10] proposed a deep learning model, for

skeleton-based physical rehabilitation exercise classification, by implementing a spike train feature on the UI-PRMD dataset. They encoded the data into spike trains as spike train features are hugely rewarding towards deep learning.

GCN-LSTM has proven useful in a wide range of fields ranging from action recognition to image captioning. Jinyin Chen et al. [11] proposed a GCN embedded LSTM for end to end Dynamic link prediction. They claimed to be the first to use GCN-LSTM for predicting dynamic links in a network. The use of GCN allows their model to learn the node structure of the network in every time frame and the utilization of LSTM allows the model to learn temporal features. Their results show the model is capable of outperforming state of the art methods. A similar model, proposed by Yanguo Huang [12] is targeted for Short-Term Traffic Flow Prediction. They Proposed a GCN-LSTM with encoder and decoder structure trained with traffic flow data. Their implementation shows GCN-LSTM can predict short term traffic flow on real data. The use of GCN allows the model to employ the spatial features and thus has a higher prediction accuracy compared to a LSTM. Meanwhile, Zhishuai Li [13] proposed a hybrid deep learning method called graph and attention-based long short-term memory network (GLA). They employed a GCN to extract spatial features of traffic over multiple observation stations. A data driven approach allows them to create the adjacency matrix. The output of the GCN is used in a LSTM to learn the temporal features and finally the use of a soft attention mechanism allows the model to predict the flow of traffic. Ting Yao [14] also explored the use of attention based encoder-decoder framework for image captioning. They presented a GCN-LSTM, which used graphs built over detected objects in images based on spatial and semantic relations. The model used a LSTM based Captioning model that employs an attention mechanism for sentence generation.

Researchers have engaged GCN-LSTM on skeleton based data proving its usability in capturing spatial-temporal data. Han Zhang [15] proposed a GCN-LSTM to automatically learn spatio-temporal features to model action sequences that concern both dimensions. Their proposed model uses a GCN that feeds into a conventional RNN unit in every time stamp. Meanwhile Zhao et al. [16] uses GCN from pose data from skeleton data for Skeleton based Action Recognition. Firstly, they propose a GCN-LSTM to capture the spatio-temporal features and then employ a Bayesian framework to capture more variation in data. The classification is done by framing a Bayesian inference problem. The benefits of the framework are expressed in comparison with other similar models.

III. Dataset

There are many skeleton-based datasets widely used for human action recognition. When it comes to physical rehabilitation exercises, two datasets are the most prominent, University of Idaho - Physical Rehabilitation Movements Data Set (UI-PRMD) [17] consisting of only skeleton data and KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical Rehabilitation (KIMORE) [6]

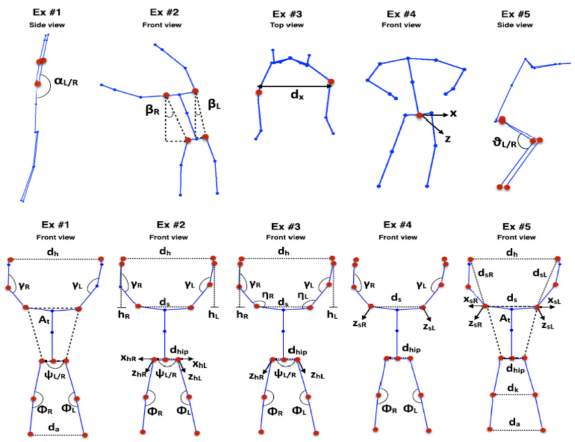

Fig. 3. Features prescribed by Physicians for the five exercises. Row one contains the Primary outcomes and Row two illustrates the Control factors [6].

dataset consisting of RGB depth video, along with skeleton joint position and orientations. But to our knowledge, no other datasets other than the KIMORE dataset have defined features and also physician's assessment or scoring.

The KIMORE dataset provides a collection of different physical rehabilitation exercises collected using a RGB-D sensor, Kinect. The data were collected for five different rehabilitation exercises, primarily specific for lower back pains which are prescribed by physicians. Along with the sensor data, this dataset comes with a set of features, which are specifically defined by physicians to evaluate and assess the quality of motion performed by the subjects. These features are then used to validate with respect to a stereophotogrammetric system to give a score to the subject's performance. The dataset also covers an assessment of the same performance by the physicians, collected through a clinical questionnaire.

The KIMORE dataset consists of a large heterogeneous population of 78 subjects, divided into 2 groups with 44 healthy subjects and 34 with motor dysfunctions, which are then further classified into three separate classes. The classes include patients with back pain, patients that suffered from strokes, and patients who suffer from Parkinson's disease. All participants perform five different exercises. The exercises are:

1) Lifting of the arms features
2) The lateral tilt of the trunk with the arms in extension
3) Trunk rotation
4) Pelvis rotations on the transverse plane
5) Squatting

The features in the KIMORE dataset are classified into two categories, Primary Outcome (PO), and Control Factor (CF). POs and CFs signify the movement of upper limbs and physical constraints during the exercises. Corresponding to the features, the dataset also provides two classes of scores, with values for PO in the range of 0 to 15 and CF in the range of 35, totaling to a range of 0 to 50.

The KIMORE dataset has a performance metric verified by

a physician therefore to validate and test our method we will use the KIMORE dataset for experimentation.

IV. Proposed Method

Although the KIMORE dataset provides us the data of 25 joint positions, orientation, and also depth video, we will utilize only the joint positions. In the future, this will allow us to easily calibrate our model for multiple similar datasets. The KIMORE paper has identified multiple handcrafted features for the exercise the patient performs.

The proposed models have a feature extraction module that feeds into a LSTM as shown in figure 2. In the first module, we extract the features using either of the two different methods. Firstly, we extracted the mentioned handcrafted features for all the exercises. Secondly, we devised a GCN-based graph encoding that will capture the features of each frame. The intuition for this second process is to analyze the effects of computer-generated features and human specified features. This module is targeted to capture the spatial features from the data.

Next, we feed the collected features into a LSTM to understand the temporal aspects of the dataset. This module will output two separate values one for PO and another for CF. Keeping the structure of the LSTM same for both the feature extraction modules will allow more understanding about the impact of the feature extraction processes.

A. Handcrafted Features -LSTM (HF-LSTM)

We start by extracting the handcrafted features mentioned in KIMORE. KIMORE provided scripts that help extract the features using simple coordinate geometry. Each exercise has a specific number of PO and CF. An illustration of the extracted features for each of the exercises is given in figure 4.

We feed the extracted handcrafted feature into a long short-term memory (LSTM). LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike the standard feed-forward neural networks, the LSTM also has feedback connections. As RNNs are known to encounter the vanishing gradient problem during training, LSTMs were developed to solve this problem. The LSTM setup most commonly used in the literature was originally described by Graves and Schmidhuber [18]. We output two scores from the LSTM to predict the scores given by the doctors.

B. GCN-LSTM

In this subsection, we discuss the concepts of graph convolutional networks and how they are structured for our model. Then we move on to elaborate on how they work with a LSTM to predict a score.

1) Graph representation of the skeleton: We define a skeleton graph using a set of nodes and a set of edges between nodes. Each node represents a body part or joint in the body.

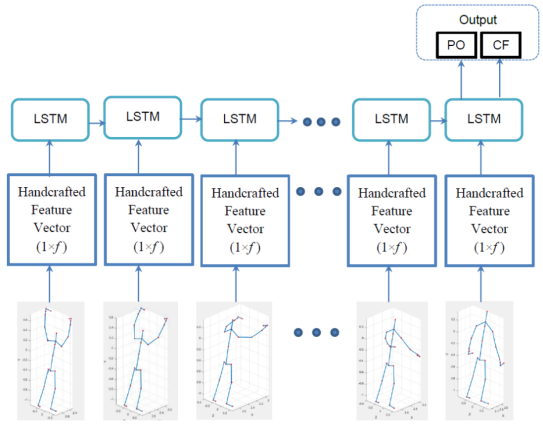

Fig. 4. HF-LSTM architecture; each frame of skeleton data, for a single exercise, is used to create a feature vector of size (1×f) where f is the number of features and forwards it to the LSTM that evaluates the features to predict the output scores.

The edges represent the connection between the body parts. For example, the wrist is connected to the elbow.

$$A_t(i,j) = \begin{cases} 1, & \text{if } (X_{ti}, X_{tj}) \in E_t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We use the common design for graphs based on anatomy as represented by Zhao et al. [16]. To be precise, we initialize an undirected graph for each time step of the exercises with $G_t = X_t, E_t$. Where, $X_t = \{X_{t1}, X_{t2}, ..., X_{tN}\}$ is the set of nodes at time t, in which each node represents a body part. N is the total number of nodes. $E_t = \{(X_{ti}, X_{tj}) : X_{ti}, X_{tj} \in X_t, X_{ti} \sim X_{tj}\}$ is the set of edges in the graph, where $X_{ti} \sim X_{tj}$ means the node i and node j are connected with an undirected edge. E_t can be specified by the adjacency matrix, $A_t \in \mathbb{R}^{(N\times N)}$.

For each node X_{ti}, the associated coordinates are 3D joint positions. Therefore, each node has a 3-dimensional coordinate and $X_t \in \mathbb{R}^{(N\times 3)}$. X_t can also be called the raw representation of the skeleton coordinates. The graph generated here acts as a brief system to specify the dependency among different body parts. We assume the graph structure does not change over each time frame, i.e., At remains the same for all t.

2) Graph Convolution Network: Initial variants of neural networks only allowed regular data or Euclidean data, whereas a large number of real-world data have an underlying non-Euclidean, graph structure. The use of graph-based data structures has led to recent improvements in machine learning with graph neural networks (GNN). In recent years many variants of GNN are being advanced, among which Graph Convolutional Network (GCN) is widely utilized. Nowadays, GCN is also considered to be one of the basic variants of graph neural networks. Similar to the convolutional layers in convolutional neural networks, the 'convolution' in GCN has the same principle. It denotes the use of multiplying the input neurons with a set of weights which are called filters or kernels. Using the graph, we consider a multi-layer Graph

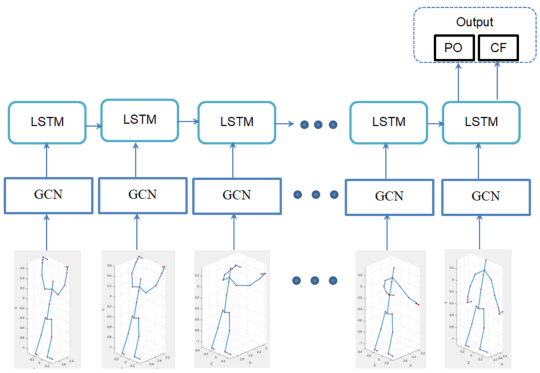

Fig. 5. GCN-LSTM architecture; each frame of skeleton data for a single exercise, is passed to a single GCN layer which extracts the features and forwards it to the LSTM that evaluates the features to predict the output scores.

Convolutional Network (GCN) with the following layer-wise propagation rule.

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}) \quad (2)$$

Here, $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph G with added self-connections, and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ its diagonal degree matrix. I_N is the identity matrix, and W_l is a layer-specific trainable weight matrix. $\sigma(\cdot)$ denotes an activation function, such as the $ReLU(\cdot) = max(0, \cdot)$. $H_l \in \mathbb{R}^{(N \times D)}$ is the matrix of activations in the last layer; $H(0) = X$.

The dot product of the adjacency matrix and node features matrix represents the sum of neighboring node features. Thus the function is iterated till $l = 2$, to collect the features of nodes that are 2 hops away. The collected feature for $H(2)$ is the feature list extracted using GCN, at a certain time frame. The feature matrix is flattened and inputted into the LSTM. Here the LSTM is the same as HF-LSTM. The LSTM takes the flattened feature vector and outputs two predicted values, PO and CF. An overview of the model is shown in figure 5.

V. Experimentation and Results

This section is divided into three subsections. We start by discussing how we implemented the proposed model. Here we have shown the tools we utilized and the specifications of our model. This section also contains the training parameters and computer specifications of the machine we executed on. Then in the second subsection, we discuss the results achieved for each model respectively. Finally, we discuss the comparisons between the two models.

A. Implementation

We implement the proposed models using Pytorch and PyTorch Geometric [19]. The handcrafted features are calculated using Matlab by importing the scripts provided by the dataset, KIMORE. The scripts have defined equations that are utilized to calculate the angles and distances from the coordinates. The GCN is implemented using the matrix multiplication defined in the above section in eqn. (2). We configure the GCN kernel size to 3, i.e. the encoded features by the GCN will comprise of the information about neighboring nodes that are 2 hops away. At any frame or time step t, the raw skeleton representation X_t is given as an input to the defined GCN. The output of the GCN is then flattened into a single vector and inputted in the LSTM. This flattened vector acts as our feature vector for the single frame. Unlike the Handcrafted features, the numbers of GCN features were not limited, thus giving us an area for experimentation. We tried to check the best possibilities by implementing 3 types of outputs for the GCN. For the second phase of the model, a single layer of LSTM units was utilized to accept the frames of the feature vector and generate an output. The output of the last time step of LSTM is used as the final image of the exercise. We feed it into a fully connected layer with ReLU activation function, because all the scores predicted are above zero. Dropout was added between the layers to avoid overfitting. The initial learning rate was set to 0.001 and ran for 50 epochs, where each epoch had a batch size of 8. As the total number of participants in the dataset is only 78. We evaluated the model using a 20% split of test data and the other 80% was used for training.

The model was implemented on an ASUS GL503GE Laptop with Intel Core i7 8th gen processor CPU, 16GB Ram, a 1TB hard disk and with an NVIDIA 1050 ti Graphics Card.

Prepossessing: Once the calculation of the features is completed, we then normalize the features using a min-max scalar transform. Secondly, we perform a similar normalization for the scores. Since we already know that the maximum score of PO is 15 and the maximum score of CF is 35, we divided the scores by their maximum value. Now all the values are ready for our next module, the LSTM to predict a score.

In this experiment, we only changed the first module, feature extraction module, and kept the LSTM specification fixed, along with the fully connected layers. The model is trained in a supervised manner. We employ the use of RMSE. The RMSE has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. Since the KIMORE dataset provides a single score for the PO and CF each, the difference between the predicted value and original value acts as an acceptable measure of error. Furthermore, as the values are squared before the average, this particularly helps to provide higher error for larger values. This helps us narrow the possibility of predicting values too far off.

B. Results

We used five-fold cross-validation to evaluate our model. Cross-validation is a statistical method to evaluate machine learning models whose goal is a prediction. A round of cross-validation partitions the data into complementary subsets, where training is performed on one set and validation testing is performed on the other. Here, to reduce variability, we perform five rounds or five-folds of cross-validation and average the predictive performance.

TABLE I
TEST RMSE FOR ALL EXERCISES USING HF-LSTM

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Ex1	0.253	0.331	0.341	0.298	0.313	0.307
Ex2	0.343	0.280	0.166	0.288	0.242	0.264
Ex3	0.318	0.383	0.302	0.396	0.301	0.340
Ex4	0.265	0.251	0.312	0.276	0.312	0.283
Ex5	0.324	0.239	0.215	0.249	0.259	0.257
Total Average RMSE:						0.290

TABLE II
TEST LOSS FOR ALL EXERCISES USING GCN- LSTM

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Ex1	0.168	0.259	0.202	0.196	0.234	0.212
Ex2	0.161	0.161	0.172	0.169	0.176	0.168
Ex3	0.180	0.221	0.161	0.186	0.234	0.196
Ex4	0.227	0.211	0.228	0.234	0.174	0.215
Ex5	0.168	0.155	0.158	0.207	0.158	0.169
Total Average RMSE:						**0.192**

1) HF-LSTM: As the scores that we hope to predict were set by the physicians to use these handcrafted features, intuitively, the use of this method should yield a very low RMSE loss. The test RMSE loss for the five exercises, achieved in the five-folds, are listed in table I. The table also mentions the average RMSE of each of the 5 exercises and also the total RMSE with the highest loss reached for exercise 3 and lowest for exercise 5.

The results have the following indications. Here, we use a LSTM network to predict a viable score for physical rehabilitation exercises with an average RMSE loss of 0.290.

2) GCN-LSTM: We experimented with different configurations of GCN by varying the number of output features. Table II shows the results obtained by the GCN which generates 100 features, as it provided the lowest testing loss. The table provides the RMSE for all of the five exercises in five-folds. Exercise 2 had the lowest RMSE of 0.168 and exercise 4 had the highest RMSE of 0.215. We can see that there is a significant improvement in comparison to the previous model as we have achieved an average RMSE loss of 0.191.

VI. CONCLUSION

Skeleton-based human movement quality assessment is a big challenge with various applications spanning from sports, talent evaluation, and rehabilitation exercise evaluation. In this paper, we have proposed two models for assessing human movement quality. Both models are capable of exploiting spatial and long-term temporal dependencies. In the first model (HF-LSTM), we have used handcrafted features (defined by physicians) that were calculated from 3D skeleton data. In another model (GCN-LSTM), we used the 3D skeleton data directly as input where Graph Convolution Network (GCN) has been used to generate necessary features. Both of the models are trained by LSTM to understand the temporal aspect of movements. The proposed models are trained and tested on the KIMORE dataset. The HF-LSTM performed well achieving an RMSE loss of 0.290 in five-folds cross validations. On the other hand, the GCN-LSTM performed outstandingly well in comparison to its prior, reaching an average RMSE loss of 0.191.

The reason for the superior performance of the GCN-LSTM is it takes advantage of the spatio-temporal characteristics of the entire skeleton data instead of only physician specified features. Furthermore, this paper also proves that both GCN-LSTM and HF-LSTM can learn spatiotemporal connections in human movement data.

REFERENCES

[1] S. R. Machlin, J. Chevan, W. W. Yu, and M. W. Zodet, "Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults," *Physical therapy*, vol. 91, no. 7, pp. 1018–1029, 2011.

[2] R. Komatireddy, A. Chokshi, J. Basnett, M. Casale, D. Goble, and T. Shubert, "Quality and quantity of rehabilitation exercises delivered by a 3-d motion controlled camera: A pilot study," *International journal of physical medicine & rehabilitation*, vol. 2, no. 4, 2014.

[3] S. F. Bassett and H. Prapavessis, "Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains," *Physical therapy*, vol. 87, no. 9, pp. 1132–1143, 2007.

[4] F. Sardari, A. Paiement, S. Hannuna, and M. Mirmehdi, "Vi-net—view-invariant quality of human movement assessment," *Sensors*, vol. 20, no. 18, p. 5258, 2020.

[5] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, 2020.

[6] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, and F. Verdini, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

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

[8] B. Crabbe, A. Paiement, S. Hannuna, and M. Mirmehdi, "Skeleton-free body pose estimation from depth images for movement analysis," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 70–78.

[9] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 280–291, 2019.

[10] F. N. Rashid and N. S. Suriani, "Spiking neural network classification for spike train analysis of physiotherapy movements," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 319–325, 2020.

[11] J. Chen, X. Xu, Y. Wu, and H. Zheng, "Gc-lstm: Graph convolution embedded lstm for dynamic link prediction," *arXiv preprint arXiv:1812.04206*, 2018.

[12] Y. Huang, S. Zhang, J. Wen, and X. Chen, "Short-term traffic flow prediction based on graph convolutional network embedded lstm," in *International Conference on Transportation and Development 2020*. American Society of Civil Engineers Reston, VA, 2020, pp. 159–168.

[13] Z. Li, G. Xiong, Y. Chen, Y. Lv, B. Hu, F. Zhu, and F.-Y. Wang, "A hybrid deep learning approach with gcn and lstm for traffic flow prediction," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1929–1933.

[14] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[15] H. Zhang, Y. Song, and Y. Zhang, "Graph convolutional lstm model for skeleton-based action recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019.

[16] R. Zhao, K. Wang, H. Su, and Q. Ji, "Bayesian graph convolution lstm for skeleton based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6882–6892.

[17] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, 2018.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005001206

[19] B. Rozemberczki, P. Scherer, Y. He, G. Panagopoulos, M. Astefanoaei, O. Kiss, F. Beres, N. Collignon, and R. Sarkar, "Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models," *arXiv preprint arXiv:2104.07788*, 2021.