Physical Fatigue Detection through EMG wearables and Subjective User Reports - A Machine Learning Approach Towards Adaptive Rehabilitation

Michalis Papakostas
michalis.papakostas@mavs.uta.edu
Heracleia Lab
University of Texas at Arlington
Arlington, Texas

Varun Kanal
varun.kanal@mavs.uta.edu
Heracleia Lab
University of Texas at Arlington
Arlington, Texas

Maher Abujelala
maher.abujelala@mavs.uta.edu
Heracleia Lab
University of Texas at Arlington
Arlington, Texas

Konstantinos Tsiakas
konstantinos.tsiakas@yale.edu
Department of Psychiatry
Yale University
Hew Haven, Connecticut

Fillia Makedon
makedon@uta.edu
Heracleia Lab
University of Texas at Arlington
Arlington, Texas

ABSTRACT

Physical fatigue due to muscle exhaustion is a symptom that can be very common in daily life. However fatigue can sometimes be suspect of more severe diseases such as multiple sclerosis and needs to be assessed appropriately. Despite the need to monitor fatigue, describing it in an objective and quantifiable manner is still an open problem due to the great levels of subjectivity involved. In this work we propose a novel method towards detecting physical fatigue. We design our approach based on objective EMG measurements and we aim to identify the presence of physical fatigue based on subjective user-reports. Based on our analysis we highlight the significance of our findings and we discuss how machine learning based modeling can become useful towards understanding fatigue and designing adaptive rehabilitation scenarios.

CCS CONCEPTS

• **Human-centered computing** → **User models**; **User studies**; **HCI theory, concepts and models**; **Ubiquitous computing**; *Interaction techniques.*

KEYWORDS

physical fatigue, EMG, physiological monitoring, user modeling, machine learning, dataset

ACM Reference Format:
Michalis Papakostas, Varun Kanal, Maher Abujelala, Konstantinos Tsiakas, and Fillia Makedon. 2019. Physical Fatigue Detection through EMG wearables and Subjective User Reports - A Machine Learning Approach Towards Adaptive Rehabilitation. In *The 12th PErvasive Technologies Related to Assistive Environments Conference (PETRA '19), June 5–7, 2019, Rhodes, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3316782.3322772

PETRA '19, June 5–7, 2019, Rhodes, Greece

ACM ISBN 978-1-4503-6232-0/19/06…$15.00
https://doi.org/10.1145/3316782.3322772

1 INTRODUCTION

Physical fatigue is one of the most common symptoms across a great variety of medical conditions, ranging from stroke and multiple sclerosis to chronic insomnia and myoskeletal injuries[2]. However, understanding, quantifying and predicting events of fatigue is a topic that remains vastly unexplored, primarily due to the subjective nature of the term. Each individual experiences fatigue in a very personal way that varies on its intensity and is affected not only by someone's physiological state but also by subjective factors such as emotion, which are very difficult to detect with certainty [12]. Our inability to capture and predict such events efficiently, can lead to negative outcomes when it comes to physical rehabilitation since it increases the chance of causing unwanted injuries and muscular exhaustion. This realization becomes even more important when it comes to autonomous rehabilitation systems and the need to design adaptive systems that match user's skills. Under that scope, understanding physical fatigue has become an area that attracts great research interest due to it's importance towards achieving effective rehabilitation [3]. During the last twenty years, numerous works have been published that proposed modeling-methods and features to capture meaningful information from EMG[6, 7]. However there is still no general truth on what is the most efficient way to model such signals [8].

In this work we present an extensive analysis and evaluation of different machine learning algorithms towards predicting physical fatigue, on a human-robot rehabilitation scenario. We exploit statistical features that have been traditionally used in EMG and/or audio analysis and we propose a post-processing method that significantly improves the results provided by the original models. We use Delsys, a non-intrusive wearable EMG sensor and we build ML models targeting user-reported labels on physical fatigue. Our analysis focuses on evaluating the robustness and generalizability of such models across different users and exercises. Due to its computational simplicity our method is ideal for real-time scenarios.

The code and the data used for this work can be downloaded for free for purposes of reproducibility and further experimentation[1].

2 DATA COLLECTION & EXPERIMENTAL SETUP

A study was conducted that involved 10 male and female subjects with a mean age of 26.3 years old. The subjects were asked to perform 3 exercises; shoulder flexion (SF), shoulder abduction (SA) and elbow extension (EE) Figure-1. These exercises were performed using the Barrett WAM arm, which is capable of applying feedback forces to the subject. The subjects were asked to hold the end-effector of the arm while performing each exercise. Two positions were important in each exercise; start position and the end position. For each exercise, the subjects would start from the start position and move the end-effector to the end position. The subjects were asked to hold the end-effector at that location so as to induce isometric contraction in the muscle. During this process, the robotic arm would provide resistive forces to the subject. EMG data were collected from the major muscles responsible for the movement. In SF and SA, EMG data were recorded from the deltoid and in EE from the triceps. The subjects were asked to hold the end-effector until they start to feel fatigued. When that occurred, they would inform the researcher conducting the experiment who would mark the time point. After the pass of almost 10 sec of the time the subject reported fatigue, the researcher would ask them to go back to the start position to complete the the exercise. Subjects were asked to perform 3 repetitions of each exercise. A short period of rest was provided between each exercise to mitigate the cascading effect of fatigue. In total we collected 10 users × 3 exercises × 3 repetitions = 90 EMG recordings.

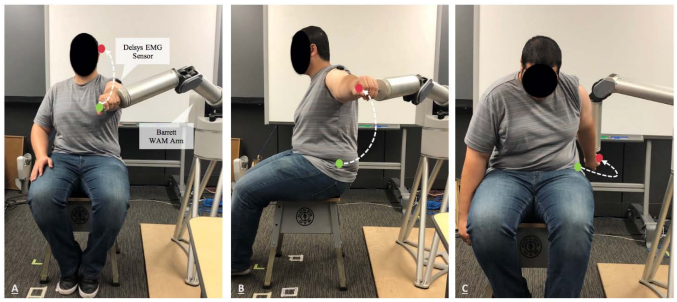

Figure 1: A- Shoulder Flexion B- Shoulder Abduction C- Elbow Extension. The green circles indicate the start positions and the red circles indicate the end positions

3 METHODOLOGY

The Delsys EMG wearable sensors, provided a sampling frequency of 1926HZ. As a first step and in order to reduce the inherent noise of the EMG recordings we filtered the signal using the median filtering technique with a window size of 11 samples. Using those filtered signals as input to our algorithm, short-term features were extracted from the time and spectral domain, which are then remodeled in a mid-term fashion. The final feature vectors extracted from the mid-term windows were used as input samples to the classification algorithms. Targeted labels were the user-reported, binary indications of fatigue (0 meaning no-fatigue and 1 meaning fatigue). Thus, a valid set of labels as provided by the user would have the following form: [0,0,0,0,0,0,...,1,1,1,1,1,1,1,1,1], where each label corresponds in a sample captured by the EMG sensors (ie. 1926 labels per second).

[1]https://github.com/MikeMpapa/MLEmg_Monitoring_Physical_Fatigue

3.1 Signal-Prepossessing

For splitting the EMG signals into short and mid-term windows, empirical window sizes were used, based on the fact that muscle fatigue changes are observed relatively slow. Short-term non-overlapping windows were extracted with a length of 0.25 sec, while overlapping mid-term windows were extracted capturing 2 sec of EMG information with a window step of 1s.

3.2 Feature Extraction

As explained in Section-3, a two-step feature extraction process was held in order to model the raw EMG information. Firstly a descriptive set of short-term features was extracted from each short-term window and then based on those features a set of statistical mid-term features was extracted to create the final feature vectors (FVs).

Based on extensive literature review on handcrafted feature extraction for effective EMG signal representation [6, 8], for every 0.25 sec short-term window we extracted the following list of features:

(1) **Minimum, Maximum, Standard Deviation & Mean** values of the time domain in a specific frame
(2) **Spectral Minimum, Maximum, Standard Deviation & Mean**
(3) **Spectral Entropy**:

$$H(X) = -\sum_{i=1}^{N} p(PS) log_{10} p(PS) \quad (1)$$

,where $PS = \frac{1}{N}|X|^2$ is the Power Spectral Density (PSD) of of spectrum X, $p(PS) = \frac{PS_i}{\sum_i PS_i}$ is the Probability Density Function of the PSD, N is the size of the spectrum and X are the observed frequencies. H(X) is the normalized spectral energies for a set of sub-frames.

(4) **Spectral Flux**:

$$FL_{i,i-1} = \sum (EN_i - EN_{i-1})^2 \quad (2)$$

, where $EN_i = \frac{X_i}{\sum X_i}$ is the normalized Discrete Fourier Coefficient at the $i_t h$ frame and X is the spectral of the signal. Spectral flux is the squared difference between the normalized magnitudes of the spectra of two successive frames.

(5) **Zero Crossing Rate** : The rate of sign-changes of the signal during the duration of a particular frame.

$$ZCR = \frac{1}{N-1} \sum_{t=1}^{N-1} (sig(x_t) - sig(x_{t-1})) \quad (3)$$

,where $sig(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{otherwise} \end{cases}$ and N is the length of the signal. ZCR indicates the rate of sign-changes of the signal during the duration of a particular frame.

(6) **Energy Entropy**:

$$H(E) = -\sum_{i=1}^{N} p(E) log_{10} p(E) \tag{4}$$

,where $E = \frac{\sum x_i^2}{Total\_E}$ are the normalised sub-frame energies, $Total\_E = \sum x_i^2$ is the total signal energy and x_i is the value of a sample within a frame or a sub-frame. This feature can be interpreted as a measure of abrupt changes.

(7) **Willson Amplitude (WAMP)**: The number of times that the difference between two consecutive amplitudes in a time segment becomes more than threshold. WAMP can be seen as an indication of muscle contraction levels.

$$WAMP = \sum_{t=1}^{N-1} f(x_i - x_{i+1}) \tag{5}$$

,where $f(x) = \begin{cases} 1 & \text{if } x > threshold \\ 0 & \text{otherwise} \end{cases}$ and N is the length of the signal. This feature is an indicator of firing motor unit action potentials (MUAP) and therefore an indicator of the muscle contraction level. [5]

These features are in general known for their ability to describe core characteristics of 1-D signals such as Accelerometer axis-based analysis or audio modeling and have been proven quite informative in the past for the specific purposes of EMG classification. Especially features stemming from the time domain such as Zero-Crossing Rate, Energy Entropy and WAMP amplitude have shown great potentials for capturing EMG based patterns. However in depth analysis of EMG feature selection is out of the scope of this paper and feature selection was mainly inspired based on the related literature and our experimental analysis.

At every step, in addition to the features extracted from the current short-term window we compute the deltas between the present set of features and the set of features extracted from its preceding window. Thus, describing each short-term frame with a set of 26 values (13 features from the current window plus 13 deltas).

For the mid-term window extraction, each set of 8 successive short-term FVs is described using the minimum, maximum, standard deviation and mean information extracted for each short-term feature. Hence, producing a final feature vector of 4*26 = 104 values.

During our experimentation other features were also evaluated like signal energy, spectral spread (ie. the second central moment of the spectrum), spectral rolloff (ie. the frequency below which 90% of the magnitude distribution of the spectrum is concentrated) and spectral centroid (the center of gravity of the spectrum). However they were omitted from our final evaluation since they didn't seem to have a significant effect on the final outcome.

3.3 Classification

For classification purposes we experimented with a set of five traditional ML algorithms that have been extensively used for signal processing and EMG modeling in particular [6, 7, 9]. More specifically we evaluated the performance of the following methods: *Linear SVM, SVM with an RBF Kernel, Gradient-Boosting (GB), Extra-Trees (ET) and Random Forests (RF).*

3.4 Post-Processing

Our initial experimentation indicated that the original methods were usually failing to correlate the EMG information to the actual labels provided by the users, as they were often achieving an Average F1 lower than 70% and in many cases just slightly higher than 50% (ie. very close to random choice). Keeping in mind that classification takes place in a mid-term window basis, this made it impossible to consistently track fatigue in a long-term sequence as the algorithm would produce labels that were very hard to interpret. For example assuming again that 0 indicates 'NO-FATIGUE' and 1 indicates 'Fatigue' a possible output sequence would look like ['0,1,1,0,0,1,0,...,0,1,1,1,0,1,0]. Thus, we developed our own post-processing method that builds upon the decisions of the initial classifiers and re-evaluates their decisions by keeping track of the N past mid-term labels assigned by the model.

In particular, as a first step we apply a median-filter of size K to the original predictions made by the classifier. Then the method gathers the successive assigned labels into groups of M. If in the N past groups, the total number of samples that have been identified as 'FATIGUE' exceeds a specific threshold, then and only then the method decides that the subject has shown signs of fatigue. Otherwise it assumes that the classification algorithm found a set of false positives and the process continues as if the subject has not been fatigued. Using this kind of post-processing the final output of our method has the following form ['0,0,0,0,0,0,0,...,1,1,1,1,1,1,1'] and provides significantly higher classification performance in all cases, as we will discuss in Section-4.

In Algorithm-1 we show the pseudo-code of the proposed post-processing technique and in Algorithm-2 we show the whole fatigue detection framework again in the form of pseudo-code. Figure-2 illustrates the overall system architecture.

In our implementation, hyper-parameters were set to $K1 = 3$, $M = 3$, $STEP = 1$, $N = 2$, $THRESH\_VAL = 0.6$ and $K2 = 11$. For reproducibility and reusability purposes our code along with the data used for the purposes of this study can be found and downloaded for free at github[2]. The hyper-parameters of each classifier were tuned using an exhaustive grid-search approach. It has to be noted that in terms of time delay's the algorithm makes a decision equal to the step-size of the mid-term frame (1s in our implementation), with only exception its first decision that takes place 2 sec after the recording has started.

4 EXPERIMENTAL RESULTS

To examine the the robustness of the proposed method in capturing subjective-fatigue, we perform 4 different types experiments. In all our experiments we evaluate our method in terms of Precision

[2]https://github.com/MikeMpapa/MLEmg_Monitoring_Physical_Fatigue

Algorithm 1 EMG Post-Processing Algorithm

$filtered\_labels = median\_filter(original\_predictions, K1)$
$group\_size = M$
$group\_step = STEP$
$thresh = THRESH\_VAL$
$prev\_window\_1, ..., prev\_window\_N = None$
$x1 = 0$
$x2 = group\_size$
while true **do**
  $current\_window = filtered\_labels[x1 : x2]$
  $t1 \leftarrow \frac{(current\_window='FATIGUE')}{group\_size} >= thresh$
  ...
  $tn \leftarrow \frac{(prev\_window\_N='FATIGUE')}{group\_size} >= thresh$
  if $(t1 \ \& \ ... \ \& \ tn) == TRUE$ **then**
    $state =' FATIGUE'$
    return $state$
  $prev\_window\_1 = current\_window$
  ...
  $prev\_window\_N = prev\_window\_(N-1)$
  $x1 = x1 + group\_step$
  $x2 = x2 + group\_step$

Algorithm 2 Fatigue Detection Framework

$filtered\_signal = median\_filter(original\_signal, K2)$
$st\_features \leftarrow st\_feature\_extraction(RAW\_EMG)$
$mt\_FV \leftarrow mt\_feature\_extraction(st\_features)$
$Prediction \leftarrow Classifier(mt\_FV)$
$Fatigue\_Prediction \leftarrow Algortithm1(Prediction)$

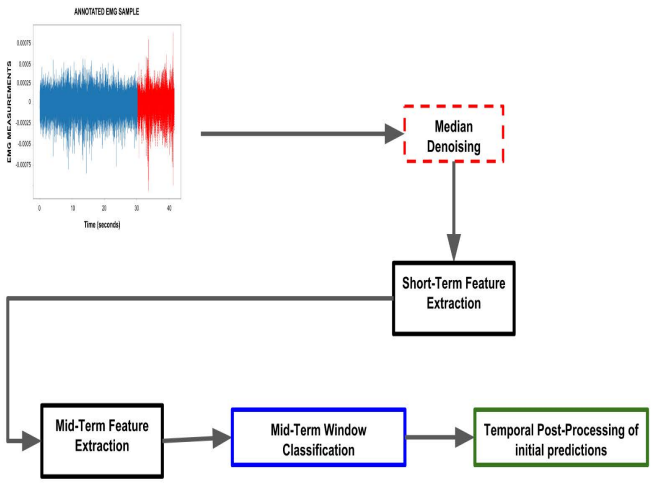

Figure 2: The overall system architecture. Blue and red EMG values correspond to NO-FATIGUE and FATIGUE ground truth labels respectively.

(Pr), Recall (Rec) and F1 measures. In all the following results 'NF' indicates the 'NO-FATIGUE' label and' F' corresponds to 'FATIGUE'.

4.1 Cross-User Evaluation (E1)

In the first experimentation we perform a *Cross-User Evaluation*. A leave-one-out cross-validation technique was applied with respect to the different users. At each step, all recordings from 9 users were used for training and all recordings of the remaining user were used for testing. We performed this process 10 times, each time using a different user for testing (Table-1). Using the proposed post-processing method, significantly improved the final results in terms of Average F1 in all cases. *Original Classification results in terms of Average F1 were: SVM = 60.6, SVM_RBF=55.4, GB=57.5, ET=54 and RF=53.7.* GB provides the best results in this evaluation, which are however directly comparable to the results provided by the SVM classifier.

Table 1: Average Performance Results on Cross-User Evaluation

	SVM	SVM-RBF	GB	ET	RF
Pr NF	70.2	75.6	64.4	62.3	58.4
Pr F	70.8	66.1	76.6	66	62.8
Rec NF	56.8	40.7	73	48.7	40.8
Rec F	81.3	89.8	68.7	77.1	77.5
F1 NF	62.8	52.9	68.4	54.6	48
F1 F	75.7	76.2	72.4	71.1	69.3
AVG F1	69.2	64.5	**70.4**	62.9	58.7

4.2 Cross-Exercise Evaluation (E2)

In the second experimentation we aimed to evaluate robustness of the proposed method across different exercises. Similarly as before a leave-one-out cross-validation was performed, but now in terms of different exercises. Thus, all samples, from all users that belong to single exercise were used for testing and all the rest were used for training. We repeated this process 3 times and we averaged the final results (Table-2). *As in the case of cross-user evaluation, post-processing significantly improved initial classification results, where the Average F1 was: SVM = 58.6, SVM_RBF=54.2, GB=54.4, ET=53 and RF=53.7.* In this scenario SVM provides by far the best classification results.

Table 2: Average Performance Results on Cross-Exercise Evaluation

	SVM	SVM-RBF	GB	ET	RF
					59.1
Pr NF	63.4	56.9	63.4	60.5	59.1
Pr F	74.3	67	69.4	68.2	67.4
Rec NF	65.5	58.3	57.8	57.4	56.5
Rec F	77.6	65.8	74.1	71	69.7
F1 NF	67.4	57.6	60.5	58.9	57.8
F1 F	75.4	66.4	71.6	69.6	68.5
AVG F1	**71.7**	62	66.1	64.2	63.1

4.3 Single User Evaluation (E3)

In this scenario we perform 10 different evaluations, each time using only the recordings that belong to a single user (ie. 3 exercises × 3 repetitions = 9 recordings). For each user the evaluation process was the same as before, where we ran the classification algorithm 9 times, each time using 8 recordings as training and the remaining as testing. Table-3 shows the averaged results across all 10, user-based evaluations. Again after post-processing the initial prediction, results were significantly improved in terms of Average F1. *Initial classification performance was: SVM = 65.2, SVM_RBF=68.5, GB=70.1, ET=72.4 and RF=71.5.*. Here in contrast to previous evaluations ET and SVM_RBF provide slightly better results than the other three classification methods.

Table 3: Average Performance Results on Single-User Evaluation

	SVM	SVM-RBF	GB	ET	RF
Pr NF	75.4	77.5	72.3	73.5	71.2
Pr F	76.2	78.5	77.9	82	79.6
Rec NF	66.6	70.3	71.1	77.8	74.7
Rec F	83.2	84.2	78.9	78.2	76.7
F1 NF	70.7	73.7	71.7	75.6	72.9
F1 F	79.6	82.3	78.4	80	78.1
AVG F1	75.1	77.5	75.1	**77.8**	75.5

4.4 Single Exercise Evaluation (E4)

Our final evaluation targets to recognise fatigue based only on samples that belong to a single exercise. Similarly as in the previous case we performed 3 different evaluations, one for each exercise. In each evaluation we used all recordings from all users that belong to the specific exercise (ie. 10 users × 3 repetitions = 30 recordings). For each evaluation we followed again a leave-one-out approach where we used 29 recordings as training and 1 as testing and we repeated the process 30 different times for each exercise (each time using a different recording for testing). Table-4 shows the averaged results across all 3 exercise-based evaluations. Post-processing results are again superior compared to the initial classifications in terms of Average F1. *Original classifier outputs without post-processing where: SVM = 61.5, SVM_RBF=65, GB=67.2, ET=68.5 and RF=66.8.*. Here all methods provide comparable results, but GB slightly outperforms the rest.

4.5 Overall Classification Improvement

In Figure-3 we show the % improvement of the classification results for each evaluation scenario after applying the post-processing temporal modeling method described by Algorithm-1. As the results indicate after the application of Algorithm-1, classification results showed significant improvements for all tested classifiers. Maximum improvement with a magnitude of 13.1% is shown for the SVM classifier for the E2 evaluation scenario while minimum improvement was 4% for the RF classifier for the E3 evaluation.

Table 4: Average Performance Results on Single-Exercise Evaluation

	SVM	SVM-RBF	GB	ET	RF
Pr NF	86.2	76.1	78.4	75.3	76.6
Pr F	71.2	76.3	76.2	74.7	76.3
Rec NF	56.5	66.5	68.2	63.4	66.1
Rec F	92.3	83.8	85.7	83.9	84.3
F1 NF	68.3	71	72.2	68.9	71
F1 F	80.4	79.9	81.1	79	80.1
AVG F1	74.3	75.4	**76.6**	74	75.5

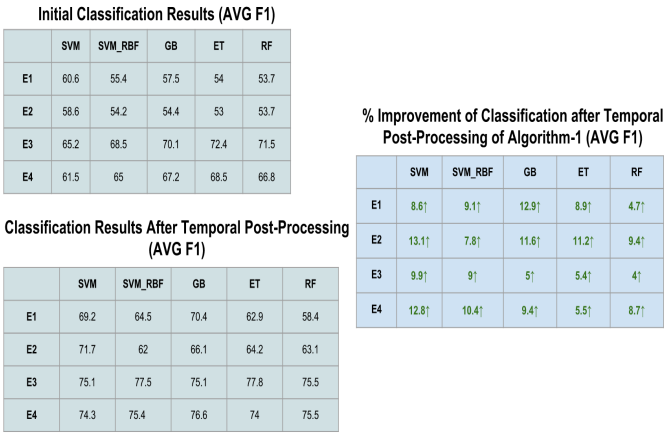

Figure 3: % Classification Improvement in terms of Average F1 after applying the temporal post-processing method described in Algorithm-1

4.6 Temporal Evaluation

To evaluate the efficiency of the proposed method in terms of temporal accuracy we performed a Paired Two One-Sided (TOST) equivalence test. Equivalence statistical tests aim to validate the fact that a difference between two sets lies within a given interval. TOST is based on the classical t-test used to test the hypothesis of equality between two means. In particular TOST performs two types of t-tests; one to verify if the difference is below a higher threshold and a second one that evaluates if the difference is higher than a lower threshold. In this work, the difference was calculated as $T_{reported\_fatigue} - T_{detected\_fatigue}$. The threshold, which was calculated through trial and error, signifies the maximum time difference between the subject feeling fatigued and the system recognising the event. There were two thresholds considered; the lower threshold indicates delayed detection while the higher signifies early detection. Table-5 illustrates the TOST results.

Despite the fact that temporal evaluation for each exercise has been reported individually, it has to be noted that no safe results can be drawn due to limited sample availability. Only the last column (*Avg Performance*) can be considered for safe evaluation purposes. The rest of the results are mainly reported for informational purposes since they can help draw useful insights. For the temporal

evaluation of the method we used only the classifiers that provided the best performance in terms of average F1 for each of the four evaluations (E1-4).

According to the results of Table-5 building models based on multiple users (E1) or on samples related to a single exercise (E4) were the most effective ones both in terms of *Success Rate* (ie. percentage of exercise sessions that the system successfully detected fatigue) and temporal accuracy. In other words these models were able to generalise their results better compared to the rest.

In the case of E1 the algorithm must have been able to depict the most generic of patterns that can eventually apply in the majority of users. However, judging from the classification results provided in Section-4.1 in most cases the model must have been making the wrong predictions almost 30% of the times, which indicates that in most scenarios the algorithm was at the limits of its temporal boundaries (ie. predictions where usually ±5 sec off).

On the other hand, in the case of E4 the model must have been able to capture similar behaviors across different users, when performing the same exercise. Along with the good performance reported in Section-4.4, such exercise-based models seem to be the best choices towards modeling physical fatigue using subjective reports, especially on users with similar physical characteristics.

In the other two training scenarios (E2 and E3) even-though temporal boundaries were relatively low (±6 sec), Success Rate was comparably low in both cases. For models based on E2 this comes along with the findings of Section-4.2 and indicates that such methods are in general unable to generalise across different exercises. According to these findings integrating exercise characteristics is very critical towards designing robust models, since they remain constant and must be followed by all users in the same way. For models designed based on E3 the low Success Rate is an observation that comes to our surprise and further analysis needs to be done in the future. A possible reason of the contradictory findings between the high average F1 reported in Section-4.3 and the low Success Rate observed in Table-5 might be due to the fact that evaluation of Section-4.3 is averaged firstly across all the sessions performed by the same user and secondly across all the users in order to get the final aggregated results. Moreover performance is reported as a total metric across all exercises. Hence, it is very possible that in most cases of E3 the algorithm provided high performance for some users and relatively low for others. Those kind of differences cannot be sufficiently represented by the aggregated results, but are easily observable through the analysis provided by Table-5. Hence models built based on E3 cannot be considered trustworthy by default since they seem to be very depended by within user variations.

5 CONCLUSION

In this work we tried to tackle the very complex problem of recognising physical fatigue based on subjective user reports and EMG information. We proposed a post-processing method that can be applied in real-time and seems to significantly improve classification results of traditional ML-based techniques in terms of fatigue-detection. Modeling subjective fatigue can be an extremely challenging problem due to the great variability between self-reports and actual EMG measurements across different users and scenarios. In addition creating an objective measure of fatigue is still very pre-mature as it is a factor that depends highly on each individual and his/hers mental and physical state. Despite the aforementioned challenges, it seems that a combination of carefully selected features and classifiers can provide promising results towards targeting self-reported fatigue and can be very useful towards understanding shared behaviors across users. Such observations are crucial towards understanding better the effects of physical fatigue in human performance and can help us draw valuable assumptions between the vague correlation of physical and cognitive fatigue [4]. Our future directions will focus firstly on incorporating the intuitive observation that fatigue is gradually increasing as time passes, which currently has not been taken into account by our proposed methodology and secondly to experiment with more sophisticated methods for temporal modeling of fatigue such as HMMs or Conditional Random Fields. Long-term goal of this research is to eventually design adaptive and personalised rehabilitation scenarios [1, 11] that refine their behavior based on user's physical and mental condition and improve the quality of current assistive technologies related to fatigue assessment [10].

ACKNOWLEDGMENT

This work is supported by NSF (CHS 1565328, PFI 1719031)

REFERENCES

[1] Maher Abujelala, Alexandros Lioulemes, Paul Sassaman, and Fillia Makedon. 2015. Robot-aided rehabilitation using force analysis. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 97.

[2] Ekaterina Dobryakova, Helen M Genova, John DeLuca, and Glenn R Wylie. 2015. The dopamine imbalance hypothesis of fatigue in multiple sclerosis and other neurological disorders. *Frontiers in neurology* 6 (2015), 52.

[3] PA Karthick, Diptasree Maitra Ghosh, and S Ramakrishnan. 2018. Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms. *Computer methods and programs in biomedicine* 154 (2018), 45–56.

[4] Michalis Papakostas, Konstantinos Tsiakas, Theodoros Giannakopoulos, and Fillia Makedon. 2017. Towards predicting task performance from EEG signals. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 4423–4425.

[5] Sang-Hui Park and Seok-Pil Lee. 1998. EMG pattern recognition based on artificial intelligence techniques. *IEEE transactions on Rehabilitation Engineering* 6, 4 (1998), 400–405.

[6] Angkoon Phinyomark, Pornchai Phukpattaranont, and Chusak Limsakul. 2012. Feature reduction and selection for EMG signal classification. *Expert systems with applications* 39, 8 (2012), 7420–7431.

[7] Angkoon Phinyomark, Franck Quaine, Sylvie Charbonnier, Christine Serviere, Franck Tarpin-Bernard, and Yann Laurillau. 2013. EMG feature evaluation for improving myoelectric pattern recognition robustness. *Expert Systems with applications* 40, 12 (2013), 4832–4840.

[8] Angkoon Phinyomark and Erik Scheme. 2018. EMG pattern recognition in the era of big data and deep learning. *Big Data and Cognitive Computing* 2, 3 (2018), 21.

[9] Angkoon Phinyomark and Erik Scheme. 2018. A feature extraction issue for myoelectric control based on wearable EMG sensors. In *2018 IEEE Sensors Applications Symposium (SAS)*. IEEE, 1–6.

[10] Akilesh Rajavenkatanarayanan, Varun Kanal, Konstantinos Tsiakas, Diane Calderon, Michalis Papakostas, Maher Abujelala, Marnim Galib, James C Ford, Glenn Wylie, and Fillia Makedon. 2019. A Survey of Assistive Technologies for Assessment and Rehabilitation of Motor Impairments in Multiple Sclerosis. *Multimodal Technologies and Interaction* 3, 1 (2019), 6.

[11] Konstantinos Tsiakas, Michalis Papakostas, Benjamin Chebaa, Dylan Ebert, Vangelis Karkaletsis, and Fillia Makedon. 2016. An Interactive Learning and Adaptation Framework for Adaptive Robot Assisted Therapy. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 40.

[12] J Cutsem Van, S Marcora, K Pauw De, S Bailey, R Meeusen, and B Roelands. 2017. The Effects of Mental Fatigue on Physical Performance: A Systematic Review. *Sports medicine (Auckland, NZ)* 47, 8 (2017), 1569–1588.

Table 5: Temporal Evaluation of the computed models. Multi-User model corresponds to the models built for E1, Multi-Exercise to E2, Single-User to E3 and Single Exercise to E4. *#Failure Samples* indicates cases (exercise sessions) where the proposed method failed to detect fatigue even though it was present according to user reports. *#Valid Samples* corresponds to the total number of exercise sessions that fatigue was successfully detected by the system. TOST was performed on the amount of samples indicated by #Valid Samples. *Success Rate* indicates the percentage of #Valid Samples over the total number of available sessions (30 sessions per exercise and 90 sessions in total); *p_upper* and *p_lower* correspond to the values of statistical significance that the system would detect fatigue within ±*bounds* sec from the time user reported fatigue. Even-though per-exercise evaluation is also reported for each model, only useful insights can be drawn but no safe and generalizable results due to the limited number of samples used for statistical analysis. *Only Average Performance evaluation (bold) can provide representative and trustworthy evaluation for the temporal performance of the proposed methodology.*

MODEL		Exercise#1	Exercise#2	Exercise#3	**Avg Performance**
Multi-User	#Failure Samples	6	4	13	**23**
	#Valid Samples	24	26	17	**67**
	Success Rate	80%	87%	57%	**74%**
	p_upper	*.019*	.024	<.001	**.015**
	p_lower	<.001	<.001	.018	**<.001**
	bounds (sec)	±10	±7	±6	**±5**
Multi-Exercise	#Failure Samples	7	9	16	**32**
	#Valid Samples	23	21	14	**58**
	Success Rate	77%	70%	47%	**64%**
	p_upper	.031	.034	.025	**.038**
	p_lower	.019	<.001	<.001	**<.001**
	bounds (sec)	±4	±13	±8	**± 6**
Single-User	#Failure Samples	15	11	9	**35**
	#Valid Samples	15	19	21	**55**
	Success Rate	50%	63%	70%	**61%**
	p_upper	.013	.037	.026	**.013**
	p_lower	<.001	.025	<.001	**<.001**
	bounds (sec)	±7	±7	±8	**±6**
Single-Exercise	#Failure Samples	11	7	6	**24**
	#Valid Samples	19	23	24	**66**
	Success Rate	63%	77%	80%	**73%**
	p_upper	.029	.019	.029	**.014**
	p_lower	<.001	<.001	<.001	**<.001**
	bounds (sec)	±9	±5	±6	**±5**