Assessment of Rehabilitation Exercises from Depth Sensor Data

Shehzan Haider Chowdhury
1720109

Murshed Al-Amin
1720069

Report submitted in partial fulfillment
of the degree of Bachelor of Science, Computer Science & Engineering
Department of Computer Science & Engineering
Independent University, Bangladesh (IUB)

Abstract

Rehabilitation exercises play a very important part in a patient's postoperative recovery and treatment of many musculoskeletal conditions. Reports shows that over 90% of all rehabilitation exercise sessions are being performed in a home-based setting [2]. In hospitals, each of the patients is monitored by a trained physician. However, as the number of patient increases this process becomes costly and infeasible. An effective way to resolve this is to provide technological support for making home-based rehabilitation. The patient stays at home and performs the exercises before the camera and the video is then transmitted to the physician allowing them provide feedback on the exercises. More intelligent systems can automatically assess the exercises performed and can just notify the patient and physician how well the exercises are being performed. In this project, we propose two machine learning-based methods to assess the quality of exercises where the data is captured by such an intelligent system. More specifically, we evaluate the exercise data provided in the KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical REhabilitation (KIMORE) dataset. The KIMORE dataset consists of skeleton data of 5 exercises from 78 patients. Skeleton data is the time series data of skeleton joint positions extracted from depth videos captured using Kinect, a motion-sensing device. It also contains the physician's rating of the quality of the exercises. For the first baseline, we use the features provided with the KIMORE dataset, validated by physicians to train a long short-term memory (LSTM) network. The average root mean square error (RMSE) loss for the first baseline is 0.290. For the second baseline, we extract features from the KIMORE skeleton data using graph convolution network (GCN) where each node represents a body part or joint in the body and the edges represent the connection between the body parts, which is used to train a LSTM network similar to the first baseline. The average RMSE loss for the second baseline is 0.191. We conclude that LSTM is more accurate at predicting the results when GCN features are used. One limitation of the project is that it evaluates only one dataset as this is the only publicly available dataset for which physician's assessments of the quality are provided. In the future, we would like to collect data from rehabilitation patients and apply our methods to that dataset. Another limitation is the lack of cheap consumer hardware which patients can easily have access to, as devices such as Kinect.

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

Attestation

We understand the nature of plagiarism, and we are aware of the university's strict policy on the matter. We certify that this is an original work by us. However, others' written work or software used in any part of this project is properly cited following internationally accepted academic guidelines.

Signature:
Name: Shehzan Haider Chowdhury

Signature:
Name: Murshed Al Amin

Evaluation Committee

Signature: ……

Name: ……

Supervisor: ……...

Signature: ……

Name: ……

Internal Examiner: …..……

Signature: ...…….

Name: ……

External Examiner: …..…….

Signature: …….

Name: ……

Convener: …….

Acknowledgment

We would like to take this opportunity to convey our special appreciation and gratitude to Dr. Amin Ahsan Ali, our respected supervisor, for his guidelines and suggestions throughout the whole project.

We would like to thank Dr. AKM Mahbubur Rahman for his valuable overview and advice. And last but not the least; we would like to acknowledge Artificial Intelligence and Cybernetics Lab Independent University, Bangladesh for its support.

Table of Content

List of Figures

List of Tables

1. Introduction

Rehabilitation exercises are a key part of a patient's postoperative 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 [14], leading to extensions of the recovery period. The use of an automated system to evaluate and provide feedback on how well the exercise was done would reduce the hassle of periodically visiting the physician while also allowing the patients to fix their own mistakes as the system would evaluate the movement just as a physician. 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.

The task of evaluating the quality of assessment of exercises falls under the category of quality of human movement assessment which in turn falls under the general category of human action recognition and 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. Similar to action classification and detection, the quality of human movement assessment is also included under action analysis. However in exercise quality assessment we are interested in both recognizing the exercise and also analyzing it, thus this falls under action analysis.

Recently work on quality of human movement assessment has gained attention resulting in various tools and devices to assist physical rehabilitation. For example, Sardari et al. [3] proposed a view-invariant method using a pre-trained convolutional neural

network (CNN) to evaluate the quality of human movement. Their model needs to be trained separately for each movement type but their performance drops when long-term occlusions occur. On the other hand, as exercises are events that are related to time series, LSTM has been proven useful by Liao et al. [4]. However, their model is validated by measuring variations in movement data without any ground truth assessment. On the other hand, Sardari et al. [3] provide a ground truth using physician's evaluation. 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. Therefore, there is still a lack of robust, lightweight systems for automatic monitoring and assessment of patient's performance.

1.1. Project Objective

The primary objectives of this project are as follows:

1. Propose a framework for assessment of rehabilitation exercises:

Our primary goal is to propose two deep learning frameworks for the assessment of rehabilitation exercises. This will enable researchers to design systems that can either automatically provide feedback to the patients without requiring the involvement of a physician or provide summary feedback to the remotely located physician who can then provide feedback to the patient.

2. Assessment of physician's prescribed features:

First we propose a framework which uses handcrafted features defined by the physicians to propose a score based on the exercise. Next, we evaluate the features and their temporal relationship to estimate a score. The physicians evaluate the exercise mostly based on angles and distances between joints.

3. Assessment of computationally generated features:

Secondly, we move to generate the features automatically, using a GCN, to look into the possibility of bridging the gap of human assessment of these exercises. Next, we evaluate the features and their temporal relationship to estimate a score.

4. Provide the code for the repository to help further research on human movement assessment:

The vast majority of researchers are more familiar with a similar problem, human action classification. Within the same substrata, human action assessment lacks the favoritism of researchers and students, possibly due to the lack of work in the field, thus leading to no benchmark to base their work on. We hope to provide a reasonable benchmark to new researchers to indulge their interests in the topic and possibly improve the results by sharing our model and code for the repository.

Further prospects of the project can be found in the "future work" of the conclusion section

1.2. Scope of the project

The aim of the project includes establishing a framework 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. The framework proposed is hoped to provide a model with an acceptable loss, to ease the process of physical rehabilitation. In the process of the study, we also looked at similar work, and how they have tried to tackle the problem. With the proposal of two frameworks, the study is also able to contrast the feature extraction processes. Therefore allowing area for research in whether physician's defined features is required for automated assessment of physical rehabilitation exercises.

1.3. Contribution of the project

The main contributions of the project are as follows:

1. We provide a new framework for the assessment of rehabilitation exercises using physician's prescribed features:

The primary goal of the report is to generate a solution to ease the Rehabilitation process of patients suffering from musculoskeletal conditions using a computer-aided framework to generate the required feedback, allowing the chances of faster recovery. The framework proposed reaches an RMSE test loss of 0.290.

2. We provide a new framework for the assessment of rehabilitation exercises using GCN Features:

The ability to use automatically generated features using GCN reduces the need for a dataset with specific features. The proposed framework reaches an average RMSE test loss of 0.191. This will allow new researchers to assess their requirements for generating new datasets.

3. Provide a contrast of automatically generated features and physician's features:

The comparison between the two frameworks will allow estimating which framework is computationally cheaper and more plausible in real-world solutions.

1.4. Roadmap

This report contains six major parts. Firstly, we discuss the related work and the groundwork that led to the embodiment of this project. Secondly, we will shed light on the variety of approaches, algorithms that researchers have tried to achieve this or similar goals. Next, we will discuss the availability of datasets and their differences to come up with the best dataset to help us train and validate our concept. Next, we will discuss the proposed method section, where the details of our framework will be put forth. Then in the experimentation section, we will discuss implementation and specification of the models and discuss the achieved results and the experimentation that led us to the conclusion. Finally, we discuss our conclusion and prospects of future work in the domain.

2. Literature Review

Action analysis has recently picked up pace in the past few years, out of which most of the research in the area is either based on physical rehabilitation, skill assessment or sports analysis. Within this subfield of study, many researchers have worked on similar goals using non-skeleton-based methods, such as CNN, mainly for sports scoring, and then applied their work on physical rehabilitation [5].

Out of the many movement assessment studies, the most similar to our subject is discussed in the following many of which 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. The Kinect device incorporates RGB cameras, infrared projectors, and detectors that mapped depth through either structured light or time of flight calculations. Figure 1 shows sample frames from KIMORE dataset for different exercises. Using image-based models, recent research has produced promising results using CNN. For example, Crabbe et al. [6] 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 for actions like walking on stairs.

Figure 1: Sample frames from KIMORE for 5 exercises. [10]

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. [4], 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. [7], 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 performed this by classifying each movement

sequence into a normal and abnormal class using a probabilistic normalcy model from their descriptors obtained from regular subjects. Next, 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. [3] 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. Firstly, view-invariant trajectory descriptor for each body joint is collected from RGB images to form a collection of trajectories for all joints. They use this in an adaptive, pre-trained 2D CNN to establish spatial relationships among the different parts of the body and predict a score for the movement quality. They applied their model on their own generated dataset, a multi-view, non-skeleton, non-mocap, rehabilitation movement dataset (QMAR), and also in a Kinect based on the KIMORE dataset. Figure 2 illustrates the data capture mechanism for QMAR dataset, where they used multiple cameras to generate a multi viewed dataset. As the KIMORE dataset is not a multi-viewed dataset, the model was limited to using a single view at which they reached a rank correlation of 0.66.

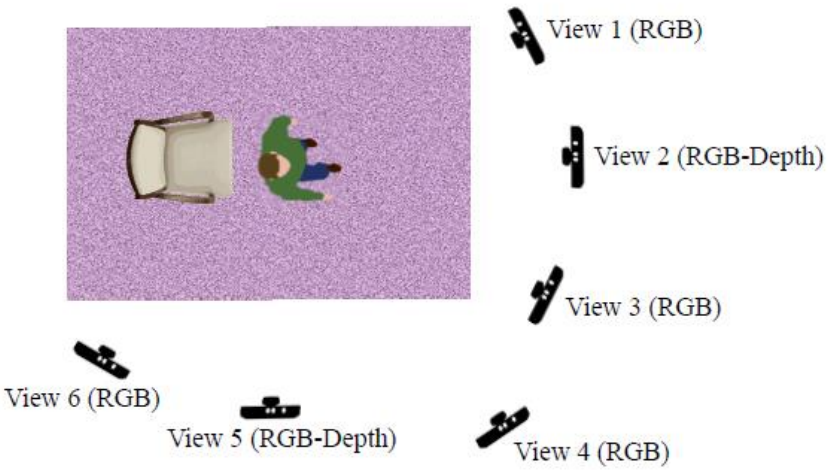


Figure 2: Typical camera views in the QMAR dataset with each one placed at a different height. [3]

Nor Rashid et al. [8], proposed a deep learning model, for skeleton-based physical rehabilitation exercise classification, by implementing a spike train feature on the UI-PRMD dataset. A spike train is a sequence of recorded times at which a neuron fires an

action potential, i.e. a sudden, fast and propagating change. They performed a spike trained analysis by encoding the data into spike trains as spike train features are hugely rewarding towards deep learning because it can visually differentiate the physiotherapy movements based on the pattern differences. They parted the data into sections of 100 frames and used the section to generate spike trains. Then they used it in a CNN to classify the movement their proposed model achieved an accuracy of 0.77 on physical rehabilitation exercise classification.

Liao et al. [4], proposed a deep learning framework for physical rehabilitation exercise assessment on a skeleton-based dataset. The proposed model is a deep Spatio-temporal neural network that positions data in a temporal pyramid and utilizes the sub-networks to process joint displacements of body parts to generate the spatial characteristics of human movements. Their main focus was on quantifying a movement performance by providing scoring for mapping the performance metrics onto an acceptable score of movement quality. Secondly, provide a deep Neural network model to generate movement quality scores using a supervised learning method. The proposed framework was applied on their own Skeleton-based physical rehabilitation dataset, UI-PRMD. Their performance metric is defined based on the log-likelihood of a Gaussian mixture model while encoding low-dimensional data representation obtained from their deep auto encoder network. They also employed distance functions, such as Euclidean, Mahalanobis distance, and dynamic time warping (DTW) for the performance metric.

3. 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) [9] consisting of only skeleton data and KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical Rehabilitation (KIMORE) [10] 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. We provide the details of the dataset in the following.

The KIMORE dataset provides a collection of different physical rehabilitation exercises collected using a RGB-D sensor, Kinect. The sensor was used to collect three different types of data input which are, RGB, depth videos, and skeleton joint positions. 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 stereo photogrammetric 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 are patients with back pain, patients who suffered from a stroke, and patients who suffer from Parkinson's disease. All the participants perform five different exercises. The exercises are mentioned below:

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

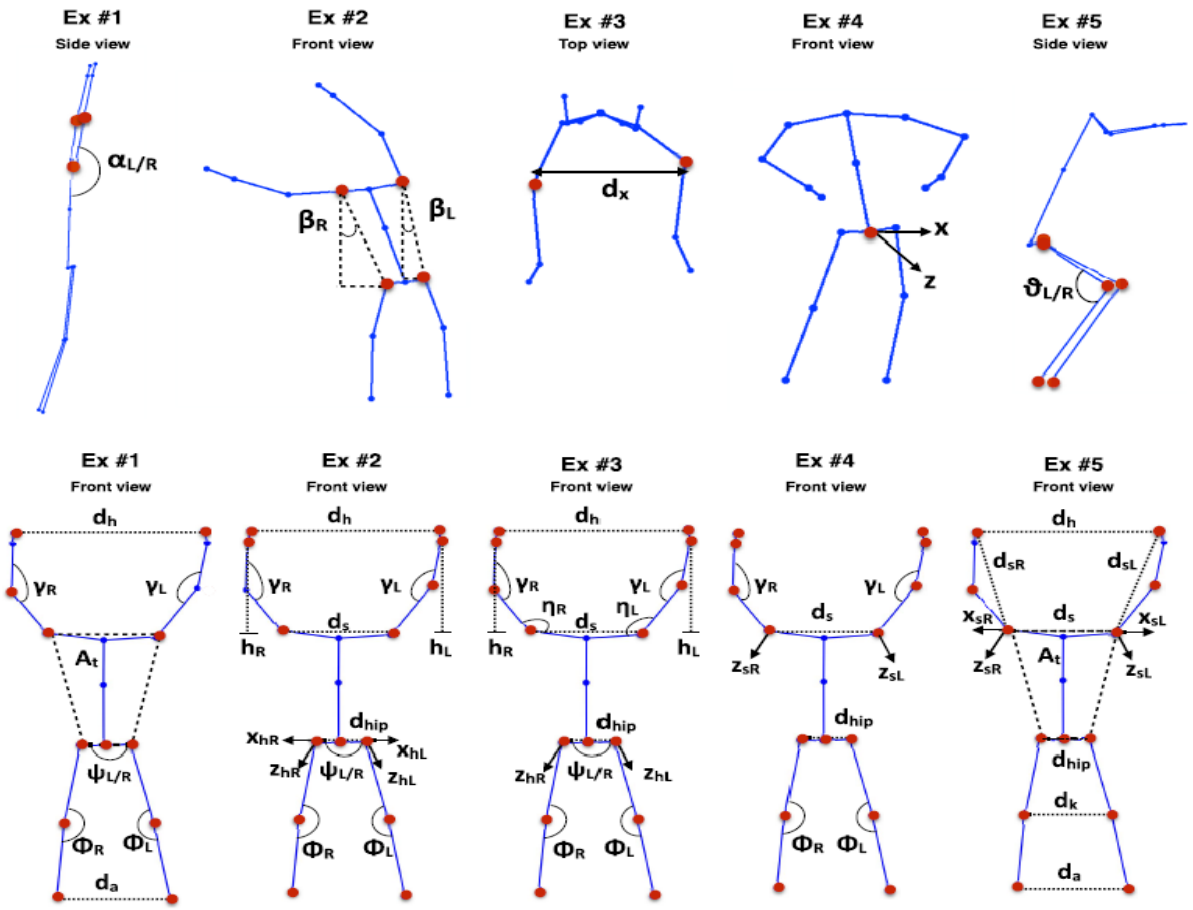


Figure 3: Features prescribed by Physicians for the five exercises. Row one contains the Primary outcomes and Row two illustrates the Control factors [10].

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 is the only skeleton-based dataset that has a performance metric verified by a physician therefore to validate and test our method we will use the KIMORE dataset for experimentation.

4. Proposed Method

Although 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's performs.

The proposed frameworks have a feature extraction module that feeds into a Temporal Score Module as shown in figure 4. In the first module, we extract the features using 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.

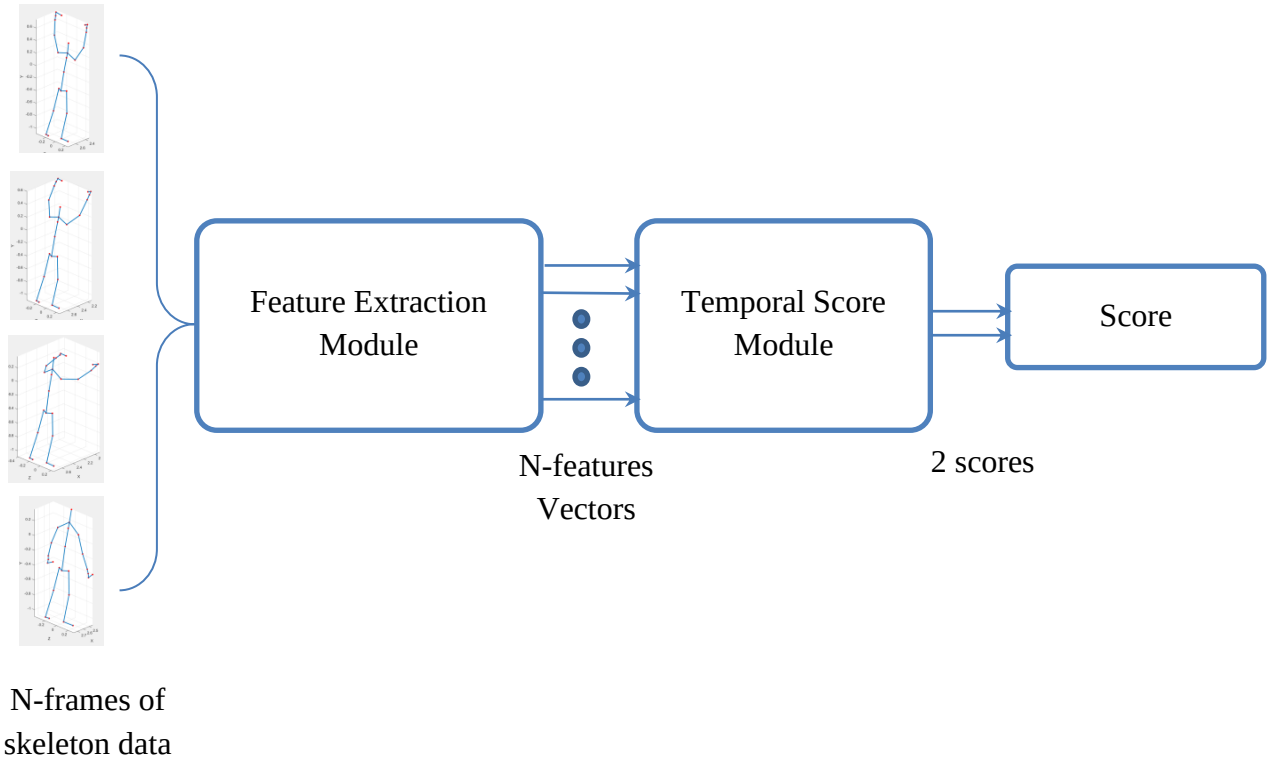


Figure 4: Proposed framework has a feature extraction module and a temporal score module. The feature extraction module calculates the features from each frame and the temporal score module uses these features to predict a score.

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 processes. The specifics about the frameworks are discussed in the following pages.

4.1. Handcrafted Features -LSTM (HF-LSTM)

In this subsection, we introduce the handcrafted features that are prescribed in the KIMORE dataset. And then we introduce the next module, the LSTM. Finally, we discuss how the handcrafted features are used in the LSTM.

4.1.1. Handcrafted features

We start by extracting the features mentioned in KIMORE. The dataset has available scripts that help to extract the features using simple coordinate geometry. An illustration of the features is given in figure 3. The extracted features for different exercises are:

Exercise 1: Lifting of the arms

Extracted Features: Angles between right/left arm and upper torso in the sagittal plane (αl/r) represent the POs. Elbow extension angles (γl/r), knee extension angles (φl/r), hip angles (ψl/r), torso area (At), hands distance (dh), ankle distance (da) are the CFs to be considered.

Exercise 2: Lateral tilt of the trunk with the arms in extension.

Extracted Features: Right and left angles between the anatomical segment defined by the hip and shoulder and the vertical axis (βl/r) in the frontal plane (x, y)are defined as POs, while elbow extension (γl/r), knee extension angles (φl/r), hip angles (ψl/r), hand distance (dh), shoulder-distance (ds), hip-distance (dhip) and the vertical distance between the wrists and the shoulders (hl/r) and the transverse plane coordinates of the hip (zhl/r , Xhl/r) normalized to zero mean, are the CFs.

Exercise 3: Trunk rotation

Extracted Features: PO is the horizontal distance between the elbows (dx), normalized with respect to the maximum variation. The elbow extension angle (γl/r), shoulder extension angles (ηl/r), knee extension angles (φl/r), hip angles (ψl/r), shoulder-distance (ds), hip-distance (dh) the distance between the wrists and the shoulders (hl/r) and the depth coordinates of the hip (zhl/r) normalized to zero mean, are the CFs.

Exercise 4: Pelvis rotations on the transverse plane

Extracted Features: POs are given by the spine base trajectories, normalized to zero mean, in the transverse plane (x, z), to ensure that the subject's position is independent of the sensor. The shoulder distance (ds), hip-distance (dh), elbow extension (γl/r), knee extension angles (φl/r) and the depth coordinates of the shoulders (zsl/r) normalized to zero mean, are the CFs.

Exercise 5: Squatting.

Extracted Features: The right and left knee angles in the sagittal plane (θl/r) are POs. Hand distance (dh), shoulder-distance (ds), hip-distance (dhip), knee distance (dk), ankle distance (da), torso area (At), the distance between hand and shoulder (dsl/r) and the transverse plane coordinates of the shoulder (zsl/r , xsl/r) normalized to zero mean, are the CFs.

4.1.2. LSTM

Long short-term memory (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. Therefore along with the ability of processing single data points, it can also process entire sequences of data, such as videos. LSTM networks are most commonly used for processing, classifying, and making predictions based on time series data sets, since there can be lags of unknown duration between important events in a time series. As RNNs are known to encounter the vanishing gradient problem during training, LSTMs was developed to solve this problem. The LSTM setup most commonly used in the literature was originally described by Graves and Schmidhuber [11]. The basic mathematical implementation of the LSTM node can be defined as shown below in eqn 1.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$
$$h_t = o_t \circ \sigma_h(c_t) \quad (1)$$

Here, $x_t \in \mathbb{R}^d$ is the input vector to the LSTM unit and $f_t \in \mathbb{R}^h$ is the forget gate's activation vector. $i_t \in \mathbb{R}^h$ input/update gate's activation vector, $o_t \in \mathbb{R}^h$ output gate's activation vector, $h_t \in \mathbb{R}^h$ hidden state vector also known as output vector of the LSTM unit and $\tilde{c}_t \in \mathbb{R}^h$ cell input activation vector, $c_t \in \mathbb{R}^h$ cell state vector. Where, $W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$ weight matrices and bias vector parameters need to be learned during training where the subscripts ***d*** and ***h*** refer to the number of input features and a number of hidden units, respectively.

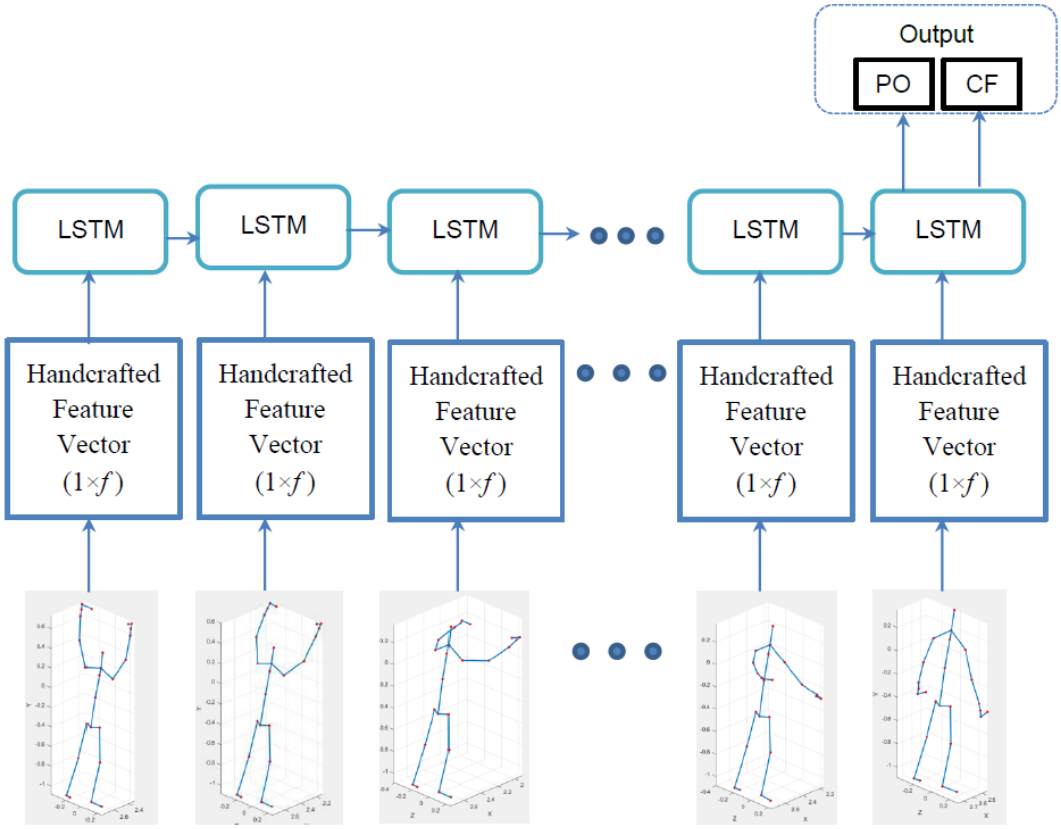


Figure **5**: HF-LSTM model Architecture; each frame of skeleton data, for a single exercise, is used to create a feature vector of size $(1 \times f)$ where f is the number of features and forwards it to the LSTM that evaluates the features to predict the output scoress

4.2. GCN-LSTM

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

4.2.1. Graph representation of the skeleton

We define a skeleton graph using a set of nodes and a set of edges that are connecting the nodes. Each node represents a body part or joint in the body. The edges represent the connection between the body parts. For example, the wrist is connected to the elbow.

We use the common design for graphs based on anatomy. 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. An example of the skeleton graph is depicted in Figure 6. E_t can be specified by the adjacency matrix, $A_t \in \mathbb{R}^{N\times N}$.

$$A_t(i,j) = \begin{cases} 1, & \text{if } (X_{ti}, X_{tj}) \in \boldsymbol{E_t} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

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., A_t remains the same for all t.

4.2.2. 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 network (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.

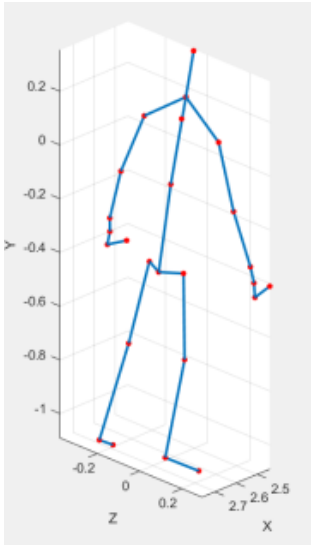


Figure 6: Skeleton Graph; consisting of the nodes (body joints), and edges connecting the joints.

Using the graph, we consider a multi-layer Graph Convolutional Network (GCN) with the following layer-wise propagation rule.

$$H_{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H_l W_l\right) \quad (3)$$

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. σ (·) denotes an activation function, such as the ReLU(·) = max(0, ·). $H_l \in \mathbb{R}^{N \times D}$ is the matrix of activations in the l^{st} 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 temporal score module of the framework. Here the same temporal score module as HF-LSTM is employed, i.e a LSTM. The detail of this module is elaborated in section 4.1.2. The LSTM takes the flattened feature vector and outputs two predicted values, PO and CF. An overview of the framework is shown in figure 7.

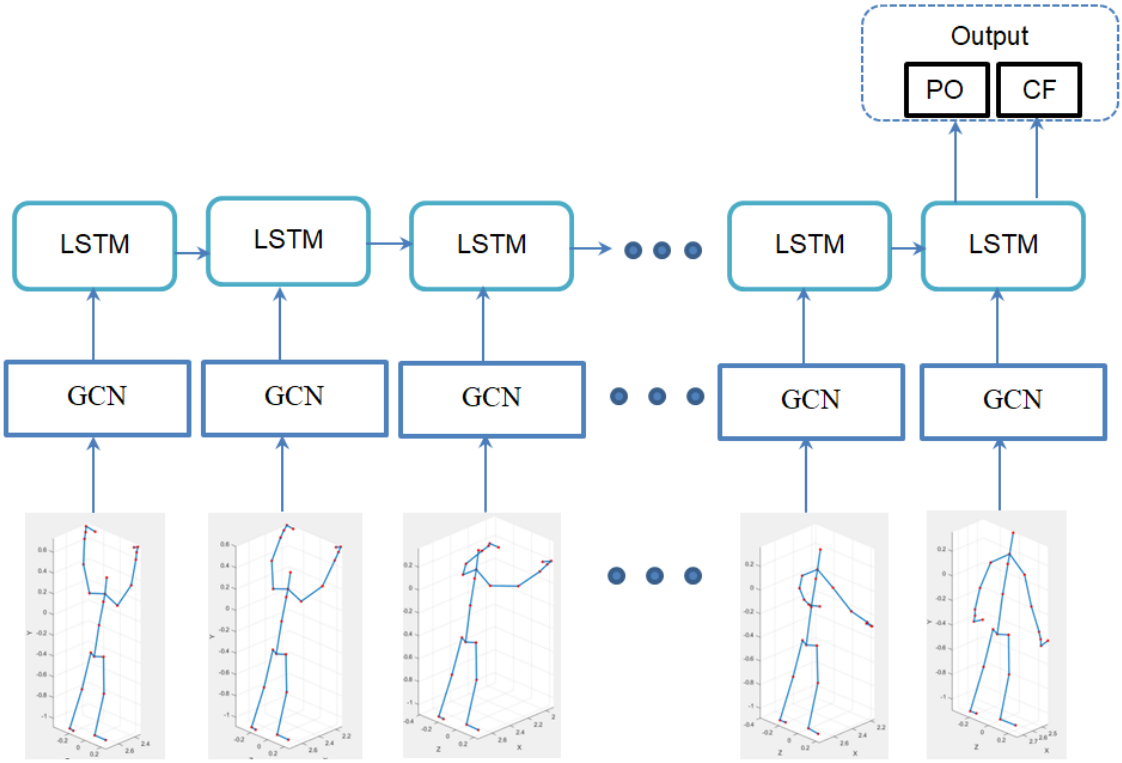


Figure 7: GCN-LSTM model Architecture; each frame of skeleton data, for a single exercise, is inputted in a single GCN layer which extracts the features and forwards it to the LSTM that evaluates the features to predict the output scores.

5. Experimentation and Results

This chapter is divided into three subsections. We start by discussing how we implemented the proposed framework. 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.

5.1. Implementation

We implement the proposed frameworks using Pytorch [12] and PyTorch Geometric [13]. 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. (3). 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 framework, 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.

- **Preprocessing:**

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 CFs is 35, we divided the scores by their maximum value. Now all the values are ready for our next module, the temporal score module to predict a score.

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

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 framework 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. The mathematical intuition is given in Eqn 4.

$$RMSE = \sqrt{\sum_{i=0}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad (4)$$

Here, n is the number of exercises in the batch, y_i is the actual score and $\hat{y}_i$ is the predicted score.

We repeat the procedure with the same setting while shuffling the dataset, and report the average to account for the random initializations. The results obtained through this process are discussed in the following section.

5.2. Results

We employed the use of 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.

5.2.1. 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 training loss and the test loss for the five exercises, achieved in the five folds are listed in table 1 and table 2, respectively. The results have the following indications. Firstly we can use a LSTM network to predict a viable score for physical rehabilitation exercises with an average RMSE loss of 0.290.

Table 1: Train for all exercises using HF-LSTM

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
Ex1	0.219	0.146	0.14	0.149	0.164	0.1636
Ex 2	0.273	0.201	0.19	0.253	0.223	0.228
Ex3	0.322	0.288	0.2496	0.21	0.326	0.27912
Ex4	0.244	0.186	0.235	0.214	0.226	0.221
Ex5	0.242	0.22	0.207	0.183	0.214	0.2132
Total Average RMSE:						**0.220984**

Table 2: Test loss 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	0.3072
Ex 2	0.343	0.28	0.166	0.288	0.242	0.2638
Ex3	0.318	0.383	0.302	0.396	0.301	0.34
Ex4	0.265	0.251	0.312	0.276	0.312	0.2832
Ex5	0.324	0.239	0.215	0.249	0.259	0.2572
Total Average RMSE:						**0.29028**

5.2.2. GCN-LSTM

We first implement this framework for three different configurations of GCN on just one dataset, to select the optimal model for our framework. The first configuration is such that the final output H_l is a matrix of size (1×25), which when flatten gives a vector of 25 features. Secondly, we try the configuration such that the final output H_l is the matrix of size (4×25), i.e. when flattened we obtain 100 features. Similarly, the last configuration has an output matrix of (16×25), resulting in a flattened vector of 400 features. The results obtained by varying the GCN are given in table 2.

Table 3: Train loss and test loss for exercise 1 using different variations of GCN-LSTM

	25 Features	**100 features**	**400 features**
Train-loss	0.193	0.179	0.175
Test-loss	0.203	0.168	0.295

As the table clearly shows, even though the GCN with 400 features has the lowest training loss, it has the highest test loss. As the GCN with 100 feature output has the highest validation accuracy, we pick this configuration for the rest of our models. The results of the five exercises for training and testing are given in table 4 and 5 respectively.

Here we can see that there is a significant improvement in comparison to the previous framework. We have achieved an average RMSE loss of 0.191.

Table 4: Train loss for all exercises using GCN- LSTM

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
Ex-1	0.179	0.198	0.188	0.184	0.191	0.188
Ex2	0.189	0.226	0.212	0.201	0.217	0.209
Ex3	0.217	0.1933	0.215	0.204	0.201	0.20606
Ex4	0.201	0.195	0.205	0.21	0.204	0.203
Ex5	0.179	0.211	0.193	0.211	0.195	0.1978
Total Average RMSE:						0.200772

Table 5: 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	0.2118
Ex2	0.161	0.161	0.172	0.169	0.176	0.1678
Ex3	0.18	0.221	0.161	0.186	0.234	0.1964
Ex4	0.227	0.2108	0.228	0.234	0.174	0.21476
Ex5	0.168	0.155	0.158	0.207	0.158	0.1692
Total Average RMSE:						0.**191992**

5.2.3. Comparison between the models:

In every exercise the GCN-LSTM has outperformed the HF-LSTM framework. A comparison between their results is illustrated in the figure 8.

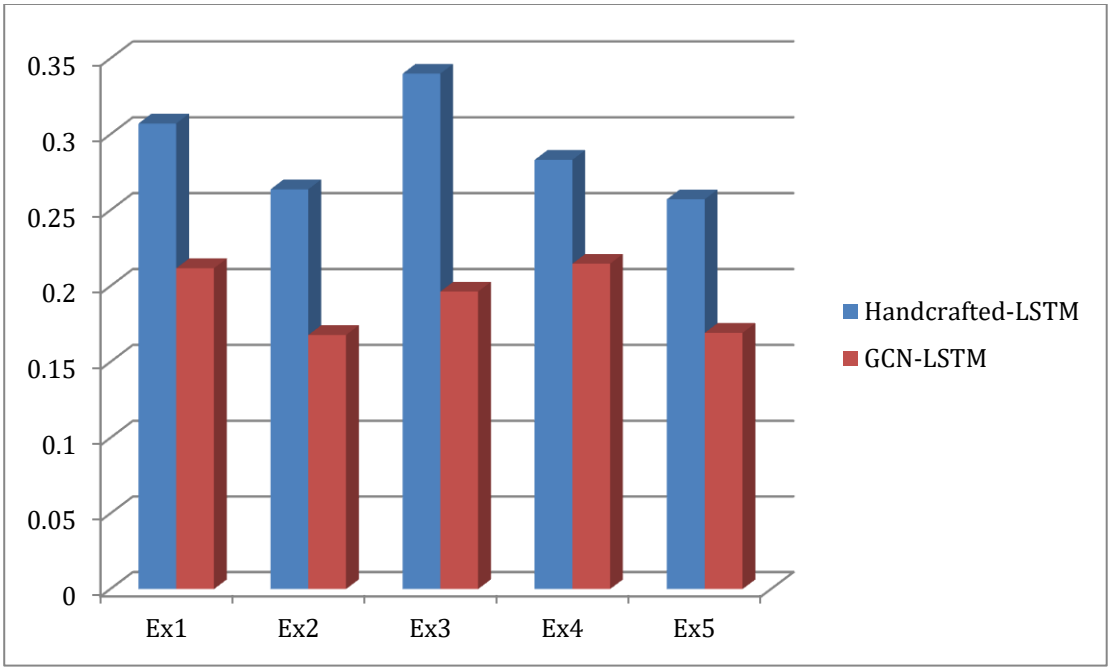
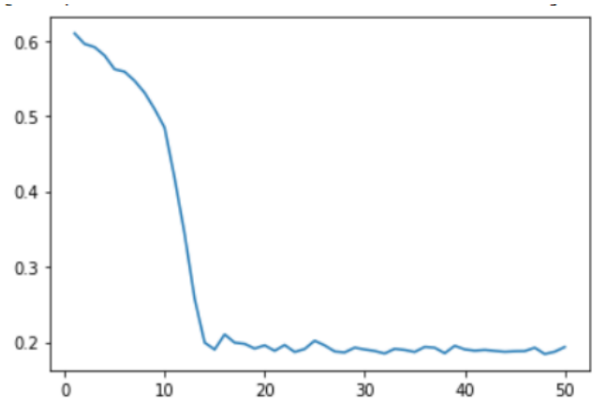


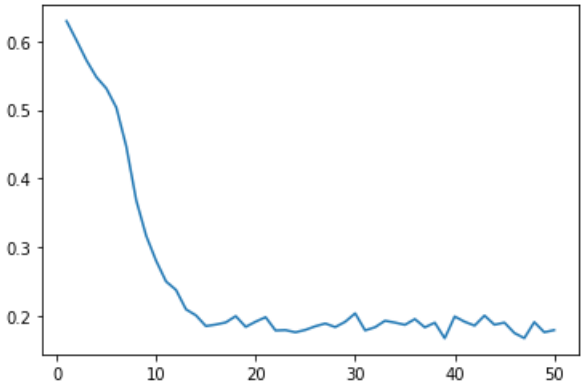
Figure 8: Comparison between test RMSE loss of handcrafted feature-LSTM and GCN-LSTM; on every exercise, The GCN-LSTM out performs the handcrafted-LSTM model.

The following figures give an example (just one fold) of the train loss achieved using the frameworks. This indicates that most of the models reached their minimum loss within 30 epochs. The extra epochs help strengthen the fact that the model has had adequate epochs to train.

HF- LSTM Exercise 1



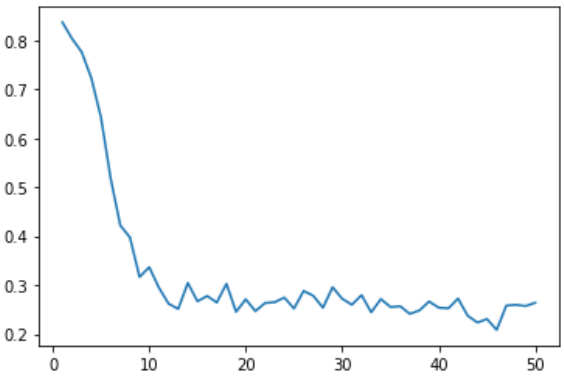
GC-LSTM Exercise 1



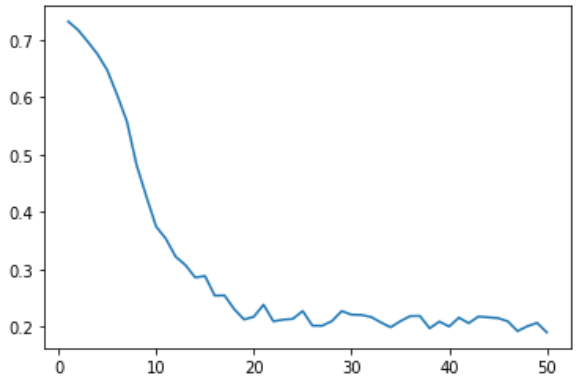
x-axis: No of Epochs | y-axis: Train RMSE Loss

Figure 9: Train loss achieved for Exercise 1 in both models in fold 1.

HF- LSTM Exercise 2



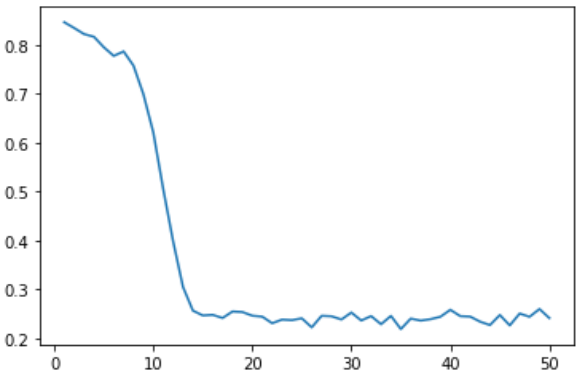
GC-LSTM Exercise 2



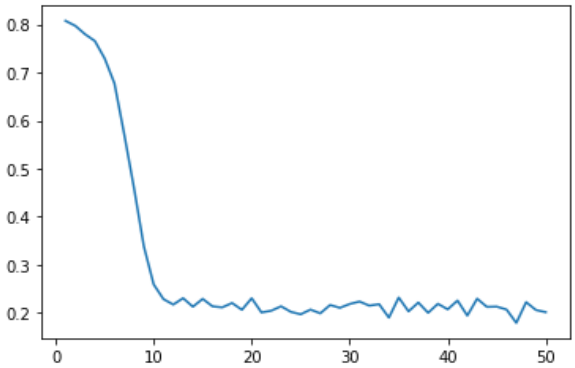
x-axis: No of Epochs | y-axis: Train RMSE Loss

Figure 10: Train loss achieved for Exercise 2 in both models in fold 1.

HF- LSTM Exercise 3



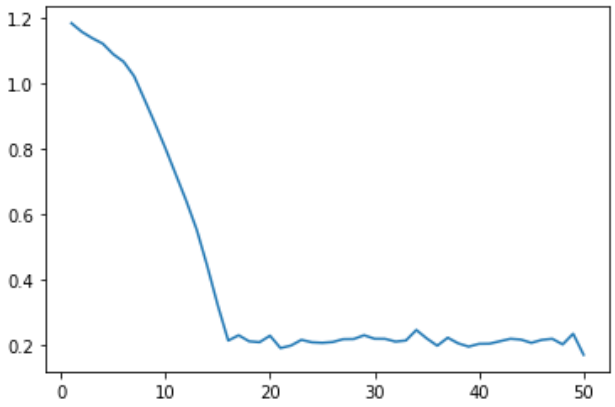
GC-LSTM Exercise 3



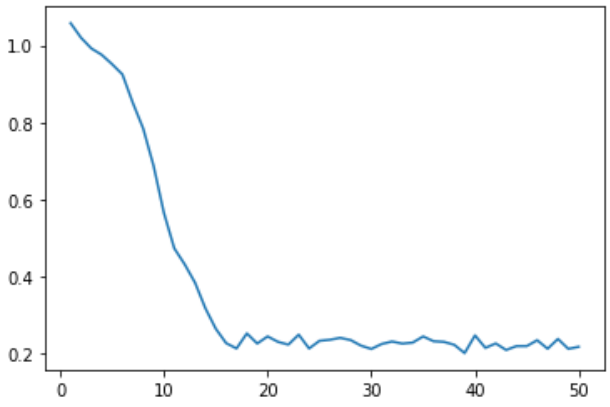
x-axis: No of Epochs | y-axis: Train RMSE Loss

Figure 11: Train loss achieved for Exercise 3 in both models in fold 1.

HF- LSTM Exercise 4



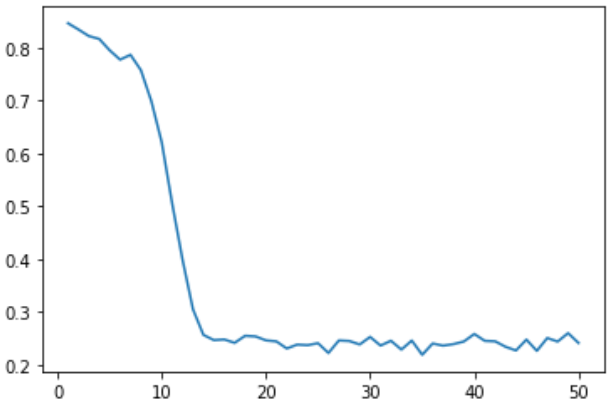
GC-LSTM Exercise 4



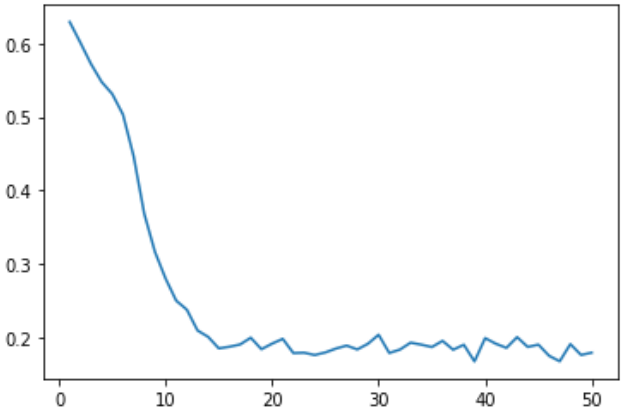
x-axis: No of Epochs | y-axis: Train RMSE Loss

Figure 12: Train loss achieved for Exercise 4 in both models in fold 1.

HF- LSTM Exercise 5



GC-LSTM Exercise 5



x-axis: No of Epochs | y-axis: Train RMSE Loss

Figure 13: Train loss achieved for Exercise 5 in both models in fold 1.

6. Conclusion

6.1. Summary

Skeleton based quality of human movement assessment is a significant challenge with numerous applications ranging from sports, skill assessment to rehabilitation exercise assessment. In this project, we have proposed two frameworks for assessing human movement quality, which can be trained using a relatively small dataset. We explicitly tackle the problem by identifying two key elements, spatial dependencies, and long-term temporal dependencies. Firstly we extract the spatial features using either of the two different modules. We proposed using either handcrafted features (defined by physicians) or using GCN to automatically generate features. The module is succeeded by a LSTM used to understand the temporal aspect of the movement. The proposed model is demonstrated on the KIMORE dataset. The HF-LSTM performed well achieving an RMSE loss of 0.290. On the other hand, the GCN-LSTM performed outstandingly well in comparison to its prior, reaching an average RMSE loss of 0.191.

This establishes the claim that an LSTM can more accurately predict the result when it can take advantage of the whole data, instead of using only the physician's specified features. Furthermore, the project also proves that both GCN-LSTM and HF-LSTM can learn spatiotemporal connections in human movement data.

6.2. Limitations

Assessment of rehabilitation exercise is fairly an underdeveloped topic and thus comes with a lot of limitations. The most significant challenge, when it comes to deep learning-based models, is the lack of large datasets. This also acts as another indicator of the lack of attention from researchers. The largest dataset, the one we used has 78 subjects, performing 5 different exercises. Since each exercise requires a separate model for evaluation, we are limited to only 78 sets of samples for both training and testing. While another skeleton dataset has a higher number of participants, they do not have the required physician's evaluation for a supervised training method. Secondly, the data in KIMORE are classified into 5 categories, expert, non-expert, patients with Parkinson's disease, patients with back pain, and patients who suffered from a stroke. The number of patients in each category is not consistent, thus the model is likely to underperform.

Another big limitation is the real-world use of the framework as it uses separate hardware, Kinect, to collect the data. The mass number of patients, whom we are trying to help, is unlikely to have a Kinect device at their home. To take advantage of our framework they would require investing in a device, a financial investment that may not be useful anyway after recovery unless they plan to play games with it.

6.3. Future Work

6.3.1. Performance improvements

- **Add new models**

The proven frameworks all had a goal to capture the essence of spatial-temporal data. Even though they were able to learn the features; there is a lot of room for improvement. When working with time series and graph data, newer variations of GCN for example, Spatio-Temporal Graph Convolutional Networks (STGCN) have shown promising results in traffic forecasting. Similarly, we plan to also implement more proven spatial-temporal models such as MS-GDN, in the hope of achieving higher accuracy.

- **Generate new Dataset**

One of the biggest challenges faced by the quality of human movement assessment is that there is no large skeleton dataset, which also comes with an acceptable movement score. Therefore we plan to generate a dataset with a larger number of participants, ranging to more exercises, that is also evaluated by physicians.

6.3.2. Usability improvements

- **Image-based**

One of the biggest limitations of our proposed models is that we use skeleton data collected by a separate device, Kinect. It is very unlikely to be existing in the household of patients. Whereas, most patients already own a camera or a device with an integrated camera in this modern era. If we can generate an acceptable model that uses a RGB camera to map movement onto skeleton joints, the usability of this framework will increase dramatically.

- **Movement recognition and assessment**

The current frameworks are limited to only assessing one exercise. The proposal of a single framework that recognizes the movement and then provides an assessment based on the exercise will improve the user experience drastically.

- **Real-time responses**

Currently, the framework only gives a score for the whole exercise. When a patient uses it, they can use the score to know their movement accuracy. But if the system gives real-time feedback to improve the accuracy of the movement, the patient can fix their motion and adjust accordingly.

References

[1]S. Machlin, J. Chevan, W. Yu and M. Zodet, "Determinants of Utilization and Expenditures for Episodes of Ambulatory Physical Therapy Among Adults", *Physical Therapy*, vol. 91, no. 7, pp. 1018-1029, 2011. Available: 10.2522/ptj.20100343.

[2]R. Komatireddy, "Quality and Quantity of Rehabilitation Exercises Delivered By A 3-D Motion Controlled Camera: A Pilot Study", *International Journal of Physical Medicine & Rehabilitation*, vol. 02, no. 04, 2014. Available: 10.4172/2329-9096.1000214.

[3]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. Available: 10.3390/s20185258.

[4]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, pp. 468-477, 2020. Available: 10.1109/tnsre.2020.2966249.

[5]H. Pirsiavash, C. Vondrick and A. Torralba, "Assessing the Quality of Actions", *Computer Vision – ECCV 2014*, pp. 556-571, 2014. Available: 10.1007/978-3-319-10599-4_36 [Accessed 22 June 2021].

[6]B. Crabbe, A. Paiement, S. Hannuna and M. Mirmehdi, "Skeleton-Free Body Pose Estimation from Depth Images for Movement Analysis", *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015. Available: 10.1109/iccvw.2015.49 [Accessed 22 June 2021].

[7]A. Elkholy, M. 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, 2020. Available: 10.1109/jbhi.2019.2904321.

[8]F. ‘Atyka Nor Rashid and N. 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. Available: 10.11591/eei.v9i1.1868.

[9]A. Vakanski, H. Jun, D. Paul and R. Baker, "A Data Set of Human Body Movements for Physical Rehabilitation Exercises", *Data*, vol. 3, no. 1, p. 2, 2018. Available: 10.3390/data3010002.

[10]M. Capecci et al., "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. Available: 10.1109/tnsre.2019.2923060 [Accessed 22 June 2021].

[11]A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", *Neural Networks*, vol. 18, no. 5-6, pp. 602-610, 2005. Available: 10.1016/j.neunet.2005.06.042 [Accessed 22 June 2021].

[12] Pytorch.org. 2021. *PyTorch*. [online] Available at: <https://pytorch.org/> [Accessed 4 July 2021].

[13]"PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models", *DeepAI*, 2021. [Online]. Available: https://deepai.org/publication/pytorch-geometric-temporal-spatiotemporal-signal-processing-with-neural-machine-learning-models. [Accessed: 22- Jun- 2021].

[14]S. 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. Available: 10.2522/ptj.20060260.

[15]K. Jack, S. McLean, J. Moffett and E. Gardiner, "Barriers to treatment adherence in physiotherapy outpatient clinics: A systematic review", *Manual Therapy*, vol. 15, no. 3, pp. 220-228, 2010. Available: 10.1016/j.math.2009.12.004.