# COMP0053 Affective Computing and Human-Robot Interaction Individual Coursework

Douglas Chiang (15055142)
hou.chiang.20@ucl.ac.uk
Group: ASCERTAIN

$21^{th}$ May, 2021

# Introduction

This report consists of two parts. The first part provides a critique of our data collection activity. The second part critiques our implemented Affective Recognition System (ARS).

# 1 Part I

## 1.1 Aim of the project

In our project, we considered the audience's real-time reactions to movie trailers of three genres in: sad, funny and horror. Then we used the data to examine the extent to which the measurements can capture the resulting emotional arousal of the audience.

If results show promising correlations between affective video content and human emotions, then the significance of this study can provide a direction for creators to create more affective contents. In terms of content platforms, this means a possibility to make more profit as they can retain customers by utilizing the results for better suggesting movies for them at an individual level.

Albeit our project only investigates the correlation between affective contents and human emotions, our ultimate goal is to provide a tailored movie rating system for a better customer experience. This is motivated by:

1. Observing the increasing supply in recent years of various video contents on different platforms such as Netflix and Amazon prime video.
2. Recognizing that different audiences have various preferences in video contents, and different contents can arouse different emotional states in the audience [1].

In addition, the result of the study can also help content creators to develop businesses in advertisement and marketing. There are studies that show positive correlations between emotional branding and customer loyalty, especially in fields like fashion and beauty products [2][3]. So content creators can profit from the demand of these areas. Another opportunity that content providing platforms may consider is to make a chatbot that can chat with humans about different video contents by incorporating natural language processing.

However, applying this study may induce consequences in the content creating field. One of such is that the creations may lean more on emotion strategies rather than pure creative contents. Another consequence is that since the sample size of our study is limited, it may not reflect variations across different ethnicity and race [4], which may cause errors when deployed in applications and may face accusations of racism. We can also foresee that videos will become more addictive since they are more emotionally arousing, which may have an issue of auto manipulation of people's mind.

## 1.2 Choice of affective states

In our study, we have chosen three affective states, namely happy, sad and horror. We chose these three states because they are quite common types of emotion arousal in movies. This works in the sense that in our study we are just investigating the feasibility of extracting emotional signals from the audience's facial reactions to movie trailers. If it is in a real application, our choice will subject to limitations. An obvious limitation is that we left out many other different affective states such as angry, relaxed, nervous, etc. Another limitation is that we haven't considered different shades of affective states. An article at BBC concluded that out of 10 different types of smile, only six of them indicates that people are having a happy moment. They also observed that we do smile when we are embarrassed, not being comfortable and in pain. Additionally, smiling can also be a result of culture. For example, smiling with the eyes is considered as having good manner and is emphasize considerably in Japanese society [5].

To improve, we can incorporate voice data of participants with our visual data. Studies showed that it is effortless to recognize anger, happiness, surprise and dislike in facial expressions. However, sadness and fear are mainly revealed in audios[6]. This may also help explain the reason why we find it hard to identify sadness when cross labelling. Additionally, we realized that most of us found sad trailers has insignificant emotion arousing effect. This may due to the unclear context when the story is presented in the trailer since the main function trailers is to attract people to watch the films. Recently, people further utilize head movement, hand gestures and body movement to predict mixed emotions [7]. This shows that to cover different mixed emotions, using more sensors and a multimodal approach is a solution and will be further discussed in Section 1.6. However, the main idea for improvement here is that we should account for more affective states, so that along with more modalities we can predict a wider range of emotions in real applications.

## 1.3 Choice of labelling

Our videos with facial reactions were subjected to self-labelling and cross-labelling. In self-labelling, we annotated our recorded reaction to one trailer of each of the three genres individually. Cross labelling followed afterwards, which involved a different member in our group to re-examine the self-labelled reactions.

However, since we are from different ethnicities and cultures, our interpretation of our groupmates' facial expression may be incorrect. Studies revealed ethnic bias when people in different ethnicity label facial expressions of people from another ethnicity [4]. To improve, it will be better to recruit people of the same ethnicity to cross label our emotions. Some may find tempting to just process the videos using OpenFace, leveraging the Face Action Units (AU) and then look up a dictionary such as Facial Action Coding System Affect Interpretation Dictionary (FACSAID) [8] to label the emotions. However, the AUs that OpenFace has is just part of all the AUs [9]. Moreover, the facial behaviours in FACSAID are just tagged with meaning agreed by experts in Facial Action Coding System. So if we are building a system for a real application, looking up FACSAID may be insufficient given the deficiency in AUs that are recognized by OpenFace and the ethnicity problem described above. To conclude, it would be better to find people of the same ethnicity to cross label our emotions. However, it will be difficult to do in practice since there are people with one ethnicity born in another country and it brings up a whole new topic of whether a person's ethnicity has a bigger effect on the

accuracy of recognizing the emotions of people from the same ethnicity than the environment.

## 1.4 Sensor selection

Since our application uses facial expressions as inputs, we need sensors that can visually capture our facial responses with respect to our movie trailers, so a camera is a natural choice. Cameras are also used extensively to capture facial expressions in emotion studies such as [10][11][12]. Combining with OpenFace, our results shows that OpenFace can place the landmark accurately on faces and eyes with a confidence level of 98%, so we can see that using cameras for detecting facial expression can indeed generate precise data for further analysis and applications.

On the other hand, we also use cameras for heart rate (HR) extraction. One may challenge our approach of using non-contact measuring methods for HR rather than contact methods such as using watches equipped with electrocardiography (ECG). However, our decision of using a non-contact method is based on two main reasons. Firstly, we want participants to immerse themselves in the experiment setting as natural as possible. In this regard, using a non-contact method makes participants feel more comfortable as there is no device attached to any parts of their bodies. The second reason is because of the COVID-19 pandemic. Since there are various travelling restrictions keeping us away from accessing the devices on campus, we cannot normalize which device the participants are going to use.

We leveraged the iPhys toolbox [13], which has five different image Photoplethysmography (iPPG) methods for HR measurement. They are green channel (GREEN), independent component analysis (ICA), chrominance-based (CHROM), "plane-orthogonal-to-skin (POS) and ballistocardiogram (BCG). A study analysed the performance of all five methods and the results show that except BCG and GREEN, the other three methods have a good performance as evaluated in Table 1 using standard deviation (SD), mean absolute error (MAE), root mean squared error (RMSE) [14].

Table 1: Performance of different iPPG methods with the UBFC-RPPG dataset[1]

| Method | SD (BPM) | MAE (BPM) | RMSE (BPM) |
|---|---|---|---|
| GREEN | 11.090 | 4.469 | 11.598 |
| BCG | 26.113 | 27.922 | 37.962 |
| CHROM | 4.454 | 3.435 | 4.614 |
| POS | 6.501 | 2.436 | 6.608 |
| ICA | 8.258 | 3.507 | 8.635 |

The study also shows that the performance evaluated in RMSE increases with a shorter distance between the participant and the camera, and a higher resolution as shown in Figure 1.

[1]UBFC-RPPG is a public dataset designed for the non-contact HR measurement task

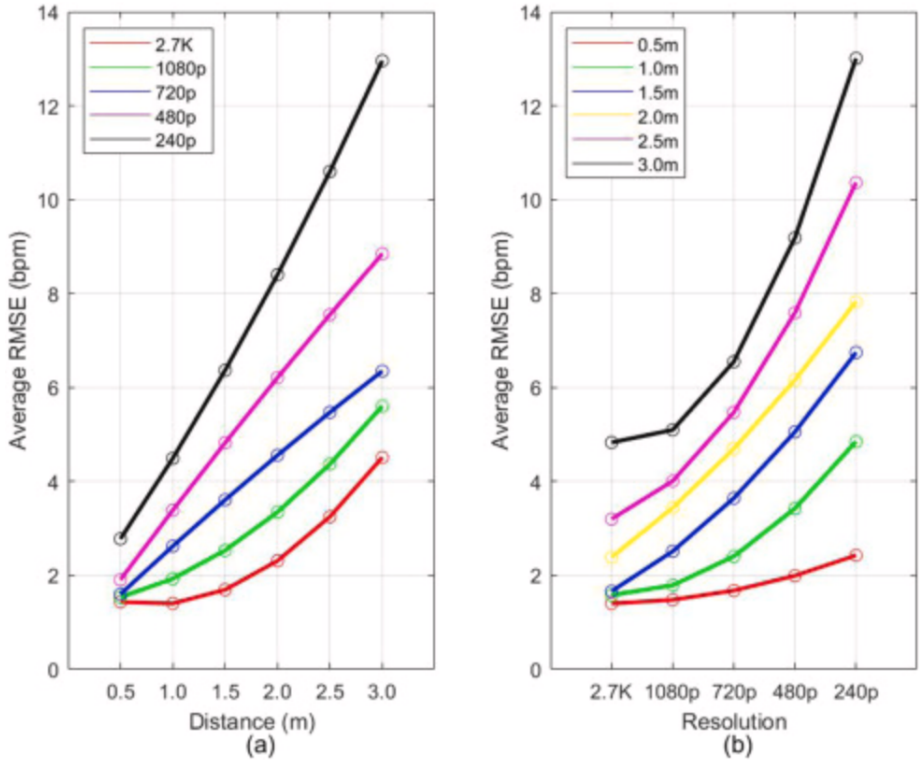


Figure 1: Performance evaluated in RMSE with respect to distance and resolution

Given the performance, capturing facial expression and measuring HR visually using camera is an effective and relevant sensor to capture emotion arousals.

## 1.5 Modality selection

To detect emotions, there are different modalities to choose from. In our project, we have chosen facial expressions and HR for the purpose. A study exploring the activity of different parts of our bodies when subjected to different emotion stimuli shows that most activities concentrate on our faces and chests [15]. Figure 2 shows the intensities of responses from different parts of our bodies in different emotional states as a heat map.

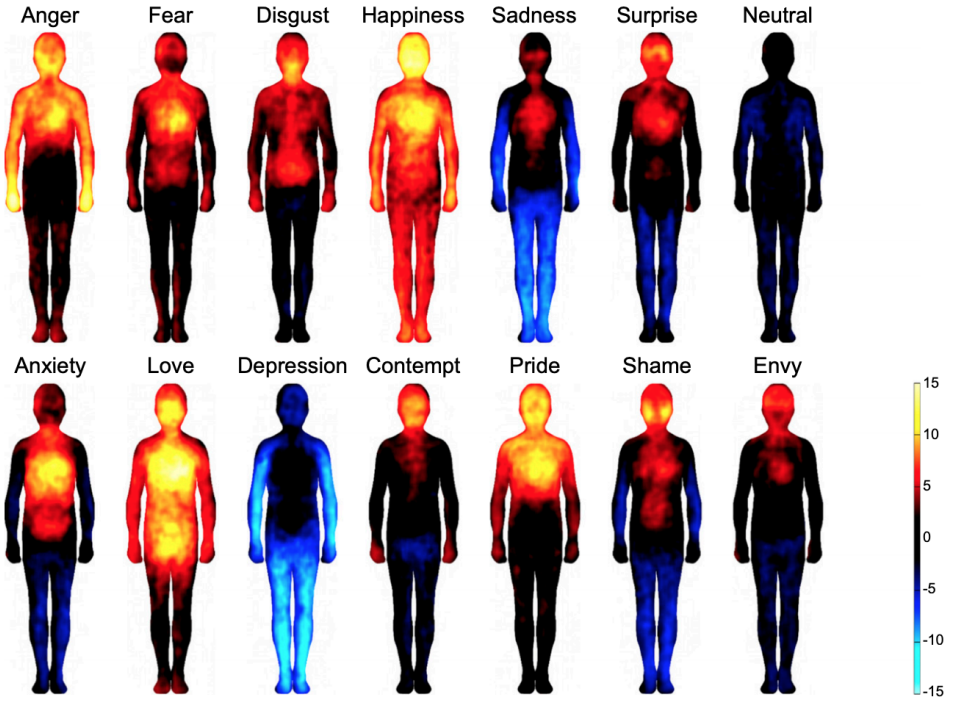


Figure 2: Body heat maps showing regions of activation when feeling different emotions[2]

Therefore, we can conclude that using facial expressions and HRs are relevant and effective modalities for emotion detection.

## 1.6 Improvements

Our data show significant variation among HR changes of different participant across the three genres of trailers. Large fluctuations of HRs can be observed when participants watching funny or sad trailers. In addition to the possibility that the data extracted contained outliers and noises, the different lighting effect from the background and head movement of different participants may also contribute to the variation since we are using a visual approach. Moreover, the resolution of our recordings are not standardized since we use different cameras. This raised another source of noise to the data as pointed out in [13].

Since this set of HR data shows few significant trends, we may need to mix different modalities. In Section 1.2, we mentioned that voice data incorporating facial expressions do perform well in predicting emotions. And since we have sad videos, adding voice data and a mic to our experiments will be a good choice. Hand gestures may also be a good modality to add since we can see in Figure 2 that hand activities are quite significant when we experience anger and happiness. However, we may need a better camera because the participants will need to sit further away from the camera so that we can capture their hands while maintaining good resolution.

[2]Colorbar displays the $t$-statistic range

# 2 Part II

In this part, we created an ARS to predict pain levels and protective behaviours at any moment of a subject in the EmoPain dataset [16].

## 2.1 Feature selection

In our study, we selected 3D coordinates of body joints and surface electromyography (sEMG) as input features, and use them to predict pain levels and protective behaviours.

A study using the EmoPain dataset [17] also selected similar features as input. However, they pre-processed the Euclidean positions to angles and energies since they are invariant to joint positions and has a better representation of body movement. For sEMG data, there are also studies like [18] [19] make use of sEMG to infer pain levels and protective behaviour.

This indicates that we should use features that are independent to the positions of joints instead of using the positional motion profile since the pain and protective behaviors should depend on the actions instead of the positions.

## 2.2 Feature Analysis

In our report, we briefly analyzed the EmoPain dataset and concluded that all the healthy patients will not exhibit any protective behaviour and unhealthy pain. On the other hand, chronic patients show pain at different levels, and they experience more pain as the difficulty of the exercise increases as in Table 2.

Table 2: 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 |

Here, we further investigate whether each feature is suitable for our predictions. Figure 3 and 4 show correlation coefficient heat maps of all the given data of P921D and P191N respectively, with all the motion coordinate data converted to the magnitude of acceleration. This is done because it makes more sense to compare pain data with the magnitude of the acceleration of joints compared to positions. Healthy participants are not shown here because they have no unhealthy pain and protective behavioural data recorded as shown in Table 2.

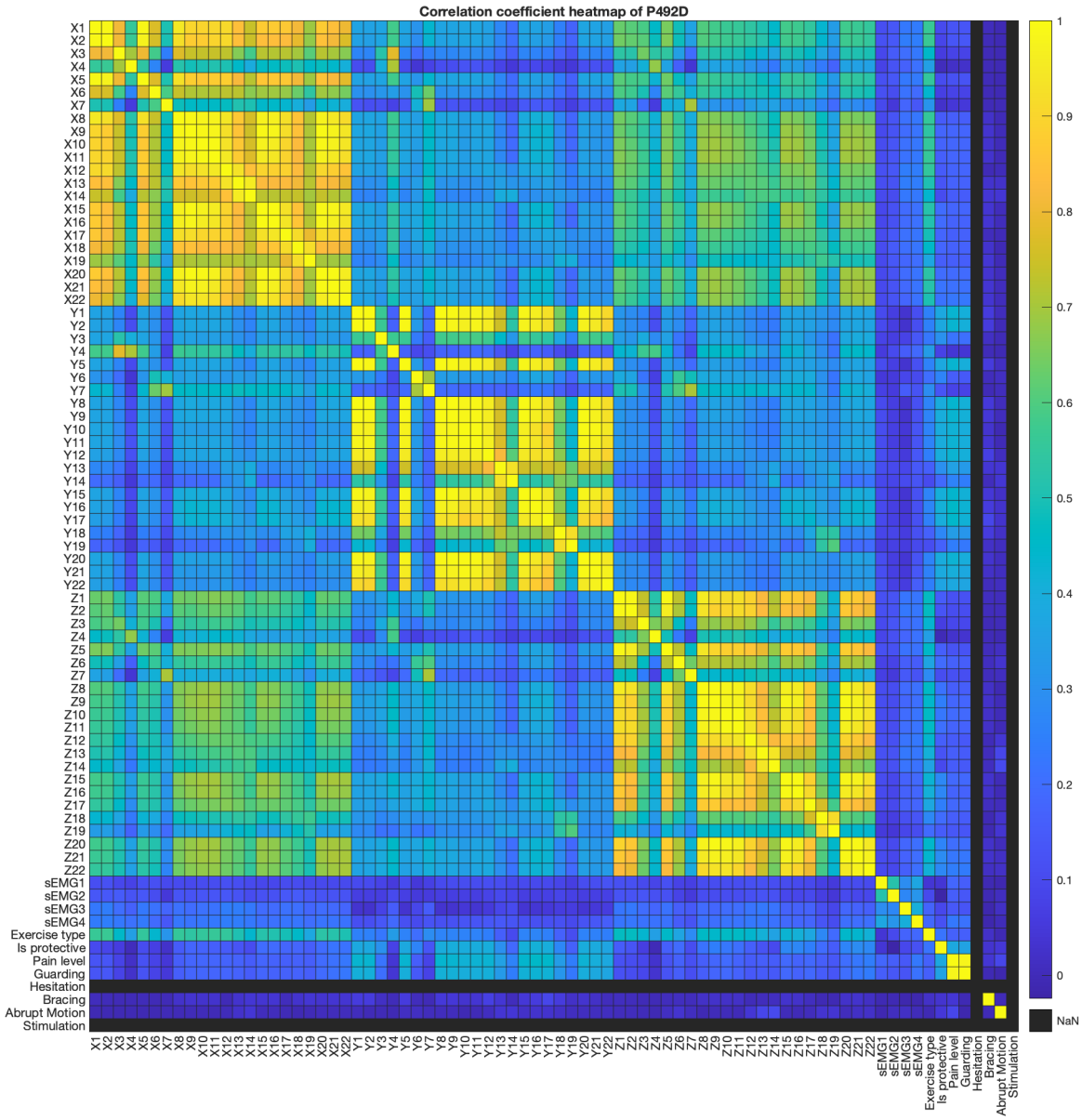


Figure 3: Correlation coefficient heat map of P492D with Euclidean positions converted to acceleration magnitudes

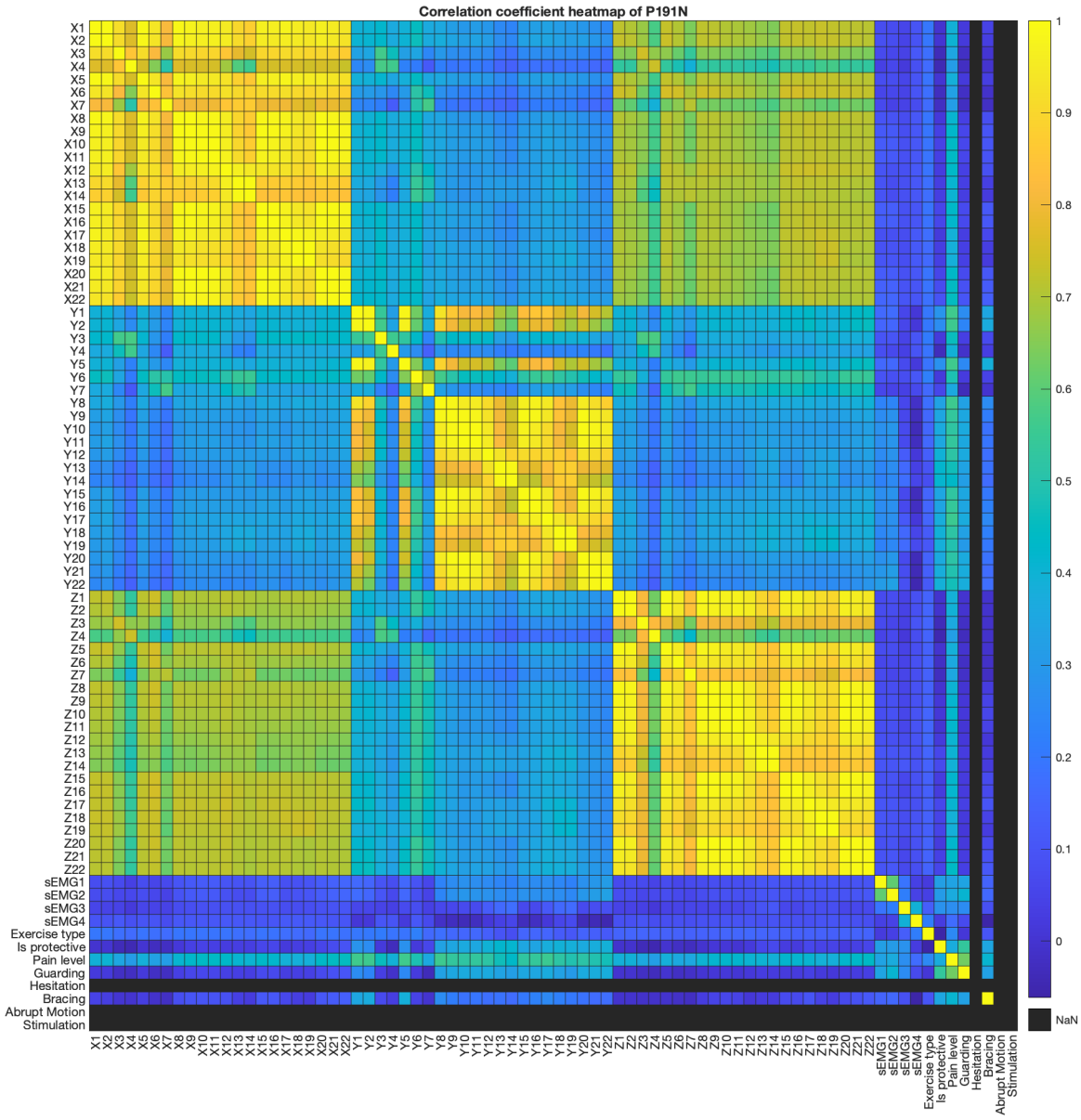


Figure 4: Correlation coefficient heat map of P191N with Euclidean positions converted to acceleration magnitudes

Noticing from the heat maps that the pain data is indeed positively correlated to our chosen features. A closer look reveals that acceleration in y-coordinates has higher correlations to the pain data, especially in Figure 3. This makes sense since we require the participants to sit down, stand up, etc. Albeit all these actions has a higher intensity of motions in the y-direction, it does not mean that pain is only associated with y-direction motions in general. If we require the participants to run, then the pain data may show more correlation with motions in other directions as well.

Since we used position information instead, we also present the correlation coefficient heat map of the same participants in Figure 5 and 6.

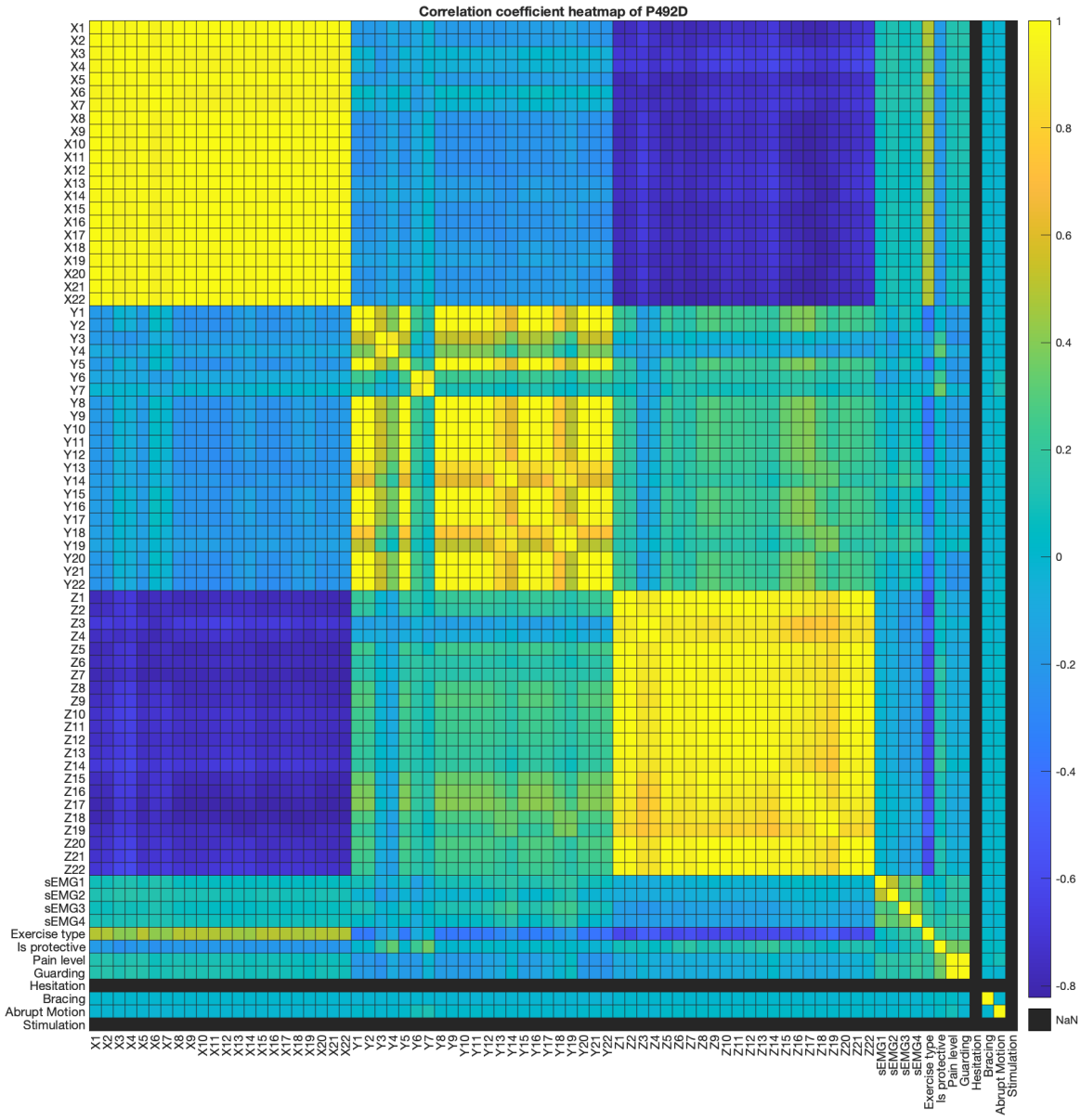


Figure 5: Correlation coefficient heat map of P492D

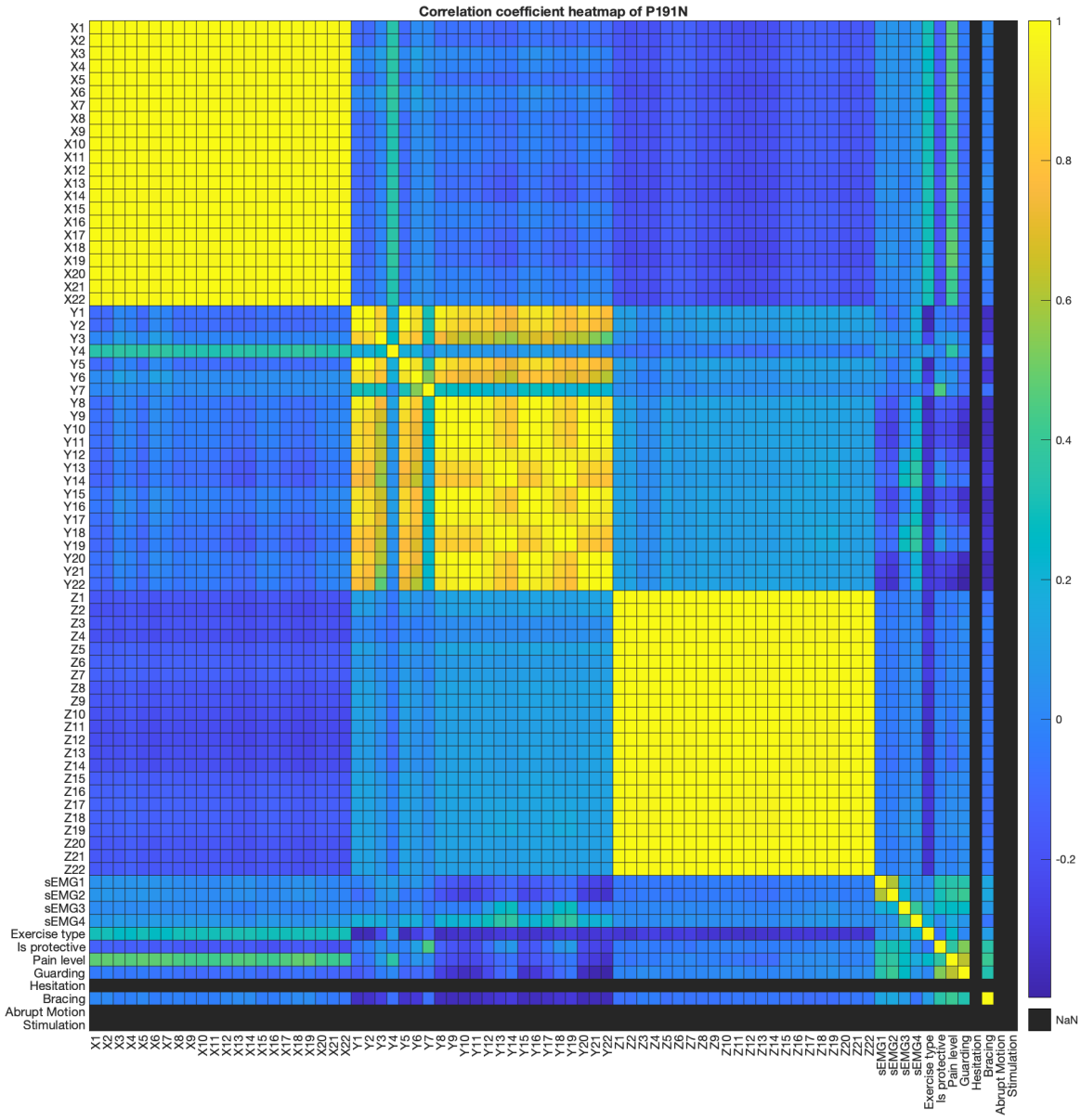


Figure 6: Correlation coefficient heat map of P191N

We observe that although the range of correlation of coefficients between the positions and pain and protective behaviours are similar, the one with accelerations gives a higher significance between the direction of the motion and the pain and protective behaviour data. This implies that accelerations are better features compared to the position ones.

## 2.3 Modelling

In our study, we fused the data into a feature vector sized 70 for every time frame and feed k steps into the Multi-layer Perceptron (MLP) model.

MLP has been a baseline architecture for time series according to [20], and can be trained on time series data as long as we use sliding window method as we did in our report. However, it is not a appropriate for our application. Since MLP ignores the fact that features are related between time steps, it violates the statistical learning principle that more information gives a better predictive performance. A better choice of model that suits our dataset is a type of Recurrent Neural Network (RNN) called Long Short Term Memory (LSTM) [21]. LSTM is known for its capability of learning dynamics in the dataset, so it is suitable for time series. There are also a number of studies considered LSTM for the EmoPain dataset [16] [22] [19]. Nonetheless, there are in general two drawbacks for LSTMs. Since LSTMs are neural networks, they usually require much more data and are difficult to interpret.

Moreover, the sliding window technique is fundamentally flawed in our context. In our analysis, we can observe that participants finish different actions in different durations. A fixed sliding window may cover just part of an action or multiple actions. Furthermore, our period is set to $\frac{1}{3}$ seconds which is actually not meaningful because people who can finish an action like sit down and stand up in $\frac{1}{3}$ basically do not exist, especially those who suffers chronic pain. On the other hand, if we increase the duration of the window, more parameters will be added to the model so there will be a higher chance of over-fitting. This is not an issue with RNN methods as the parameters are shared between frames.

Additionally, from Table 2 we observed that the portion of frames that are tagged with pain and protective behaviour are much less than that of the healthy ones. This means the model trained will lean more on predicting no-pain then pain. To solve this, one may consider balanced accuracy when we train the model [23], or train the model on different classes of the dataset.

## 2.4 Optimization

In our model, we used Adam Optimizer as it consistently gives a high training accuracy. Adam is known for its high efficiency and low memory requirements. In addition, Adam can also do well on large, non-stationary and sparse and noisy data. Moreover, it's interpretation is intuitive and require little tuning [24]. However, compared to Adam, other optimization algorithm like stochastic gradient descent (SGD) tends to concentrate more on datapoints and generalize in a better manner. So the performance of Adam is dependent on the data provided and the trade-off between speed and generalization [25].

## 2.5 Evaluation

We evaluate the our model using mean squared error (MSE). Since we are dealing with a classification problem, using MSE may not be an appropriate method. In the model, we use the `tanh(` activation function, which is non-convex with MSE function, and differentiating the activation function `tanh()` gives a large region with derivative equals to zero. If we insisted to use MSE as lost function with binary data, minimization is not guaranteed [26]. Therefore, we should use loss functions appropriate for binary and multiclass classification like categorical cross entropy and KL Divergence [27]. This can be used because if we treat the true labels and the predicted labels as two different distributions, we can measure the difference between these

two distributions by function like KL divergence. Furthermore, minimizing KL divergence is the same as minimizing cross entropy since KL divergence can be expressed in terms of entropies.

To conclude, using MLP with MSE loss function and sliding window on imbalanced data is not sufficient for our application and we should investigate into models that is capable of learning dynamics of the data like LSTM and use a loss function suitable for multi class classification like KL divergence. To compensate for data imbalance, we should consider using balanced accuracy or train the model on different classes of the dataset.

# References


[1] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, Jun. 2006. [Online]. Available: https://doi.org/10.1109/tcsvt.2006.873781

[2] L. Malär, H. Krohmer, W. D. Hoyer, and B. Nyffenegger, "Emotional brand attachment and brand personality: The relative importance of the actual and the ideal self," *Journal of Marketing*, vol. 75, no. 4, pp. 35–52, Jul. 2011. [Online]. Available: https://doi.org/10.1509/jmkg.75.4.35

[3] Y.-K. Kim and P. Sullivan, "Emotional branding speaks to consumers' heart: the case of fashion brands," *Fashion and Textiles*, vol. 6, no. 1, Feb. 2019. [Online]. Available: https://doi.org/10.1186/s40691-018-0164-y

[4] J. E. Kilbride and M. Yarczower, "Ethnic bias in the recognition of facial expressions," *Journal of Nonverbal Behavior*, vol. 8, no. 1, p. 27–41, 1983. [Online]. Available: http://dx.doi.org/10.1007/BF00986328

[5] Z. Gorvett, "There are 19 types of smile but only six are for happiness," 2017. [Online]. Available: https://www.bbc.com/future/article/20170407-why-all-smiles-are-not-the-same

[6] L. D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No.97TH8237)*. IEEE. [Online]. Available: https://doi.org/10.1109/icics.1997.647126

[7] A. S. Patwardhan, "Multimodal mixed emotion detection," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*. IEEE, Oct. 2017. [Online]. Available: https://doi.org/10.1109/cesys.2017.8321250

[8] J. C. Hager, "Description of facsaid - a tool for interpreting facial expressions." [Online]. Available: https://web.archive.org/web/20110520164308/http:/face-and-emotion.com/dataface/facsaid/description.jsp

[9] T. Baltrusaitis, "Action units," 2019. [Online]. Available: https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units

[10] H. Meyer, P. Wei, and X. Jiang, "Intelligent video highlights generation with front-camera emotion sensing," *Sensors*, vol. 21, no. 4, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1035

[11] B. T. Nguyen, M. H. Trinh, T. V. Phan, and H. D. Nguyen, "An efficient real-time emotion detection using camera and facial landmarks," in *2017 Seventh International Conference on Information Science and Technology (ICIST)*. IEEE, Apr. 2017. [Online]. Available: https://doi.org/10.1109/icist.2017.7926765

[12] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha, and A. Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," *CoRR*, vol. abs/2003.01062, 2020. [Online]. Available: https://arxiv.org/abs/2003.01062

[13] D. J. McDuff and E. B. Blackford, "iphys: An open non-contact imaging-based physiological measurement toolbox," *CoRR*, vol. abs/1901.04366, 2019. [Online]. Available: http://arxiv.org/abs/1901.04366

[14] R. Song, S. Zhang, J. Cheng, C. Li, and X. Chen, "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Computers in Biology and Medicine*, vol. 116, p. 103535, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519303944

[15] L. Nummenmaa, E. Glerean, R. Hari, and J. K. Hietanen, "Bodily maps of emotions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 2, pp. 646–651, 2014. [Online]. Available: https://www.pnas.org/content/111/2/646

[16] J. Egede, T. A. Olugbade, C. Wang, S. Song, N. Berthouze, M. F. Valstar, A. C. de C. Williams, H. Meng, M. H. Aung, and N. D. Lane, "EMOPAIN challenge 2020: Multimodal pain evaluation from facial and bodily expressions," *CoRR*, vol. abs/2001.07739, 2020. [Online]. Available: https://arxiv.org/abs/2001.07739

[17] C. Wang, T. A. Olugbade, A. Mathur, A. C. De C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data," in *Proceedings of the 23rd International Symposium on Wearable Computers*, ser. ISWC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 225–230. [Online]. Available: https://doi.org/10.1145/3341163.3347728

[18] Y. Li, S. Ghosh, and J. Joshi, "PLAAN: Pain level assessment with anomaly-detection based network," *Journal on Multimodal User Interfaces*, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s12193-020-00362-8

[19] X. Yuan and M. Mahmoud, "ALANet:Autoencoder-LSTM for pain and protective behaviour detection," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, Nov. 2020. [Online]. Available: https://doi.org/10.1109/fg47880.2020.00063

[20] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, Mar. 2019. [Online]. Available: https://doi.org/10.1007/s10618-019-00619-1

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[22] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. de C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, Oct. 2016. [Online]. Available: https://doi.org/10.1109/taffc.2015.2462830

[23] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, Aug. 2010. [Online]. Available: https://doi.org/10.1109/icpr.2010.764

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[25] N. Sharma, "Exploring optimizers in machine learning," *HEARTBEAT*, 2020. [Online]. Available: https://heartbeat.fritz.ai/exploring-optimizers-in-machine-learning-7f18d94cd65b

[26] R. Khan, "Why using mean squared error(mse) cost function for binary classification is a bad idea?" *towards data science*, 2019. [Online]. Available: https://towardsdatascience.com/why-using-mean-squared-error-mse-cost-function-for-binary-classification-is-a-bad-idea-933089e90df7

[27] C. Choy, "Regression vs. classification: Distance and divergence," Jan 2018. [Online]. Available: https://chrischoy.github.io/research/Regression-Classification/