Finzel, Bettina; Rieger, Ines; Kuhn, Simon; Schmid, Ute

Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition

Primary publication

Finzel, Bettina; Rieger, Ines; Kuhn, Simon; Schmid, Ute (2023): „Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition“. In: A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A.M. Tjoa, E. Weippl (Ed.), Machine learning and knowledge extraction, Cham, Switzerland: Springer Nature Switzerland, pp. 31–44, doi: 10.1007/978-3-031-40837-3_3.

Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition

Bettina Finzel(✉), Ines Rieger, Simon Kuhn, and Ute Schmid

Cognitive Systems, University of Bamberg, Bamberg, Germany
{bettina.finzel,ines.rieger,ute.schmid}@uni-bamberg.de

Abstract. Research in the field of explainable artificial intelligence has produced a vast amount of visual explanation methods for deep learning-based image classification in various domains of application. However, there is still a lack of domain-specific evaluation methods to assess an explanation's quality and a classifier's performance with respect to domain-specific requirements. In particular, evaluation methods could benefit from integrating human expertise into quality criteria and metrics. Such domain-specific evaluation methods can help to assess the robustness of deep learning models more precisely. In this paper, we present an approach for domain-specific evaluation of visual explanation methods in order to enhance the transparency of deep learning models and estimate their robustness accordingly. As an example use case, we apply our framework to facial expression recognition. We can show that the domain-specific evaluation is especially beneficial for challenging use cases such as facial expression recognition and provides application-grounded quality criteria that are not covered by standard evaluation methods. Our comparison of the domain-specific evaluation method with standard approaches thus shows that the quality of the expert knowledge is of great importance for assessing a model's performance precisely.

Keywords: Convolutional Neural Networks · Explainable Artificial Intelligence · Facial Expressions · Explanation Evaluation · Robustness

1 Introduction

Deep learning approaches are successfully applied for image classification. However, the drawback of these deep learning approaches is their lack of robustness in terms of reliable predictions under small changes in the input data or model parameters [10]. For example, a model should be able to handle out-of-distribution data that deviate from the training distribution, e.g., by being blurry or showing an object from a different angle. However, often models produce confidently false predictions for out-of-distribution data. These can get unnoticed,

The work presented in this paper was funded by grant DFG (German Research Foundation) 405630557 (PainFaceReader).

A. Holzinger et al. (Eds.): CD-MAKE 2023, LNCS 14065, pp. 31–44, 2023.
https://doi.org/10.1007/978-3-031-40837-3_3

as deep learning models are per default a black box approach with no insight into the reasons for predictions or learned features.

Therefore, explainers can be applied that visually explain the prediction in order to enhance the transparency of image classification models and to evaluate the model's robustness towards out-of-distribution samples [33,35]. Visual explanations are based on computing the contribution of individual pixels or pixel groups to a prediction, thus, helping to highlight what a model "looks at" when classifying images [36].

An important aspect is that both, robustness and explainability, are enablers for trust. They promote reliability and ensure that humans remain in control of model decisions [13]. This is of special interest in decision-critical domains such as medical applications and clinical assistance as demonstrated by Holzinger et al. [14] and Finzel et al. [9]. As robustness and explainability are therefore important requirements for application-relevant models, measures that help to assess the fulfillment of such requirements should be deployed with the models. With respect to visual explanations as a basis to rebustness and explainability analysis, it is worth noting that visualizations express learned features only qualitatively. In order to analyze a model's robustness more precisely, quantitative methods are needed to evaluate visual explanations [3,28,34]. These quantitative methods, however, do not provide domain-specific evaluation criteria yet that are tailored to the application domain.

In this work, we propose a framework for domain-specific evaluation and apply it to the use case of facial expression recognition. Our domain-specific evaluation is based on selected expert knowledge that is quantified automatically with the help of visual explanations for the respective use case. The user can inspect the quantitative evaluation and draw their own conclusions.

To define facial expressions, one psychologically established way is to describe them with the Facial Action Coding System (FACS) [7], where they are categorized as sets of so-called Action Units (AUs). Facial expression analysis is commonly performed to detect emotions or pain (in clinical settings). These states are often derived from a combination of AUs present in the face [22,23]. In this paper, we analyze only the AUs that are pain- and emotion-relevant. The appearance and occurrence of facial expressions may vary greatly for different persons, which makes it a challenging and interesting task to recognize and interpret them reliably and precisely. A substantial body of research exists to tackle this challenge by training deep learning models (e.g., convolutional neural networks) to classify AUs in human faces from images [11,29,31,40]. Our approach is the first that adds a quantitative evaluation method to the framework of training, testing and applying deep learning models for facial expression recognition. Our research contributes to the state-of-the-art as follows:

– We propose a domain-specific evaluation framework that allows for integrating and evaluating expert knowledge by quantifying visual explanations. This increases the transparency of black box deep learning models for the user and provides a domain-specific evaluation of the model robustness.

- We show for the application use case of facial expression recognition that the selection and quality of expert knowledge for domain-specific evaluation of explanations has a significant influence on the quality of the robustness analysis.
- We show that the domain-specific evaluation is especially beneficial for challenging use cases such as facial expression recognition based on AUs. AUs are a multi-label classification problem with co-occurring classes. We provide a quantitative evaluation that facilitates analyzing AUs by treating them separately.

This paper is structured as follows: First, the related work gives an overview on similar approaches, then our evaluation framework is presented in Sect. 3 in a step by step manner, explaining the general workflow as well as the specific methods applied for the use case of facial expression recognition. Section 4 presents and discusses the results. Finally, we point out directions for future work and conclude our work.

2 Related Work

Work related to this paper mainly covers the aspect of explaining image classifiers and evaluating the generated explanations with respect to a specific domain. Researchers have developed a vast amount of visual explanation methods for image classification. Among the most popular ones are LIME, LRP, GradCAM, SHAP and RISE (see Schwalbe and Finzel (2023) for a detailed overview and references to various methods [35]). There already exist methods and frameworks that evaluate multiple aspects of visual explanations, e.g., robustness, as provided for example by the Quantus toolbox that examines the impact of input parameter changes on the stability of explanations [12]. Hsieh et al. [15] present feature-based explanations, in particular pixel regions in images that are necessary and sufficient for a prediction, similar to [6], where models are evaluated based on feature removal and preservation. Work that examines the robustness of visual explanations of models applied in different domains, was published for example by Malafaia et al. [28], Schlegel et al. [34] and Artelt et al. [3]. However, these methods do not provide evaluation criteria tailored to the application domain itself. For this purpose, XAI researchers have developed a collection of *application-grounded* metrics [35,42].

Application-grounded perspectives may consider the needs of explanation recipients (explainees) [39,42] and an increase in task performance for applied human-AI decision making [16] or the completeness and soundness of an explanation [21], e.g., with respect to given metrics such as the coverage of relevant image regions [19]. Facial expression recognition, which is the application of this work, is usually a multi-class problem. For multiple classes, application-grounded evaluation may also encompass correlations between ground truth labels of different classes and evaluating, whether learned models follow these correlations [32]. In this work, we focus on evaluating each class separately and whether

visual explanations, generated by explainers for image classification, highlight important image regions.

A review of state-of-the-art and recent works on techniques for explanation evaluation indicates that defining important image regions by bounding boxes is a popular approach. Bounding boxes can be used to compute, whether visual explanations (e.g., highlighted pixel regions) cover important image regions, for classification as well as object detection [17,24,31,35]. In terms of robustness, the robustness of a model is higher if the highlighted pixel regions are inside the defined bounding boxes. With respect to the aforementioned definition of robustness [10], a robust model should show an aggregation of relevance inside the bounding boxes even when out-of-distribution data is encountered. However, bounding boxes are not always suitable to set the necessary boundaries around the important image regions. This can lead to a biased estimation of the predictive performance of a model as bounding boxes usually define areas larger than the region of interest. If models pay attention to surrounding, irrelevant pixels, a bounding box based evaluation may miss this. Hence, as the explanation itself can be biased, the explanation is not robust, which is an important feature of explainability methods [2].

Using polygons as an alternative to bounding boxes is therefore an important step towards integrating domain-specific requirements into the evaluation of explanations to make them more robust. Domain-specific evaluations have not yet been sufficiently discussed across domains, nor broadly applied to the very specific case of facial expression recognition.

In this work, we therefore thoroughly define regions for facial expressions and evaluate the amount of positive relevance inside the defined regions compared to the overall positive relevance in the image (see Sect. 3.5). Instead of using bounding boxes that are very coarse and that might contain class-irrelevant parts of the face as well as background (see Fig. 2), we compute polygons based on class-relevant facial landmarks according to AU masks defined by Ma et al. [27]. We compare a standard bounding box approach with our polygon-based approach for evaluating two state-of-the-art models on two different data sets each and open a broad discussion with respect to justifying model decisions based on visual explanations for domain-specific evaluation. Our domain-specific evaluation framework is introduced in Sect. 3.1.

3 Materials and Methods

The following subsections describe the components of our framework (see Fig. 1 for step numbering), starting with the data sets and evaluated models, and followed by the heatmap generation, and finally our method to quantitatively evaluate the visual explanations by using domain-specific information. Please note that the following paragraphs describe one possible selection of data sets, models, visual explanation method, and explanation evaluation. The framework can be extended or adapted to the needs of other application and evaluation scenarios.

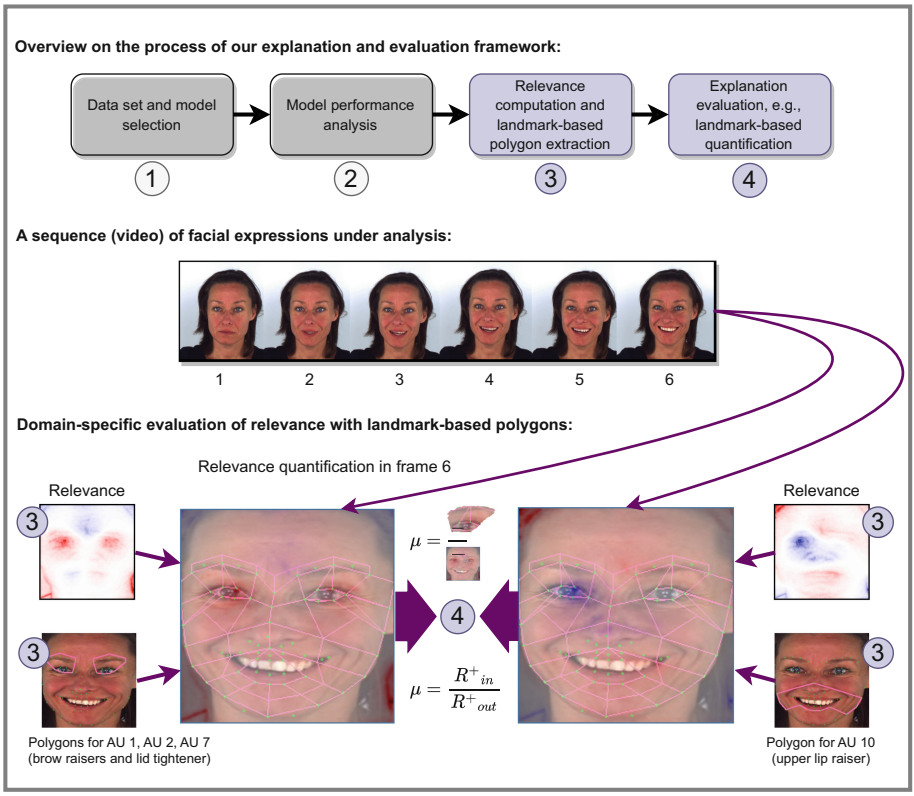


Fig. 1. This figure shows an overview on the components of our framework with exemplary illustrations for the use case of facial expression recognition. The framework processes 4 steps. First, it allows for a flexible and configurable data set and model selection (step 1). Secondly, it analyzes the model's performance with respect to correct predictions (step 2). In step 3, relevance is computed that gets attributed to each pixel by layer-wise relevance propagation. In the same step, polygons are derived from pre-defined domain knowledge in the form of facial landmarks. The aggregation of relevance inside the resulting polygonal image regions gets quantified by our evaluation approach in step 4. For our domain-specific evaluation approach, we consider the positive relevance values computed in step 3 (see red pixel regions in the heatmap-based illustration of relevance). For each image in a video sequence (here: frame 6), we evaluate the aggregation of relevance within the polygons of all predicted AUs (here: AU1, AU2 and AU7, see on the left side of the figure, and AU10, see on the right side of the figure). This is done by dividing the positive relevance aggregation within the region(s) of interest by the total positive relevance within an image (as defined by Eq. 2). With our domain-specific evaluation, a well-performing and robust model would detect positive relevance only within the defined polygonal regions. Deviations of this expectation can be easily uncovered with our framework. (Color figure online)

3.1 Evaluation Framework Overview

Figure 1 presents an overview on the components of our proposed domain-specific evaluation framework. The evaluation framework closes the research gap of providing application-grounded evaluation criteria that incorporate expert knowledge from the facial expression recognition domain. Our framework is intended as a debug tool for developers and as an explanation tool for users that are experts in the domain of facial expression analysis.

In the first step, the data set and a trained classification model, e.g., a convolutional neural network (CNN), is selected. In this paper, we apply two trained CNNs for the use case of facial expression recognition via AUs. In step 2, the model performance is evaluated on images selected by the user with a suitable metric (e.g., F1 score). A visual explanation is generated in step 3, which computes a heatmap per image and per class. The heatmaps display the relevance of the pixels for each output class and can be already inspected by the expert or developer. In the fourth and most crucial step, the domain-specific evaluation based on the visual explanation takes place. By applying domain specific knowledge, it is possible to quantify the visual explanation. For our use case, the user evaluates models with respect to their AU classification using landmark-based polygons that describe the target region in the face. The following subsections describe the four steps in more detail.

3.2 Step 1: Data Set and Model Selection

For the domain-specific evaluation, the Extended Cohn-Kanade [25] (CK+) data set (593 video sequences) and a part of the Actor Study data set [37] (subjects 11–21, 407 sequences) are chosen. The CK+ and Actor Study data set were both created in a controlled environment in the labatory with actors as study subjects.

We evaluate two differently trained models, a model based on the ResNet-18 architecture [30] and a model based on the VGG-16 architecture [38]. They are both CNNs. A CNN is a type of artificial neural network used in image recognition and processing that is specifically designed to perform classification on images. It uses so-called multiple convolution layers to abstract from individual pixel values and weights individual features depending on the input it is trained on. This weighting ultimately leads to a class decision. In the CNNs we use, there is one predictive output for each AU class. AU recognition is a multi-label classification problem, so each image can be labelled with more that one AU, depending on the co-occurrences.

While the ResNet-18 from [31] is trained on the CK+ data set as well as on the Actor Study data set, the VGG-16 is trained on a variety of different data sets from vastly different settings (e.g., in-the-wild and in-the-lab): Actor Study [37] (excluding subjects 11–21), Aff-Wild2 [20], BP4D [41], CK+ [25], the manually annotated subset of EmotioNet [5], and UNBC [26]. We use the same training procedure as in [29] to retrain the VGG-16 without the Actor Study subjects 11–21, which is then our testing data.

With the two trained models we can compare the influence of different training distributions. Furthermore, we apply the domain-specific evaluation with respect to training and testing data. By inspecting explanations for the model on the training data, the inherent bias of the model is evaluated that can arise for example by overfitting on features of the input images. By evaluating the model on the testing data, we can estimate the generalization ability of the model.

The dlib toolkit [18] is used to derive 68 facial landmarks from the images. Based on these landmarks and the expert knowledge about the regions of the AUs, we compute the rectangles and polygons for the evaluation of generated visual explanations.

3.3 Step 2: Model Performance Analysis

For evaluating the model performance, we use the F1 score (Eq. 1), the harmonic mean of precision and recall with a range of [0,1], whereas 1 indicates perfect precision and recall. This metric is beneficial if there is an imbalanced ratio of displayed and non-displayed classes, which is the case for AUs [29]. The ResNet-18 is evaluated with a leave-one-out cross validation on the Actor Study data set, and the performance of the VGG-16 is evaluated on the validation data set, and additionally on the testing part of the Actor Study (subjects 11–21).

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{1}$$

3.4 Step 3: Visual Classification Explanations

We apply layer-wise relevance propagation (LRP) [4] to visually identify the parts of the image which contributed to the classification, i.e., to attribute (positive and negative) relevance values to each pixel of the image. "Positive relevance" denotes that the corresponding pixel influenced the CNN's decision towards the observed class. "Negative relevance" means it influenced the decision against the observed class. For a given input image, LRP decomposes a CNN's output with the help of back-propagation from the CNN's output layer back to its input layer, meaning that each pixel is assigned with a positive or negative relevance score. These relevance scores can be used to create heatmaps by normalizing their values with respect to a chosen color spectrum [4].

We choose the decomposition scheme from Kohlbrenner et al. [19] based on the implementation provided by the *iNNvestigate* toolbox [1]. For the ResNet-18 we select *PresetB* as the LRP analyzer and for the VGG-16 network we select *PresetAFlat*, since these configurations are usually best working for the respective network architectures [19].

3.5 Step 4: Domain-Specific Evaluation Based on Landmarks

As a form of domain-specific knowledge, polygons enclosing the relevant facial areas for each AU are utilized (see Fig. 2). Each polygon is constructed based

on a subset of the 68 facial landmarks to enclose one region. The regions are defined similar to Ma et al. [27].

As motivated earlier, the selection and quality of domain-specific knowledge is of crucial importance. Figure 2b shows a coarse bounding box approach of Rieger et al. [31] and Fig. 2c shows our fine-grained polygon approach exemplary for the AU9 (nose wrinkler). We can see that for b) also the background is taken into account, which makes the quantitative evaluation error-prone. This shows for the use case of AUs, being a multi-class multi-label classification problem, the importance of carefully defining boundaries, so that ideally one boundary only encloses class-relevant facial areas per AUs, which is where our polygon approach aims at.

For our evaluation approach, we consider only positive relevance in heatmaps, since these express the contribution of a pixel to the target class, e.g., a certain AU. However, the evaluation of the aggregation of negative relevance inside boundaries would also be possible, but is not considered here.

For quantitatively evaluating the amount of relevance inside the box or polygon, we use the ratio μ of the positive relevance inside the boundary (R_{in}) and the overall positive relevance in the image (R_{tot}) (Eq. 2). To make our approach comparable, we use the same equation as Kohlbrenner et al. [19].

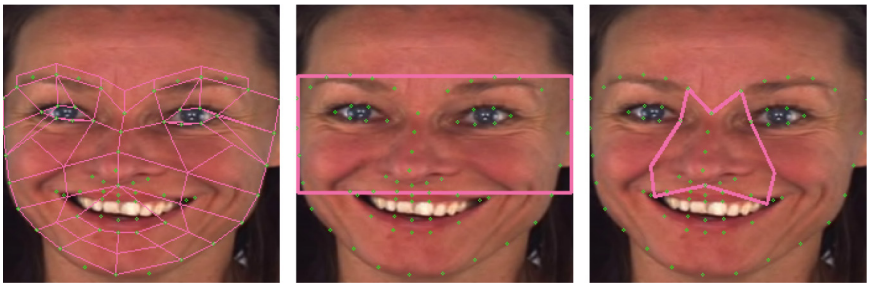
$$\mu = \frac{R_{in}}{R_{tot}} \tag{2}$$

The μ-value ranges from 0 (no positive relevance inside the boundary) to 1 (all positive relevance inside the boundary). High μ-values indicate that a CNN based its classification output on the class-relevant parts of the image. This means, that for a μ-value above the value of 0.5, the majority of relevance aggregates inside the boundaries.

For our evaluation, we consider only images for which the ground truth as well as the classification output match in the occurrence of the corresponding AU.

4 Results and Discussion

Table 1 shows the overview of the performance and domain-specific evaluation of the VGG-16 model. The performance on the validation data set differs greatly for some AUs (e.g., AU10, or AU14), which can be explained by the big array of different training data sets. Henceforth, the data distribution of the Actor Study is not predominantly represented by the trained model. We may keep in mind that the Actor Study is a posed data set, so some facial expression can differ in their visual appearance from the natural ones. However, when looking at the average μ_{poly}-values of the polygon boundaries, we can see a correlation of the higher μ-values with the validation performance for some AUs. For example, the model displays a good performance on the validation data set for AU10, but a significantly lower one on the testing data set. However, in comparison, the μ_{poly}-value is the highest of all evaluated AUs. A similar pattern can be found for example for AU14. Since we only use the correctly classified images for our



(a) 68 facial landmarks [27] (b) AU9 bounding box [31] (c) AU9 polygon (ours)

Fig. 2. The domain-specific knowledge for evaluating the heatmaps are facial landmarks. Exemplary image with emotion *happy* and highlighted region for Action Unit 9 (AU9) (nose wrinkler). AU region boundaries are pink and facial landmarks green dots. (Color figure online)

domain-specific evaluation, we can interpret that the model can locate the region for AU10 or AU14, but that there are probably many out-of-distribution images for these AUs in the testing data set, hence making a good model performance difficult. We can also observe that for instance for AU25, there is a strong performance on both the validation and testing data set, but a low μ-value, which can indicate that the model did not identify the expected region as important.

Table 2 shows a comparison between our polygon approach μ_{poly} with the standard bounding box approach μ_{box} [31] for the ResNet-18. The bounding box approach μ_{box} yields overall higher μ-values than the polygons (μ_{poly}), which is expected since the boxes enclose a larger area than the polygons. This can also indicate that the coarse boxes contain pixels that get assigned with relevance by the ResNet-18, although they are not located in relevant facial areas, hence highlighting once more the importance of the quality of the domain-specific knowledge. Our polygons enclose in contrast to the bounding boxes only class-relevant facial areas. Looking closely at the AUs, we can see that although the μ_{box} is high for AU4, it has also the highest difference to μ_{poly} for both the data sets CK+ and Actor Study. We can therefore assume a high relevance spread for AU4, which is ultimately discovered by applying the fine-grained polygon approach. In contrast, AU10 looses the least performance for both data sets concerning the μ-value, but displays also the lowest F1 value, which can indicate that although the AU is not accurately predicted in a lot of images, the model has nonetheless learned to detect the right region for images with correct predictions.

Overall, the μ-values are low for all classes, indicating a major spread of relevance outside of the defined boxes. Some of the relevance may be outside of the polygons due to a long tail distribution across the image with a lot of pixels having a low relevance value. This can lead to low μ-values for all polygons. When comparing the μ_{poly} with the μ_{box} approach, it is apparent that the μ_{box}-values are higher compared to the μ_{poly}-values, and only μ_{box}-values reach an average μ-value above of 0.5 across data sets. Both findings show the need for a

Table 1. Classification performance and domain-specific evaluation of the VGG-16 model. The performance is measured by the F1 score on the validation and testing data set respectively. The domain-specific evaluation is measured with the average μ-values of the polygones on the testing data set. The testing data set is the Actor Study data set, subjects 12–21. Best results are in bold.

AU	F1 score		av. μ_{poly}
	validation	testing	
1	0.61	**0.71**	0.125
2	0.42	0.58	0.155
4	0.51	0.56	0.114
6	0.72	0.50	0.137
7	0.79	0.44	0.092
10	**0.83**	0.19	**0.239**
12	0.77	0.58	0.220
14	0.77	0.13	0.221
15	0.62	0.13	0.215
17	0.67	0.43	0.008
23	0.49	0.05	0.026
24	0.63	0.19	0.037
25	0.66	0.69	0.036

Table 2. Comparison of our approach μ_{poly} with the standard bounding box approach μ_{box} [31] for ResNet-18. Highest values are in bold.

AU	F1	CK+			Actor Study		
		μ_{box}	μ_{poly}	$\lvert\mu_{box} - \mu_{poly}\rvert$	μ_{box}	μ_{poly}	$\lvert\mu_{box} - \mu_{poly}\rvert$
04	0.68	**0.579**	0.118	**0.461**	0.458	0.061	**0.397**
06	0.55	0.432	0.144	0.288	0.360	0.087	0.273
07	0.62	0.499	0.097	0.402	0.417	0.038	0.379
09	0.48	0.492	0.195	0.297	0.293	0.084	0.209
10	0.32	0.350	**0.339**	0.011	0.197	0.196	0.001
25	**0.83**	0.413	0.211	0.202	0.457	0.175	0.282
26	0.60	0.330	0.143	0.187	0.493	0.201	0.292
27	0.57	0.463	0.203	0.260	**0.545**	**0.223**	0.322

domain-specific evaluation with carefully selected expert knowledge in order to assess a model's performance as good as possible but also the precision of used visual explainers with respect to the spread of relevance.

Furthermore, our approach emphasizes the general need of an evaluation beyond classification performance of models. Although the models display high F1 scores for most of the classes, the relevance is not in the expected areas.

A limitation of our evaluation results is that they do not consider μ-values normalized according to the size of regions, although our approach allows such an extension in principle. This is an important aspect, since the areas for each AU are differently sized in relation to the overall image size. This means that some AU boundaries may be more strict on the relevance distribution than others and may penalize the model's performance thereof. For that we suggest a weighted μ-value calculation, optimally with respect to the overall relevance distribution in an image, e.g., based on thresholding the relevance [8].

5 Conclusion

In this paper, we present an approach for domain-specific evaluation of visual explanation methods in order to enhance the transparency of CNNs and estimate their robustness as precisely as possible. As an example use case, we applied our framework to facial expression recognition. We showed that the domain-specific evaluation can give insights into facial classification models that domain-agnostic evaluation methods or performance metrics cannot provide. Furthermore, we could show by comparison that the quality of the expert knowledge is of great importance for assessing a model's performance precisely.

References

1. Alber, M., et al.: iNNvestigate neural networks! J. Mach. Learn. Res. **20**(93), 1–8 (2019)
2. Alvarez-Melis, D., Jaakkola, T.: On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049 (2018)
3. Artelt, A., et al.: Evaluating robustness of counterfactual explanations. In: Proceedings of Symposium Series on Computational Intelligence, pp. 1–9. IEEE (2021). https://doi.org/10.1109/SSCI50451.2021.9660058
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), 01–09 (2015). https://doi.org/10.1371/journal.pone.0130140
5. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5562–5570. IEEE (2016). https://doi.org/10.1109/cvpr.2016.600
6. Dhurandhar, A., et al.: Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Proceedings of Advances in Neural Information Processing Systems, vol. 31, pp. 590–601 (2018)
7. Ekman, P., Friesen, W.V.: Facial Action Coding Systems. Consulting Psychologists Press (1978)
8. Finzel, B., Kollmann, R., Rieger, I., Pahl, J., Schmid, U.: Deriving temporal prototypes from saliency map clusters for the analysis of deep-learning-based facial action unit classification. In: Proceedings of the LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR. CEUR Workshop Proceedings, vol. 2993, pp. 86–97. CEUR-WS.org (2021)

9. Finzel, B., Tafler, D.E., Thaler, A.M., Schmid, U.: Multimodal explanations for user-centric medical decision support systems. In: Proceedings of the AAAI Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN). CEUR Workshop Proceedings, vol. 3068. CEUR-WS.org (2021)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009
11. Hassan, T., et al.: Automatic detection of pain from facial expressions: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 1815–1831 (2019)
12. Hedström, A., et al.: Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond. J. Mach. Learn. Res. **24**(34), 1–11 (2023)
13. Holzinger, A.: The next frontier: AI we can really trust. In: Kamp, M., et al. (eds.) ECML PKDD 2021. CCIS, vol. 1524, pp. 427–440. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93736-2_33
14. Holzinger, A., et al.: Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf. Fusion **79**, 263–278 (2022). https://doi.org/10.1016/j.inffus.2021.10.007
15. Hsieh, C., et al.: Evaluations and methods for explanation through robustness analysis. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021). OpenReview.net (2021)
16. Jesus, S.M., et al.: How can I choose an explainer?: an application-grounded evaluation of post-hoc explanations. In: Proceedings of Conference on Fairness, Accountability, and Transparency (FAccT 2021), pp. 805–815. ACM (2021). https://doi.org/10.1145/3442188.3445941
17. Karasmanoglou, A., Antonakakis, M., Zervakis, M.E.: Heatmap-based explanation of YOLOv5 object detection with layer-wise relevance propagation. In: Proceedings of International Conference on Imaging Systems and Techniques, (IST), pp. 1–6. IEEE (2022). https://doi.org/10.1109/IST55454.2022.9827744
18. King, D.E.: Dlib-ML: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
19. Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., Lapuschkin, S.: Towards best practice in explaining neural network decisions with LRP. In: Proceedings of International Joint Conference on Neural Networks (IJCNN 2020), pp. 1–7. IEEE (2020). https://doi.org/10.1109/IJCNN48605.2020.9206975
20. Kollias, D., Zafeiriou, S.: Aff-wild2: extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
21. Kulesza, T., Stumpf, S., Burnett, M.M., Yang, S., Kwan, I., Wong, W.: Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, pp. 3–10. IEEE Computer Society (2013). https://doi.org/10.1109/VLHCC.2013.6645235
22. 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)
23. Kunz, M., Meixner, D., Lautenbacher, S.: Facial muscle movements encoding pain-a systematic review. Pain **160**(3), 535–549 (2019)
24. Lin, Y., Lee, W., Celik, Z.B.: What do you see?: evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

(KDD 2021), Virtual Event, Singapore, pp. 1027–1035. ACM (2021). https://doi.org/10.1145/3447548.3467213
25. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J.M., Ambadar, Z., Matthews, I.A.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, pp. 94–101. IEEE Computer Society (2010). https://doi.org/10.1109/CVPRW.2010.5543262
26. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.A.: Painful data: the UNBC-McMaster shoulder pain expression archive database. In: Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, pp. 57–64. IEEE Computer Society (2011). https://doi.org/10.1109/FG.2011.5771462
27. Ma, C., Chen, L., Yong, J.: AU R-CNN: encoding expert prior knowledge into R-CNN for action unit detection. Neurocomputing **355**, 35–47 (2019). https://doi.org/10.1016/j.neucom.2019.03.082
28. Malafaia, M., Silva, F., Neves, I., Pereira, T., Oliveira, H.P.: Robustness analysis of deep learning-based lung cancer classification using explainable methods. IEEE Access **10**, 112731–112741 (2022). https://doi.org/10.1109/ACCESS.2022.3214824
29. Pahl, J., Rieger, I., Seuss, D.: Multi-label learning with missing values using combined facial action unit datasets. In: The Art of Learning with Missing Values Workshop at International Conference on Machine Learning (ICML 2020) abs/2008.07234 (2020). https://arxiv.org/abs/2008.07234
30. Rieger, I., Hauenstein, T., Hettenkofer, S., Garbas, J.: Towards real-time head pose estimation: exploring parameter-reduced residual networks on in-the-wild datasets. In: Wotawa, F., Friedrich, G., Pill, I., Koitz-Hristov, R., Ali, M. (eds.) IEA/AIE 2019. LNCS, vol. 11606, pp. 123–134. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22999-3_12
31. Rieger, I., Kollmann, R., Finzel, B., Seuss, D., Schmid, U.: Verifying deep learning-based decisions for facial expression recognition. In: 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2020), Bruges, Belgium, pp. 139–144 (2020). https://www.esann.org/sites/default/files/proceedings/2020/ES2020-49.pdf
32. Rieger, I., Pahl, J., Finzel, B., Schmid, U.: CorrLoss: integrating co-occurrence domain knowledge for affect recognition. In: Proceedings of the 26th International Conference on Pattern Recognition (ICPR 2022), pp. 798–804. IEEE (2022)
33. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
34. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of XAI methods on time series. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW 2019), Seoul, Korea (South), pp. 4197–4201. IEEE (2019). https://doi.org/10.1109/ICCVW.2019.00516
35. Schwalbe, G., Finzel, B.: A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min. Knowl. Discov. 1–59 (2023). https://doi.org/10.1007/s10618-022-00867-8
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7

37. Seuss, D., et al.: Emotion expression from different angles: a video database for facial expressions of actors shot by a camera array. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019), Cambridge, United Kingdom, pp. 35–41. IEEE (2019). https://doi.org/10.1109/ACII.2019.8925458
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA (2015). https://arxiv.org/abs/1409.1556
39. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021). https://doi.org/10.1016/j.inffus.2021.05.009
40. Werner, P., Martinez, D.L., Walter, S., Al-Hamadi, A., Gruss, S., Picard, R.W.: Automatic recognition methods supporting pain assessment: a survey. IEEE Trans. Affect. Comput. **13**(1), 530–552 (2022). https://doi.org/10.1109/TAFFC.2019.2946774
41. Zhang, X., et al.: BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. **32**(10), 692–706 (2014)
42. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics **10**(5), 593 (2021)