Using Explainable AI to Identify Differences Between Clinical and Experimental Pain Detection Models Based on Facial Expressions

Pooja Prajod(✉), Tobias Huber, and Elisabeth André

Chair for Human-Centered Artificial Intelligence, Augsburg University, Augsburg, Germany
{pooja.prajod,tobias.huber,elisabeth.andre}@uni-a.de

Abstract. Most of the currently available pain datasets use two types of pain stimuli - people with clinically diagnosed conditions (e.g. surgery) performing tasks that cause them pain (we call this clinical pain) and pain caused by external stimuli such as heat or electricity (we call this experimental pain). In high-risk domains like healthcare, understanding the decisions and limitations of various types of pain recognition models is pivotal for the acceptance of the technology. In this paper, we present a process based on Explainable Artificial Intelligence techniques to investigate the differences in the learned representations of models trained on experimental pain (BioVid heat pain dataset) and clinical pain (UNBC shoulder pain dataset). To this end, we first train two convolutional neural networks - one for each dataset - to automatically discern between pain and no pain. Next, we perform a cross-dataset evaluation, i.e., evaluate the performance of the heat pain model on images from the shoulder pain dataset and vice versa. Then, we use Layer-wise Relevance Propagation to analyze which parts of the images in our test sets were relevant for each pain model. Based on this analysis, we rely on the visual inspection by a human observer to generate hypotheses about learned concepts that distinguish the two models. Finally, we test those hypotheses quantitatively utilizing concept embedding analysis methods. Through this process, we identify (1) a concept which the *clinical* pain model is more strongly relying on and, (2) a concept which the *experimental* pain model is paying more attention to. Finally, we discuss how both of these concepts are involved in known pain patterns and can be attributed to behavioral differences found in the datasets.

Keywords: Automatic pain detection · Explainable Artificial Intelligence · Cross-database evaluation · Facial expression recognition

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847926 MindBot and from the DFG under project number 392401413, DEEP.

1 Introduction

Expressing pain is an important social component as it triggers social reactions such as empathy and care [3]. In clinical practice, recognizing pain facial expression helps in pain diagnosis and eliminates the need for verbalization. This is especially important for patients who cannot provide verbal pain reports like people with dementia, infants, and ventilated patients [3]. So, assessing facial expressions is crucial for many healthcare applications and also a valuable skill for medical staff [3,18]. In many of these cases, caregivers have to routinely monitor the pain levels of a patient for optimal pain management. However, this is often not possible due to practical issues like a lack of available clinical staff. As a consequence, there has been an increasing interest in developing methods to automatically detect pain from facial expressions.

Many works [22,26] have proposed models for automatic pain recognition. These models are shown to achieve good performances but are usually trained and tested on the same dataset. Cross-database evaluations are not very common. According to Othman et al. [16], this could be because well-trained models tend to perform poorly on other databases. This was observed in [6], where the authors found that a model trained using one pain dataset performed poorly on another dataset. Real-world scenarios like hospitals or nursing homes will inevitably be different from the settings used to collect the training datasets. Therefore, cross-database evaluation is important to ensure the robustness of the models and verify that they are not learning database-specific patterns.

Another drawback of many state-of-the-art pain recognition models is that they come in the form of deep neural networks whose decisions are usually incomprehensible to humans. The research area of Explainable Artificial Intelligence (XAI) aims to make such "black-box" models more understandable. Explainability is crucial for deploying pain recognition models to support therapy, as patients and therapists should be able to understand the system's decisions in order to achieve successful treatment [3]. Petyaeva et al. [17], for instance, showed that briefly training medical staff on pain observation scales and therefore increasing the staff's understanding of the scales, led to more frequent and more confident use of the scales in everyday care. This indicates that similar improvements can be expected from increasing the comprehensibility of automatic pain recognition models. A second benefit of increasing the explainability of pain recognition models is that it can help to mitigate some of the problems of the models that do not perform well on cross-database evaluations. On the one hand, understanding the reasoning and biases of models trained on different datasets can help to create less biased models and datasets in the future. On the other hand, pain recognition models are often employed in ensembles of several pain recognition models. Here, it is crucial to understand the reasoning and biases of each model to judge which models can be trusted in a given situation.

As a first step to solve the aforementioned problems, we study two pain recognition models in this paper - one trained on clinical pain images from the UNBC shoulder pain database and the other on experimental pain images from the BioVid heat pain database. In addition to the typical within-database

evaluation, we perform cross-database evaluation of these models. Furthermore, we utilize XAI techniques to find more detailed differences between the models' reasoning and biases. As far as we know, this is the first time that XAI is used for cross-database evaluation of pain models. First, we generate saliency maps that highlight which areas of input images were important for the decision of each model. Based on a manual inspection of these saliency maps, we formulate hypotheses about the concepts learned by the models. Then, we use concept embedding analysis to test our hypotheses. Finally, we discuss our results in light of underlying behavioral differences between the datasets.

2 Related Work

2.1 Cross-database Evaluation of Pain Datasets

Cross-database evaluations involve evaluating a model on samples from different databases than the one used for training. Such evaluations are crucial for building robust models and testing the generalization capabilities of a model. In [16], the authors used the BioVid heat pain database and the X-ITE pain database (thermal pain) to train pain recognition models. They trained pairs of pain recognition models using facial activity descriptors based on random forest classifiers and CNNs. They performed cross-database evaluation of the models using the two datasets. They found that models trained using both methods performed well in cross-database evaluations. It can be noted that both BioVid and X-ITE databases use the same stimuli (varying temperatures) to induce pain.

Dai et al. [6] studied various combinations of emotion datasets and the UNBC shoulder pain dataset to train a real-time pain detection model. They also tried CNNs and SVMs based on Action Units (AUs) - visible indicators of the activity of individual facial muscles. In addition to the within-dataset evaluations, they tested the models on datasets consisting of posed emotion and pain expressions. They found that even though the CNN models performed extremely well in within-database evaluations, all posed pain images were classified as no-pain. They concluded that CNNs learned database-specific features which enabled them to predict pain in subjects from the UNBC database. The only model that performed well in real-time posed pain recognition was an AU-based SVM trained on AffectNet and UNBC shoulder pain images. They tested this model on images of 20 randomly chosen participants from the BioVid heat database. The model performed poorly which prompted the authors to examine their test images. They found that many of the participants closed their eyes for most parts of the experiment. Some even had closed eyes for the entire experiment. Since closed eyes are an indicator of pain in the UNBC dataset, they attribute the poor performance of the model to this difference in behavior between datasets.

2.2 Explainable Artificial Intelligence

In scenarios where neural networks are used to support medical therapy, patients and therapists must be able to understand the system's decisions in order to

achieve successful treatment. A common way of analyzing the predictions of neural networks trained on images is the creation of so-called *saliency maps* that highlight how important each pixel of an input image was for the prediction. For the specific use-case of pain recognition, Weitz et al. [23] applied and compared two different saliency map methods: Layer-wise Relevance Propagation (LRP) [15] and Local Interpretable Model-agnostic Explanations (LIME) [19]. While the authors found that the salience maps generated in their work provide some initial insights into the reasoning of the network, they concluded that saliency maps in their current form are often ambiguous and hard to interpret for end-users. Besides saliency maps, another promising explanation approach is the generation of counterfactual images that show how an input image could be modified to change the network's prediction. For medical applications, Mertes et al. [13] generated such counterfactual explanations for a pneumonia detection network and found in a user study that those counterfactual explanations were easier to interpret than LRP and LIME saliency maps. However counterfactual explanations still heavily rely on the final interpretation by human users.

To reduce the amount of interpretation that has to be done by the user, recent work on *concept embedding analysis* investigates which human-comprehensible concepts were learned by a given network. Bau et al. [4] show that semantic concepts are often embedded in individual neurons of the latent space of a neural network. For example, Khorrami et al. [8] demonstrate that certain neurons in the final convolutional layer of a network trained to analyze facial expressions learned to recognize specific AUs. To extend this method to concepts that might be embedded in multiple neurons within the latent space, Kim et al. [9] trained a binary linear classifier that takes as input the output of an intermediate layer of the network and learns to recognize a predefined concept. If the linear classifier can recognize the concept, then it is likely that the concept is embedded in the intermediate layer that acts as input to the linear classifier. They tested their approach on multiple image classification networks and a network for predicting diabetic retinopathy. A common challenge for the aforementioned concept embedding analysis techniques is that the potential concepts have to be externally identified by human experts. To mitigate this effort, Prajod et al. [18] utilized LRP saliency maps to facilitate the identification of potential concepts.

3 Approach

3.1 Datasets

UNBC-McMaster Shoulder Pain Expression Database [12] - This database contains image sequences of 25 participants with shoulder pain performing a range of arm movements. Each image is annotated with a Prkachin and Solomon Pain Intensity(PSPI) score on a scale of 0 (no pain) to 15 (extreme pain). Since these images are from a video, many of them are similar. We remove the redundant images through the approach followed in [26]. For each image sequence, whenever the pain intensity doesn't change for more than five consecutive images, we keep only the first image. From the down-sampled dataset, we

reserve images belonging to four participants, who gave consent to publish their images, as test set. The images belonging to one randomly chosen participant are used for validation and the images of the remaining 20 participants for training.

BioVid Heat Pain Database [21] - We use the part A of this database that contains short videos showing facial expressions of 87 participants reacting to heat pain stimuli. Each participant has 20 short videos (5.5s long) for each of five conditions (no pain + four pain intensities). In [25], the authors found that the initial two pain intensities failed to trigger a facial response in many participants and use only the highest intensity for discerning pain vs. no pain. Based on their findings, we only consider videos that are labeled as baseline (no-pain) and highest pain level. The authors also found that the facial activity for the highest pain level starts at around 2s and peaks around the 4s mark. So, we choose the frame at 4s in the video as a representative image. As suggested by the creators of the BioVid dataset [1], we exclude 20 participants who did not have a visible reaction to the stimuli. Among the remaining 67 participants, 15 participants were reserved for testing, five participants were used for validation and the images of the remaining participants formed the training set.

3.2 Pain Training

After selecting representative images from the videos or image sequences (see Sect. 3.1), the resulting datasets are relatively small with around 1000–2000 images. This is typical for pain datasets [22], but deep learning usually requires larger amounts of training data. To mitigate this problem, deep learning models are often trained using *transfer learning*. This involves re-using the knowledge that the model learned for a specific task A, for training an adjacent task B. In this paper, we adopt the transfer learning process from [18]. The idea is to fine-tune an emotion recognition model to discern between pain and no-pain images. We use the same emotion recognition model as Prajod et al. [18] which is a VGG-16 based convolutional neural network (CNN) trained on the Affect-Net dataset [14]. This dataset consists of 420299 face images that are manually annotated with 11 emotions. As described by Prajod et al. [18] we remove images belonging to 'None', 'Uncertain' and 'Non-face'. Afterward, we modify the prediction layer of the model to predict pain vs. no-pain. To train the model for pain recognition, we fine-tune all layers of this model. We train two pain models - one trained on clinical pain images (UNBC shoulder pain dataset) and one on experimental pain images (BioVid heat pain dataset). Both models are trained using SGD optimizer (learning rate = 0.01) and focal loss [11] given by:

$focal\_loss = (1 - p_t)^{\gamma} \times cross\_entropy\_loss$. The variable p_t is the predicted probability of a sample belonging to its true class (t) and we set the hyperparameter $\gamma = 2$.

Before passing an image through our models, we detect and crop the face using OpenCV. We also scale them to the default VGG16 input dimensions (224×224). While training both the models, we use Keras data augmentation options: rotation ($[-25^o, 25^o]$), height shift ($[-10\%, 10\%]$), width shift

([−10%, 10%]), shear ([−10%, 10%]), zoom ([−10%, 10%]) and horizontal flip. We train the models using NVIDIA GeForce GTX 1060 6GB GPU.

Unlike the BioVid dataset, the UNBC dataset is imbalanced. So, for the clinical pain model, we use weighted focal loss. We follow the weighting scheme proposed in [5], where weights of classes are computed as (we set the hyperparameter $\beta = 0.99$): $weighted\_loss = \frac{1-\beta}{1-\beta^{samples\_per\_cls}} \times focal\_loss$.

3.3 Cross-database Evaluation

Many works propose automatic pain recognition models that achieve good performances on the datasets they are trained on. However, cross-dataset evaluations are less explored. One reason for this might be that well-trained models may not perform well on other datasets [16]. So, in addition to the typical performance evaluation, we perform a cross-dataset evaluation of both our models and determine the generalization capabilities of these models. The generalizability of a model is particularly important when deploying it in real-world applications. First, we evaluate the performance of the models in terms of f1-score, recall, precision, and accuracy. Since the UNBC dataset is not balanced, we compute the macro-average of these metrics - compute the metric for each class and average them. In this step, the evaluation is within-database i.e., training and test images come from the same database. After the within-database evaluation, we perform the cross-database evaluation. Here we compute the same performance metrics as before, but for the test set derived from the other database. If the model's cross-database performance is comparable to its initial performance, we say the model learned generic pain features and not dataset-specific features.

3.4 Visual Analysis

After determining if the models learned generic pain features, we explore the differences in the features that were relevant for each model. We follow the technique proposed in [18] to visually inspect these differences. To this end, we generate saliency maps highlighting the areas of the input image which, according to each model, are indicators of pain (i.e. were important for the pain prediction neuron). For generating the saliency maps, we use the iNNvestigate [2] implementation of LRP with the z-rule for fully connected layers and the z^+-rule for convolution layers. This composite LRP method is relatively robust to sanity checks [20] and retains the conservation property of LRP which states that the relevance values of all pixels sum up to the prediction value. The conservation property is important for an accurate comparison of different saliency maps. We use our test sets from both the UNBC and BioVid datasets as input images. For each input image, we obtain two saliency maps - one from the clinical pain model and the other from the experimental pain model. To better highlight the differences between these models, we subtract the raw saliency maps from each other and normalize the differences between 0 and 1. With this method, we obtain saliency maps that highlight the areas that the experimental pain model payed more attention to than the clinical pain model and vice versa. We

manually inspect these images to derive hypotheses about potential differences in the concepts that were relevant for the clinical and experimental pain models.

3.5 Concept Embedding Analysis

In this section, we describe our method of verifying the hypothesis that the concepts identified with the method described in Sect. 3.4 actually distinguish the models trained on clinical and experimental pain. To this end, we follow the approach of Prajod et al. [18] and Kim et al. [9] and train binary linear classifiers on the output of an intermediate layer of each model to investigate how well these concepts are embedded. We use our test sets of both UNBC and BioVid datasets as training data for the linear classifiers. For each image in the combined test set, two of the authors manually annotate whether the concept is present or not. Afterward, we check where the two annotators disagree. Those images are additionally labeled by a third annotator, who is not involved with the paper, and the final label is chosen by majority vote.[1] The authors have experience with facial affect recognition models which was sufficient for annotating the specific concepts we identified in our experiment. By computing the output of the last pooling layer of each pain model for all images within our combined test set, we obtain an experimental pain and a clinical pain feature-set that represent the latent space of the respective pain recognition models. We then train a linear Support Vector Machine (SVM) on the task of detecting the concept candidate on each of those two feature sets. For this training, we use 2-fold cross-validation and compute the average f1-score of the two folds. This training process is repeated for 500 iterations using different random seeds for fold image selection and weight initialization. Finally, we run a paired t-test between the 500 averaged F1 scores of each feature-set. This comparison method is suggested in [7] for five iterations as 5×2 cross validation paired t-test and we extend it to 500 iteration as suggested by Kim et al. [9]. The result of this test shows whether there is a significant difference between the performance of SVMs trained on the clinical pain feature-set and the SVMs trained on the experimental pain feature-set. If there is a significant difference then it is likely that there is a difference in the quality of the embedding of the concept candidate between the latent spaces of the two pain models.

4 Results

As described in Sect. 3.2, we train two pain recognition models based on a clinical pain dataset (UNBC shoulder pain) and an experimental pain dataset (BioVid heat pain). Tables 1 and 2 show the results of within-database and cross-database evaluations of the clinical pain model and the experimental pain model, respectively. The within-database accuracy of both our models are comparable to other papers (clinical pain: 85% [6], experimental pain: 66% [16]).

[1] The final annotations are available upon request to the authors.

Table 1. Performance of clinical pain model

Test images	Precision			Recall			F1-score			Accuracy
	No pain	Pain	Avg.	No pain	Pain	Avg.	No pain	Pain	Avg.	
Clinical pain	0.74	0.87	0.80	0.71	0.88	0.80	0.72	0.88	0.80	0.83
Experimental pain	0.74	0.56	0.65	0.28	0.90	0.59	0.41	0.69	0.55	0.59

Table 2. Performance of experimental pain model

Test images	Precision			Recall			F1-score			Accuracy
	No pain	Pain	Avg.	No pain	Pain	Avg.	No pain	Pain	Avg.	
Clinical pain	0.48	0.87	0.68	0.80	0.61	0.71	0.60	0.72	0.66	0.67
Experimental pain	0.65	0.80	0.73	0.87	0.54	0.70	0.74	0.64	0.69	0.70

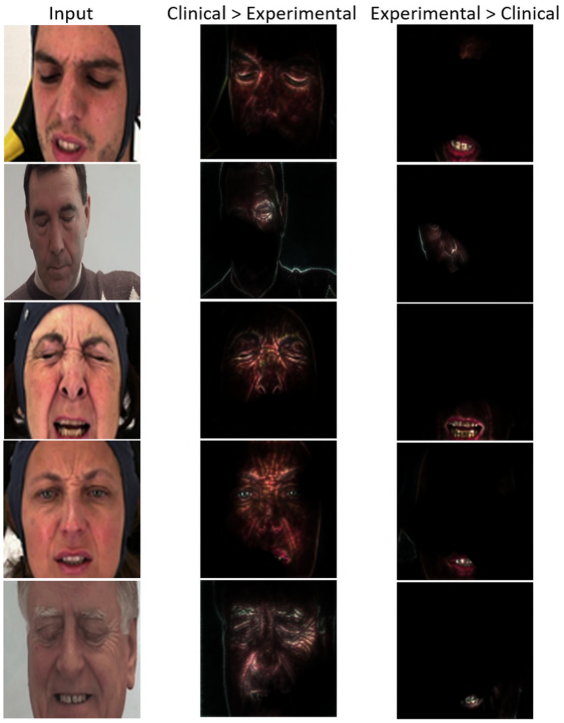

Fig. 1. Saliency maps of some images belonging to our experimental pain (BioVid) and clinical pain (UNBC) test sets. Each row shows the original input image and the result of subtracting the saliency maps generated for both models. The images under 'Clinical > Experimental' highlight the areas that the clinical pain model pays more attention to than the experimental pain model. The images under 'Experimental > Clinical' highlight the areas that are more relevant for the experimental pain model than for the clinical pain model.

Looking at the within-database evaluation, the clinical pain model yields better overall performance and seems better at recognizing pain images. However, the performance of the clinical pain model is considerably lower in cross-database evaluation. The experimental pain model has comparable performance on both datasets. To draw insights about the differences in the features that were relevant for the two models, we generate saliency maps for both models and manually analyze them as described in Sect. 3.4. Figure 1 shows some of the input images and the corresponding saliency map differences. We noticed that the clinical pain model pays more attention to the eye area, especially on closed eyes. In contrast, the experimental pain model pays attention to the mouth area, especially on visible teeth. So, we hypothesize that the clinical pain model is more biased towards closed eyes whereas, the experimental pain model is biased towards detecting visible teeth. We use concept embedding analysis (see Sect. 3.5) to test our hypothesis. We choose closed eyes and visible teeth as the concepts for our analysis. For analyzing the concept of closed eyes, we divide the images from test sets into two sets - images where the participants closed both their eyes and images where they did not. We use the clinical pain model and experimental pain model as feature extractors. We trained pairs of SVMs (one on clinical pain features and the other on experimental pain features) to recognize the concept of closed eyes. We found that the SVMs trained using clinical pain model features significantly outperformed the ones trained on experimental pain model features (clinical mean $F1 = 81.6\%$, experimental mean $F1 = 78.38\%$, t-statistic: 92.2, p-value: < 0.001). We follow the same procedure to analyze the concept of visible teeth. This time the images are divided into two sets based on whether the teeth were at least partially visible or not. In this case, the SVMs trained using experimental pain features were significantly better in discerning visible teeth images (clinical mean $F1 = 73.47\%$, experimental mean $F1 = 82.54\%$, t-statistic: -131.43, p-value: < 0.001).

To ensure that our findings are indeed based on differences in the datasets and not due to our specific models, we repeated fined-tuning the models and the concept embedding analysis four more times (using different seeds). In all the iterations, the models differed significantly on closed eyes and visible teeth.

5 Discussion

One interesting finding is that the clinical pain model doesn't perform well in cross-database evaluation, although it performs well on the clinical pain dataset. In contrast, the experimental pain model performs well on both datasets. It can be seen from Table 1 that misclassification of no-pain images from the experimental pain dataset is a key reason for the drop in performance of the clinical pain model. As the results of our concept analysis show, the clinical pain model pays more attention to the eye area, especially the closed eyes. In [24], the authors studied various facial activity descriptors from the BioVid database to predict pain. They found that eye closure is less relevant in predicting pain than other features. They attributed this to their observation that some participants

close their eyes even during no-pain videos. When we annotated the test sets for concept embedding analysis, we noticed that around 20% of the no-pain images from the experimental pain test set were annotated as closed eyes. This could be a plausible reason for the misclassification of no-pain images by the clinical pain model. Our result is also in line with [6], where the authors observed that it was difficult for a model trained on the UNBC dataset to recognize pain in inputs from the BioVid dataset. They chose the videos of 20 random participants from the BioVid database and found that many of them closed their eyes for most of the experiment. In contrast, the participants from UNBC database look at the camera and usually close their eyes while in pain. The authors attribute the poor performance of their model to this difference in behavior. Our results reinforce their hypothesis by, for the first time, analyzing the trained models themselves through XAI and empirically showing that the model trained on the UNBC dataset is paying more attention to closed eyes.

The experimental pain model pays more attention to the mouth area, especially the visibility of teeth (see Fig. 1). This concept can be associated with the pain pattern of 'open mouth' - one of the four facial pain patterns identified in [10]. They associate AUs 25, 26, 27 with the open mouth pattern. However, as noted in [24], these AUs are absent in the calculation of PSPI scores. The clinical pain dataset (UNBC database) is annotated based on PSPI scores whereas the experimental pain dataset (BioVid database) is annotated based on the temperature applied. Therefore, it is plausible that an image in the clinical pain dataset with an open mouth is labeled as no-pain (if PSPI AUs are absent). Moreover, from our manual annotations, we found that around 90% of visible teeth images were from the experimental pain dataset. While the total number of images with visible teeth is low our results show that this bias is reflected in the trained models. Hence, future works that use the BioVid dataset and medical personal that employ models based on this dataset should be aware of this bias.

6 Conclusion

In this paper, we explored the differences between models trained on clinical and experimental pain datasets. We used the UNBC shoulder pain database for clinical pain facial expressions and the BioVid heat pain database for experimental pain expressions. Using these datasets, we trained a clinical pain and an experimental pain model. In addition to the typical within-database evaluations, we evaluated the models on cross-database test sets. We found that the clinical pain model performed poorly on cross-database evaluation whereas the experimental pain model performed similarly on both datasets. This prompted us to use XAI techniques to explore the features that each model prioritizes in its predictions. We found that the concept of closed eyes is more important for clinical pain models and an open mouth with visible teeth is important for experimental pain models. We also found that these differences are rooted in the difference in the behavior of the participants in these datasets. Knowing these biases will aid researchers and medical personal when working with these datasets or when they employ models trained on those datasets.

The insights from this work show the potential merits of cross-database evaluations of pain recognition models with both performance metrics as well as XAI techniques. However, one limitation of our work is that we only tested one specific pair of pain datasets. People in real-life scenarios express various emotions other than pain. Therefore, our next step is to train models that can recognize other emotions like anger, fear, etc., along with pain and investigate those models in XAI-assisted cross-database evaluations based on different pain and emotion datasets. It will also be interesting to study the generalization performance of the models for different levels of pain (e.g. trace, weak, and strong pain).

References

1. The biovid heat pain database. http://www.iikt.ovgu.de/BioVid.html. Accessed 18 July 2021
2. Alber, M., et al.: iNNvestigate neural networks! J. Mach. Learn. Res. **20**(93), 1–8 (2019)
3. André, E., Kunz, M.: Digitale gesichts- bzw. schmerzerkennung und ihr potential für die klinische praxis. In: Digitalisierung und Gesundheit. G.IP - Gesundheitsforschung. Interdisziplinäre Perspektiven, Nomos Verlagsgesellschaft mbH & Co. KG (to appear)
4. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 3319–3327. IEEE Computer Society (2017)
5. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)
6. Dai, L., Broekens, J., Truong, K.P.: Real-time pain detection in facial expressions for health robotics. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 277–283. IEEE (2019)
7. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. **10**(7), 1895–1923 (1998)
8. Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition? In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 19–27 (2015)
9. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp. 2668–2677. PMLR (2018)
10. Kunz, M., Lautenbacher, S.: The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain. Eur. J. Pain **18**(6), 813–823 (2014)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
12. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: the UNBC-McMaster shoulder pain expression archive database. In: Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops, pp. 57–64. IEEE (2011)

13. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: Ganterfactual - counterfactual explanations for medical non-experts using generative adversarial learning. CoRR abs/2012.11905 (2021)
14. Mollahosseini, A., Hassani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2019)
15. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R.: Layer-wise relevance propagation: an overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 193–209. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_10
16. Othman, E., Werner, P., Saxen, F., Al-Hamadi, A., Walter, S.: Cross-database evaluation of pain recognition from facial video. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 181–186. IEEE (2019)
17. Petyaeva, A., et al.: Feasibility of a staff training and support programme to improve pain assessment and management in people with dementia living in care homes. Int. J. Geriatr. Psychiatry **33**(1), 221–231 (2018)
18. Prajod, P., Schiller, D., Huber, T., André, E.: Do deep neural networks forget facial action units?-exploring the effects of transfer learning in health related facial expression recognition. arXiv preprint arXiv:2104.07389 (2021)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
20. Sixt, L., Granz, M., Landgraf, T.: When explanations lie: why many modified BP attributions fail. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event, pp. 9046–9057 (2020)
21. Walter, S., et al.: The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE International Conference on Cybernetics (CYBCO), pp. 128–131. IEEE (2013)
22. Wang, F., et al.: Regularizing face verification nets for pain intensity regression. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1087–1091. IEEE (2017)
23. Weitz, K., Hassan, T., Schmid, U., Garbas, J.U.: Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. tm-Technisches Messen **86**(7–8), 404–412 (2019)
24. Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., Traue, H.C.: Automatic pain assessment with facial activity descriptors. IEEE Trans. Affect. Comput. **8**(3), 286–299 (2016)
25. Werner, P., Al-Hamadi, A., Walter, S.: Analysis of facial expressiveness during experimentally induced heat pain. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 176–180. IEEE (2017)
26. Zhao, R., Gan, Q., Wang, S., Ji, Q.: Facial expression intensity estimation using ordinal information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3466–3474 (2016)