PhysiQ: Off-site Quality Assessment of Exercise in Physical Therapy

HANCHEN DAVID WANG, Vanderbilt University, USA

MEIYI MA, Vanderbilt University, USA

Physical therapy (PT) is crucial for patients to restore and maintain mobility, function, and well-being. Many on-site activities and body exercises are performed under the supervision of therapists or clinicians. However, the postures of some exercises at home cannot be performed accurately due to the lack of supervision, quality assessment, and self-correction. Therefore, in this paper, we design a new framework, PhysiQ, that continuously tracks and quantitatively measures people's off-site exercise activity through passive sensory detection. In the framework, we create a novel multi-task spatio-temporal Siamese Neural Network that measures the absolute quality through classification and relative quality based on an individual's PT progress through similarity comparison. PhysiQ digitizes and evaluates exercises in three different metrics: range of motions, stability, and repetition. We collect and annotate 31 participants' motion data with different levels of quality. Evaluation results show that PhysiQ recognizes the nuances in exercises, works with different numbers of repetitions, and achieves an accuracy of 89.67% in detecting levels of exercise quality and an average R-squared correlation of 0.949 in similarity comparison.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Activity Quantitative Assessment, HAR, Neural Network, Physical Therapy

1 INTRODUCTION

Physical therapy (PT) is a process in which patients regain their strength through exercises after surgery, incidents, or illness. It benefits patients by reducing pain, improving mobility, preventing further injury, and improving muscle balance. Patients undergo challenges, endeavors, and struggles with lasting benefits with well-prescribed instruction and supervision. Currently, there are more than 5.1 billion Years Lived with Disability (YLDs)[1] growth per year [23]. In the U.S., there are 38,800 physical therapy clinics operating and an average of 150 patients in each clinic each week, with approximately 300 million sessions for patients each year [43]. Usually, patients require extensive sets of exercise to return to regular activities. However, they usually have limited time in clinics with supervision under physical therapists, and are required to perform exercises by themselves **at home**. Nevertheless, patients have very little knowledge of how well they perform without monitoring or supervision. Moreover, they do not have the flexibility and convenience to set up a camera to self-monitor [46]. Therefore, having a quantitative measurement of the quality of exercises with wearable devices for patients and therapists is crucial to support the patients get their wellness back.

---

[1]This measures the impact of an illness before it resolves or leads to death

Authors' addresses: Hanchen David Wang, hanchen.wang.1@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, Tennessee, USA, 37240; Meiyi Ma, Vanderbilt University, 2301 Vanderbilt Place, Nashville, Tennessee, USA, 37240, meiyi.ma@vanderbilt.edu.

208

1.1 Motivation

To improve effective rehabilitation, self-efficacy, self-motivation, social support, intentions, and previous adherence to physical therapies can help patients perform exercises in home-based physical therapy [14]. However, there is a considerable gap existing between how patients perform self-monitored offsite therapeutic and clinically supervised exercises. Patients and their therapists have no effective and convenient way to track exercises quantitatively at home.

Human Activity Recognition (HAR) in wearable devices is a prevalent research topic that includes many day-to-day locational, behavioral, and planning recognition. For instance, handwashing [50], finger gestures [11], eating behavior [5], and daily activities (writing, cooking, and cleaning) [3] improve to recognize activities through wearable technologies. However, although these works meet the need for daily human activities, a limited attempt exists to help therapeutic rehabilitation for patients.

Moreover, existing works on the quality of exercise utilize vision-based devices, such as cameras, K2 Kinect [17, 36], to track the quality of exercises and provide simple feedback. However, users must set them up physically and potentially interfere with the occlusion of the camera. Moreover, calibrating and adjusting the vision-based devices are costly and inconvenient for injured or immobile people. Therefore, there is a need in providing portable and wearable devices to measure the quality of exercises.

Lastly, attempts on wearable devices to track quality, such as calorie intake [21] and gait authentication [37], suggest the endeavors to use wearable sensors to track the quality of exercises. Ghanashyama et al. attempt to use deep learning models to recognize and count the repetitions of exercises using a single sensor [39]. However, the qualitative pieces of information are not therapeutic metrics to help rehabilitation.

In summary, there is a high demand to improve how patients and users quantitatively measure their exercises offsite with meaningful feedback with wearable devices.

1.2 Challenges

There are three significant challenges in designing such a framework. First, how to digitize physical metrics is an open question. To the best of our knowledge, there is no existing systems or models measuring the quality of an activity, and existing models are not sophisticated enough to directly return a quantitative measurement. Secondly, the quality of exercise varies for different people of different ages, heights, weights, and gender. For example, a tall person has a long traveling distance of shoulder abduction exercise because of his height and arm lengths. Suppose someone of average height performs the same PT exercise with the same range of motion, the model is supposed to be able to tell the difference and similarity even though their heights are different. Similarly, suppose an elder performs differently than a teenager in a PT exercise; even though their objective quality differs, quality should remain the same if they raise their arm to 90 degrees compared to 60 degrees. Lastly, there are no existing datasets with annotated quality of exercise.

1.3 Contributions

Targeting these challenges, this paper introduces a novel framework, PhysiQ, that continuously tracks and quantitatively measures people's off-site exercise activity through passive sensory detection. The framework is robust and general to handle users who perform exercises with different speeds, positions, and postures. We summarize our main contributions below:

- To the best of our knowledge, this is the first framework for quantitative measurement of exercises using a smartwatch. The framework identifies and digitalizes three key exercise metrics of *range of motion*, *stability*, *repetition*, which are built upon the muscular system for understanding the functionality of the skeletal system.

- We create a novel multi-task spatio-temporal Siamese Neural Network that measures both absolute quality and relative quality based on an individual's PT progress through similarity comparison. It enables patients to understand the quality of their offsite exercises over time.
- We build an application collecting users' motion data in a smartwatch and giving explainable feedback with recommendations based on their quality of exercises in real-time.
- We collect and annotate 31 participants' motion data with different levels of stability, range of motion, and repetition, in three shoulder exercises, which are shoulder abduction, external rotation, and forward flexion.
- We perform extensive evaluation using real user's data. Results show that our framework performance outperforms the baselines by 47.67% on average in R-Squared for all exercises and all three metrics. We also provide insights of how user's behaviors influence the framework through a user experience study.

1.4 Paper Organization

In the rest of the paper, we discuss the related work in Section 2. We present our framework PhysiQ in Section 3. Next, we show how we collect the exercise data with different quality in Section 4, and evaluation results in Section 5. Furthermore, we present a survey and discuss user experience using our app and implications in Section 6, followed by a discussion and summary in Section 7 and Section 8, respectively.

2 RELATED WORK

In this section, we present existing literature on state-of-the-art measuring the quality of activity and applications, and deep learning methods for these applications through similarity comparison and other methods.

2.1 Measuring Quality of Activity

Though there does not exist a field of such, Human Activity Quality Recognition (HAQR) is a critical research question for real-world application. Therefore, we have gathered and scrutinized related works in state-of-the-arts. For example, one quality in exercises is repetition counting. Work, such as [47], focuses on multi-modality to provide a more accurate result of repetitions counting and exercise recognition. A fascinating work done by Radhakrishnan et al. focuses on using in-ear devices such as wireless headphones fusing with inertial measurement units to quantify insights and feedback in gym exercises [40].

Additionally, one work uses smart speakers to analyze the duration, intensity, continuity, and smoothness of exercises at home [56]. However, such work requires people to have knowledge about exercises and some level of proximity to the actual devices. Additionally, IMUTube introduces how to simulate virtual IMU data from video [26]. Kwon et al. utilizes video to simulate IMU through a number of off-the-shelf computer vision and graphics techniques. Furthermore, Radhakrishnan et al. suggest a system that uses a magnetic accelerometer sensor. The device is mounted on the weight stack of a gym machine to infer exercise behavior using multiple machine learning models to identify the person, amount of weights, type of exercise, and mistakes [41].

Like the quality of exercises, sleep quality analyzes humans' brain activity through Electroencephalography (EEG) signal. It categorizes the level of sleep activity in Rapid Eye Movement (REM), Non-REM, S2 (light sleep), and S3 (slow-wave sleep). Additionally, polysomnography is considered the standard methodology for detecting the sleep pattern with carefully analyzed records of epochs [12]. There are two leading technologies for sleep monitoring: portable and contact. Portable devices such as smartwatches or smartphones seamlessly collect users' data through passive means. Contact devices are medical-level tools to collect reliable data through less passive means [12].

Portable devices require stationary devices to record the daily routine of the subjects [6, 30, 31, 44, 48]. These works include signals of accelerometer data to approximate respiration rate and heart rate through, for example,

sound recording to validate sleep time and duration, and electrocardiogram (ECG) signal to proximate heart rate and distinguish threshold for sleeping and waking. Contact devices, which require direct contact with the subjects, are the cases of PPG, EEG, and Actigraphy. These assessment technologies have the main advantage of accurately sampling the human body's physiological and mechanical signals. However, such devices are perceived as obtrusive due to their limited portability and transparency. Sleep quality recognition and detection using contact devices have been well studied. One research targets whether having additional information such as age and sleep stage information can distinguish abnormality using the deep learning method on EEG signal data [49]. Other research focuses on developing an automatic sleep staging method in EGG signals, in which the author proposes multi-epoch methods to segment and encodes the feature and re-concatenate to predict its sleep stage [29]. Lastly, one aims to develop a sleep scoring toolbox with the competency of multi-signal processes, feature extraction, and classification and prediction but only using a simple logic-feature-based method to differentiate sleep stage [57].

2.2 Deep Learning Models for Quality Measurements

Siamese Neural Network (SNN), utilizing two identical networks, is commonly applied to compare if two subjects are same or not. Typical tasks include image classification, object recognition, and object tracking [13, 16, 18, 28, 45, 51]. However, it only returns a binary result (i.e., True or False) without a quantitative measurement.

Furthermore, recognizing similarities in 1-D signal data, such as radar, speech, and natural language process (NLP) [2, 15, 32, 35], has also gained many usages in different applications. An intriguing application stands out using semantic similarity between sentences, supplementing recurrent neural networks with synonymic encoding. Mueller et al. use LSTM to encode the different length inputs with its positional encoder to analyze the semantic similarity with an outstandingly high Pearson correlation of 0.8822 [33]. Furthermore, A deep dive evaluation of the SNN proposes a residual module to reduce learning biases caused by padding [58]. The determinant factors, such as stride, padding, and receptive field size (the size of the region that produces the feature), are crucial to the performance of the SNN. Additionally, Lawrence et al. suggest an innovative way of utilizing spatial and temporal aspects from videos to recognize human emotions is inspiring. [27]. Lastly, Agrawal et al. use an inventive way to calculate articles' similarity distance for political stance using discrete labeling of relatedness [1].

Additionally, it is worth mentioning a self-supervised method called contrastive learning. It is a technique to learn the general features of a dataset without labels by telling what data points are similar or dissimilar. The reason why this is important to us is that the similarity of the task is to recognize the similarity. However, the goal might differ for contrastive learning due to their limited dataset or no handcrafted labeling. Several interesting works related to SimCLR, SIMCLRv2, and MoCo (momentum contrast) are fine-tuning with a few labeled examples to achieve high accuracy [8, 10, 19]. These works are essential in unsupervised learning fields and provide highly accurate and efficient solutions. Additionally, an audio similarity work is being introduced to assign high similarity to the audio segment from the same recording while assigning low similarity to different segments from di qfferent recordings [42].

In summary, these systems and applications measure the quality of words, speech, sleep, and emotion. However, none of them standardizes the metrics of exercises through the muscular system. PhysiQ recognizes therapeutic exercises and digitalizes the exercises to provide feedback through the metrics and compare exercises using deep learning methods.

3 SOLUTIONS

We build PhysiQ to continuously monitor users' off-site exercise activity and quantitatively measure the quality.

(a) Exercise Qualitative Feedback on Patients' GUI

(b) TO-DO list on Patients' GUI
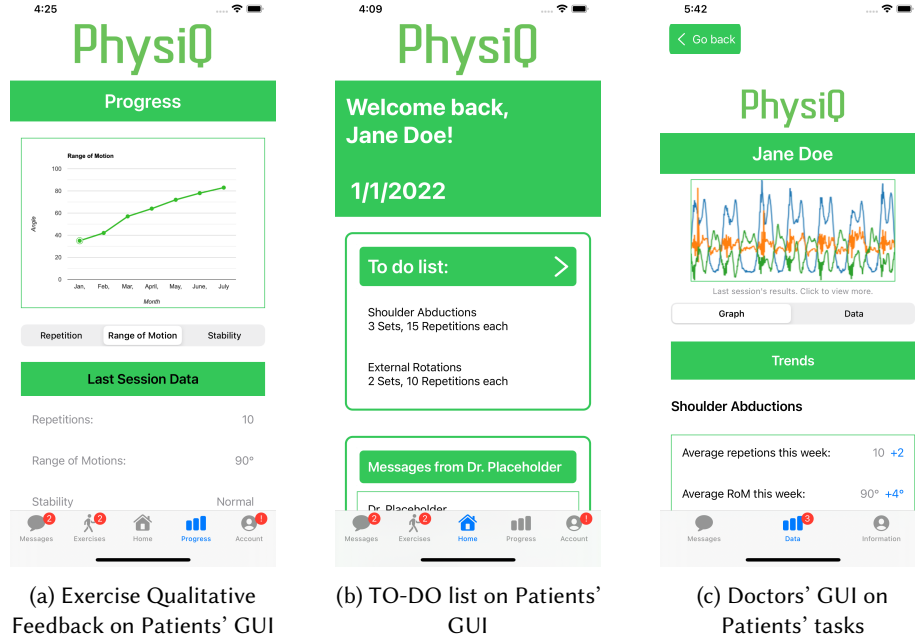
(c) Doctors' GUI on Patients' tasks

Fig. 1. PhysiQ on Smartphone Graphical User Interface (GUI) for patients and doctors. Leftmost GUI demonstrates the patients' progress tab and how they make progress throughout the months; rightmost GUI demonstrates the doctor' tab to monitor patients' results and how much they improve or aggravate with visualization of the data.

We particularly measure the quality of exercises in three metrics of *range of motion*, *stability*, and *repetition*. The framework is shown in Fig. 2. PhysiQ apps run on three platforms. Users first perform activities wearing a smartwatch. The PhysiQ app on the smartwatch extracts the sensory data and syncs the data with the smartphone and cloud in real-time. Then, the model on the cloud generates scores for the exercises in different manners, such as range of motion and stability. Based on the scores, the model sends feedback, such as, "The score for a range of motion of this exercise is 150, which means you did a good job on range of motion. " and recommendations, such as "You could do 3 more repetitions!" on the PhysiQ app on user's smartphone. Meanwhile, it also uploads the user's progress to the cloud for the therapist to review. We present some of the graphical user interfaces (GUI) of our app on the phone in Figure 1. In the rest of this section, we first formalize our problem, then present how we digitalize the exercise metrics on sensory data, and finally elaborate on the details of our multi-task spatio-temporal SNN-based quality measurement model.

3.1 Problem Formulation

We formalize the problem as follows: given the smartwatch with built-in IMU sensors, it returns the input of IMU sensory data $\mathbb{X}_i^j$ by j-th participant and i-th sample. Each data $\mathbb{X}^j$ contains T^j number of samples $\{x_1^j, ..., x_{T^j}^j\}$. Each x_i^j is comprised of 3-axis of accelerometer data (A^x, A^y, A^z) and 3-axis of gyroscope data (G^x, G^y, G^z). Our model outputs a similarity score s.

We further divide our problem into three folds for three metrics:

(1) Our problem input is x_i^j and output is a set of $X' = \{x'\}$ with x' represents one repetition and n represents total repetitions.
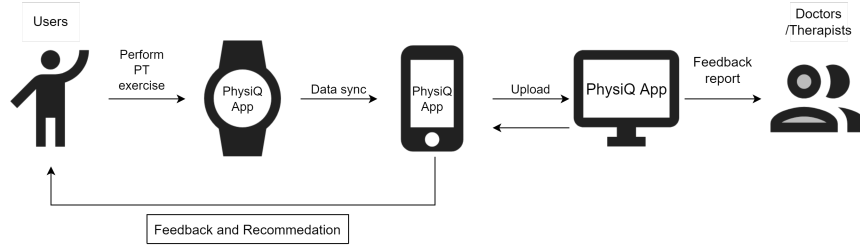
Fig. 2. Overview of PhysiQ framework

(2) In *range of motion* metrics, we divide it into an absolute and relative values. First, we formalize our problem under the same problem of input sensory data, $x' \in \mathbb{X}_i^j$. In absolute value, for each x', we define $y_{rom}(x')$ as the absolute score of *range of motion*. Secondly, in relative value, we define the relative value, $s_{rom}(x_i', x_{i*}')$, as similarity score under *range of motion* metrics, where $i*$ is the anchor exercise's index.

(3) For $x' \in \mathbb{X}_i^j$, We define the relative value, $s_{stability}(x_i', x_{i*}')$, as similarity score of *stability* metrics.

3.2 Digitalizing Exercise Metrics and Ground Truth

3.2.1 Repetition. We implement a novel energy function to segment the exercise's repetitions. We calculate the energy as shown below:

$$E(i) = \frac{1}{f_s + 1}\left(h(i) + \sum_{n=-T}^{T} \sqrt{h(n+i)}\right), \text{where } f_s = 50Hz, T = \frac{f_s \lambda N}{2000} \tag{1}$$

$$h(i) = |A^x(i) * W^x| + |A^y(i) * W^y| + |A^z(i) * W^z| \tag{2}$$

In the formula above, i is the positional index to calculate the energy throughout the signals, and N is the actual length of the signals in 10 repetitions. As shown in Fig. 3, we use this method of calculation to find the cutting position to semi-automatically segment the signal of 10 repetitions to the numbers of 1 repetition. Noted, since some exercises have a low and high amplitude signal, we use hyper-parameters W^x, W^y, W^z, and λ weights to adjust the smoothness of the energy. The purpose of the energy is to merge two peak from previous and next repetition to form a significant energy level to identify the cutting position. As a result, a good hyper-parameters are required to pre-process the segmentation of the data.

3.2.2 Range of Motion. After using the energy method to segment the signal of participants' exercises, we annotate each exercise according to its labels and positions. In our case, we have three potential labels: *range of motion*, *stability*, and *repetition*. We modify our method to generate *stability* using the Equation 3 as our ground truth for stability. *Range of motion* metric is collected through participants' exercises under the supervision's of experimenters. We examine the participants' range of motion as they performed and verify with recorded videos. *Repetition* is labeled based on the number of repetitions merged together. We target one particular exercise to build our framework and use two additional exercises to verify its competency: shoulder abduction (SA), external rotation (ER), and forward flexion (FF). We explain the process in Section 5. Due to different muscle activation of *range of motion*, we examine shoulder abduction as our initial exercise to design our metrics and framework. Additionally, we classify shoulder exercises into two main categories: half arm span (HAS) and half-half arm span (HHAS). HAS contains exercises that require both forearm and arm into motion. HHAS only requires forearm or arm into motion.

(a) Energy Plot with Accelerometer Data

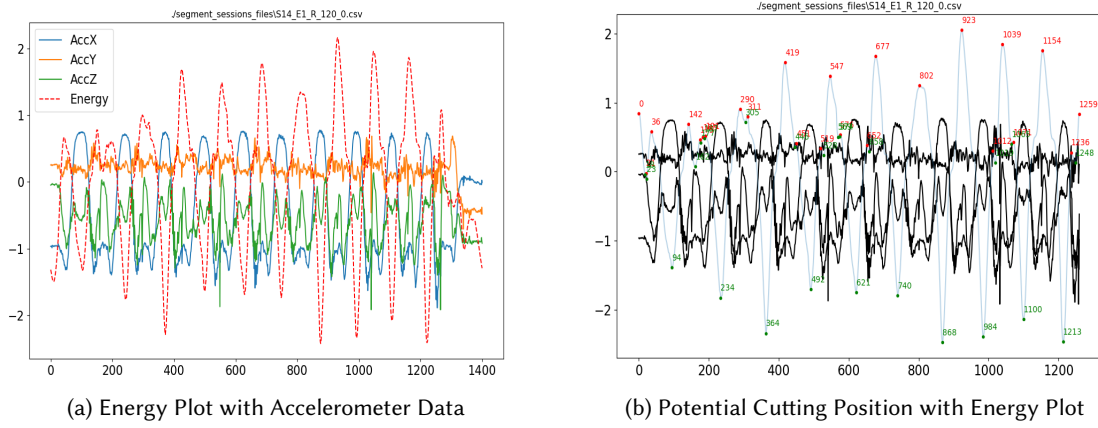(b) Potential Cutting Position with Energy Plot

Fig. 3. These two figures show how the energy plot suggests the cutting positions for the repetition of 10 exercises. For Fig. 3a, the energy is plotted in a red dashed line. The specialist manually does the exercises' beginning and ending cut-off procedure. In Fig. 3b, we change the original data colors to black and emphasize the color of the cutting position for segmentation and energy plot for visualization purposes.
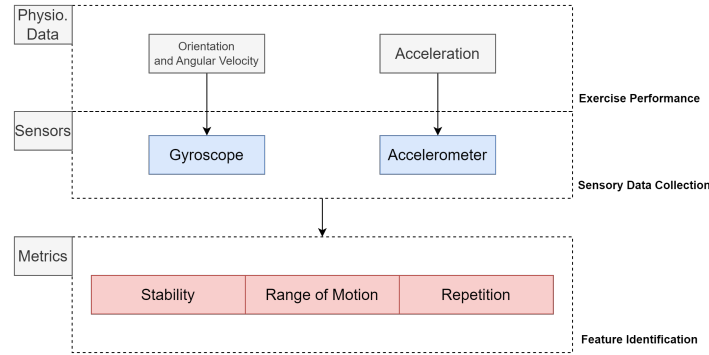

Fig. 4. Exercises Metrics: we first identify the type of exercise to perform and which sensory data is collected. Based on the signal data gathered from participants, we measure them against our metrics. The framework assesses the quality of exercises based on the *range of motion*, *stability*, and *repetition*.

In shoulder abduction exercise, it involves the glenohumeral joint and scapulothoracic articulation in different *range of motion*. At first 20 to 30 degrees of motion, subjects do not use the scapulothoracic joint motion, and the supraspinatus tendon should be the only muscle helping during this. Deltoid muscles are activated to support from 30 to 120 degrees of range of motion. Lastly, beyond 120 degrees, a full abduction is considered when the arm is externally rotated with the humerus activating. Different range with different muscles activation inspires us to finalize five different *range of motion* and *stability* (we mentioned on how to create stability in Section 4) as our categories to understand the quality of exercises with Fig. 5 [52–55].

Range of Motion for HAS Exercise Muscular Activation

- 30 degree ROM: supraspinatus muscles
- 60 degree ROM: deltoid muscles

(a) Shoulder anatomy, front view

(b) Serratus anterior muscle
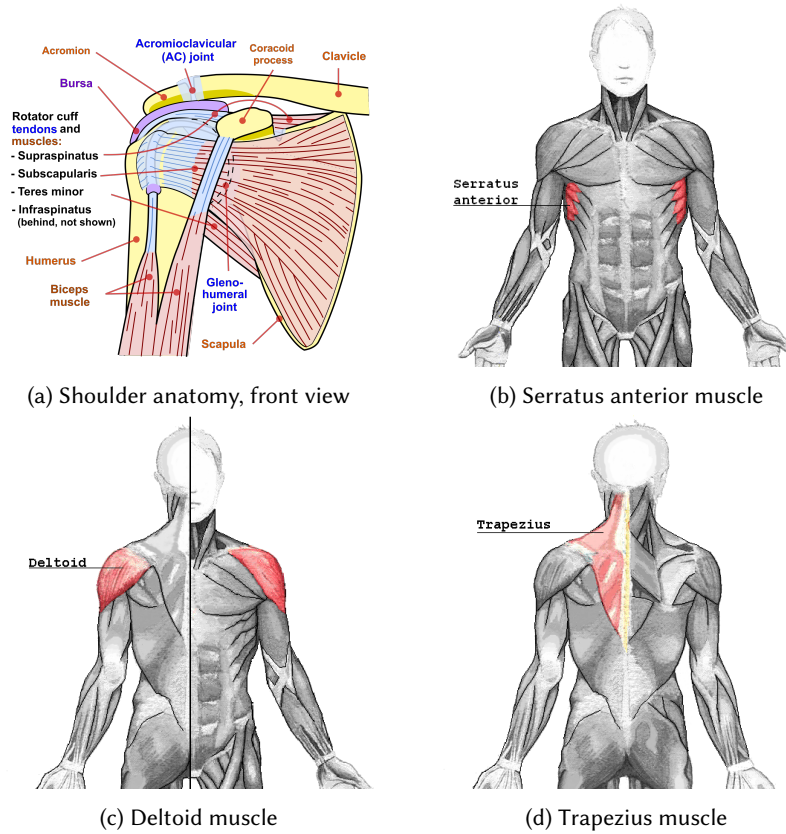

(c) Deltoid muscle

(d) Trapezius muscle

Fig. 5. This is a visualization of how shoulder with joints, tendons, and muscles work in a close look [52–55].

- 90 degree ROM: deltoid muscles, transition to trapezius muscle
- 120 degree ROM: trapezius and serratus anterior muscle
- 150 degree ROM: trapezius, serratus anterior muscle, and humerus activation

In external rotation exercise, the rotator cuff is used to perform this exercise. The rotator cuff consists of four muscles that stabilize the shoulder in external rotation, infraspinatus, teres minor, supraspinatus, and subscapularis as shown in Fig. 5. As in external rotation, the infraspinatus muscle stabilizes the shoulder joint and acts as the prime mover in this exercise [22].

Range of Motion for HHAS Exercise Muscular Activation

- 45 degree ROM: minimally infraspinatus muscle
- 90 degree ROM: infraspinatus muscle, posterior deltoid
- 150 degree ROM: infraspinatus muscle, posterior deltoid, and all other muscles in rotator cuff.

Similarly, in forward flexion exercise, the muscle activation includes supraspinatus, infraspinatus, and anterior deltoid. We classify *range of motion* into 5 similarly to shoulder abduction because these two exercises are considered half arm span exercise.
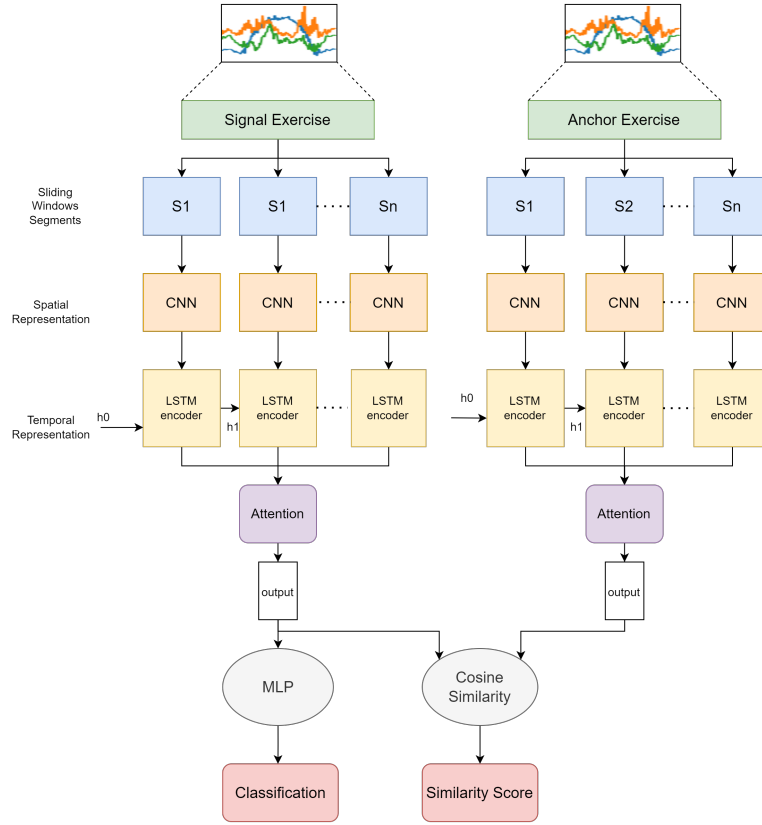
Fig. 6. Multi-task Spatio-temporal SNN: First of all, we have 2 one repetition exercises feed into our network as signal and anchor exercise. Signal exercise can be compared against anchor and vice versa. After Sliding Windows Segmentation, Spatial and Temporal encoding, we feed our model into an attention mechanism. Finally, we compare these two hidden features representation and output its similarity score using cosine similarity. Additionally, the hidden feature of the signal exercise is processed into a MLP network to get result of classification based on the metrics of *range of motion*, *stability*, or *repetition*.

3.2.3 *Stability.* Initially, we defined our ground truth of stability using resistance bands relative to the participants' strength. For example, if the participant is strong, we progressively find his or her maximum strength with the resistance band and create two different instability based on that, labeled as two classes. However, this method does not guarantee that stability correlates with the levels of resistance bands. Furthermore, as we visualize and evaluate it, we do not find the difference between the two different resistance bands. Therefore, we define the stability metrics using a low pass filter and coefficient of variation through a mathematical methodology. By doing so, we can measure stability of all the exercises performed by participants. Low pass filter is used to differentiate what is human motion and signal noise since our IMU sensors captures at 50 Hz. As suggested by Khusainov et al., human activity frequencies are between 0 and 20 Hz, and 98% are below 10 Hz [25]. Therefore, for maximal capturing of stability, we use 20 Hz as our parameter for the low pass filter function.

$$instability(S) = Tanh\left(\mid CV(\text{lps}(S, \ f=20)) \mid\right), \text{where } S = [A^x, A^y, A^z, G^x, G^y, G^z].T \tag{3}$$

$$CV(S) = \frac{\mu_S}{\sigma_S} \tag{4}$$

In the formula above, the lps represents the trivial low pass filter function we used to filter some noises.
Stability Muscular Activation

- 0.0 stability: stable and perform normal exercises
- in-between 0.0-1.0 stability: rotator cuff
- 1.0 stability: (unstable) supraspinatus, infraspinatus, teres minor, and subscapularis

3.3 Spatio-temporal Feature Representation

As shown in Fig. 6, we develop a method of combining spatial and temporal representation to recognize the shape of signals of exercises [4, 7, 34, 38]. We use attention mechanism as a method of message passing to understand relationships in different the hidden features of sliding windows segments in one repetition.

To discern and recognize the details of the signals and find the correlation between metrics and exercises, we tackle the problem in two different aspects: time and space. The exercises performed by users are the input of temporal signals. The signals result in a pattern in space to depict different angles of the exercises.

We build a CNN-based spatial encoder as:

$$CNN(w) = (\sigma(sum(W_1 \odot w) + b_1), \ldots, \sigma(sum(W_n \odot w) + b_n)), \tag{5}$$

where $W_1, W_2, \ldots, W_n$ are learnable weights matrices, $b_1, b_2, \ldots, b_n$ are biases, $\odot$ is element-wise multiplication, sum is element-wise summation, and σ is the activation function such as ReLu. CNN is capable of effectively interpreting spatial information and transforming it into a hidden pattern. It has the potential to compress information to represent into a smaller space, and it is very effective to compress sliding windows of signals. This provides a feature extraction mechanism across windows with specific weight matrices and biases. What is important in spatial encoding is by using a feature extraction method, our model can interpret the importance of each window given its features. Additionally, a max pooling is applied in our CNN model to reduce the signal's dimension.

However, understanding the features in each window is not enough to distinguish any temporal knowledge due to the property of time series, such as trends, and seasonal or non-seasonal cycles. Next, we employ Long Short Term Memory networks (LSTM) for temporal encoding as [20]:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{7}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{8}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{9}$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \tag{10}$$

$$h_t^r = o_t tanh(c_t), \tag{11}$$

where, U W, and b are weights and biases that is not time dependent, σ is a sigmoid activation function, and i, f, o are input, forget, and output gates respectively. Lastly, c and h are the cell and hidden states vector given the time t.

Next, attention mechanism on a set of queries, keys of dimension d_k, and values are calculated using the matrix of output as shown below:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

Instead of performing one attention function, we apply multi-head attention where head is the number of paralleled attention mechanism. Multi-head attention allows our model to attend information with different representation at different locations in time and space. We use H as the number of head in multi-head attention:

$$MultiHead(Q, K, V) = concat(head_1, \ldots, head_h)W^O \tag{13}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{14}$$

In our work, we use $H = 16$ heads of paralleled attention layers. Additionally, we use $d_{model} = 256$; therefore, each $d_k = d_v = d_{model}/H = 16$.

As a result, we describe our model, PhysiQ. It takes advantage of the fact that every sliding windows w of size k x 6 can be interpreted as a frame in time. Using this, we feed each window w into the CNN to output h^c as z^c x 1. Additionally, we feed h^c into LSTM to get h^r with a size of z^r x 1. The LSTM creates a sequence of hidden states $[h_0^r, \ldots, h_n^r]$, acted similar to a positional encoding to understand the temporal information. We pass the sequence into the attention mechanism returning our hidden representation, A that has both temporal spatial information, and relational message passing knowledge.

Algorithm 1 PhysiQ Framework Encoder

1: Def: $A = ENCODER(e)$, s.t. exercise segment e passes in, return A, where A has n number of hidden presentation as $h_0^r, \ldots, h_{n-1}^r$
2: **Input**: e, exercise segment
3: **Output**: A, hidden representation of one exercise segment
4: $w_0, w_1, \ldots, w_{n-1} = SLIDE(e)$
  $SLIDE$ takes an exercise e and return W, where W is the size of n x k x 6, n number of sliding windows.
5: $W = w_0, w_1, \ldots, w_{n-1}$
6: **for** each w_i in W **do**
7:   $h_i^c = CNN(w_i)$
8:   $h_i^r = LSTM(h_i^c)$
9: **end for**
10: $H = [h_0^r, h_1^r, \ldots, h_{n-1}^r]$
11: $A = Attention(H, H, H)$
12: return A

Algorithm 2 PhysiQ Similarity Comparison

1: Def: $s_{ij} = SIMILARITY(e_i, e_j)$, where s is the similarity score and e is the signal segment
2: **Input**: e_i, e_j, a pair of exercise segments
3: **Output**: s_{ij} is the similarity score between a pair of exercises
4: $A_i = ENCODER(e_i)$
5: $A_j = ENCODER(e_j)$
6: $s_{ij} = Cosine(A_i, A_j) = A_i \cdot A_j / ||A_i|| * ||A_j||$
7: return s_{ij}

3.4 Similarity Comparison

The supervised learning framework has recently improved dramatically on 1-D and 2-D healthcare signal processing tasks. However, it does not leverage the framework to understand the relationship between the inputs. Significantly, how can patients improve without an anchor comparison from the previous performance? By comparing their day-to-day performance, PhysiQ understands whether participants enhance their performance based on the result of their exercises and how the patients are improving. To address this issue, we utilize the Siamese Neural Network, a type of contrastive learning framework that can extract useful features from data itself without the need for large handcrafted labels. On top of that, we design a data collection strategy to gather multi-modality of data for future evaluation analysis.

Siamese Neural Network (SNN) is a neural network that shares and contains two identical networks with the same configuration and the sharing of the weights. The identical model is used to find the similarity between two inputs. At the same time, the advantages of the SNN are more robust to the class imbalance in the data, learning a tremendous hidden and embedding deeply semantic similarity; however, it does not necessarily output probabilities but the distance between classes. We re-design the network and make it to fit the problem of reference comparison in Fig. 4. We formulate a regressive distance between two exercises as the ground truth label to train SNN with a prior assumption of a maximum of the *range of motion* of R in shoulder abduction, as shown below:

$$s_{rom}(m_a, m_b) = 1 - |\frac{m_a}{R} - \frac{m_b}{R}| \tag{15}$$

With the Equation 3 to get the signal's *stability*, we can measure the similarity:

$$s_{stability}(S_a, S_b) = 1 - |instability(S_b) - instability(S_b)| \tag{16}$$

Lastly, with the assumption of maximum of repetition M, we can measure the similarity of *repetition*:

$$s_{repetition}(r_a, r_b) = 1 - |\frac{r_a}{M} - \frac{r_b}{M}|, \tag{17}$$

where r_a, r_b represent the number of repetitions of signal and anchor exercise.

The SNN is very popular and used to solve various problems in research. However, to accommodate our specific problem, we aim to improve our accuracy by carefully designing our feature extraction in spatial and temporal aspects. Therefore, we adapt the SNN for our signal comparison because there is an underlying knowledge and information that the deep learning method interprets and understands. Additionally, we leverage the temporal encoding, spatial encoding, and attention mechanism to generalize the model with our metrics. Additionally, we use cosine similarity as our similarity measurement because of its ability to differentiate orientation distance between two encoded exercise features.

4 DATA COLLECTION

4.1 Type of Exercises

Selecting well-represented exercises is very crucial to our problem because the exercise itself should have the quality of repetitiveness, singularity (a defined beginning and ending), reconstructiveness (an exercise that can reshape part of a person's body as an improvement), and representativeness (can cover many parts of the muscles). Therefore, considering these factors, we conduct our evaluation of these three exercises, as shown in Fig. 7:

- Shoulder Abduction: The shoulder abduction is an exercise that requires subjects to stand straight with both hands tucking on the side of the legs as the starting point. Subjects perform the exercise by raising the instructed arm to a certain degree of motion and straightening the arm. Once the subjects reach the stopping point, they drop their arm steadily, as it is similar to raising their arms.

(a) Shoulder Abduction
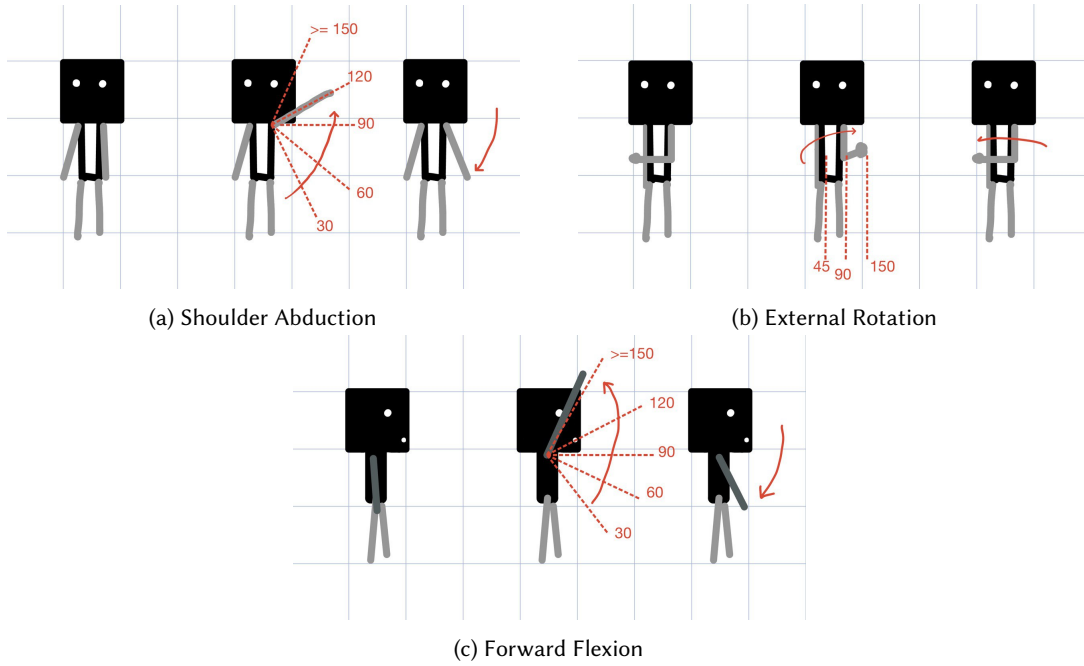
(b) External Rotation

(c) Forward Flexion

Fig. 7. Three Exercises: The exercise on the top left is one repetition of shoulder abduction, the exercise on the top right is one repetition of external rotation, and the exercise on the bottom is forward flexion. Also noted that we define greater equal than 150 degree of motion as 150 on the shoulder abduction and forward flexion.

- External Rotation: The external rotation has two components. First, the upper arm (bicep and tricep) tucks in the armpit while rotating the arm externally and keeping the lower arm raised to about 90 degrees, perpendicular to the chest or abdominal. Secondly, the arm should move horizontally and stop at a certain degree of motion or to the full ROM, where the subject's shoulder should feel a sense of blocking by the joints.
- Forward Flexion: The forward flexion is similar to shoulder abduction but different in direction. Subjects perform the exercise by raising the arm slowly forward, reaching the point of ROM, then consistently lowering the arm to the beginning position.

We simplify our problem as a preliminary examination from the description we mentioned above. Each exercise has its advantage and weakness. As a result, shoulder abduction, external rotation, and forward flexion have aspects that we look for in exercises, including subjects not needing to lie down and wear a smartwatch on the leg to perform exercises in initial modeling. Additionally, we choose shoulder abduction as our first exercise to test our framework, PhysiQ, because shoulder abduction has all the factors we considered: repeating in sets, singular in the beginning and ending, reconstructing people's bodies, and targeting many muscles around the shoulder. It requires extensive time and resources to find the participants during the pandemics, and the quality of data is varied based on the subjects because we minimize how much we tell the subjects to perform the exercises while giving enough details on reaching different stop points of degrees in motions and instability. In order to perform the metric of *range of motion*, *stability*, and *repetition*, we see how robust our model is and in what scenario it can handle and fail.

4.2 Data Statistics

In total, we have 1550 segmented one repetition exercise of *range of motion (ROM)* and 1170 segmented one repetition exercise of *stability* for shoulder abduction. At the same time, we have 600 segmented one repetition exercises of *ROMs* for external rotation. Additionally, the third exercise, forward flexion, has 650 segmented one repetition data. In total, we have collected 31 participants for all data collection in different evaluation periods. In additional, we have 31 participants from shoulder abduction, 24 participants from external rotation, and 11 participants from forward flexion. The metrics of *range of motion* is labeled as we collect the data, and the *stability* is generated using our method as shown in Equation 3. The metric of *repetition* is utilized through our energy segmentation and merged based on the number of repetitions for evaluation. We can form any number of *repetition* by combining the adjacent neighbors of segmented repetitions.

In SNN, we create pairs of inputs for similarity comparison in one repetition exercise of *range of motion*, *stability*, and *repetition*. Therefore, we define the problem as no comparison between subjects, only the comparison of exercises within one particular subject at a time, because with the presumably perfect anchor exercise, it is a relative measure of the similarity of the signal exercise on a particular user.

Table 1. Participants information

Attribute	Male	Female	avg	std
Gender	17	14	N.A.	N.A.
Age	18-25	18-44	22.32	4.77
Height (cm)	167.6-190.5	149.8-182.9	173.6	10.27
Weight (kg)	54.4-108.8	45.3-88.9	67.31	15.90
Previously Shoulder Injures	2	1	N.A.	N.A.

5 EVALUATION

We evaluate PhysiQ from three perspectives: model performance in exercises, generalizability in different metrics, and importance of components in the model. We compare the overall performance of the framework with different types of baseline and how the model performs in different exercises and metrics. We explain the design and set up in Section 5.1, 5.2. Additionally, we test our model performance and generalizability in Section 5.4. Lastly, we evaluate our model's components in Section 5.5 and Section 5.6.

5.1 Implementation

We implement the PhysiQ application in consumer-level IOS Apple Watch with an automatically connected application on iPhone. PhysiQ uses a built-in accelerometer and gyroscope to collect sensory motion data and analyze the data quantitatively. The maximum sampling rate that we choose is 50 Hz, because through our analysis of a single exercise, we observed that the Fourier frequency is mostly below 10 Hz, and other papers also support this observation [24]. Additionally, one of our metrics is stability, and such core motion of the body should be captured more cautiously with a higher frequency rate. Thus, 50 Hz is what we decide to use. Lastly, once the users have performed exercises, the result automatically synced from the watch to the smartphone through Xcode WCSessions.

5.2 Training and Testing Dataset

Once the data is recorded, we segment the data according to our energy function. We used weights W_x, W_y, W_z for each accelerometer x, y, and z, and λ as an additional hyper-parameter. Next, we split the dataset and apply

standard scales for all exercise segments; the scaler is an axis-wise scaler standardized on the current axes (x, y, and z for both accelerometer and gyroscope sensory data). There are two splitting methods for evaluation we employed. The first one is leave-one-person-out cross-validation (LOOCV). LOOCV is a cross-validation approach that treats each subject as a "test" set. This type of k-fold cross validation has the k value as the number of participants. LOOCV separates the models from seeing the validation/testing set. As a result, the model does not see the distribution of validation subjects (participant), and we can analyze the generalizability of our model on 34,000 training samples and about 3,000 validating and testing samples, of total 31 participants. In ROM and stability, we have a total of 37000 data samples. In repetition, we perform a repetition comparison among 1, 2, and 3 repetition, with a total of 280,000 data samples in shoulder abduction, 15,000 in external rotation, and 8,000 in forward flexion. For efficiency, we randomly choose 10% of the data in shoulder abduction for the overall evaluation for 10 times. Secondly, we perform a standard 70% 10%, and 20% respectively on training, validation, and testing splits as our overall evaluation. We randomly extract these splits in each subject in 70, 10, and 20 fashions in normal splits. By having both evaluations, we should know how our model works in real-world scenarios and how well our model can perform.

5.3 Evaluation Metrics

We evaluate the performance using three measure methods in all experiments, i.e., R-squared, Mean Square Error (MSE), and Mean Absolute Error (MAE). R-squared, the coefficient of determination, is how close the data are fitted in the model or the percent of variation explained by the model. MSE measures the mean of the squares of the errors, meaning it calculates the average squared difference between predicted and target values. Finally, MAE measures how far predicted values are from observed values with their absolute difference. Equations shown below:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i (y_i - \bar{y})}, \tag{18}$$

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{19}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, \tag{20}$$

where n is the number of dataset, $\hat{y}_i$ is the predicted value, y_i is the ground truth value, and $\bar{y}$ is the mean value. $\hat{y}_i$, in our case, is the similarity score of two inputs of segments, and y_i is our ground truth of similarity score based on *range of motion*, *stability*, or *repetition*. To test our model with the baselines, we evaluate the experiments on our local machine with a CPU of AMD Ryzen 9 5950X with a 16-Core processor (3.40 GHz), RAMs of 64 GB (3200 MHz), and a GPU of Nvidia GeForce RTX 3090. The operating system is Windows 10 Pro. Additionally, we envision to deploy our deep learning model into cloud server to support API call directly from our mobile application.

5.4 Performance

5.4.1 Siamese Similarity. We explore different ways of quantitatively measuring our exercise metrics. There are two necessary baseline models: RNN and CNN because RNN is known for arbitrary input with time-varied data, and CNN learns well in spatial and temporal features [47]. Additionally, we choose SimCLR as third baseline because its architecture enables useful representation learning in contrastive learning [9]. Thus, we provide and twist the networks such as SimCLR [9], VGG, and RNN as our baselines to further understand our problem. In SimCLR, the original paper proposed using the hidden features in the final layer of ResNet as a representation to compare with a different image (or augmentation of the same images) to learn the underlying knowledge
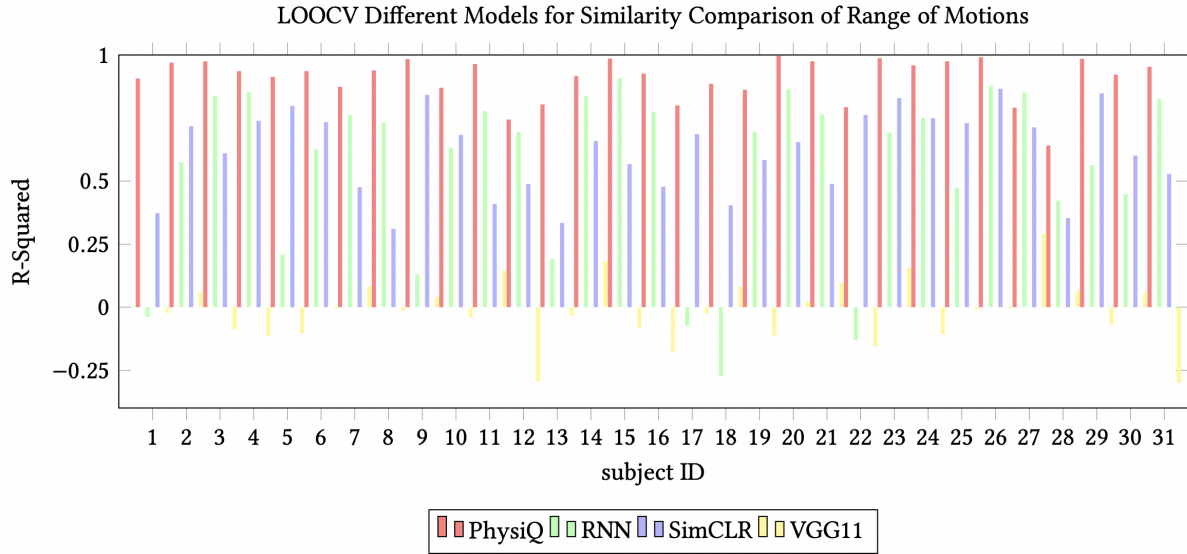
Fig. 8. This figure demonstrates that our framework, PhysiQ, outperforms many other architectures by capturing the spatiotemporal information. As shown above, the red line represents our framework, PhysiQ with R-Squared result on y-axis. Noted, the higher the R-Squared, the better the result as it explains how fitted our model is given the unseen data. Moreover, interestingly, as shown in green line, vanilla RNN captures the temporal information very well but it has problems of generalizability as in subject ID 1 while some of subjects the RNN model can predict very well.

of images with a contrastive loss. However, in our problem, we return a label with a value between 0 and 1. We simplify the output procedure in SimCLR by having a Cosine Similarity as the final layer to compare the two images (in our case, two segments of signals) and add a Sigmoid activation function to return such a result. Additionally, since SimCLR utilizes ResNet in image comparison, we use their structure with a 1-D convolutional layer instead of 2-D to perform an equivariant operation as well as to max pooling procedure. Moreover, VGG is applied with similar 1-D convolution and 1-D max pooling to work with signals. Lastly, we also used a vanilla RNN as our baseline. We use the outputs of the last sliding window as hidden representation and feed into a similar output procedure as described earlier in SimCLR.

There is a potential drawback between half arm span and half-half arm span exercises. Because of the limited external rotation motion, the model might have a harder time distinguishing all three metrics. But overall, our model in all three exercises outperforms the baseline extensively in the metrics of *range of motion*, *stability*, and *repetition*, as shown in Table 2, Table 3, Table 4. Moreover, because external rotation does not have as many waypoints of motion as forward flexion (external rotation only has three, but forward flexion has five), its results are not as good as shoulder abduction and forward flexion in *stability*. Additionally, RNN and SimCLR can have a related good performance throughout the metrics but have difficulties with higher accuracy because of the low complexity of the models. Similarly, in Figure 8, the baselines have inconsistent performance throughout subjects, but PhysiQ outperforms them and consistently results well.

5.4.2 Classification. Next, we evaluate the PhysiQ's classification component on the quality of activity, i.e., how accurate can our model predict on *ROMs*. Thus, our baselines for the models are some of the best classification models that modern machine learning and deep learning can achieve: CNN with Multi-Layer Perceptrons (MLP),

Table 2. The overall evaluation on exercise **shoulder abduction** similarity comparison of *range of motion*, *stability*, and *repetition*.

	ROM				Stability				Repetition			
	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG
MSE	**0.00217**	0.0153	0.0138	0.0442	**0.00454**	0.0143	0.0149	0.0561	**0.00716**	0.0145	0.0180	0.106
MAE	**0.0310**	0.0927	0.0914	0.175	**0.0514**	0.0930	0.0966	0.184	**0.0690**	0.0941	0.102	0.272
R-Square	**0.949**	0.634	0.676	-0.0341	**0.791**	0.342	0.314	-0.0645	**0.869**	0.736	0.668	-0.931

Table 3. The overall evaluation on exercise **external rotation** similarity comparison of *range of motion*, *stability*, and *repetition*.

	ROM				Stability				Repetition			
	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG
MSE	**0.0117**	0.0174	0.0129	0.0192	**0.00166**	0.0121	0.00220	0.00291	**0.0105**	0.0277	0.0305	0.119
MAE	**0.0882**	0.0990	0.0937	0.122	**0.0250**	0.0537	0.0309	0.0339	**0.0749**	0.134	0.140	0.285
R-Square	**0.757**	-0.0404	0.226	0.0184	**0.423**	-3.412	0.217	-0.0236	**0.805**	0.488	0.437	-1.298

Table 4. The overall evaluation on exercise **forward flexion** similarity comparison of *range of motion*, *stability*, and *repetition*

	ROM				Stability				Repetition			
	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG	PhysiQ	SimCLR	RNN	VGG
MSE	**0.00215**	0.00963	0.00988	0.0454	**0.00602**	0.0176	0.0155	0.0545	**0.00276**	0.0121	0.0121	0.0597
MAE	**0.0366**	0.0771	0.0673	0.179	**0.0594**	0.103	0.0948	0.190	**0.0411**	0.0888	0.0888	0.200
R-Square	**0.950**	0.775	0.772	-0.0515	**0.883**	0.657	0.700	-0.0450	**0.9513**	0.785	0.785	-0.0472

LSTM with MLP, and Logistic Regression (Linear). Because classification and similarity comparison are two different problems, we target our problem using a different baseline but similar structure overall. CNN with MLP utilizes the 1-D Convolutional Neural Network to capture spatial knowledge and use the hidden representation to go through an MLP network to classify *ROMs* and *stability*. Similarly, LSTM with MLP has a similar approach, except a sliding windows segmentation is used to create a temporal representation, which feeds into the fully connected layer as MLP. Lastly, Logistic Regression is simply a fully connected model that considers all the points (dimensional flattening) in the signal and predicts the labels.

As a result, we provide confusion matrices for all baselines and our model for classification of *range of motion*. As shown in Figure 10a, 11a, and 12a, compared against the other 3 baselines, our model demonstrates the best performance overall with a minimal inaccuracy on higher degree *range of motion* exercises. The difficulties in differentiating between 30 and 60 degrees happen because of limited supervision and little self-reflection (such as a mirror), and all subjects might not perform homogeneously in the exercises. As a result, this might leads to inconsistency in the model to understand the difference between 30 degrees and 60 degrees ROM.

5.5 Parameters Evaluation

5.5.1 Sliding Windows. To evaluate the accuracy of our proper performance, we use different sliding window segmentation of samples to test our model results. We test the performance on leave-one-subject-out cross-validation (LOOCV). By having sliding windows that are 50, 100, and 150. The model decreased its performance when the sliding windows increased while keeping the same step size. Increasing the sliding window size decrease the number of time, or redundancy, for the model to see. Having smaller sliding windows helps the model to understand the quality of exercises. At the same time, as we increase the step size (the size of overlapping), the model also performs better but sees a drop in the testing dataset. An extensive redundancy overfits the training distribution and does not generalize it well. As a result, we choose a sliding window of 50 with a step size of 15
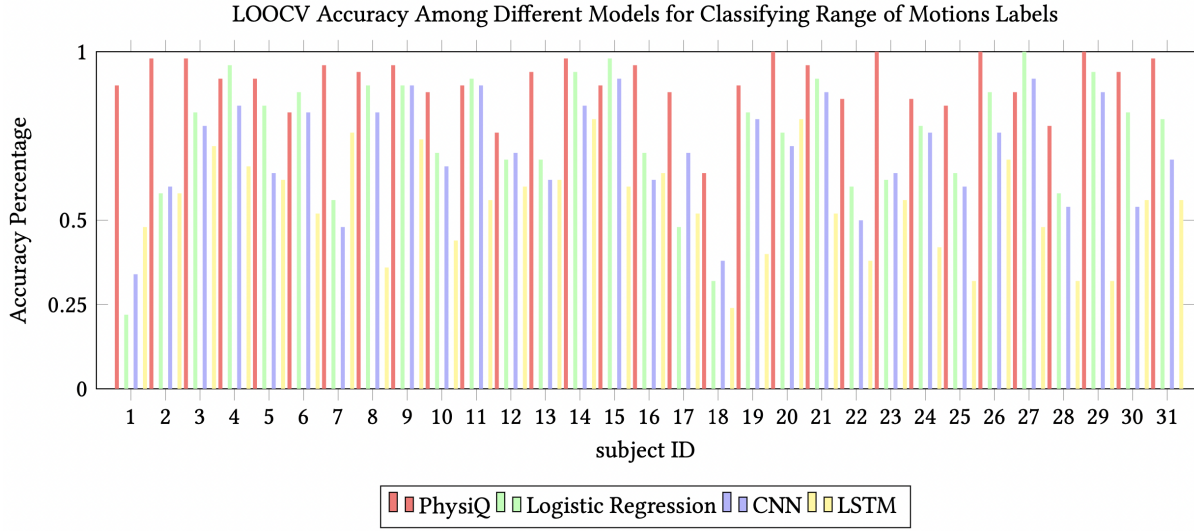
Fig. 9. This figure shows the accuracy of our model with output of classification of *range of motion*. As showed in this figure, our PhysiQ (shown in red) performs mostly better than the other models in this diagrams with a percentage accuracy on y-axis.
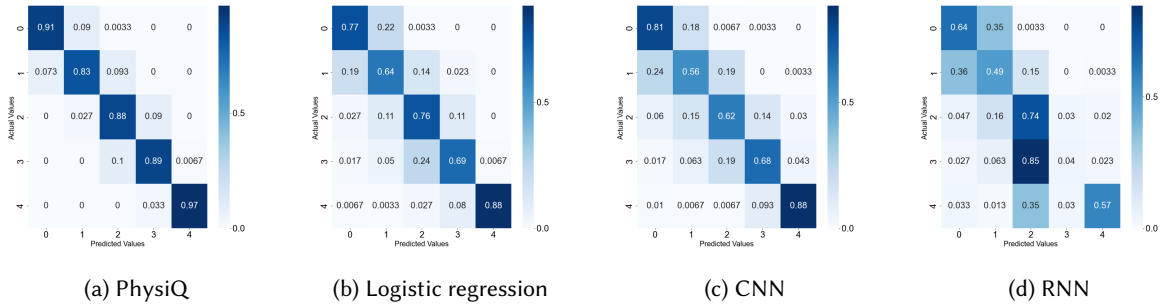

Fig. 10. This is the confusion matrices for classification of *range of motion* in **shoulder abduction** with PhysiQ and baselines. Shoulder abduction has a class of 5 from 30, 60, 90, 120, and 150, which is labeled from 0 to 4

for training, validating, and testing, which alike to our data collection rate of 50 Hz. Having some redundancy helps the model to generalize the relationship between sliding windows.

5.5.2 Repetition Size. In repetition size, we have a different size in similarity comparison. We have each ID number of segmented exercises for each subject. Adjacent ID number represents that they can be concatenated together to output different. We use the same number of repetition sizes in training and testing. The model has a slight difference in accuracy (3 - 5 percent) in bigger repetitions as length increases. However, this is also affected by the hyperparameters that we choose. Without changing the step size and window size, the model can still perform well because our model generalizes very well. Moreover, we also see a performance increase when
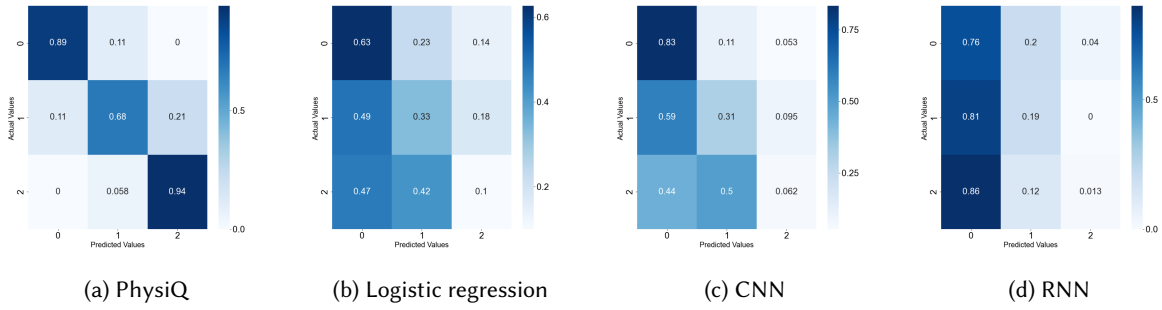
Fig. 11. This is the confusion matrices for classification of *range of motion* in **external rotation** with PhysiQ and baselines. The external rotation has a class of 3 from 45, 90, and 150, which is labeled from 0 to 2
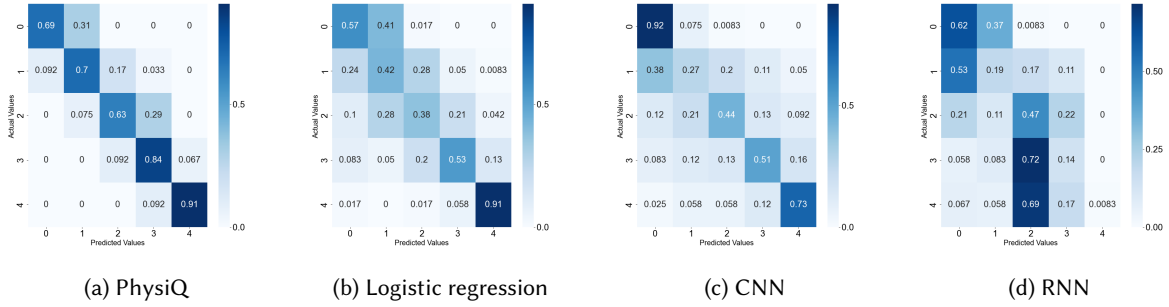

Fig. 12. This is the confusion matrices for classification of *range of motion* in **forward flexion** with PhysiQ and baselines. Forward flexion has a class of 5 from 30, 60, 90, 120, and 150, which is labeled from 0 to 4

adding dropout regularization deep LSTM network. As a result, the performance of four repetitions becomes 0.921 in the R-Squared correlation.

5.5.3 Padding. We also evaluate the effect of front padding and back padding. We hypothesize that there should be no difference between front and back padding. However, when the padding is at the end of the temporal sequence, the model does not learn, resulting in an average R-Squared score of 0. Moreover, the results remain unchanged throughout the epochs. This non-learning behavior happens possibly due to the fact that many of the segments are varied significantly. In that cases, many of the windows remain zero and do not help the model to learn. At the same time, as we attempt the model without padding or minimize the amount of padding, we are facing the issue that the temporal model could cheat the result because there is a correlation between higher *range of motion* exercises and the length of the data. In contrast, lower ROM exercises can have a shorter overall length. This relation could result in an accumulating effect of shortcuts for the deep learning model to learn and cheat. So using constant front padding for all one repetition exercises is consistent and performs similarly in a different number of repetitions.

5.6 Ablation Study

5.6.1 Removal of Hidden Representations. After removing the **temporal representation**, LSTM encoder, part of PhysiQ, the model still performs relatively well with an MSE 0.007168, MAE 0.0665, and R-Squared 0.8346 with 256 hidden features. At the same time, we did the same procedure for **spatial representation**. After removing the CNN encoder from PhysiQ, the model performs relatively well with an MSE 0.00873, MAE 0.063, and 0.8208 in R-Squared with one hidden temporal layer and no dropout regularization. Interestingly, in the usages of both spatial and temporal models, our model can achieve 0.92 R-Squared while removing either of them drops significantly by around 10 percent. The decrease happened due to the reasoning that the signal data contain both spatial and temporal information, and a single model can only capture part of the information with the help of an attention mechanism.

5.6.2 Dropout. Testing the dropout is also essential to see how the model is generalized. This dropout is applied in every layer, including the spatial encoder (CNN) and temporal encoder (LSTM). We used LOOCV to compare our results with 0%, 20%, and 50% dropout rates to see if the difference is significant. We test our model with hidden features of 256 for temporal and spatial encoding, two LSTM layers, 16 heads of attention, and a batch size of 1024. The model should not have a significant difference between 20% and 50% but possibly a slight performance boost in 20% or 50% compared to 0%. Our model is meant to generalize well across participants and should not overfit the training distribution; the average result of 0 percent dropout across the participants is 89.66%. The average result with 20 percent dropout is 89.73%. The average result with 50 percent dropout is 87.87%. Interestingly, the 20% dropout rate generates a slightly better results than 50%. This is because 20% has already created its maximal regularization for the model, and increasing the dropout rate does not benefit from regularization. On the other hand, a much higher dropout rate could result in a higher variance in the model, resulted degradation in performance. Having only a 20% dropout rate creating the maximum of regularization makes our model more convincing in terms of generalizability.

5.6.3 Attention. We test the importance of attention in our model. Attention is applied to selectively focus on more critical aspects of the input sequence. In our model, attention is a final layer to measure the relationship between each hidden state. With that in mind, we remove the attention layer to compare results in the metrics of the *range of motion* in shoulder abduction, external rotation, and forward flexion. Without the attention layer, the performance of LOOCV in shoulder abduction drops to a 0.892 R-squared correlation from 0.908. On exercise external rotation, the performance of LOOCV drops to a 0.560 R-squared correlation from 0.675. Meanwhile, the performance of LOOCV on forward flexion has an insignificant increase from 0.815 to 0.830. In summary, if the motion data is distinguishable and representative, the attention layer does not have a huge impact. However, when exercise has a limited motion, such as half-half arm span exercise external rotation, the attention layer is more impactful.

6 A SURVEY ON USER EXPERIENCE

To investigate PhysiQ's application in practice, we surveyed the participants who used our system for data collection. Users wear an Apple Watch with a connected iPhone with our PhysiQ apps installed during data collection. In the beginning, users have the exercise instruction on the phone to start. PhysiQ on the phone visualizes the signal to see the repetition quality and feedback to the users (as shown in Fig. 1).

After collecting the data, we distribute a questionnaire to all users. The questionnaire asks six questions on different scales. We designed the survey questions for three purposes. First, we get users' feedback on our current design (Q1). Secondly, we explore users' preferences in alternative platforms and recommendations to improve our current design (Q2 and Q3). Last, we gather users' demographic information (presented in Section 4) and ask behavioral questions (Q4-Q6). We study the correlation between users' behaviors and the performance of

our algorithms in measuring their activities. We attach the questions and summary of results below because we believe our findings are valuable for our future work and the research community.

- **Q1: How do you like the current feedback and recommendation system?**
  - scale of 1 to 5, with 1 being the lowest, and 5 being the highest
- **Q2: If we have a different platform, what platform will you like to have recommendation system of the exercises on?**
  - Smartwatch
  - Smartphone
  - Smartglasses such as VR, AR glass
- **Q3: During what period of the exercise do you like such a feedback? For example, for today's exercise session, you have 5 different exercises to perform and for each exercise you have 10 repetitions of 5 sets?**
  - After a set of exercise (during this particular exercise, after 10 repetition)
  - After the particular exercise of 5 sets
  - After the entire exercise session (after 5 exercises)
- **Q4: During the time of collecting your exercise data, do you drink coffee or any caffeinated drinks regularly? If so, how often?**
  - Every day
  - A few times a week
  - About once a week
  - A few times a month
  - Once a month
  - Less than once a month
- **Q5: During the time of collecting your exercise data, how many hours do you sleep?**
  - Time in hour
- **Q6: This question is regarding your medical history and you do not need to specify the medication. At the time of collecting your exercise data, do you know and take any medication at the time that would affect your ability to perform exercise or activity?**
  - Yes or no

Out of 31 participants, we have 27 participants who complete the follow-up survey. Overall, we have an average of 4.26 rating on how much the users like our system. Interestingly, we have 48.1% of the users who want to have a recommendation on a smartwatch, 51.9% of the users on a smartphone, and no one wants to use smart glass to get recommendation feedback. Additionally, 59.3% of our users want to have their feedback after a set of exercises, 22.2% after the particular exercises, and 14.8% after an entire session. Lastly, one participant suggests they should be able to see the feedback anytime they want.

Moreover, we have 18.5%, 33.3%, 14.8%, 18.5%, 3.7%, and 11.1%, respectively, on the frequency of coffee consumption based on the answer order above. At the same time, on average, our participants sleep 7.63 hours with a minimum of 6 and a maximum of 9. Lastly, we only have 1 participant possibly on medication that affected their performance overall, but we do not find that medication was affecting the performance of our model testing on this participants' exercises.

7 DISCUSSION

7.1 Applications

7.1.1 Physical Therapy Home Assessment. . This application is intended to work for patients and people injured, postoperative, or mentally traumatized. While performing at-home exercises, our application can provide real-time feedback and assess the quality of the exercises to provide better interaction and supervision while clinics are not accessible. At the same time, we believe our model is capable of analyzing and predicting people who might suffer from a different illness or potential injuries. One example is that people with heavy usage of handcrafting might suffer from carpal tunnel syndrome, which is caused by pressure on the median nerve.

7.1.2 Daily Exercises Assessment. . This application can also support working with people who enjoy exercise. People can benefit from it by closely assessing how they perform specific repetitions of exercises. Moreover, this application could also apply to people who are playing sports. We envision that our model can eventually support people who play sports like tennis to predict a player's direction, speed, or posture.

7.2 Assumption and Limitation

This paper discusses the potential of using a smartwatch to support users assess their exercises with a quantitative measure of the quality. However, there are a few assumptions made to support this application. First of all, we assume that the users' posture should have some level of decency. For example, suppose users perform the exercise of shoulder abduction while bending their neck or wiggling around as not in a straight form. In that case, the application likely can still give a good score on the exercises simply because the users might still perform the exercise correctly. Because the smartwatch cannot capture all the movement within the body, we do not have the luxury of analyzing all the posture. Secondly, we assume that the primary segmentation is performed upon accelerometer data, and its energy is only extracted from the accelerometer. Thirdly, we assume there is 5 level of quality of exercises in *range of motion* in shoulder abduction. These are our metrics to digitalize from body movement to computational numbers. The five levels can vary based on different professionals' metrics, and our goal is to standardize a metric in our model that can measure and understand.

We notice that the energy plot does not have a clear pattern in some exercises for some subjects. Such limitation is due to our little knowledge regarding the data collection. Applying our energy plot when collecting data from the users could be more rigorous and use it as a checkpoint to verify if the subjects have correctly performed the exercises. In the future work, we will improve our framework to better handle above assumptions and limitations.

8 SUMMARY

We develop an innovative system, PhysiQ, to quantitatively digitalize the quality of exercises through new metrics on a commodity smartwatch. We scrutinize and verify that different metrics and exercises have unique characteristics that can be recognized and understood by our deep learning model. By developing such a model based on Siamese Neural Network with additional spatiotemporal representation encoding, our model can achieve 95 percent R-square correlation and 90 percent accuracy in classification. Moreover, a comprehensive evaluation and user studies are performed to show the effects on our metrics in range of motion, stability, and repetition. The end goal is to improve the prediction and assessment of people who needs therapy to improve their quality of life. In addition, we envision that current technologies and relevant professions can be benefited from it by expanding usability, generalizability, and model learnability. By combining deep learning and the physical therapy method, we believe that our framework is the tool to lead people to improve their quality of life.

ACKNOWLEDGMENTS

REFERENCES

[1] KA Agrawal, Delenn Chin, and Kevin Chen. 2017. Cosine siamese models for stance detection. *tech. rep.* (2017).

[2] Yassine Benajiba, Jin Sun, Yong Zhang, Longquan Jiang, Zhiliang Weng, and Or Biran. 2019. Siamese networks for semantic pattern similarity. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 191–194.

[3] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[4] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. 2020. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1342–1350.

[5] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. 2018. Auracle: Detecting eating episodes with an ear-mounted sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–27.

[6] Liqiong Chang, Jiaqi Lu, Ju Wang, Xiaojiang Chen, Dingyi Fang, Zhanyong Tang, Petteri Nurmi, and Zheng Wang. 2018. SleepGuard: Capturing rich sleep information using smartwatch sensing data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–34.

[7] Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. 2019. Multi-agent attentional activity recognition. *arXiv preprint arXiv:1905.08948* (2019).

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]* (June 2020). http://arxiv.org/abs/2002.05709 arXiv: 2002.05709.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.

[11] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[12] Antonino Crivello, Paolo Barsocchi, Michele Girolami, and Filippo Palumbo. 2019. The meaning of sleep quality: a survey of available technologies. *IEEE access* 7 (2019), 167374–167390.

[13] Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 459–474.

[14] Rosie Essery, Adam WA Geraghty, Sarah Kirby, and Lucy Yardley. 2017. Predictors of adherence to home-based physical therapies: a systematic review. *Disability and rehabilitation* 39, 6 (2017), 519–534.

[15] Ameya Govalkar and Kiran George. 2021. Siamese Network Based Pulse and Signal Attribute Identification. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 0176–0179.

[16] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. 2017. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE international conference on computer vision*. 1763–1771.

[17] Reza Haghighi Osgouei, David Soulsby, and Fernando Bello. 2020. Rehabilitation Exergames: Use of Motion Sensing and Machine Learning to Quantify Exercise Performance in Healthy Volunteers. *JMIR Rehabil Assist Technol* 7, 2 (18 Aug 2020), e17289. https://doi.org/10.2196/17289

[18] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. 2018. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4834–4843.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[21] Ghulam Hussain, Bander Ali Saleh Al-rimy, Saddam Hussain, Abdullah M Albarrak, Sultan Noman Qasem, and Zeeshan Ali. 2022. Smart Piezoelectric-Based Wearable System for Calorie Intake Estimation Using Machine Learning. *Applied Sciences* 12, 12 (2022), 6135.

[22] Jun-Hyeok Jang and Jae-Seop Oh. 2014. Changes in shoulder external rotator muscle activity during shoulder external rotation in various arm positions in the sagittal plane. *Journal of Physical Therapy Science* 26, 1 (2014), 135–137.

[23] Tiago S Jesus, Michel D Landry, and Helen Hoenig. 2019. Global need for physical rehabilitation: systematic analysis from the global burden of disease study 2017. *International journal of environmental research and public health* 16, 6 (2019), 980.

[24] Rinat Khusainov, Djamel Azzi, Ifeyinwa Achumba, and Sebastian Bersch. 2013. Real-Time Human Ambulation, Activity, and Physiological Monitoring: Taxonomy of Issues, Techniques, Applications, Challenges and Limitations. *Sensors* 13, 10 (Sept. 2013), 12852–12902. https://doi.org/10.3390/s131012852

[25] Rinat Khusainov, Djamel Azzi, Ifeyinwa E Achumba, and Sebastian D Bersch. 2013. Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations. *Sensors* 13, 10 (2013), 12852–12902.

[26] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[27] Divina Lawrance and Suja Palaniswamy. 2021. Emotion recognition from facial expressions for 3D videos using siamese network. In *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, Vol. 1. IEEE, 1–6.

[28] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. 2016. Learning by tracking: Siamese CNN for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 33–40.

[29] Fan Li, Rui Yan, Reza Mahini, Lai Wei, Zhiqiang Wang, Klaus Mathiak, Rong Liu, and Fengyu Cong. 2021. End-to-end sleep staging using convolutional neural network in raw single-channel EEG. *Biomedical Signal Processing and Control* 63 (2021), 102203.

[30] Meiyi Ma, Ridwan Alam, Brooke Bell, Kayla de la Haye, Donna Spruijt-Metz, John Lach, and John Stankovic. 2017. Mˆ2G: a monitor of monitoring systems with ground truth validation features for research-oriented residential applications. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 10–18.

[31] Milad Asgari Mehrabadi, Iman Azimi, Fatemeh Sarhaddi, Anna Axelin, Hannakaisa Niela-Vilén, Saana Myllyntausta, Sari Stenholm, Nikil Dutt, Pasi Liljeberg, Amir M Rahmani, et al. 2020. Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. *JMIR mHealth and uHealth* 8, 11 (2020), e20465.

[32] Gabriel Mittag and Sebastian Möller. 2020. Full-reference speech quality estimation with attentional siamese neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 346–350.

[33] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[34] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM international symposium on wearable computers*. 100–103.

[35] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. 148–157.

[36] Nikolay Neshov, Agata Manolova, Krasimir Tonchev, and Ognian Boumbarov. 2019. Detection and Analysis of Periodic Actions for Context-Aware Human Centric Cyber Physical System to Enable Adaptive Occupational Therapy. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Vol. 2. 685–690. https://doi.org/10.1109/IDAACS.2019.8924300

[37] Ioannis Papavasileiou, Zhi Qiao, Chenyu Zhang, Wenlong Zhang, Jinbo Bi, and Song Han. 2021. GaitCode: Gait-based continuous authentication using multimodal learning and wearable sensors. *Smart Health* 19 (2021), 100162.

[38] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–16.

[39] Ghanashyama Prabhu, Noel E. O'Connor, and Kieran Moran. 2021. A Deep Learning Model for Exercise-Based Rehabilitation Using Multi-channel Time-Series Data from a Single Wearable Sensor. https://doi.org/10.1007/978-3-030-70569-5_7

[40] Meera Radhakrishnan and Archan Misra. 2019. Can earables support effective user engagement during weight-based gym exercises?. In *Proceedings of the 1st International Workshop on Earable Computing*. 42–47.

[41] Meera Radhakrishnan, Archan Misra, and Rajesh K Balan. 2021. W8-Scope: Fine-grained, practical monitoring of weight stack-based exercises. *Pervasive and Mobile Computing* 75 (2021), 101418.

[42] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.

[43] Rafael Salazar. 2019. 2019 survey results: Outpatient PT & OT Clinicians & Clinic owners. https://rehabupracticesolutions.com/2019-survey/

[44] Wilhelm Daniel Scherz, Daniel Fritz, Oana Ramona Velicu, Ralf Seepold, and Natividad Martínez Madrid. 2017. Heart rate spectrum analysis for sleep quality detection. *EURASIP Journal on Embedded Systems* 2017, 1 (2017), 1–5.

[45] Jianbing Shen, Xin Tang, Xingping Dong, and Ling Shao. 2019. Visual object tracking by hierarchical attention siamese network. *IEEE transactions on cybernetics* 50, 7 (2019), 3068–3080.

[46] John A Stankovic, Meiyi Ma, Sarah Masud Preum, and Homa Alemzadeh. 2021. Challenges and Directions for Ambient Intelligence: A Cyber Physical Systems Perspective. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 232–241.

[47] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[48] Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. 2017. Sleepmonitor: Monitoring respiratory rate and body position during sleep using smartwatch. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–22.

[49] KG Van Leeuwen, H Sun, M Tabaeizadeh, AF Struck, MJAM Van Putten, and MB Westover. 2019. Detecting abnormal electroencephalograms using deep convolutional networks. *Clinical neurophysiology* 130, 1 (2019), 77–84.

[50] Fei Wang, Xilei Wu, Xin Wang, Jianlei Chi, Jingang Shi, and Dong Huang. 2021. You can wash better: Daily handwashing assessment with smartwatches. *arXiv preprint arXiv:2112.06657* (2021).

[51] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4854–4863.

[52] Wikipedia. 2021. Serratus Anterior Muscle. https://en.wikipedia.org/wiki/Serratus_anterior_muscle.

[53] Wikipedia. 2022. Deltoid Muscle. https://en.wikipedia.org/wiki/Deltoid_muscle.

[54] Wikipedia. 2022. Shoulder. https://en.wikipedia.org/wiki/Shoulder.

[55] Wikipedia. 2022. Trapezius. https://en.wikipedia.org/wiki/Trapezius.

[56] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit+: Personalized Fitness Monitoring via Audio Signals on Smart Speakers. *IEEE Transactions on Mobile Computing* (2021).

[57] Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi, and Fengyu Cong. 2019. An automatic sleep scoring toolbox: multi-modality of polysomnography signals' processing. In *International Conference on Signal Processing and Multimedia Applications*. SCITEPRESS Science And Technology Publications.

[58] Zhipeng Zhang and Houwen Peng. 2019. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4591–4600.