COMP0053 Group coursework

Alexander Smirnov (20169806)
ucaba27@ucl.ac.uk

Andras Nagy (20148372)
andras.nagy.20@ucl.ac.uk

Chiang Hou Nam (15055142)
hou.chiang.20@ucl.ac.uk

Renjie Zhou (17000505)
renjie.zhou.20@ucl.ac.uk

Tim Andersson (20167996)
ucabta2@ucl.ac.uk

Xingjian Lu (20063057)
xingjian.lu.20@ucl.ac.uk

26^{th} April 2021

1 Introduction

This report is comprised two parts. The first part describes a proof of concept experiment for an emotion-based movie trailer rating system. Our study showed that it was feasible to extract emotional signals from movie trailer viewers using laptop cameras in home setting.

The second part examines the application of machine learning techniques to pain recognition. Our analysis has shown that it is possible to achieve a high (over 90%) validation accuracy rate using a version of Multilayer Perceptron adapted to time series data.

2 Part 1 - Proof of concept for an emotion-based movie trailer rating system

The underlying hypothesis of this study was that it is possible to extract emotional signals from the audience's facial reactions to movie trailers. These emotional signals can subsequently be used for various useful and interesting applications, such as a more tailored movie trailer rating system, making for a better consumer experience.

In this proof of concept experiment, we have considered the audience's real-time reactions to movie trailers of three genres, namely sad, funny, and horror. We then examined the extent to which our measurements can capture the resulting emotional arousal (or lack of) in the audience.

The first step of this process was to to collect the user reaction data - this is described in section 2.1. The next step was labelling and other data processing for the purposes of qualitative analysis and subsequent machine learning model training for emotion recognition. This step is described in section 2.2.

2.1 Experiments and data storage

2.1.1 Experiment set-up

As described in the data collection protocol attached in appendix, the participants of the experiment are the 6 group members from our group. Each of us recorded his or her facial expressions when watching the pre-selected 9 videos, each around 2-3 minutes long. Each genre has 3 movie trailers. In between every two movie trailers, there is neutral video clip to for the audience to calm down to a baseline emotional status. This is to make sure that the emotional reaction to each video is not affected by the lingering emotional change due to the previous movie trailer.

All participants watched the same playlist, with the fixed ordering for the 9 movie trailers. All participants also watch the neutral video clip for the same duration of time of 2 minutes between every 2 movie trailers. These were fixed for each participant to control the potentially impacting factors to reduce their influence over the results of the experiments.

2.1.2 Data transfer and storage

The recordings, together with the self-labelling, are then uploaded to a OneDrive folder that is accessible exclusively to the group members with their UCL emails. The recordings are then processed by the software OpenFace and the toolbox iPhys. OpenFace processes the videos and extracts the key coordinates of the face markers and orientations for every frame of the video. iPhys toolbox extracts the continuous heart rate signal from the images of the face in the videos.

After the processing, the videos will be deleted from the OneDrive folder as soon as the permission from the UCL reviewers of this coursework is granted. This will ensure that no recordings of the participants' faces are stored for longer than required. The processed data, i.e. the numerical features extracted from the videos, will also be renamed using random numbers instead of the participants' names to further increase the subjects' anonymity. Permissions from the participants were acquired should any facial images be put in this report for demonstration purpose to show how OpenFace and iPhys toolbox work.

All group members have obtained the UCL-certified GDPR training and all data collection, storage and transfer activities were conducted in line with GDPR guidelines. All group members gave their consent to participating in the study and were granted an option to terminate their involvement at any time. None of the group members exercised this option.

2.2 Data processing and analysis of the results

2.2.1 Data labelling

In order to examine the effectiveness of our data collection methodology, the videos with facial reactions were subjected to two types of labeling:

- Self-labelling
- Cross-labelling

Under self-labelling, each participant annotated their recorded reactions to one film video from each of the three genres. The self-labelling was supplemented with cross-labelling, whereby a different member of the group examined the reactions that were self-labelled by other participants. This information allows us to examine the consistency (or lack of) between people's interpretation of the same facial expression.

The resulting labelling files are available in Appendix A.

2.2.2 Analysing facial landmarks with OpenFace

OpenFace is an open source facial behaviour analysis toolkit.[1] The tool allows to examine facial behaviour using facial landmark motion, head pose, facial expressions and eye gaze, and facial action units all together. The tool defines 68 facial landmarks, with the sets of point distribution trained separately and patched later for eyes, lips and eyebrows.[2] In particular, we used the `nn4.small.v1` version of the model,[3] which has achieved a 93% accuracy on the Labeled Faces in the Wild(LFW) benchmark.

Due to time constraints, only the 3 labelled videos from each participant were processed using OpenFace. Specifically, the portion of reaction video that corresponds to the selected cross-labeled reactions were trimmed out, converted to .avi format, and then processed with the FeatureExtraction function in OpenFace. Through this processing, we have obtained the csv files with features for each processed recording. In each file, each frame is converted into a sequence of the x, y and z coordinates of the gaze positions of both eyes, landmarks of both eyes and landmarks of the rest of the face. There are 56 landmarks for each eye and 68 landmarks for the entire face. Together with the gaze orientations, the head pose and action units, these make up the features extracted from each frame. Figure 1 shows two example frames processed by OpenFace. The red dots are the landmarks identified by the algorithm on the face of the participant. The green arrows indicate the orientation of the gaze. Finally, the blue cube shows the pose of the head of the participant. As shown in the figure, these features are continuously re-computed for each frame of the video.

[1] T. Baltru˘saitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in 2016 IEEE Winter Conference on Ap- plications of Computer Vision (WACV), 2016, pp. 1–10.

[2] Ibid.

[3] http://cmusatyalab.github.io/openface/models-and-accuracies/#accuracy-on-the-lfw-benchmark

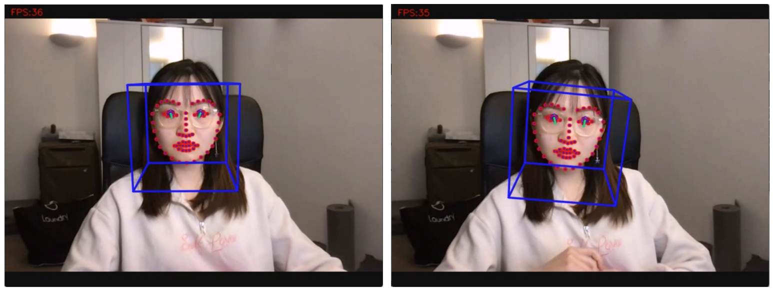

Figure 1: Examples of frames in video processed by OpenFace

Our examination of the csv files generated by OpenFace suggests that the software can reasonably accurately capture the facial and eye landmarks throughout the video. A random selection of frames across different participants exhibit a consistent 98% confidence level for the location of the landmarks.

The processed results are saved in the OneDrive accessible only to the group members and any further university members whom it may concern, such as TAs of the module.

2.2.3 Analysing heart rate with iPhys toolbox

Iphys toolbox, an open source imaging-based physiological measurement toolbox, was used to measure the heart rate of the participants throughout the experiment.[4] This toolbox contains many of the most frequently used computational methods. In this project the Green-Channel Method from Verkruysse et all (2008)[5] was employed as the green channel from the three TGB colour channels contains the strongest pulsatile signal. This method calculates the processed Blood Volume Pulse (BVP) for the entire given time period and then gives one final estimated Pulse Rate (PR) from processed BVP time series using peak in periodogram.

In order to make the measurements more representative, the code of the toolbox was modified. A code was created where a 10 seconds wide time window is slided through the desired time period of the given input video, and thereafter, the program outputs a heart rate value for every 10 seconds of the video.

Again, due to time constraints, only the 3 labelled videos from each participant were processed using iPhys to get the continuous heart rate data from the images. The plots for the respective 3 videos for 6 participants are shown below.

[4] T. Baltruˇsaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in 2016 IEEE Winter Conference on Ap- plications of Computer Vision (WACV), 2016, pp. 1–10.

[5] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." Opt. Express, vol. 16, no. 26, pp. 21434– 21 445, Dec 2008. [Online]. Available: http://www.opticsexpress.org/ abstract.cfm?URI=oe-16-26-21434

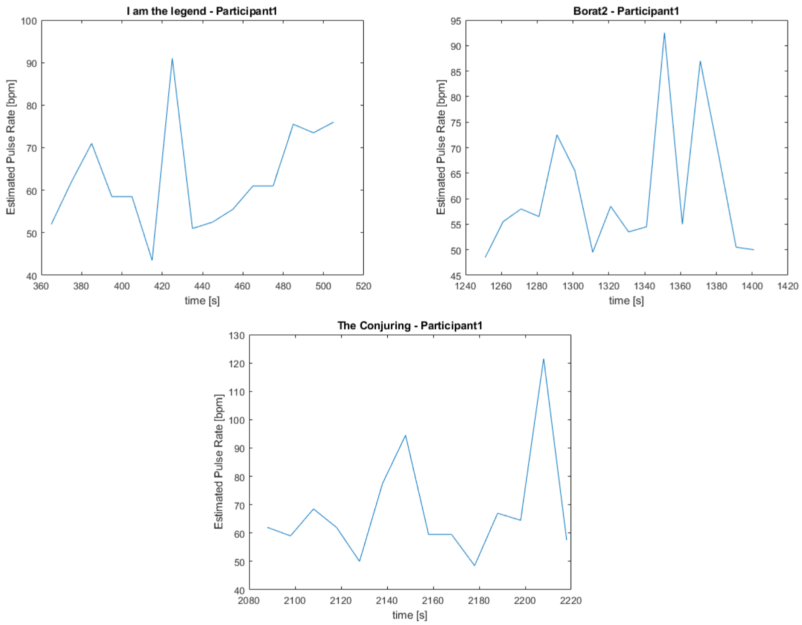

Figure 2: Heart rate plots for participant 1

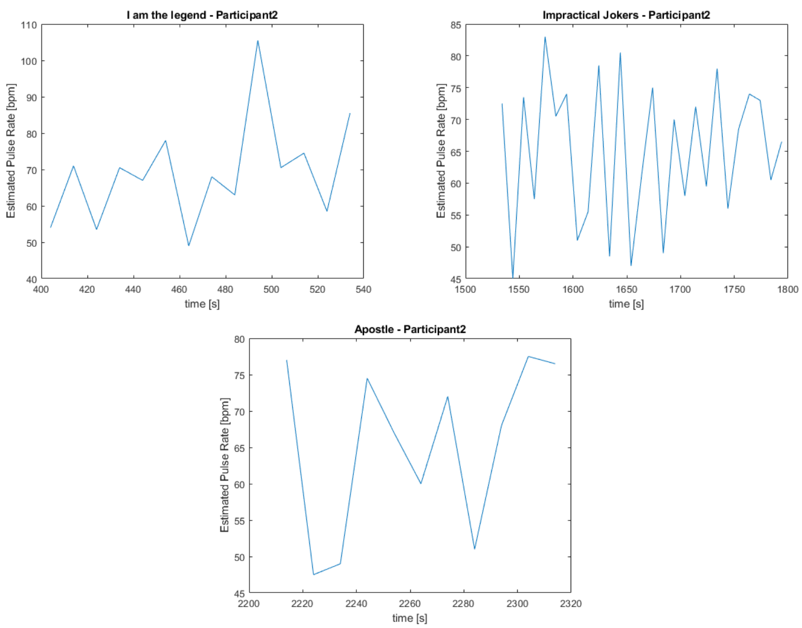

Figure 3: Heart rate plots for participant 2

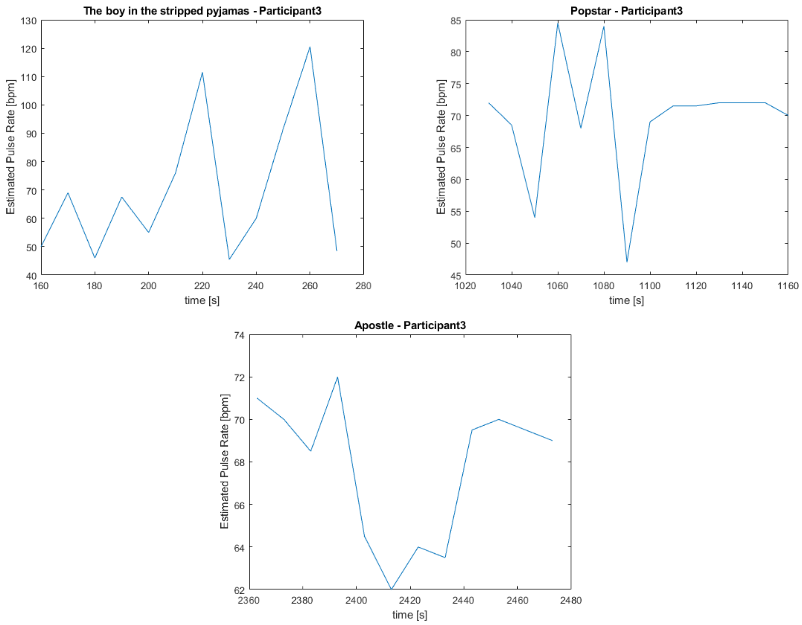

Figure 4: Heart rate plots for participant 3

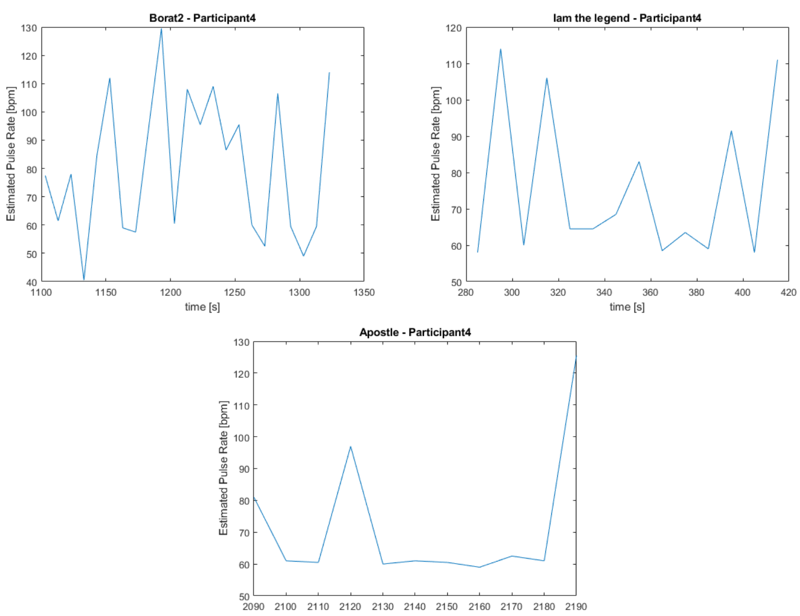

Figure 5: Heart rate plots for participant 4

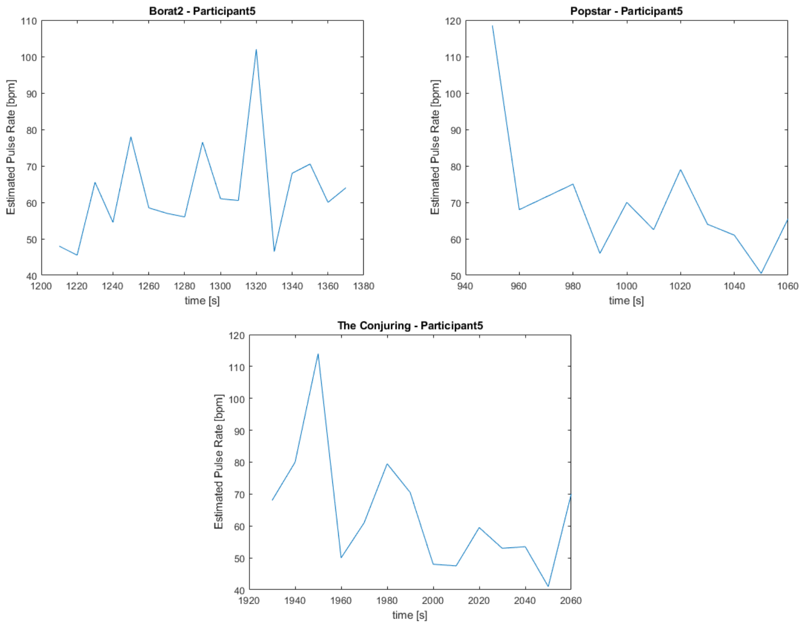

Figure 6: Heart rate plots for participant 5

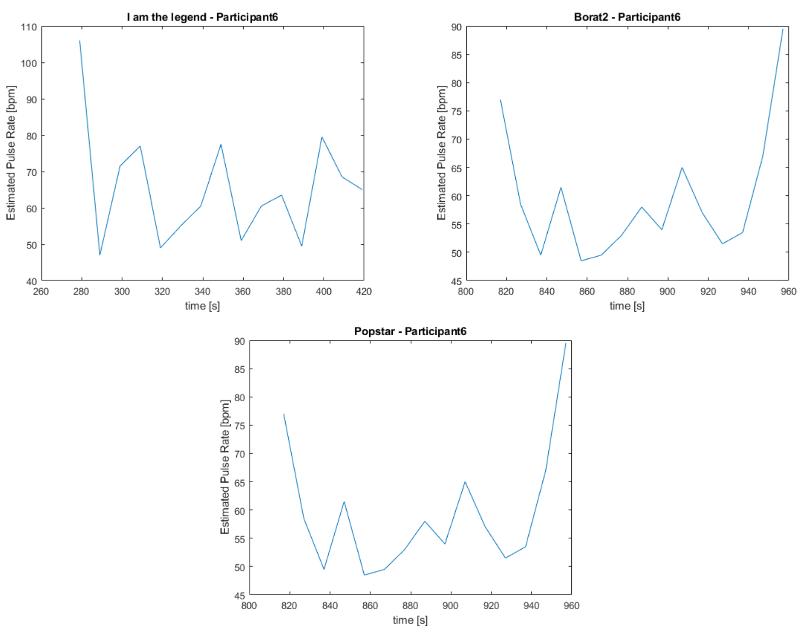

Figure 7: Heart rate plots for participant 6

As shown in the figures from 2 to 7, there is quite a big variation among the heart rate changes of different participant across the three genres of the movie trailers. Most participants have the biggest fluctuations in the heart rate when watching horror movie trailers, i.e. Apostole and The Conjuring. However, the pattern is not observed in some of the participants. For example, the heart rate of participant 2 stays roughly around the normal heart rate level with an average of 70 bpm while watching he trailer of *Apostole* as shown in 3. More surprisingly, the heart rate of participant 5 decreases along the course of watching the trailer of *The Conjuring* as shown in 6.

Meanwhile, relatively large fluctuations are observed for the heart rate of the participants when watching the trailers of funny movies trailers or clips, such as *Popstar* or *Boart2*, with some going above 110 bpm as observed from participant 4 shown in 5. Similar pattern can be observed for the heart rate changes to watching sad movie trailers as well.

As we don't have enough data samples to draw a conclusion on the commonly observed pattern, we can't conclude on which kind of patterns is more probable to be the outliers. Moreover, different lighting across venues used by different participants may affect the results of using iPhys toolbox, since it is an image-based heart rate detection toolbox. Furthermore, the results could be subject to noise, as it can be seen in 2. A drastic change in the estimated heart rate from 70 bpm to 120 bpm and back to around 60 bpm within a few seconds is more likely to result from measurement noise than from the true heart rate. The data will be much more accurate if continuous monitoring of heart rate via contact sensors is used.

3 Part 2 - Machine Learning for Affective Computing using EmoPain Dataset

3.1 Summary of results

In this report we have examined two tasks: (i) a binary prediction of whether or not a test subject exhibits protective behaviour at any given moment; and (ii) a multiclass prediction of the type of protective behaviour exhibited by the subject, if any. In both tasks, the prediction models presented in this report have consistently achieved a validation accuracy of over 94% across 100 validation datasets randomly sampled from the EmoPain dataset.

Figure 8 presents the validation accuracy for the binary protective behaviour prediction model.

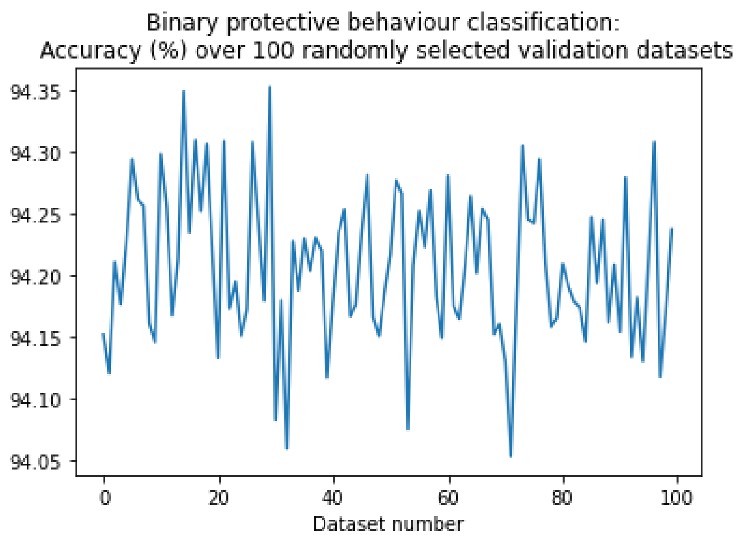

Figure 8: Validation accuracy for binary protective/non-protective behaviour classification

It can be seen that the model achieves a consistent validation accuracy of 94% across all samples. The multiclass prediction, in turn, achieves an even better accuracy of 98%, as illustrated in Figure 9.

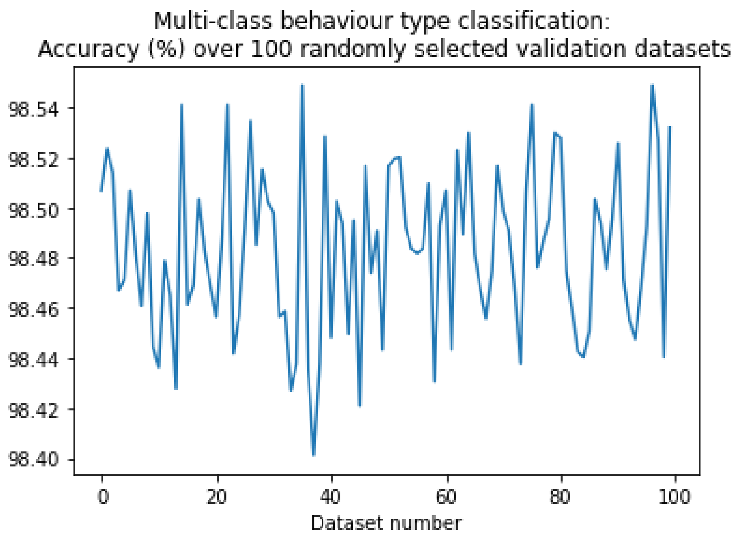

Figure 9: Validation accuracy for multi-class protective behaviour prediction

The rest of this section describes how the results above were achieved, including the data employed for this study, the data processing procedures carried out before the predictor training, the architecture of the underlying machine learning model, and the values of the associated hyperparameters.

3.2 Data

3.2.1 Description of features

The data comprises of features extracted from footage of 35 test subjects performing specified actions.[6] In particular, each subject was asked to perform the following nine actions:[7]

- One leg stand
- Reach forward
- Bend
- Sit to stand
- Stand to sit

[6]We are aware that slide 2 of the EmoPain Dataset slide deck by Chongyang Wang lists 22 healthy participants and 18 participants with chronic lower-back pain. However, in the dataset shared with us, only 35 files were available.

[7]The labels of the data include an additional category of 'other', which denotes the transition between the activities listed below.

- Sitting still
- Standing still
- Walking

The features extracted from each video feed include (i) the 3-D coordinates of 22 body joints over time and (ii) output of four surface electromyography (sEMG) sensors placed on the subject's back. Figure 10 and Figure 11 illustrates the location of the tracked joints on the body and the sEMG sensors on the subjects' back respectively.

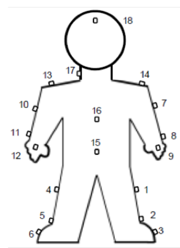

Figure 10: Tracked joints

Figure 11: Location of the sEMG sensors on a subject's back

In summary, for every test subject, the collected features include 70 time series[8] with a duration between 176.15 and 405.42 seconds.

3.2.2 Description of labels

For this dataset, continuous labels were generated by experts observing the videos of the test subjects. The labels for each frame specify the type of protective behaviour (if any) exhibited by the test subject, as well as the level of pain (if any) experienced during the activity. Table 1 shows the pain level and the types of protective behaviour identified in the underlying data collection study, which has 7 labels in total.

[8]Time series of X, Y ,and Z coordinates for each of the 22 joints, plus time series for each of the four sEMG sensors.

Table 1: Labels

Labels	Value	Meaning
Exhibits protective behaviour	0	Not active
	1	Active
Guarding or Stiffness	0	Not active
	1	Active
Hesitation	0	Not active
	1	Active
Support or Bracing	0	Not active
	1	Active
Abrupt motion	0	Not active
	1	Active
Rubbing or Stimulation	0	Not active
	1	Active
Pain Level	-1	Not recorded
	0	Healthy
	1	Low
	2	High

To learn more about the experiments, we created animations to recreate the participants' motions over the course of the experiments. We also included the information of different pain level, protective behaviour and exercise type to the animation. To do this, in addition to display status messages in the legend, we also change the color and marker type to illustrate different protective behaviour and pain level respectively as described in Table 2 and Table 3.

Table 2: Pain level markers

Pain Levels	Marker shape
Not reported	circle
Healthy	square
Low level pain	upper triangle
High level pain	diamond

Table 3: Protective behaviour colors

Protective behaviours	Marker colors
Guarding	green
Hesitation	magenta
Bracing	cyan
Abrupt motion	blue
Stimulation	red
More than 1 behaviours	black
No behaviour	orange

Figure 12 and 13 illustrate participant P300D transitioning from sit-to-stand in the animation.

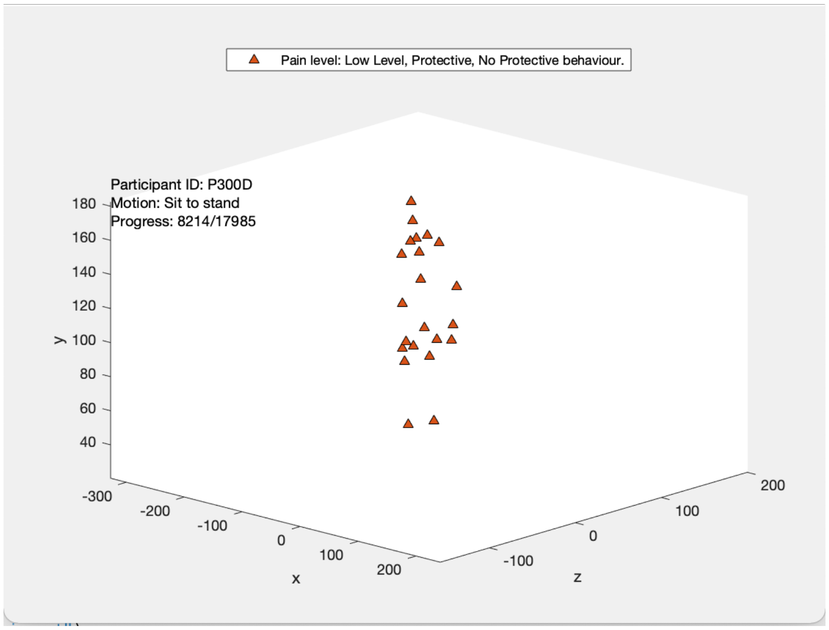

Figure 12: Sit-to stand: Initiation

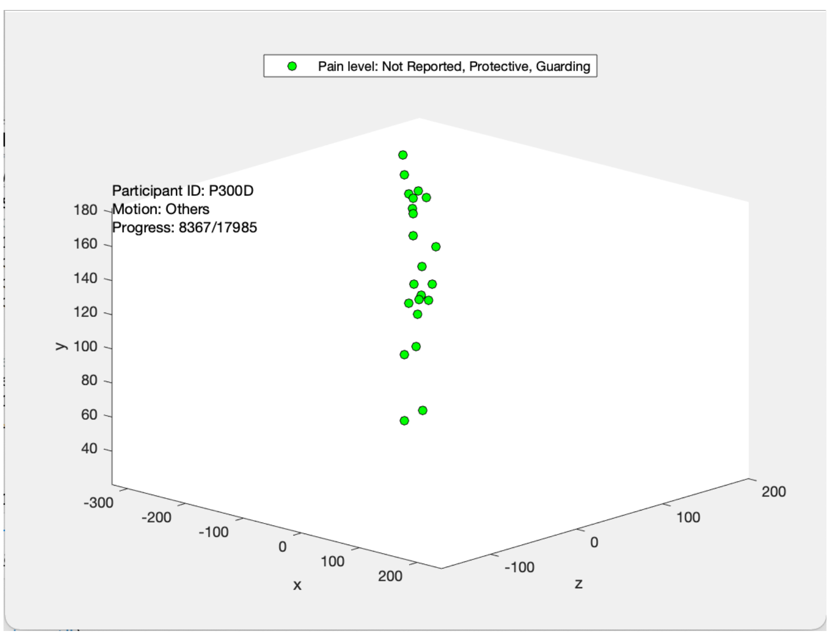

Figure 13: Sit-to stand: Completion

We can see that this subject experienced low level pain when s/he tried to stand up and after s/he stood up s/he was guarding to keep her/himself from falling or slipping. Besides visualizing the experiment, we also visualize the output of the four sEMG sensors along with the protective behaviours. One of the sEMG profile in consistent with P300D is shown in Figure 14.

An overview of Figure 14 shows that this participant suffers from both low and high level pain during almost every transition from one pose to another. With a closer look we can see that s/he suffers high level pain when s/he is trying to work out those relatively demanding exercises, such as bending and one leg stand.

Turning to the protective behaviours chart, we can see that protective behaviours usually occur around the time when the participant experiences pain. For instance, between time steps 13,000 and 15,000, we can see that s/he is guarding, bracing and stimulating him/herself when s/he is asked to sit down and stand up repetitively.

One more thing that we can observe from the figure is that compared to the other three sEMGs, the sEMG for left upper trapezius has a higher amplitude at almost all the times, except the time period when s/he tried to bend. This may suggest the chronic pain hinder the participant from using muscles around his/her left and right lumber paraspinal and right upper trapezius. Therefore, s/he was only relying on the muscles around left upper trapezius most of the time. In contrast, the sEMG outputs from participants with no chronic back pain exhibit similar trends regardless of amplitude in right and left lumber paraspinal. An example of this is subject C67D shown in Figure 15. As we look into the exercise list we can see that any healthy individuals can perform the movements with minimum effort from the upper body, which means it is normal to expect rather inactive trends of sEMG for right and left trapezius. This suggests that participant P300D needed to use his/her upper body muscles to help him/herself finish the exercise under the pain s/he was suffering.

For the full sets of sEMG profiles of all participants and the program that run animations, please refer to our MATLAB program.

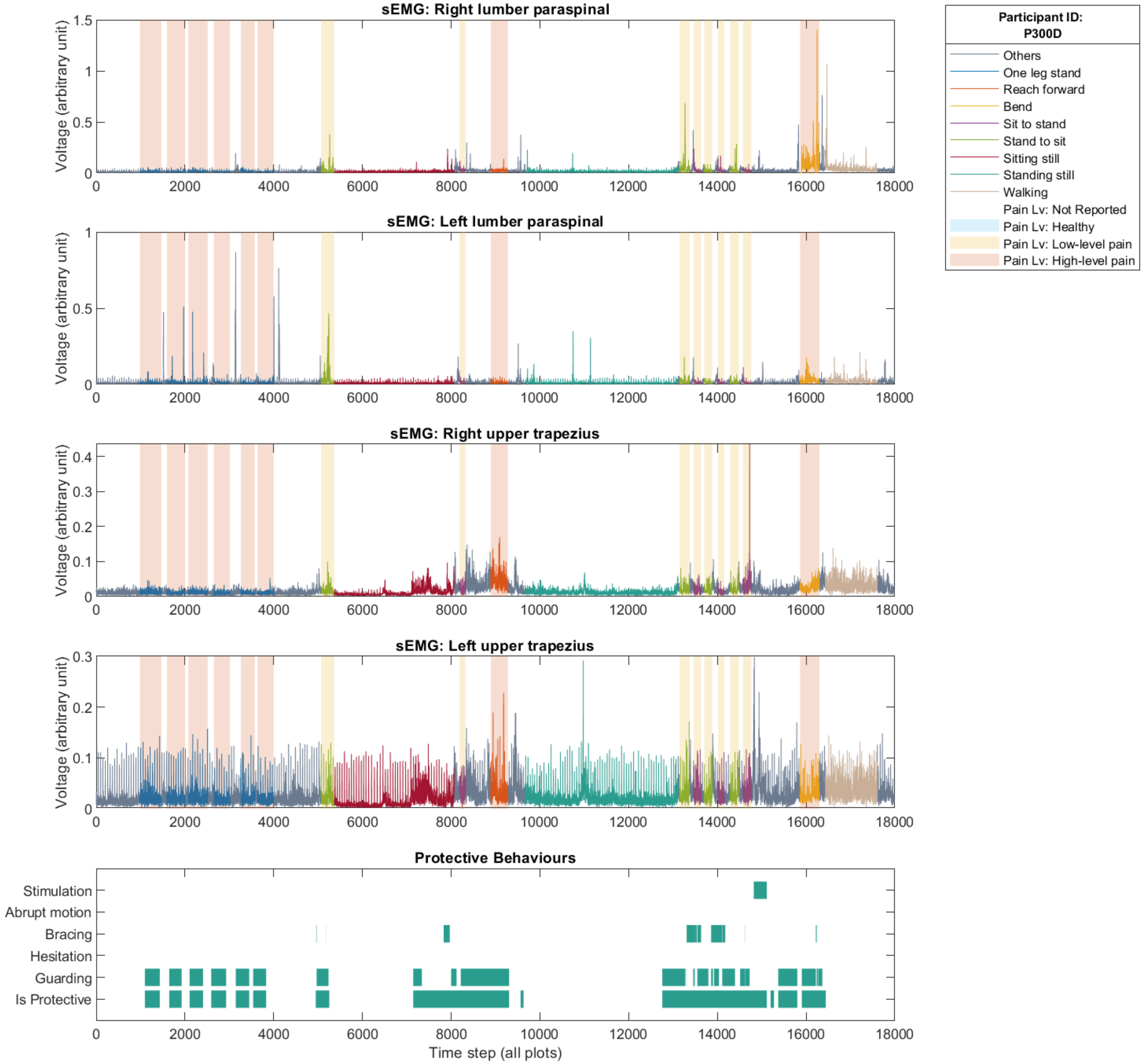

Figure 14: sEMGs of participant P300D

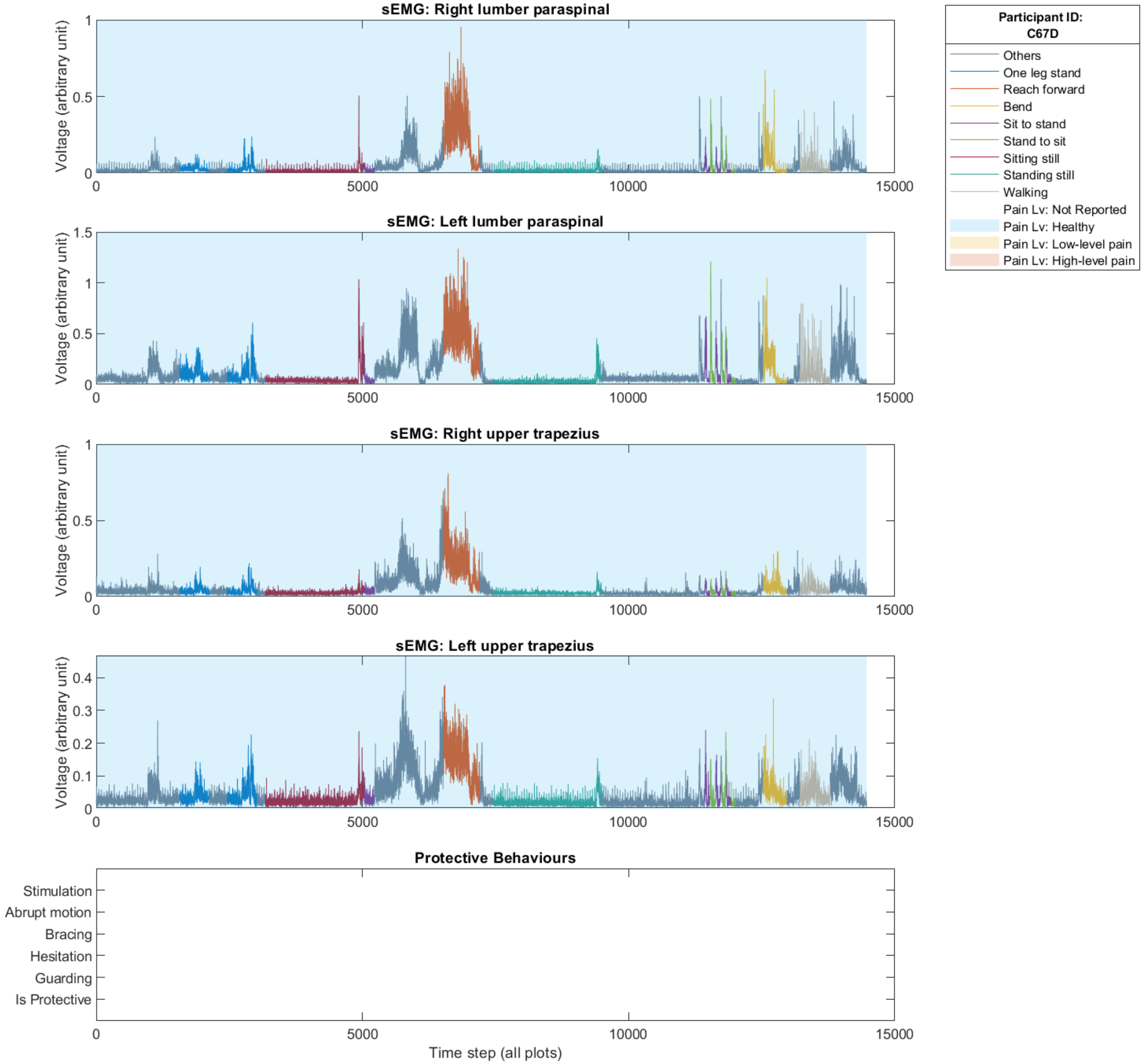

Figure 15: sEMGs of participant C67D

Along with the introduction to the features and labels, we now aggregate all 35 experiments and categorized into four groups as in Table 4, namely:

- Healthy control participant doing normal exercise (Group CN)
- Healthy control participant doing difficult exercise (Group CD)
- Chronic pain participant doing normal exercise (Group PN)
- Chronic pain participant doing difficult exercise (Group PD)

Table 4: Participants in four different categories

Group CN	Group CD	Group PN	Group PD
C382N	C382D	P113N	P113D
C709N	C544D	P191N	P191D
C788N	C56D	P299N	P299D
C202N	C93D	P492N	P300D
C256N	C202D	P531N	P336D
	C256D	P699N	P492D
	C67D	P890N	P699D
		P921N	P921D
		P342N	P342D
		P487N	P487D
		P54N	P54D
		P649N	

Here we try give a high level idea of the properties of different categories. Firstly, we can see an increasing trend in the duration of the experiments presented in Table 5. This suggest that people who suffers chronic pain will usually take longer to finish the given set of exercise, and it also takes more time to work out exercises in the difficult set than in the normal set.

Table 5: Duration of different experiment groups

	Group CN	Group CD	Group PN	Group PD
Duration in $mean \pm std$ (s)	178.27 ± 17.89	213.32 ± 18.01	263.08 ± 64.36	275.83 ± 50.26

Next, with the knowledge that we have nine types of exercise, four pain levels, two values for protective activities and six types of protective behaviour, we created a frequency table for all 432 combinations for each experiment which subsequently normalized over the duration of that experiment. So that all the fractions related to each protective behaviour sum to one. Next we average the table over the all experiments and then normalize over all six protective activities. After that, we sum all numbers related to the followings:

- No recorded pain

- Healthy
- Low level pain
- High Level pain
- Protective behaviour triggered

We then get the tables for each category shown in Table 6

Table 6: Probability of triggering protective behaviour and each pain level

	Protective behaviour triggered	No recorded pain	Healthy	Low level pain	High Level pain
Group CN	0.0000	0.0000	1.0000	0.0000	0.0000
Group CD	0.0000	0.0000	1.0000	0.0000	0.0000
Group PN	0.0525	0.5047	0.2539	0.1563	0.0851
Group PD	0.0718	0.5194	0.1965	0.2073	0.0769

Under this matrix, we can see that for normal participants there was no record for triggering protective behaviour and the pain level was indicated as "Healthy" at all times.
For the participants suffering chronic pain, we can see non-zero entries for all categories. As we move from the group doing normal exercises to the group doing difficult exercises, we can see that along with a slight decrease in No recorded pain, High level pain and a noticeable drop in "Healthy" level of pain, there is an increase in Protective behaviour triggered and Low level pain. This suggest that the difficult set of exercise was indeed more challenging for participants suffering chronic pain since there is an increase in protective behaviour triggered. Albeit some may argue that the number in high level pain slightly dropped, if we sum Low level pain and High level pain together, we can see that the chances that chronic pain people actually do suffer more pain in the difficult set of exercise than that in normal set of exercise. This is more convincing if we also look at the drop of "Healthy" pain level at this point.

3.2.3 Pre-training data processing

The raw data was subjected to two operations before being used in the classifier training.

- Standardisation
- Batching

In this study we have standardised each feature by using the linalg.norm function of the numpy library.[9] Such standardisation decreases the likelihood of the optimisation algorithm getting stuck in a local minimum, as well as

[9] For more details see https://numpy.org/devdocs/reference/generated/numpy.linalg.norm.html

makes the training process significantly faster. See function normalise_data for the python implementation of the standardisation.[10] The batching involves randomly sampling a subperiod from the raw data time series. Different batch sizes, from one to 20 frames, have been tested to analyse the effect of batch size on accuracy and loss. The overall training sample consists of 100 thousand batches. Figure 16 illustrates how batches are sampled from raw data. Note that the batch width is not drawn to scale.

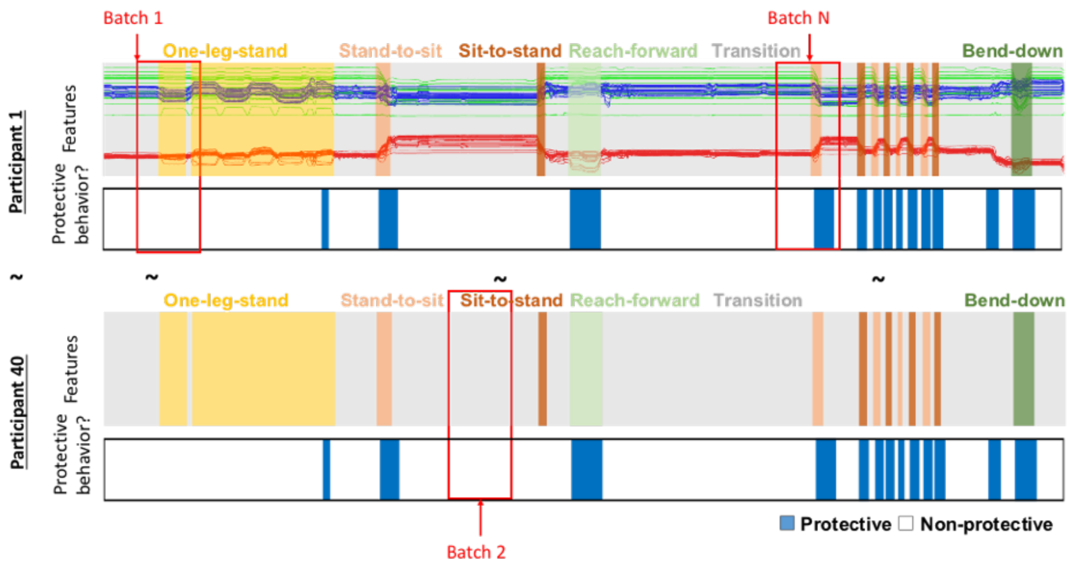

Figure 16: Illustration of the batching process.

[10]Cell 2 of the notebook accompanying this report.

3.3 Model architecture

3.3.1 Binary Protective/Non-Protective Behaviour Classification: Model Description & Data Pre-processing

The model paradigm used for the task of binary classification of protective behaviours on the EmoPain dataset is that of the methods described in Campolucci et all (1999)[11] and Roy (2020)[12]. The algorithm is an adaptation of a Multi-layer Perceptron (MLP) to time series data. Figure 17 illustrates the algorithm graphically for a single training instance, i.e. a single batch.

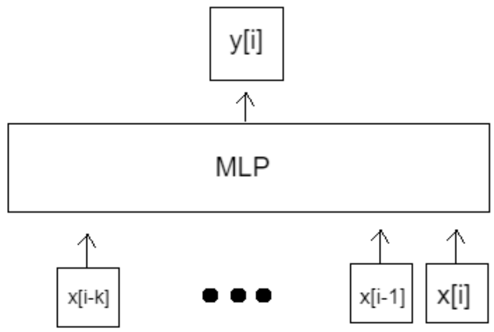

Figure 17: Model architecture used for binary prediction on EmoPain dataset

This paradigm essentially makes the MLP 'online', and in practice, dependent on the task at hand and the hardware/software in use, could be used in real-time and trained in close to real-time. The model is implemented using the Tensorflow Keras library, along with numpy and other basic libraries. Function make_model found in the notebook accompanying this report presents the python implementation of the model architecture.[13] The next subsection elaborates on the hyperparameters of the model.

[11]Campolucci, Paolo Uncini, Aurelio Piazza, Francesco Rao, Bhaskar. (1999). On-line learning algorithms of locally recurrent neural networks. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council. 10. 253-71. 10.1109/72.750549, available at https://www.researchgate.net/publication/5598777_On-line_learning_algorithms_of_locally_recurrent_neural_networks/citation/download

[12]Roy, B., R. (2020), Multi-Layer Perception Time Series, Data driven investor, 6 October, available at https://medium.datadriveninvestor.com/multi-layer-perception-time-series-a6261ab95ceb

[13]See cell 3 of the notebook.

3.3.2 Binary Protective/Non-Protective Behaviour Classification: Hyperparameter optimisation

The model employed in this study has six hyperparameters:

- Number of hidden layers
- Number of nodes per hidden layer
- Activation functions
- Regularisation technique
- Optimisation algorithm
- Learning rate

The rest of this subsection provides more detail on the values of these hyperparameters adopted in the modelling.

The number of hidden layers was set at four, given that the addition of the fifth layer did not provide a noticeable improvement in the model's predictive performance. The number of nodes was selected such that the size of the network shrinks approximately linearly with each additional hidden layer - the methodology for calculating the number of nodes for each layer is presented in table 7.[14]

Table 7: Number of nodes in each hidden layer

Features per frame	[A]	71
Frames per training instance	[B]	7
Features per training instance	[C] = [A]x[B]	497
Number of nodes in	[D]	[E] = [C]x[D]
Hidden layer 1	85%	422
Hidden layer 2	50%	249
Hidden layer 3	30%	149
Hidden layer 4	15%	75

Initial experimentation on a random sample of training instances showed that the Adam optimizer consistently yielded a higher training accuracy. Other algorithms considered include optimizers available in the Keras package.[15]

Our analysis revealed that the learning rate does not have a material impact on the effectiveness of the model training, as long its value is no higher than 0.1. Therefore, in this study we have relied on the learning rate of 0.1, since setting

[14]Note that the number of features per frame includes the 70 features described in section 3.2.1, plus a bias feature.

[15]Namely, SGD, RMSprop, Adadelta, Adagrad, Adamax, Nadam and Ftrl. See https://keras.io/api/optimizers/

the learning rate to an unnecessarily low value results in wasted computation time.

Activation function was selected through cross-validation performed over 100 thousand randomly drawn training instances. The analysis showed that tanh yields the highest average validation accuracy, when compared to other activation functions available in Keras.[16] The python implementation of the cross-validation analysis can be found in cell 6 of the notebook accompanying this report. All of the models considered in this study were regularised with a weight dropout, with the proportion of dropout set to 20%. This allows to mitigate overftting - a common problem of neural networks. All models considered during the hyperparameter optimisation process,as well as the final model, were trained over ten epochs, as prior experimentation showed that this was sufficient for the convergence of the training loss.

3.3.3 Multi-Class Protective Behaviour Type Classification

The multi-class protective behaviour type predictive model was based on the same methodology as the binary classification model described above. Python implementation of the model can be found in the notebook, accompanying this report, cells 14-25.

[16] Namely, the other functions considered include logistic, ReLu, eLu, SeLu, and Softsign. All layers were assumed to have the same activation function.

A Appendix

COMP0053 Data Collection Protocol

Alexander Smirnov
Andras Nagy
Chiang Hou Nam
Renjie Zhou
Tim Andersson
Xingjian Lu

March 2021

1 Objectives

We are planning to collect the emotional response from the audience by taking the footage of them watching the pre-selected trailers.

2 Preparation for data collection

At the start of the experiment, each participant is reminded of the rules of the experiment and is asked to conduct the following checks.

First, the participant has to be alone in the room while conducting the experiment. Second, the room has to be in a quiet location, to the extent that it is feasible within the field environment. Both of these conditions ensure that the participant is not distracted during the experiment. This also prevents accidental capturing of data about other individuals. For the same reason, if the participant would like to conduct experiment on someone else's premises, they need to obtain prior agreement from the dwellers of those premises.

If all of the above checks are satisfied, the participant checks that the required sensory equipment is available.

The equipment used in this study includes:

- a laptop, to show the video materials
- a camera, potentially built into the laptop, to record the participant's facial image which will later be used to analyse heart rate and facial expression

Once ready, each participant initialises their video recording. The video monitoring continues until the end of the experiment, including the breaks between viewings (see Section 3).

3 Data collection process

Each participant is asked to watch three pairs of video clips, with a three-minute break in between. The first video in the viewing is selected from a list of pre-selected neutral videos. Measurements from the viewing of neutral videos will be used to establish the baseline of heart rate and facial expression for every participant of the study. Such an approach is consistent with Shu et al. [1]

Next, each participant will be shown a film trailer, randomly chosen (without replacement) from a pool of pre-determined trailers (see Appendix 1: Pool of Trailers). The average duration of each trailer is around 2.5 minutes, which is consistent with duration of video clips used in Soleymani et al. [2]

Following the viewing of a trailer, each participant takes a five-minute break, including watching the three-minute neutral clip for returning back to ground emotional state, after which the procedure is repeated until every participant has viewed all of the trailers in the trailer pool.

The video monitoring is terminated five minutes after the participant watched the last trailer in the pool.

4 Data transfer and storage

The video files are then saved locally, with pre-generated participant numbers as identifiers.

We will be only recording data from the group members, and store and share the videos within the group locally for local processing using OpenFace software and iPhys toolbox later onwards. Therefore, the video data we will collect is only from our team and the access will be strictly restricted within the group.

After recording individual response to the trailers, the participant will watch the reaction videos of the other group members and label their reactions. If needed, the participant can refer to the original trailer for ground truth emotion labelling.

Following the labelling process, we shall process the videos with OpenFace software and iPhys toolbox. The OpenFace software [3] will extract the coordinates of the key face markers and face orientation for every frame of the video and the iPhys toolbox [4] will extract continuous heart rate signal from the video.

[1] Shu, Lin; Yu, Yang; Chen, Wenzhuo; Hua, Haoqiang; Li, Qin; Jin, Jianxiu; Xu, Xiangmin. 2020. "Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet" Sensors 20, no. 3: 718. https://doi.org/10.3390/s20030718

[2] The authors utilised video clips with an average duration of 1.5 minutes. See M. Soleymani, J. Lichtenauer, T. Pun and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42-55, Jan.-March 2012, doi: 10.1109/T-AFFC.2011.25.

[3] In particular, we use the nn4.small1.v1 version of the model, which has achieved a 93% accuracy on the LFW benchmark. See the following link for more details: `http://cmusatyalab.github.io/openface/models-and-accuracies/#accuracy-on-the-lfw-benchmark`

[4] McDuff, D. and Blackford, E., 2019, July. iphys: An open non-contact imaging-based physiological measurement toolbox. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6521-6524). IEEE.

Once the processing is complete, the processor permanently deletes the video of the participant. At this point, all data becomes fully anonymous. All of the video and heart rate data exchange required for the processing described above will be carried out exclusively over UCL OneDrive.

5 GDPR and ethical issues

For this data collection task, the participants will be only from our group, i.e. all the group members. All of the us have granted a consent to take part in the study. As the data collector as well as the participant, we have all passed the required UCL GDPR training course. The video trailers were also pre-screened by an expert panel, to identify content that could be considered emotionally disturbing. No such content was identified for this study. The participants also have the right to terminate their involvement in the experiment at any time. All data collection and processing was done in accordance with the latest GDPR requirements.

Appendix 1: Pool of Trailers

Each of the participants has selected three trailers in three genres: sad, funny, and scary. None of the participants have seen any of trailers of the respective films – the trailers were identified through web search. This is done with the aim to achieve the most spontaneous reactions.

In this experiment, we have selected the following trailers:[5]

- Sad
  - The boy in the striped pyjamas, main trailer
  - Brain on fire, main trailer
  - The light between oceans, first trailer
  - What happened on September 11 (2019), official trailer
  - Beautiful boy, first trailer
- Funny
  - Popstar: Never stop never stopping, second trailer
  - Borat 2, main trailer
  - The nice guys, main trailer
  - Zack and Miri make a porno, main trailer
  - The hustle, main trailer
- Scary
  - The conjuring, main trailer
  - Apostle, main trailer
  - IT Chapter 2 (2019), first trailer
  - Sinister 2, first trailer
  - Orphan, first trailer

[5] The list below contains two spare trailers in each category. The final choice of trailers will be made at a later stage.