2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)

A Curriculum Learning Approach for Pain Intensity Recognition from Facial Expressions

Adria Mallol-Ragolta[1], Shuo Liu[1], Nicholas Cummins[1] and Björn Schuller[1,2]
[1] Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] GLAM – Group on Language, Audio & Music, Imperial College London, UK
adria.mallol-ragolta@informatik.uni-augsburg.de

***Abstract*— The high prevalence of chronic pain in society raises the need to develop new digital tools that can automatically and objectively assess pain intensity in individuals. These tools can contribute to an optimisation of clinical resources, as they offer cost-effective solutions for early detection, continuous monitoring, and treatment personalisation by utilising Artificial Intelligence techniques. In this work, we present our contribution to the Pain Intensity Estimation from Facial Expressions task of the EMOPAIN 2020 Challenge. Specifically, we compare the performance of Recurrent Neural Networks trained with standard or Curriculum Learning (CL) approaches to predict the pain intensity level of individuals reported in an 11-point scale from facial expressions. The results obtained using the test partition support the use of CL-based approaches in the automatic prediction of pain from facial features. The best model trained using a CL approach achieved a Concordance Correlation Coefficient (CCC) of 0.196 in the test partition, while the model trained using a standard approach, without CL, achieved a CCC of 0.174. In terms of CCC, these results respectively represent an improvement of 0.136 and 0.114 on the best results of the baseline system reported by the Challenge organisers using the test partition.**

I. INTRODUCTION

Chronic pain is a major public health concern [11]. A study conducted in 2003 reveals that the prevalence of chronic pain among European adults is 19 % [5]. Similarly, the prevalence of US adults suffering from chronic pain was estimated in 2006 to be 20.4 % [7]. Therefore, there is a real need to develop digital tools for the automatic detection and recognition of pain. These solutions can contribute to early detection, continuous monitoring, and treatment personalisation for improving patients' health state and wellbeing.

Automatic pain detection has been widely investigated in the literature; for a recent review, the reader is referred to [25]. Targeted populations have included adults [15], children [28], and neonatal babies [6]. These studies have been undertaken in a variety of different conditions, e. g. , in laboratory [15] or *Intensive Care Unit* (ICU) settings [1].

Previous research in this area suggests that facial expressions can be used as a reliable indicator of pain [27], [19], [15]. In this direction, researchers have investigated the use of hand-crafted features [22], features extracted using deep learning [21], and even the fusion of both [9] to encode the information inferred from facial expressions for the automatic detection or recognition of pain. Machine learning techniques such as *Support Vector Machines* (SVM) [17], [12], [20] or Random Forests [26], and deep learning techniques such as *Convolutional Neural Networks* (CNN) and *Recurrent Neural Networks* (RNN) [18], [24], [13] were explored in this context.

Herein, we present our contribution to the Pain Intensity Estimation from Facial Expressions task of the EMOPAIN 2020 Challenge[1] [8]. In this work, we investigate the use of specific *Action Units* (AUs) of the *Facial Action Coding System* (FACS) [10] to train a RNN able to predict the pain intensity on individuals reported on a continuous 11-point scale. Specifically, we focus our analysis on the performance comparison of systems trained using standard or *Curriculum Learning* (CL) [4] approaches. To the best of the authors' knowledge, this paper represents the first time a CL approach has been considered in the automatic prediction of pain from facial expressions.

The rest of the paper is laid out as follows. Section II introduces the dataset utilised, while Section III describes the methodology followed. Section IV then presents the results obtained from the experiments performed. Finally, Section V concludes the paper and highlights some potential future work directions.

II. EMOPAIN DATASET

The data used in the current task of this Challenge belongs to the EmoPain dataset [2]. This dataset provides fully annotated multimodal data from individuals with *Chronic Lower Back Pain* (CLBP) with high resolution multiple-view facial videos, head-mounted and room audio signals, full-body 3D motion capture, and electromyographic signals from back muscles of both CLBP and healthy control participants (cf. Table II).

For the Challenge [8], only the features extracted from the facial videos have been made available to participants. The available facial features include facial landmarks, head pose, *Histogram of Oriented Gradient* (HOG) features, action unit intensity values and occurrence extracted with OpenFace [3], and deep-learnt feature representations extracted using VGG-16 [23] and ResNet-50 [14] pre-trained models.

[1]https://mvrjustid.github.io/EmoPainChallenge2020/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), and from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902. This joint undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

TABLE I

SUMMARY WITH THE TOTAL NUMBER OF VIDEOS AND FRAMES AVAILABLE IN BOTH TRAIN AND VALIDATION PARTITIONS OF THE EMOPAIN DATASET, IN ADDITION TO THE TOTAL NUMBER OF FRAMES ANNOTATED WITH ALL POSSIBLE PAIN INTENSITY SCORES.

	# Videos	# Frames	Pain Intensity Instances										
			0	1	2	3	4	5	6	7	8	9	10
Train	66	869 452	646 634	39 694	31 032	61 148	41 286	17 122	16 958	9 140	3 734	626	2 078
Validation	48	607 928	475 717	20 731	31 697	25 613	20 765	15 416	7 425	9 972	198	176	218
$\sum$	114	1 477 380	1 122 351	60 425	62 729	86 761	62 051	32 538	24 383	19 112	3 932	802	2 296

TABLE II

EMOPAIN DATASET CHARACTERISTICS REGARDING THE NUMBER OF CLBP AND HEALTHY CONTROL PARTICIPANTS AVAILABLE IN THE TRAIN, VALIDATION AND TEST PARTITIONS.

	CLBP Participants	Healthy Participants	$\sum$
Train	8	11	19
Validation	3	6	9
Test	3	5	8
$\sum$	14	22	36

TABLE III

DISTRIBUTION OF THE ANNOTATIONS SELECTED IN THE THREE DIFFERENT LEVELS FOR THE THREE DIFFERENT CONFIGURATIONS INVESTIGATED DURING NETWORK TRAINING WITH A CURRICULUM LEARNING APPROACH.

	First level	Second level	Third level
C1	[1,2,9,10]	[1,2,3,4,7,8,9,10]	[1,2,3,4,5,6,7,8,9,10]
C2	[1,2,9,10]	[1,2,3,4,7,8,9,10]	[0,1,2,3,4,5,6,7,8,9,10]
C3	[0,9,10]	[0,1,2,7,8,9,10]	[0,1,2,3,4,5,6,7,8,9,10]

Pain intensity levels on the recorded individuals were annotated in a continuous 11-point scale. On this scale, scores of zero are assigned to healthy participants, while scores of one to ten, both inclusive, are assigned to CLBP participants. The distribution of the annotations in the reported scale from both training and validation sets is summarised in Table I.

III. METHODOLOGY

This section describes the methodology followed in this work. Section III-A details the data conditioning performed, while Section III-B presents the architecture of the Neural Network implemented with the two different approaches considered. Finally, Section III-C details the post-processing applied to the predictions before the actual performance evaluation of the trained models.

A. Data Conditioning

Based on previous research suggesting the reliability of facial expressions as pain indicators [27], [19], [15] and despite the different facial feature sets available to the participants of the Challenge, we decided to focus this analysis exclusively on the use of the facial features representing the intensities of the *Facial Action Units* (FAUs) AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, and AU45. Although multiple cameras with different points of view were used during data collection, in this work, we treat the FAUs extracted from the different data sources equally.

In order to capture the dynamics of the facial expressions that contribute to the current pain intensity level, we select a window of features to model each pain intensity annotation, i. e. , in a many-to-one prediction manner. Specifically, features from previous and current 290 frames, which approximately corresponds to 5 minutes of the video signal, are used to predict the current annotation. Zero-padding on the features is used when accessing previous frames to the actual beginning of the recording. This strategy can contribute to an augmentation of the available data.

Pain intensity annotations are clearly imbalanced (cf. Table I) and, as a consequence, could risk the models' overfitting to the most populated class. In order to overcome this issue during training, we downsample the generated windows of features, without changing the information encapsulated inside, for every video separately. Specifically, we select as many samples as the least populated class for all reported pain intensity annotations in each video and their corresponding windows of features. For healthy participants, which only have 0-score pain intensity annotations, we select 1 % of the samples in each video.

B. Neural Network Architecture

To model the time dependencies of the facial expressions in the automatic prediction of pain, we implement a RNN with two consecutive *Long Short-Term Memory* (LSTM) RNN networks, with a dropout probability of 20 % between them, and followed by a fully connected layer. The implementation has been performed using the framework PYTORCH [16]. The first LSTM receives 17 features and learns an embedded representation of the input features in a 128-dimensional space. The second LSTM RNN receives this representation and learns a second embedded representation in a 64-dimensional

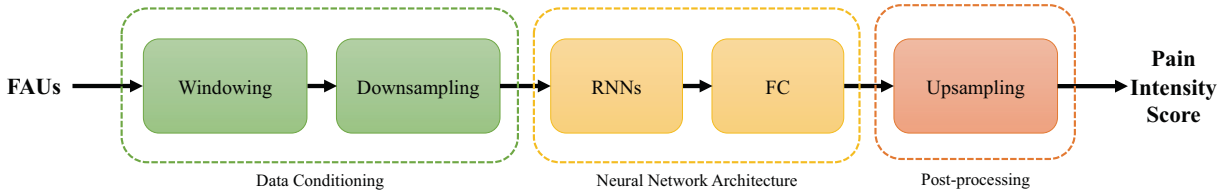

Fig. 1. Diagram of the baseline approach. Input data is firstly conditioned to predict the current pain intensity score in a many-to-one manner, and downsampled to overcome the imbalanced data. The resulting windowed features are then fed into a recurrent neural network followed by a fully connected layer to predict the current pain intensity score. The predicted scores are upsampled before model assessment to undo the effect of the downsampling performed in the data conditioning stage.

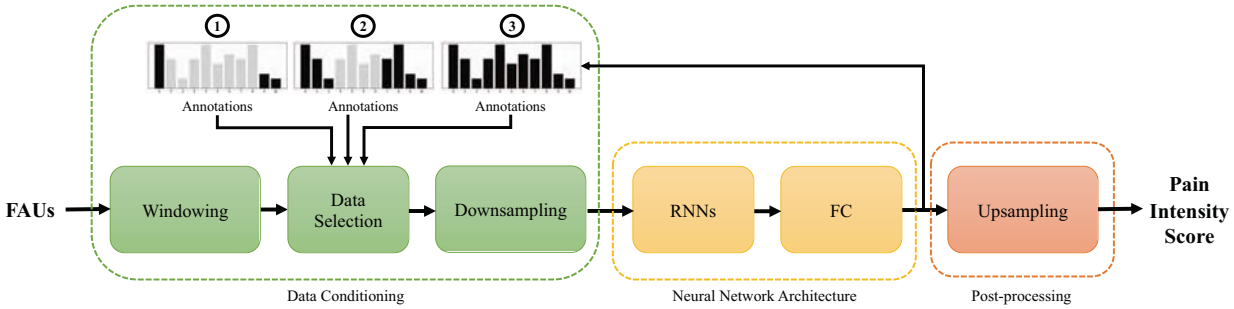

Fig. 2. Diagram of the Curriculum Learning approach. In this approach, as opposed to the baseline one, annotations are divided into different levels, starting from the most extreme annotations and progressively adding the intermediate and most complex ones. During training, each subset of the annotations and their corresponding windows of features are progressively fed to the recurrent neural network to learn a correlation between pain intensity scores and sequences of facial action units.

space. Finally, the fully connected layer receives the last 64-dimensional embedded representation to output a single value corresponding to the pain intensity label. The network is trained using the Concordance Correlation Coefficient (CCC) as the loss function to optimise with Adam as the optimiser, which uses a learning rate of 0.001. Network parameters are updated using a batch size of 16. Models are trained with a maximum of 50 epochs and an early-stopping method to stop training when the loss has not improved for 10 consecutive epochs.

1) Baseline Approach: This section describes the standard approach that we use as a baseline for our experiments (cf. Figure 1). Based on the data conditioning described in Section III-A, we use all the available downsampled samples as a whole to train the network. This is a standard approach with no particularities.

2) Curriculum Learning Approach: This approach uses a CL methodology [4], which is based on progressively adding complexity to the system to benefit the generalisation capabilities of the trained models. The models are trained iteratively with subsets of the annotations, and their corresponding features, starting from the easiest annotations to recognise to the most difficult ones.

In the context of our problem, this approach can be implemented by starting the network training with extreme annotations such as 0, 9, and 10, and then progressively adding the intermediate annotations. For this study, the annotations have been organised in three different groups or levels. Furthermore, in order to investigate performance differences in the organisation of the labels, we test three different configurations: C1, C2, and C3. The organisation of the labels for these three different configurations is summarised in Table III.

The procedure implemented to train neural networks with a CL approach is as follows (cf. Figure 2). First, we select the windows of features corresponding to the annotations belonging to the first level of labels, downsample them as described in Section III-A, and feed them to train the network. After convergence, either because of a local minimum in the validation partition or if the number of epochs allowed for training is exceeded, we select the windows of features corresponding to the second level and resume the training. As the last step and after model converge on the second level of annotations, we select the windows of features corresponding to the third and final level of labels and resume the training for the last time.

C. Post-processing

As the annotations do not change drastically over time, to speed up the time required to compute the outputs of the trained models, we downsample the data belonging to the validation and test partitions by selecting 1 in every 36 samples, which approximately corresponds to 38 seconds of the video signal. Due to the nature of the annotations, we aim to minimise the information loss by using this downsampling factor. The predictions are then replicated, so they are upsampled to match the length of the original

TABLE IV
MEAN ABSOLUTE ERROR (MAE), ROOT MEAN SQUARED ERROR (RMSE), PERSON CORRELATION COEFFICIENT (PCC), AND CONCORDANCE CORRELATION COEFFICIENT (CCC) COMPUTED BETWEEN THE ACTUAL AND PREDICTED PAIN INTENSITY SCORES IN BOTH VALIDATION AND TEST PARTITIONS WITH THE DIFFERENT APPROACHES INVESTIGATED.

Approach	Partition	MAE	RMSE	PCC	CCC
Baseline Approach	Validation	1.761	2.446	.175	.163
	Test	1.641	2.058	.187	.174
CL Approach - C1	Validation	2.048	2.600	.168	.125
	Test	3.269	3.881	.071	.026
CL Approach - C2	Validation	2.072	2.608	.203	.147
	Test	1.600	2.122	.216	.196
CL Approach - C3	Validation	2.711	3.183	.196	.097
	Test	–	–	–	–

annotations. Finally, the performance of the trained models is assessed by computing the *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Pearson Correlation Coefficient* (PCC), and *Concordance Correlation Coefficient* (CCC) between the actual and predicted pain intensity scores.

IV. EXPERIMENTAL RESULTS

The results obtained on the validation partition highlight the significance the configuration of the annotations in the different levels have in the system performance when using CL approaches (cf. Table IV). The model trained with C3 configuration, which uses data from healthy participants from the beginning of training, obtained the worst performance with a CCC of 0.097. This result highlights the importance of using balanced data to train neural networks, as overusing the most common class might bias the model performance. The C1 configuration improves the performance of the trained model, with a CCC of 0.125. Our strongest CCC score is obtained using the C2 configuration, which achieved a CCC of 0.147.

Analysing these results, one can claim that the assignation of the annotations in the different levels impacts the system performance. Furthermore, one can argue that it is important to use samples from all classes represented in the dataset, and that imbalanced data might negatively impact system performance. In terms of MAE, RMSE and CCC, models trained with CL approaches obtain worse results than the baseline, which does not use CL. Nevertheless, the C2 and C3 configurations surpass the baseline in terms of PCC. Our results are not directly comparable to those reported by the Challenge organisers [8], as we have evaluated different features. Nevertheless, our results do not beat the best baseline system of the Challenge organisers, which achieved a CCC of 0.180 using the validation partition.

As participants in the Challenge, we were able to submit three different models for evaluation with the test partition. We submitted the model trained using the baseline approach, and the CL models trained using C1 and C2 configuration labels, as these are the models that achieved the highest CCC scores when assessed using the validation partition. The models submitted to the Challenge use data from both train and validation partitions for training.

Analysing the results obtained using the test partition (cf. Table IV), we can observe that the model trained with CL and C1 configuration obtained the worst result, with a CCC of 0.026. The baseline model, which does not use CL, achieved a CCC of 0.174, and was surpassed by the model trained using CL and C2 configuration with a CCC of 0.196. Furthermore, the C2 configuration model surpasses the baseline approach in terms of MAE and PCC. Despite not being directly comparable, our best model also surpasses the baseline system of the Challenge organisers evaluated using the test partition, which achieved a CCC of 0.060.

V. CONCLUSIONS AND FUTURE WORK

This paper outlined our contribution to the Pain Intensity Estimation from Facial Expressions task of the EMOPAIN 2020 Challenge, in which we investigated the automatic recognition of pain from a specific set of FAUs. We focused this study on analysing the effect of training neural networks with CL approaches. The results obtained using both validation and test partitions on models trained with CL approaches highlight the importance of organising the annotations in the different levels of training, as they impact the overall system performance. Although the model trained without CL obtained the best CCC score using the validation partition (0.163), the best CCC score using the test partition was achieved by a model trained with CL (0.196). Therefore, we can conclude that the use of CL in the automatic prediction of pain from facial expressions is advantageous.

Further directions to keep on this research include the use of other facial features to study the effect that different feature sets with different dimensionalities have on the system performance. Adding levels in the labels' configuration in order to increase the progressive training of the networks can also be investigated, in addition to the use of upsampling strategies to overcome the data imbalance.

REFERENCES

[1] A. Ashraf, A. Yang, and B. Taati. Pain Expression Recognition Using Occluded Faces. In *Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition*, Lille, France, 2019. IEEE. 5 pages.

[2] 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. d. 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*, 7(4):435–451, 2016.

[3] T. Baltrušaitis, P. Robinson, and L. Morency. OpenFace: an Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Lake Placid, NY, USA, 2016. IEEE. 10 pages.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, Montreal, QC, Canada, 2009. ACM.

[5] H. Breivik, B. Collett, V. Ventafridda, R. Cohen, and D. Gallacher. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain*, 10(4):287–333, 2006.

[6] S. Chen, F. Luo, X. Chen, J. Yan, Y. Zhong, and Y. Pan. A Video Database of Neonatal Facial Expression based on Painful Clinical Procedures. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6565–6568, Berlin, Germany, 2019. IEEE.

[7] J. Dahlhamer, J. Lucas, C. Zelaya, R. Nahin, S. Mackey, L. DeBar, R. Kerns, M. Von Korff, L. Porter, and C. Helmick. Prevalence of Chronic Pain and High-Impact Chronic Pain Among Adults – United States, 2016. *Morbidity and Mortality Weekly Report*, 67(36):1001–1006, 2018.

[8] J. Egede, S. Song, T. Olugbade, C. Wang, H. Meng, M. Aung, N. D. Lane, A. C. D. C. Williams, M. Valstar, and N. Bianchi-Berthouze. EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions, 2020. arXiv preprint arXiv:2001.07739.

[9] J. Egede, M. Valstar, and B. Martinez. Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 689–696, Washington, DC, USA, 2017. IEEE.

[10] P. Ekman and W. V. Friesen. *The Facial Action Coding System (FACS): A technique for the measurement of facial actions*. Consulting Psychologists Press, 1978.

[11] D. S. Goldberg and S. J. McGee. Pain as a global public health priority. *BMC Public Health*, 11:770, 2011.

[12] Z. Hammal and J. F. Cohn. Automatic Detection of Pain Intensity. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 47–52, Santa Monica, California, USA, 2012. ACM.

[13] M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O. K. Andersen, E. G. Spaich, and T. B. Moeslund. Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 250–257, Xi'an, China, 2018. IEEE.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, Nevada, USA, June 2016. IEEE.

[15] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Proceedings of the 9th International Conference on Automatic Face and Gesture Recognition and Workshops*, pages 57–64, Santa Barbara, CA, USA, 2011. IEEE.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019.

[17] Patrick, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[18] L. L. Presti and M. L. Cascia. Boosting Hankel matrices for face emotion recognition and pain detection. *Computer Vision and Image Understanding*, 156:19–33, 2017.

[19] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *PAIN*, 139(2):267–274, 2008.

[20] N. Rathee and D. Ganotra. Multiview Distance Metric Learning on facial feature descriptors for automatic pain intensity detection. *Computer Vision and Image Understanding*, 147:77 – 86, 2016.

[21] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Transactions on Cybernetics*, 2017. preprint, 11 pages.

[22] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh. An Approach for Automatic Pain Detection through Facial Expression. *Procedia Computer Science*, 84:99–106, 2016.

[23] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015. ICLR. 14 pages.

[24] J. Soar, G. Bargshady, X. Zhou, and F. Whittaker. Deep Learning Model for Detection of Pain Intensity from Facial Expression. In M. Mokhtari, B. Abdulrazak, and H. Aloulou, editors, *Proceedings of the 16th International Conference on Smart Homes and Health Telematics, Designing a Better Future: Urban Assisted Living*, pages 249–254, Singapore, Singapore, 2018. Springer.

[25] R. A. Virrey, C. D. S. Liyanage, M. I. bin Pg Hj Petra, and P. E. Abas. Visual data of facial expressions for automatic pain detection. *Journal of Visual Communication and Image Representation*, 61:209–217, 2019.

[26] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue. Automatic Pain Recognition from Video and Biomedical Signals. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pages 4582–4587, Stockholm, Sweden, 2014. IEEE.

[27] A. C. d. C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439 – 455, 2002.

[28] X. Xu, K. D. Craig, D. Diaz, M. S. Goodwin, M. Akcakaya, B. T. Susam, J. S. Huang, and V. R. de Sa. Automated Pain Detection in Facial Videos of Children Using Human-Assisted Transfer Learning. In *Proceedings of the International Workshop on Artificial Intelligence in Health*, pages 162–180, Stockholm, Sweden, 2018. Springer.