

**Deep-learned faces of pain and emotions:  
elucidating the differences of facial expressions  
with the help of explainable AI methods = Tief  
erlernte Gesichter von Schmerz und Emotionen:  
Aufklärung der Unterschiede von  
Gesichtsausdrücken mithilfe erklärbarer  
KI-Methoden**

**Katharina Weitz, Teena Hassan, Ute Schmid, Jens-Uwe Garbas**

**Angaben zur Veröffentlichung / Publication details:**

Weitz, Katharina, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas. 2019.  
“Deep-learned faces of pain and emotions: elucidating the differences of facial  
expressions with the help of explainable AI methods = Tief erlernte Gesichter von  
Schmerz und Emotionen: Aufklärung der Unterschiede von Gesichtsausdrücken  
mithilfe erklärbarer KI-Methoden.” *tm - Technisches Messen* 86 (7-8): 404-12.  
<https://doi.org/10.1515/teme-2019-0024>.

**Nutzungsbedingungen / Terms of use:**

**licgercopyright**

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under the following conditions:

**Deutsches Urheberrecht**

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publizieren>



Katharina Weitz\*, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas

# Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods

Tief erlernte Gesichter von Schmerz und Emotionen: Aufklärung der Unterschiede von Gesichtsausdrücken mithilfe erklärbarer KI-Methoden

<https://doi.org/10.1515/teme-2019-0024>

Received February 28, 2019; accepted May 29, 2019

**Abstract:** Deep neural networks are successfully used for object and face recognition in images and videos. In order to be able to apply such networks in practice, for example in hospitals as a pain recognition tool, the current procedures are only suitable to a limited extent. The advantage of deep neural methods is that they can learn complex non-linear relationships between raw data and target classes without limiting themselves to a set of hand-crafted features provided by humans. However, the disadvantage is that due to the complexity of these networks, it is not possible to interpret the knowledge that is stored inside the network. It is a black-box learning procedure. Explainable Artificial Intelligence (AI) approaches mitigate this problem by extracting explanations for decisions and representing them in a human-interpretable form. The aim of this paper is to investigate the explainable AI methods Layer-wise Relevance Propagation (LRP) and Local Interpretable Model-agnostic Explanations (LIME). These approaches are applied to explain how a deep neural network distinguishes facial expressions of pain from facial expressions of emotions such as happiness and disgust.

**Keywords:** Explainable artificial intelligence, deep learning, emotion recognition, pain recognition.

**Zusammenfassung:** Tiefe neuronale Netze werden erfolgreich für die Objekt- und Gesichtserkennung in Bildern und Videos verwendet. Die derzeitigen Ansätze sind jedoch nur begrenzt in der Praxis, zum Beispiel zur Schmerzerkennung, verwendbar. Der Vorteil von Deep Learning

**\*Corresponding author: Katharina Weitz**, Fraunhofer IIS, Intelligent Systems Group, Am Wolfsmantel 33, 91058 Erlangen, Germany; and University of Bamberg, Cognitive Systems Group, An der Weberei 5, 96047 Bamberg, Germany, e-mail:

katharina-blandina.weitz@stud.uni-bamberg.de

**Teena Hassan, Jens-Uwe Garbas**, Fraunhofer IIS, Intelligent Systems Group, Am Wolfsmantel 33, 91058 Erlangen, Germany

**Ute Schmid**, University of Bamberg, Cognitive Systems Group, An der Weberei 5, 96047 Bamberg, Germany

Methoden liegt darin, dass sie in der Lage sind, komplexe, nichtlineare Zusammenhänge zwischen Rohdaten und Zielklassen zu lernen, ohne dass sie auf händisch durch Menschen generierte Merkmale angewiesen sind. Der Nachteil dieser Netzwerke besteht darin, dass sie sehr komplex sind und daher für Menschen schwer zu verstehen ist, warum das Netz zu seiner Entscheidung gekommen ist. Man bezeichnet diese Netzwerke deshalb auch als black-boxes. Methoden der erklärbaren künstlichen Intelligenz (AI) nehmen sich diesem Problem an, indem sie Erklärungen für Entscheidungen generieren und diese für Menschen in einer interpretierbaren Form darstellen. Das Ziel dieses Artikels ist es, die erklärbaren AI Methoden Layer-wise Relevance Propagation (LRP) und Local Interpretable Model-agnostic Explanations (LIME) zu nutzen, um die Entscheidungen eines tiefen neuronalen Netzes zu erklären, dass schmerzhaftes Gesichtsausdrücke von Freude und Ekel darstellenden Gesichtern unterscheidet.

**Schlagwörter:** erklärbare künstliche Intelligenz, Deep Learning, Emotionserkennung, Schmerzerkennung.

## 1 Introduction

Facial expressions are one of the most important human non-verbal signals in interacting with other people and thus contribute to the emergence and maintenance of social relationships [9]. One of the tasks of facial expressions is to communicate emotions [2, 8]. This is of particular importance when people are unable to express themselves using speech (e. g., because of illness, accidents or congenital disabilities). For this reason, nursing staff in clinics and care facilities are required to observe patients closely in order to be able to read their emotions and take action, if necessary. Due to the already significantly high number of patients, especially in nursing homes, and the prognosis that more and more people will be cared for in such facilities in the future, it would be beneficial to deploy a system

to support the nursing staff to monitor a patient's facial expressions and alert them when a pain episode is detected. Additionally, humans often have problems in differentiating between pain and other facial expressions [3, 6]. Therefore, in addition to the (classical) exploration of emotions in a psychological context, research into a technical solution for distinguishing emotions and pain has gained greater importance in the last decade. A system, which uses explainable AI methods to describe how pain differs from other emotions, can be used to train nursing staff to improve their ability to recognise pain correctly.

Towards this goal, in this paper we examine and apply the LRP [4] and the LIME method [21] to explain the decisions made by a deep Convolutional Neural Network (CNN) that is trained to distinguish facial expressions of pain, happiness, and disgust.

## 2 Related work

One of the deep learning architectures that has been successfully applied to image processing applications is CNN [15], which processes images in a hierarchical manner [19]. Compared to approaches based on explicit facial activity descriptors [27], the features of deep learning do not have to be handcrafted. Instead, the system learns the features by itself by projecting information from bitmaps into so-called convolutional layers. They can learn non-linear relationships to model dependencies among the features [23]. One disadvantage is that deep learning approaches require a lot of sample data to extract features [14]. This problem can be reduced by data augmentation methods such as flipping or rotation. The other disadvantage is that, due to its complexity, it is no longer comprehensible for humans, what the network has learned and what it bases its predictions on. In a practical application, it could be shown that these systems were not going to be accepted, because people do not blindly trust a system which they do not understand [12]. Therefore, techniques that make the black-box learning comprehensible to humans are necessary. One of these techniques for explaining the black-box deep learning is called Layer-wise Relevance Propagation (LRP), which explains the network's decisions by pixel-wise decomposition [4]. In facial image analysis, LRP can be used to explain which pixels were important for the decision of the network. For this, LRP decomposes the output of the model's decision function  $f$  (e. g., a classification result), given an input  $x$  [13]. This decomposition provides relevance values  $R_p$  for each component  $p$  of  $x$

such that  $\sum_p R_p = f(x)$ . When applying LRP, different parameters can be set to improve the resulting heatmap. In some cases, the resulting relevance scores for each pixel generated by LRP can take on unbounded values [4]. To adjust and stabilize the relevance scores, an  $\varepsilon$  value can be used. Additionally,  $\alpha$  and  $\beta$  values can be applied for stabilization. Besides the stabilizing effect,  $\alpha$  and  $\beta$  values can be used to visualize positive and negative relevance-activations of pixels [4]. With different values for  $\alpha$  and  $\beta$ , the strength of the influence of positive ( $\alpha$ ) and negative ( $\beta$ ) portions can be controlled [4, 17]. Besides these parameters, Kohlbrenner [11] showed that a 'preset' variant of the LRP algorithm achieves optimal results in the calculation of relevance maps. Using the preset approach, the relevance scores for all neurons of the lowest (first) layer are uniformly distributed to the input neuron instead of using the  $\alpha$  and  $\beta$  values [13]. To control the resolution of the heatmaps generated by LRP, Bach et al. [5] describes an approach for 'mapping influence cut-off point'. This point describes the moment from which the forward mapping function of the classifier no longer influences relevance propagation, since only the receptive field of the classifier is relevant. The cut-off at this point is called the 'flat' rule. The reference of a receptive field is adapted from neuroscience [10]. In a CNN, the convolutional and pooling layers are inspired by the biological receptive field [14].

Another approach to visualize predictions of a classifier is LIME [21]. This approach differs from the previously described LRP architecture in that it can be applied to different machine learning classifiers, whereas LRP is optimized for deep learning architectures and needs to be adapted for other machine learning approaches. In order to visualize the prediction of different classification methods for a given image, LIME learns an interpretable model locally around the prediction. The generated explanation is therefore not a description for the entire model, but for the instance (e. g. image shown) that is presented to the model. For this purpose, LIME divides the image, denoted as  $x \in \mathbb{R}^d$ , to be classified into superpixels using a segmentation algorithm. LIME then creates a permuted dataset of the original image by greying out random superpixels. LIME uses a binary vector  $x' \in \{0, 1\}^d$  as interpretable representation of the image classification. 1 stands for an original superpixel, while 0 stands for a greyed out superpixel. The images of the permuted dataset are then presented to the classifier. For the resulting prediction of the classifier,  $K$  features (superpixels) are extracted, which generate the maximum likelihood for the predicted class. The selection of the  $K$  features is calculated using a variation of the Lasso algorithm [7] and then the weights are learned

using the least squares method. This procedure is named K-LASSO [21].

### 3 Research questions

Answers should be found for the following research questions for automatic pain classification to become applicable in real-life settings:

- **Predictive performance:** How well can facial expressions of pain be automatically distinguished from those of disgust and happiness using self-learned spatial features?
- **Decision interpretation:** How can the decisions made by the model be presented to people in a comprehensible and transparent way?
- **Feature explanation:** How do the self-learned features differ for the facial expressions of pain and those of disgust and happiness?

This paper<sup>1</sup> would like to provide answers of the questions above. For this, a pre-trained VGG-Face model [18] implemented with the Keras framework was finetuned to distinguish pain from happiness and disgust. For the finetuning, images of the BioVid dataset<sup>2</sup> [26] were used. Then, the Keras implementation of the LRP approach<sup>3</sup> [1] was used to generate heatmaps at pixel level to illustrate which pixels were relevant for the classification by the VGG-Face model. For LIME, the Keras implementation<sup>4</sup> of Ribeiro et al. [20] was used. In a further step, heatmaps were generated from images of the UNBC-McMaster shoulder pain expression archive database [16] and Actorstudy dataset<sup>5</sup> in order to examine the generalization performance of the network.

### 4 Material & procedure

The procedure of this study consisted of the following steps: First, data preparation was done on the BioVid

dataset [26]. After that, the VGG-Face CNN model was finetuned for the three-class problem of distinguishing pain from happiness and disgust.

In the data preparation step, frames were extracted from the video sequences of pain, happiness, and disgust in the BioVid dataset [26]. The BioVid dataset contains video sequences of participants who feel pain and emotions (happiness, sadness, anger, disgust, fear, and neutral). Pain was induced on the right arm using a thermode. For the emotion induction, the International Affective Picture System (IAPS) was used. The IAPS pictures were shown to the participants. To finetune the VGG-face model, part A (pain stimulation without facial EMG) of the BioVid dataset was used for the classification of pain. For the classification of disgust and happiness, frames from video sequences in part D (posed pain & basic emotions) of the BioVid dataset were used. The video sequences for the pain condition are each 5 seconds long (24,012 frames), the video sequences for the emotions are each 1 minute long (114,076 frames for disgust and 112,575 frames for happiness). The dataset was balanced by manually selecting  $3 \times 10^7$  frames from each of the happiness and disgust sequences. One subject in the condition ‘disgust’ turned away from the camera and talked to the study leaders and showed no disgust expression. This subject was removed from the dataset. In Table 1, the amount of frames selected for each class after the data cleaning steps is provided. This subset of the BioVid dataset was then used to finetune VGG-Face. For the implementation, Tensorflow (version 1.8) and Keras (version 2.2.0) were used. These are two open source platforms that provide tools and libraries for machine learning. After that, the explainable AI methods LRP and LIME were applied to generate visualisations for decision interpretation and feature explanation. For LRP, the Keras implementation from [1] was adapted. For LIME, also a Keras implementation<sup>6</sup> was used.

**Table 1:** Extracted BioVid data after data cleaning steps.

Part	Name	Subjects	Frames
Part A	Pain intensity 3	87	12,006
	Pain intensity 4	87	12,006
Part D	Disgust	75	24,075
	Happiness	75	24,075

<sup>6</sup> <https://github.com/marcotcr/lime>

<sup>1</sup> This paper is based on the master’s thesis of the first author submitted on August, 31, 2018 to the University of Bamberg. Online link: <https://www.uni-bamberg.de/en/cogsys/research/theses/advised-theses/>

<sup>2</sup> <http://www.iikt.ovgu.de/BioVid.html>

<sup>3</sup> <https://github.com/albermax/innvestigate>

<sup>4</sup> <https://github.com/marcotcr/lime>

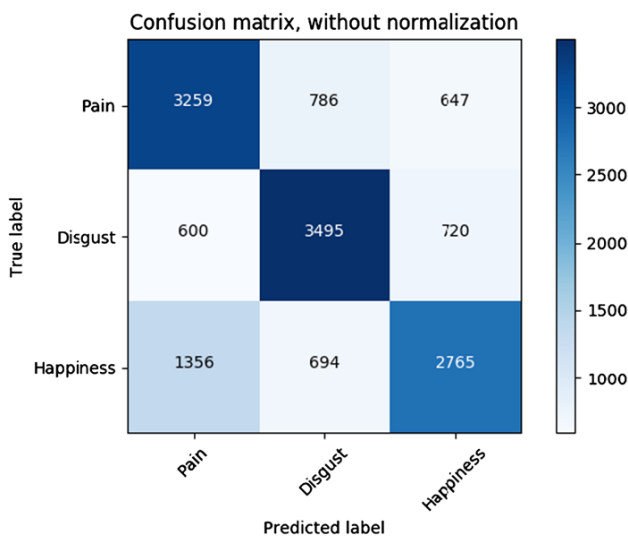
<sup>5</sup> Unpublished facial expression dataset from Intelligent Systems Group, Fraunhofer IIS, Erlangen.

## 5 Results

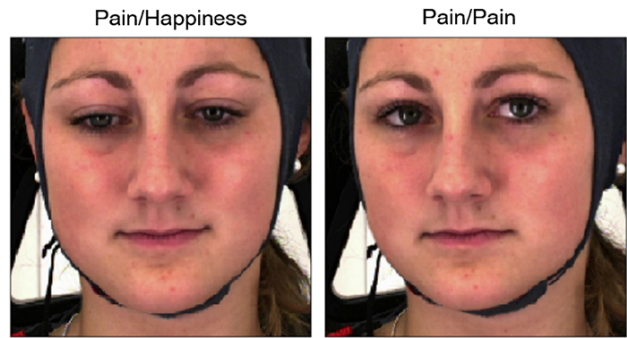
The VGG-Face CNN was fine-tuned and tested using 5-fold cross-validation. Here, 4 folds were used for training the model and the remaining fold was used to test the model. The process was repeated four times using a different fold each time for testing. The best performing test fold had an accuracy of 0.67 and was used for generating explanations using the LRP method. In Table 2, the class-wise performance of the best fold is presented. When looking at the confusion matrix (see Figure 1), it becomes clear that the CNN had problems to classify happy faces as happy. 28 % of the happy images were classified as pain. To take a closer look at this problem, the LRP approach was used. It was used to get an insight into the pixel-related areas of the image which were important for decisions of pain and happiness. To gain this insight, two test images from the pain category were selected from the BioVid test fold. In Figure 2, the first image displays the subject experiencing pain intensity 3, and in the second image the same person experiencing pain intensity 4. The first label above the image refers to the true class, and the second label to the predicted class.

**Table 2:** Class-wise results of the best performing fold.

	Precision	Recall	F1-score	#Images
Pain	0.62	0.69	0.66	4692
Disgust	0.70	0.73	0.71	4815
Happiness	0.67	0.57	0.62	4815
Average/Total	0.67	0.66	0.66	14322



**Figure 1:** Confusion matrix (without normalization) for the best test fold of the 5-fold cross-validation. 28% of the images showing happiness were classified as pain.

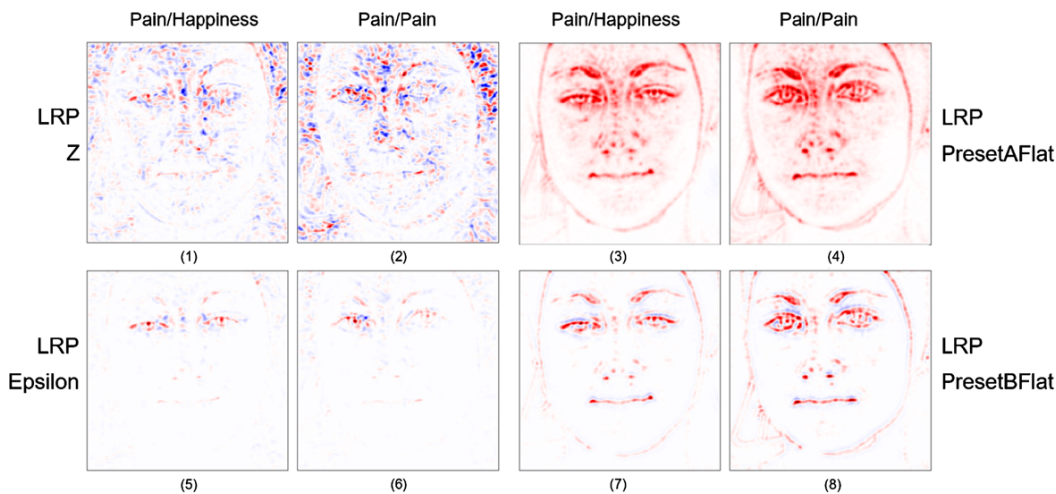


**Figure 2:** Original image from the test fold of the BioVid dataset. First label indicates the true class, second label indicates the predicted class.

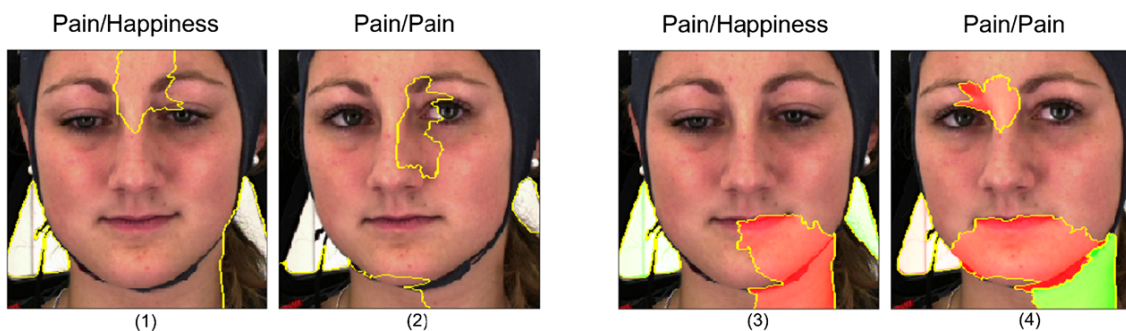
In Figure 3, the visualization generated using LRP with different parameters are shown. In LRP-Z, the basic LRP approach without stabilizers is applied. Here, a face is roughly recognizable. The noise due to the absence of stabilizers is present. In comparison to the basic LRP approach (LRP-Z), much less noise is represented using the LRP-Epsilon method. In the visualizations using LRP-PresetAFlat and LRP-PresetBFlat, red pixels indicate a positive contribution to the predicted class, and blue pixels indicate a negative contribution. In comparison to the LRP-Z heatmap, the visualization of preset-flat variants are much more detailed and clearer. In the two preset variants, it can be observed that highly positive pixel values (represented by a higher intensity of redness) are important for the decision of the CNN. It becomes apparent that with the increase of the  $\alpha$  value (LRP-PresetAFlat), the positive pixel values become more prominent. With the increase of the  $\beta$  value (LRP-PresetBFlat), the intensity of the blue pixels becomes more visible. When looking at the LRP PresetAFlat visualization, it can be seen that mostly the same areas in the face, namely the eyes, the nose and the mouth contribute to the classification of happiness and pain. This could be an indication why the accuracy of the CNN is not very high. When looking at the LRP PresetBFlat, slight differences in the contribution of negative pixels for the classification are visible. For pain, more negative pixels around the nostrils and on the lower side of the eyebrows are detectable on the heatmaps.

In comparison, the visualizations generated with LIME show no fine-grained details but instead, coarse-grained details in the form of the importance of certain superpixels. On image (1) and (2) of Figure 4, the five most important positive superpixels are presented.<sup>7</sup> It can be

<sup>7</sup> In the case of neighbouring relevant superpixels, LIME represents these as a related area.



**Figure 3:** Visualizations for applying LRP method with different parameters on two pain images. First label indicates the true class, second label indicates the predicted class. The heatmap generated with the basic LRP approach (LRP-Z) is displayed in subimages (1) and (2). Subimages (3) and (4) display the heatmap generated with the LRP-PresetAFlat variant. Subimages (5) and (6) display the heatmap generated with an  $\varepsilon$  stabilizer applied to LRP-Z. Subimages (7) and (8) are the results of applying LRP-PresetBFlat. The visualizations correspond to the predicted class.



**Figure 4:** Visualizations for applying the LIME method on two pain images. First label indicates the true class, second label indicates the predicted class. The superpixel generated in subimages (1) and (2) represent the five most relevant superpixels. Subimages (3) and (4) represent the five superpixels which are most relevant to improve the classification (green) or make the classification result worse (red). The visualizations display the predicted class.

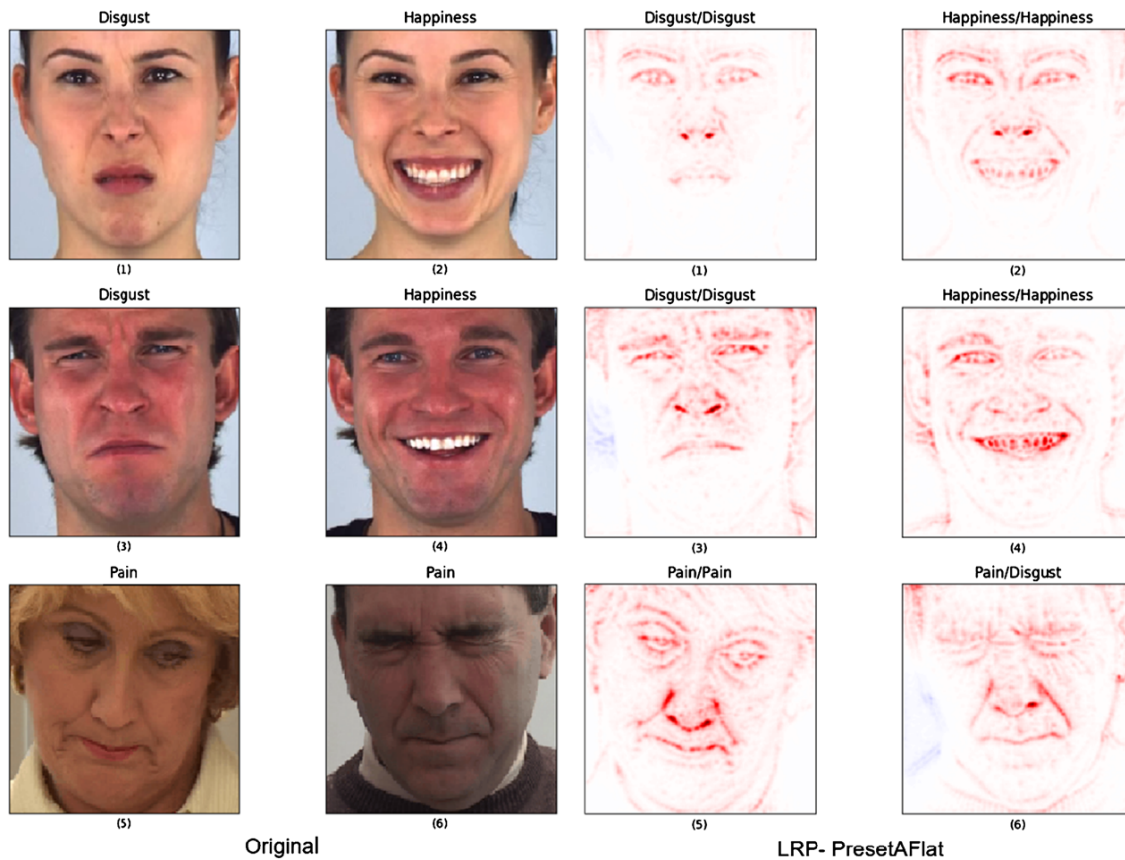
seen that, in (1) the area between the eyes is important. Important superpixels can also be found outside the face. On subimage (2), parts of the nose and eye are important. Here too, areas outside the face are displayed as relevant for the network. The subimages (3) and (4) of Figure 4 show the five most important positive or negative superpixels, represented as red or green superpixels, respectively. Red superpixels stand for areas of the image that make the classification worse, green superpixels for areas that are conducive to the classification. Again in subimage (4), it can be seen that part of the neck is highlighted as supportive for the classification.

Besides test images from the BioVid dataset, images from the UNBC-McMaster shoulder pain expression archive database for pain were used for visualization using the LRP method (see right part of Figure 5) and the

LIME method (see Figure 6). For happiness and disgust, images from the Actorstudy dataset were used. The classification results for the subimages in Figure 5 showing emotion (from (1) to (4)) and the subimage (5) showing pain are correct. A misclassification can be seen in subfigure (6). Here the pain image is classified as disgust.

The visualizations using the LRP-PresetAFlat approach are shown on the right part in Figure 5. Here it can be seen that for happiness, the eyes and the mouth are important areas for the classification. For disgust, the focus lies on the nose and the eyes. This could be a reason that the pain image (subimage 6) was misclassified as disgust. For pain, the nostrils seem to be important.

When looking at the visualizations generated by LIME (see Figure 6), the visualizations of the emotional expressions seem to be consistent with what one would expect.



**Figure 5:** Left: Input images 1–4 from Actorstudy dataset and images 5 & 6 from the UNBC-McMaster shoulder pain expression archive database (©Jeffrey Cohn) to visualize LRP approach. Right: Visualizations for applying LRP PresetAFlat method. The visualizations display the predicted class. First label refers to the true class, second label refers to the predicted class.

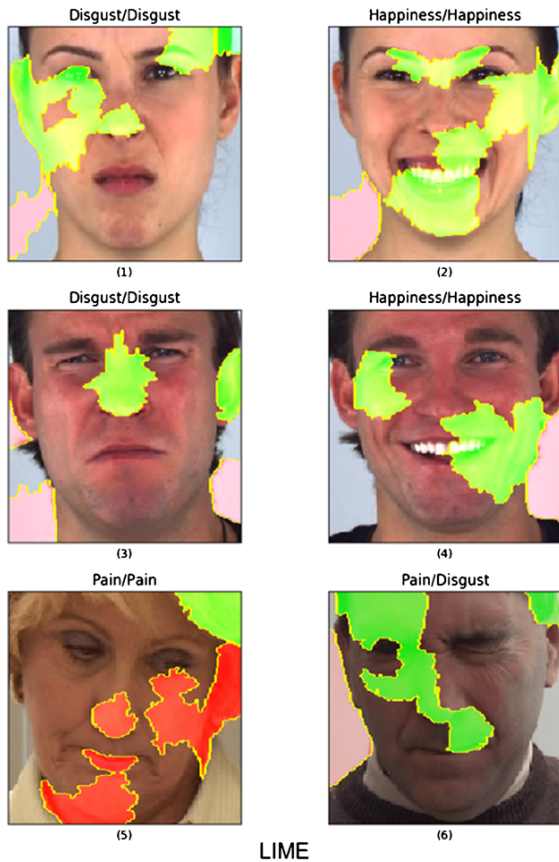
For disgust, areas of the nose are relevant, for happiness, the mouth shaped for laughter and areas around the eyes are important. It can be seen that areas of the background are not conducive to the emotion classifications. The visualizations of the pain images are less coherent. In subfigure (5) of Figure 6 it can be seen that areas of the face deteriorate the classification, while an area of the hair improves the classification of pain. In subfigure (6), the classification of disgust is promoted due to the wrinkled nose. The areas of the hair are relevant for the classification.

## 6 Discussion

For the topic of **predictive performance**, this paper shows that the CNN could distinguish images of pain, disgust, and happiness only with an accuracy of 67%. Above all, happy faces were often misclassified as faces of pain.

For the part of **decision interpretation**, LRP is a helpful tool to generate a fine-granular heatmap of relevant

pixels. The usage of LRP with its various parameters allows a wide range of adjustments. The results presented here for the categorization of pain, disgust, and happiness represent only an initial step into the research of making decisions of black-box systems comprehensible for humans. Lapuschkin et al. [13] already investigated the application of LRP for the recognition of age and gender from images of faces. They could show that the visualizations of relevant pixels allow an interpretation of the relevant facial areas to classify age and gender. However, when looking at facial expressions of happiness, pain, and disgust it becomes clear that pixel activation alone cannot yet provide a clear difference between the predicted classes for the human eye. Therefore, for the topic of **feature explanation**, the relevant features for the classification are not easy detectable by humans. The visualizations generated by LIME are coarse-granular compared to LRP. This makes it easier to identify relevant areas on the face. The size of the superpixels varies from image to image due to the segmentation algorithm used. Therefore, the relevant superpixels differ from image to image. This makes it difficult



**Figure 6:** Visualizations by applying LIME method. The visualizations display the predicted class. First label refers to the true class, second label refers to the predicted class.

to compare the individual images. In the LIME visualizations, it becomes clear that the network pays attention to areas outside the face. Additionally, in the pain images of the UNBC-McMaster shoulder pain expression archive database, areas of the hair are considered relevant. This partly explains the poor classification capabilities of the fine-tuned VGG-Face CNN. The classification accuracy of 67 % must be taken into account when looking at the visualizations.

Montavon et al. [17] describe some practical recommendations to improve the visualizations generated by the LRP method: using dropout as regularization technique, preferring sum pooling, instead of max pooling and not using too many fully connected layers in the network (whereas no definition is given for what is meant by ‘many’). In the case of LIME, one could try different segmentation algorithms to generate superpixels to improve the visualizations.

Nevertheless, additional information is needed for a clearer interpretation [24]. Future research approaches may focus on the implementation of such additional infor-

mation sources. Additional sources of information could, for example, take the form of linguistic information, the form of uncertainty formulations (e. g., pixel activations for happiness have an uncertainty value of 20 out of 100, while pixel activations for pain have an uncertainty value of 90 out of 100) or the form of paralinguistic information such as loudness, pitch, laughter, sighs, and crying. Linguistic information can on the one hand emphasize the focus of relevant areas (e. g., ‘In this image, the eyes are important for the classification of happiness’). On the other hand, linguistic information can also help to clarify specific characteristics of features (e. g., ‘The lids have to be tightened for the classification of pain’) [22]. Due to the special requirements in clinics and care facilities, where verbal expressions of patients in form of speech are often not possible, paralinguistic information besides facial expressions are relevant. A promising progress of deep learning approaches in the field of paralinguistic recognition tasks is detectable, especially in the task of recognizing affective states of disabled persons and infants [25]. The combination of linguistic information to understand visual explanations with multimodal information like paralinguistics constitute an approach to develop an informative and interpretable system. Only when such a system is archived, can a comprehensive application in real-life be considered.

## References

1. Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. investigate neural networks! *arXiv preprint arXiv:1808.04260*, 2018.
2. Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
3. Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229, 2012.
4. Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7):e0130140, 2015.
5. Sebastian Bach, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek. Controlling explanatory heatmap resolution and semantics via decomposition depth. In *Proceedings of the International Conference on Image Processing*, pages 2271–2275. IEEE, 2016.
6. Sheryl Brahmam, Chao-Fa Chuang, Frank Y Shih, and Melinda R Slack. Machine recognition and representation of neonatal

- facial displays of acute pain. *Artificial Intelligence in Medicine*, 36(3):211–222, 2006.
7. Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
  8. Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
  9. Chris Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3453–3458, 2009.
  10. David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154, 1962.
  11. Maximilian Hans Kohlbrenner. On the stability of neural network explanations, Apr 2017. Bachelor's Thesis.
  12. H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. Explainable artificial intelligence for training and tutoring. Technical report, University of Southern California Marina del Rey CA Institute for Creative Technologies, 2005.
  13. Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek. Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the International Conference on Computer Vision*, pages 1629–1638, 2017.
  14. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.
  15. Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990.
  16. Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 57–64. IEEE, 2011.
  17. Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2017.
  18. Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, pages 1–12, 2015.
  19. Chirag Ravat and Shital A Solanki. Survey on different methods to improve accuracy of the facial expression recognition using artificial neural networks. In *Proceedings of the National Conference on Advanced Research Trends in Information and Computing Technologies*, volume 4, 2018.
  20. Marco Tulio Ribeiro, Singh Sameer, and Carlos Guestrin. Lime. <https://github.com/marcotcr/lime/>, 2017.
  21. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.
  22. Ute Schmid. Inductive programming as approach to comprehensible machine learning. In *Proceedings of the 7th workshop on dynamics of knowledge and belief (DKB-2018) and the 6th workshop KI & Kognition (KIK-2018), co-located with 41st German conference on artificial intelligence*, volume 2194, 2018.
  23. Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
  24. Michael Siebers and Ute Schmid. Please delete that! why should i? *KI – Künstliche Intelligenz*, 2018.
  25. Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth André. Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant? In *Proceedings of Interspeech 2018*, pages 147–151, 2018.
  26. Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *Proceedings of the International Conference on Cybernetics*, pages 128–131. IEEE, 2013.
  27. Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299, 2017.

## Bionotes



### Katharina Weitz

Fraunhofer IIS, Intelligent Systems Group,  
Am Wolfsmantel 33, 91058 Erlangen,  
Germany  
University of Bamberg, Cognitive Systems  
Group, An der Weberei 5, 96047 Bamberg,  
Germany  
[katharina-blandina.weitz@stud.uni-bamberg.de](mailto:katharina-blandina.weitz@stud.uni-bamberg.de)

Katharina Weitz received a Master of Science in Psychology and a Master of Science in Computing in the Humanities (Applied Computer Science) at the University of Bamberg, Germany. She is currently working at the University of Augsburg at the chair for Human-Centered Multimedia. She is interested in machine learning topics in the field of social robotics and virtual agents. The influence of explainability and transparency of intelligent systems on people's trust is a central point of her research activities. She supports a human-centered usage of artificial intelligence and delves into ethical issues. In addition to her research activities, the communication of research knowledge to the general public in the form of lectures, workshops and exhibitions is an important concern to her.



**Teena Hassan**

Fraunhofer IIS, Intelligent Systems Group,  
Am Wolfsmantel 33, 91058 Erlangen,  
Germany

Teena Hassan received her Bachelor of Technology degree in Computer Science and Engineering from Cochin University of Science and Technology in Kerala, India, in the year 2006. After graduation, she worked as a Project Engineer in the Telecom/Datacom domain. In 2014, she received her Master of Science degree in Autonomous Systems from the Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin. After graduation, she joined Fraunhofer IIS, in Erlangen, where she conducted research in the field of automatic analysis of facial action units, with a special focus on modeling facial muscle motions, fusing multiple sources of facial expression information, and modeling uncertainty in measurements. Her research interests include facial expression analysis, sensor noise modeling, sensor fusion, and state estimation. She is currently a Research Associate at the Bielefeld University, conducting research on interaction architectures for social robots.



**Ute Schmid**

University of Bamberg, Cognitive Systems  
Group, An der Weberei 5, 96047 Bamberg,  
Germany

Ute Schmid holds a diploma in psychology and a diploma in computer science, both from Technical University Berlin (TUB), Germany. She received her doctoral degree (Dr. rer.nat.) in computer science from TUB in 1994 and her habilitation in computer science in 2002.

From 1994 to 2001 she was assistant professor (wissenschaftliche Assistentin) at the AI/Machine Learning group, Department of Computer Science, TUB. Afterwards she worked as lecturer (akademische Rätin) for Intelligent Systems at the Department of Mathematics and Computer Science at University Osnabrück. Since 2004 she holds a professorship of Applied Computer Science/Cognitive Systems at the University of Bamberg. Research interests of Ute Schmid are mainly in the domain of comprehensible machine learning, explainable AI, and high-level learning on relational data, especially inductive programming, knowledge level learning from planning, learning structural prototypes, analogical problem solving and learning. Further research is on various applications of machine learning (e. g., classifier learning from medical data and for facial expressions) and empirical and experimental work on high-level cognitive processes. Ute Schmid dedicates a significant amount of her time to measures supporting women in computer science and to promote computer science as a topic in elementary, primary, and secondary education.

**Jens-Uwe Garbas**

Fraunhofer IIS, Intelligent Systems Group, Am Wolfsmantel 33,  
91058 Erlangen, Germany

Dr. Garbas received the Dipl.-Ing. and Dr.-Ing. (summa cum laude) degrees in electrical engineering from Friedrich-Alexander University Erlangen-Nuremberg, Germany, in 2004 and 2010, respectively. In 2010 he joined Fraunhofer Institute for Integrated Circuits IIS, where he was appointed head of the group Intelligent Systems 2011 and deputy head of department electronic imaging in 2012, respectively. He is responsible for industrial and public research project as well as software licensing in the area of real-time computer vision, affective computing and facial analysis.