Affective Human-Robot Interaction with Multimodal Explanations

Hongbo Zhu(✉), Chuang Yu, and Angelo Cangelosi

The University of Manchester, Manchester, UK
{hongbo.zhu,chuang.yu,angelo.cangelosi}@manchester.ac.uk

Abstract. Facial expressions are one of the most practical and straightforward ways to communicate emotions. Facial Expression Recognition has been used in lots of fields such as human behaviour understanding and health monitoring. Deep learning models can achieve excellent performance in facial expression recognition tasks. As these deep neural networks have very complex nonlinear structures, when the model makes a prediction, it is not easy for human users to understand what is the basis for the model's prediction. Specifically, we do not know which facial units contribute to the classification more or less. Developing affective computing models with more explainable and transparent feedback for human interactors is essential for a trustworthy human-robot interaction. Compared to "white-box" approaches, "black-box" approaches using deep neural networks, which have advantages in terms of overall accuracy but lack reliability and explainability. In this work, we introduce a multimodal affective human-robot interaction framework, with visual-based and verbal-based explanation, by Layer-Wise Relevance Propagation (LRP) and Local Interpretable Mode-Agnostic Explanation (LIME). The proposed framework has been tested on the KDEF dataset, and in human-robot interaction experiments with the Pepper robot. This experimental evaluation shows the benefits of linking deep learning emotion recognition systems with explainable strategies.

Keywords: Explainable robotics · Facial Expression Recognition (FER) · eXplainable Artificial Intelligence (XAI) · Human-Robot Interaction (HRI)

1 Introduction

Facial expression is a critical non-verbal communication strategy, and human emotions can be expressed through facial expressions, which can be read and interpreted by emotional AI technology [7,27]. Face expression detection is significant for patients with specific diseases or congenital disabilities [13], especially when they cannot express their thoughts through words and actions. In this case, real-time facial emotion detection needs to be performed to take corresponding medical measures for the patient.

Supported by University of Manchester and UKRI Node on Trust (EP/V026682/1).

F. Cavallo et al. (Eds.): ICSR 2022, LNAI 13817, pp. 241–252, 2022.
https://doi.org/10.1007/978-3-031-24667-8_22

The advancement of AI poses challenges for humans to trace model results, especially in the field of deep learning. It is difficult for data scientists and even engineers who write AI algorithms to explain what is happening inside the models and how these AI models come to specific results [1]. XAI is proposed to address this dilemma, which is a set of methods and processes that enable users to understand the output of AI models [2]. AI developers need to have a comprehensive understanding and awareness of the working mechanism, to monitor whether the working process of the model complies with regulations, thereby reducing legal and security risks and gaining the user trust.

In this work, we explored how explainable methods (namely LRP and LIME) could make facial emotion recognition more transparent and trustworthy with visual and verbal explanations. In the visual interpretation extraction part, LRP was utilized to provide a visual explanation on a CNN-based emotion classifier. For the verbal interpretation extraction part, Openface [4] was used to recognise face action units and calculate the related intensity. Then LIME was employed to analyse the contribution of each Action Unit(AU) for model prediction.

The pipeline of our model is shown in Fig. 1. Firstly, the Pepper robot predicts the facial emotion states of the interactor during HRI. Then, Pepper verbalises the predicted emotion as linguistic feedback and shows the heatmap generated from the LRP model as explainable visual feedback. In addition, the robot can give more detailed emotion recognition feedback to increase interaction transparency. This multimodel explanation feedback can help the human interactor understand the robot's internal emotion recognition process, facilitating human-robot trust.

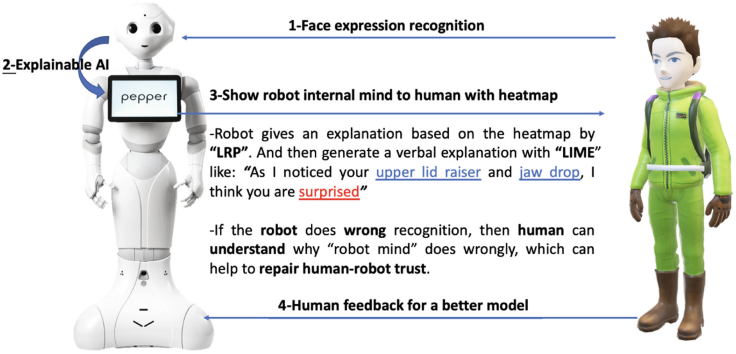

Fig. 1. The proposed multimodel explanation framework

All in all, our paper contributions are as follows:

– We retrained a deep learning model of VGG16 to perform emotion recognition task on KDEF dataset, and LRP was utlized to highlight the crucial pixel-features of the input image and generate heatmap-based explanation.

– We made use of Openface to detact AU and calcualted the corresponding intensity, then random forest was used perform emotion prediction. Finally AU-based explanation was generated by LIME.
– The proposed multimodal explainable method was tested on Pepper robot, the generated heatmap is shown on the screen of chest. Verbal explanation based on AU from LIME is given at the same time. Finally, trust is constructed and the feedback from the interactor can be used to improve the facial expression recognition (FER) model.

2 Related Works and Background

The Deep learning-based model and Facial Action Coding Systems (FACS) based model are two mainstream methods for facial emotion recognition [28]. Compared with traditional ML methods, deep learning-based black-box methods have higher accuracy but usually lack reliability and interpretability due to the complex network structure. Explainable AI is proposed to solve this challenge. Common explainable methods are backpropagation-based Layer-Wise Relevance Propagation (LRP) [3] and perturbation-based Local Interpretable Model-agnostic Explanations (LIME) [17]. The main goal of these methods is to find activation regions in the DL model and highlight the parts of the input image that have a decisive influence on the classifier's decision. While these methods account for the contribution of the input image at the pixel level, they do not give an explanation at the facial action unit level. Facial Action Coding System (FACS) [6] is a standard of most FER models for estimating and recognising AUs. It is based on the activation of facial muscles during facial expressions. These activations are represented by AUs. Action units (AUs) [22] were mostly used as features, feeding classifiers to recognize emotions. AUs are defined as subtle facial muscle movements. According to the physiological distribution of facial muscles and related characteristics, the movements of different facial muscles can be classified into different AUs [23].

Numerous interpretable techniques have been deployed to explain the dynamics process of AI models. We explored backpropagation-based and perturbation-based explainable methods and use them to develop our multimodel explanation architecture for FER in Human-Robot interaction.

2.1 Backpropagation-Based Explanation

Backpropagation is an internal algorithm common across neural network architectures. It is used to calculate the gradient of the loss function with regard to the weights of the connection between the layers of the network, and understand the correlation between the input and output to a network [15]. As for backpropagation-based explainable methods, attributions are calculated by backpropagating once or more times through the network.

Layer-Wise Relevance Propagation is a backpropagation-based interpretable method [14]. It calculates importance scores in a layer-by-layer approximation of

backpropagation, which does not interact with the training process of the network and can be easily applied to an already trained DNN model. LRP provides an intuitive human-readable heatmap of input images at the pixel level. It uses the network weights and the neural activation created by the forward-pass to propagate the output back through the network from the predicted output to the input layer [16]. The heatmap is used to visualize the contribution of each pixel to the prediction. The contribution of the intermediate neuron or each pixel is quantified to relevance value R, representing how important the given pixel is to a particular prediction.

2.2 Perturbation-Based Explanation

The perturbation-based XAI method modifies the input of the model to investigate which parts of the input elements are more critical for the model's predictions [8]. Specifically, the disturbance is generated by occluding some pixels or replacing some words in the sentence, then observing the changes in the output. If the input after the disturbance significantly changes the output, it is considered that "the cause behind the disturbance" is very significant. The perturbation-based interpretable methods are generally applicable to the vast majority of deep learning models [18].

Local Interpretable Model-agnostic Explanations is a commonly used post-hoc perturbation-based explainable model [12]. It can generate instance-based explanations for the model predictions. For a given input sample Y, LIME generates perturbed data near Y. The weights of the perturbed data are calculated according to how close they are to the sample Y. LIME then trains an interpretable sparse linear model on the perturbed dataset as a local approximate classifier. In contrast to most backpropagation-based algorithms that need to use the internal information of the classification model to generate explanations, LIME generates explanations without accessing the model's internals. [8].

2.3 Emotion Recognition for HRI

Unlike FER in human-computer interaction, the position of the face relative to the camera is relatively fixed. Robotic emotion recognition is related to environmental factors, making it an extremely challenging task to enable the robot to understand emotions [25]. Emotional robots have many real-life applications, and studies have shown that in treating children with autism, they are more inclined to interact with robots than humans [21]. Therefore, building a robotic system with emotional intelligence will help detect the emotional state of autistic children in real-time, thereby providing more efficient treatment. By dynamically interacting with the external environment, emotional robots can also learn better adaptability and flexibility [26].

3 Methodology

In this paper, explainable emotion recognition was explored in HRI through backpropagation-based model for visual explanation and perturbation-based model for verbal explanation, which will be introduced below.

3.1 Visual-Based Explanation

As for visual-based explanation, we explored explainable facial emotion recognition with LRP. This can explain inputs' relevance for a certain prediction, typically for image processing. Using LRP, the robot can extract the facial heatmap that highlights sufficient parts of the facial pixels most responsible for the emotion prediction task. Face recognition with VGG16 is used in this paper. VGG16 [5] is a simple and widely used convolutional neural network model for image classification tasks. In this work, we reused the pre-trained image classification VGG16 model to fine-tune it for our face emotional recognition task.

The input images of this model are with a fixed size of $224 * 224$. The image is passed through a series of convolutional layers, following three fully-connected layers with different depths. As our task has seven emotions, the final fully-connected layer was changed to seven dimensions, as shown in Fig. 2. The whole pre-trained model is further trained on the facial emotional dataset for facial emotion recognition. During this new training, the previous layers of the pre-trained VGG16 are kept fixed.

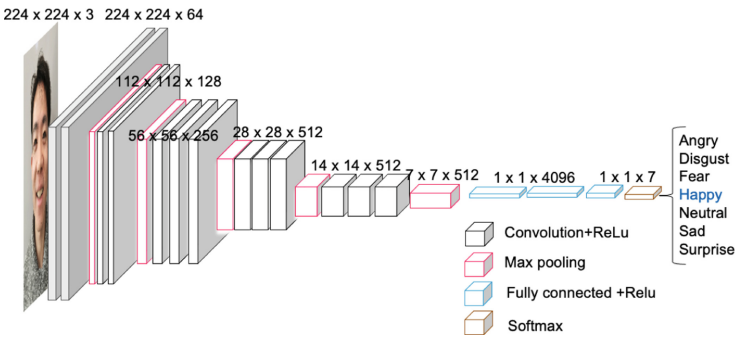

Fig. 2. The architecture of modified VGG16.

Layer-Wise Relevance Propagation (LRP) can explain the relevance of inputs for a certain prediction, typically for image processing. So we can see which part of the input images, or more precisely which pixels, most contribute to a specific prediction. As a model-specific method, LRP generally assumes that the classifier can be decomposed into several layers of computation [24]. In the forward pass process, the image goes through a convolutional neural network for feature extraction from the image. Then, these extracted features are input

to a classifier with a fully connected neural network and a softmax layer which gives the final prediction. At this point, we are interested in why the model gets that prediction. LRP goes in reverse order over the layers, we have visited in the forward pass and calculates the relevant scores for each of the neurons in the layer until we arrive at the input again. We can then calculate the relevance for each pixel of the input image. The positive relevant scores indicate how much contributions the pixels make to the model prediction, and the negative values mean these pixels would speak against it, which leads to the heatmap result.

When the LRP model is applied to the trained neural network, it propagates the classification function $f(x)$ backward in the network through pre-defined propagation rules from the output layer to the input layer. Let j and k be neurons at two continuous layers. The propagating relevance scores $R_k^{(l+1)}$ at a given layer $l+1$ onto neurons of the lower layer l is achieved by applying the following rule [20]:

$$R_j^{(l)} = \sum_k \frac{Z_{jk}}{\sum_{j'} Z_{j'k}} R_k^{(l+1)} \tag{1}$$

The quantity Z_{jk} models how much importance the neuron j has contributed to making the neuron k relevant.

$$Z_{\mathrm{jk}} = \mathrm{x}_{\mathrm{j}}^{(l)} \mathrm{w}_{\mathrm{jk}}^{(l,l+1)} \tag{2}$$

The relevance of a neuron is calculated according to Formula 1, which can calculate the relevance R for a neuron j in layer l. So our current layer is l, and the output layer becomes $l+1$. The calculation for neuron j now works as follows. For each neuron j in the layer l, we calculate the activation based on the neuron j. And the activation is calculated according to Z_{jk}. It simply multiplies the input for the neuron j in our current layer, with the weight that goes into the neuron k in the next layer. This input x comes from passing the pixel values through the previous layers, showing how strong the activation is between these neurons. Intuitively, if there is a high value, it means that the neuron was very important for the output. So we interpret this fraction as a relative activation of a specific neuron, compared to all activations in that layer. Finally, we multiply the relevant score of the neuron in the next layer with this relative value to propagate the relevance of the next layer backwards. The propagation procedure will not stop until it reaches the input layer.

3.2 Verbal-Based Explaination

According to the facial action units (AUs) that make up an expression, FACS[1] divides the face into upper and lower parts and subdivides the facial action units into different AUs to encode facial emotions, shown in Fig. 3. Openface [4] can detect action units and identify the corresponding intensity of each activated AU as an open source software. To explain the relationship between action units and

[1] https://imotions.com/blog/facial-action-coding-system/.

emotion, LIME was used to calculate and visualize the contribution of each AU to the predicted emotion in our work. Each AU represents a facial behaviour generated with an anatomically distinct facial muscle group [10]. The combination of AUs can produce most facial expressions, and the goal of facial AU recognition is to detect AU and calculate AU intensity for each input face expression. Here Openface was used in our work to detect and estimate the intensity of AUs from input images, which is shown in Fig. 4.

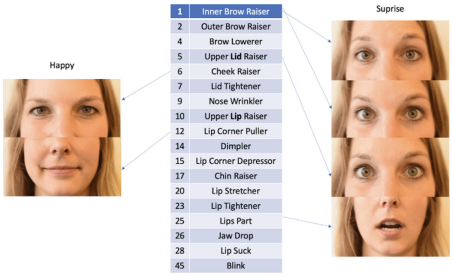

Fig. 3. The illusration of AUs.

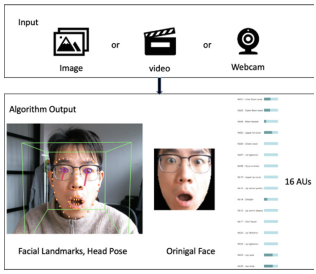

Fig. 4. The input and output of Openface.

Local Interpretable Model-agnostic Interpretation (LIME) aims to explain any black-box model by creating a local approximation, which can approximate the original model in the vicinity of an individual instance. It works on almost any input format, such as text, tabular data, images or even graphs.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g) \tag{3}$$

The idea behind LIME is quite intuitive. For instance, we know the properties of input data point x in a tabular format. In this optimization formula above, the complex model is denoted with f and the simple model or the local model is denoted with g. In this simple model, small g comes from a set of interpretable models which are denoted with a capital G, here capital G is a family of sparse linear models, such as linear regression.

The first loss term L try to find an approximation of the complex model f by the simple model g in the neighbourhood of our data point x. In other terms, we want to get a good approximation in the local neighbourhoods. The third argument π here defines the local neighbourhoods of that data point and is some sort of proximity measure.

The second loss term Ω is used to regularize the complexity of our simple surrogate model for linear regression. For instance, a desirable condition could be to have many zero-weighted input features, so ignoring most of the features and just including a few makes our explanations simpler for the decision tree. It makes sense to have a relatively small depth that stays comprehensible for humans. So overall, this Ω is a complexity measure, and as this optimization

problem is a minimization problem, we are trying to minimize Ω. In summary, this loss function says that we look for a simple model g.

To minimize those two-loss terms, it should approximate the complex model in that local area and stay as simple as possible. In the first step, we simply generate some new data points in the neighbourhood of our input data point. More specifically, we randomly generate data points everywhere, but they will be weighted according to the distance to our input data point. As we are just interested in the local area around our input. These data points are generated by perturbations. This can be achieved by sampling from a normal distribution with the mean and standard deviation for each feature. Then we get the prediction for these data points using our complex model f.

$$\mathcal{L}\left(f, g, \pi_x\right) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z)\left(f(z) - g\left(z'\right)\right)^2 \tag{4}$$

We minimize the first loss term by getting the highest accuracy on that new data set using a simple linear model for linear regression. For instance, we minimize the sum of square distances between the predictions and the ground truth. Then a loss function is used to optimize the linear model. It's basically the sum of squared distances between the label, which comes from the complex model f and the prediction of the simple model g [9]. Additionally, the proximity π is added to weight the loss, according to how close a data point is. Here exponential kernel is used as a distance metric, so we can think of this like a heatmap. The points that are close to our input data points are weighed the most. That is how we ensure that the model is locally faithful. The second loss term Ω is used to make sure that our model stays simple. In LIME, a sparse linear model is used. In practice, this can be achieved by using a regularization technique. This way we ensure to get a simple explanation with only a few relevant variables. In summary, LIME fits a linear interpretable model in that local area, which is a local approximation of a complex model.

4 Model Evaluation and Results

4.1 KDEF Datasets and Pre-processing

The Karolinska Directed Emotional Faces (KDEF) [11] dataset consists of 4900 facial expression photos with 70 individuals (half males and half females, ages from 20 to 30). Each person imitates seven different facial emotions and, each facial expression is recorded from five camera views. In this paper, we only use the front face photos in our experiment as our robot mainly interacts with a human user in a front view. Some examples are as shown in Fig. 5. That means we used one-fifth of the dataset, 980 pictures in total, so each emotion subset contains 140 front view images for each expression. The face images were rescaled to a standard 224 * 224 pixels and three colour channels, to fit the input format of the classification model. And we randomly split the front-face dataset into the trainning part, validation part and testing part with a ratio of 700:140:140.

Fig. 5. Sample of the KDEF dataset

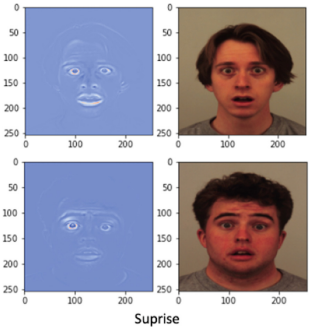

Fig. 6. Visual based explanation

4.2 Multimodal Explanation

In the affective HRI, the robot will not only recognize the human interactor emotion but also provide the multimodal explainable feedbacks, including visual feedback with explainable heatmap that illustrates emotion recognition contribution extracted from LRP model and verbal feedback with understandable robot speech to explain the face AU activation for emotion recognition.

Based on the pre-trained VGG16 model in the visual explanation part, our face emotion recognition model is further trained on an Nvidia RTX 2080Ti graphic card. We set the batch size to 32 and used Adam as the optimization algorithm, with the learning rate of 0.00001. After 250 epochs of training, the model achieves a classification testing accuracy of 91.4% on the KDEF dataset. The predicted result and model parameters were fed to LRP, and then the pixel wize contribution was calculated and shown on the heatmap for an explanation. For example, the comparison of the two heatmap images in Fig. 6 shows that the robot uses similar feature pixels in two different faces. This means that when VGG16 classifies the face as a 'surprise' emotion, the robot relies more on feature pixes near the eyes, nostrils and lips to make its prediction, which is in line with theories of human emotion perception and cognition [19].

In the verbal explanation part, we use Openface to extract the activation of 16 AUs used for emotion recognition with random forest. Finally, the AUs-based explanation chart was generated by LIME, as shown in Fig. 7. The blue bar indicates positive contribution while the orange bar indicates the negative contribution of surprise prediction. According to the histogram, AU26 (Jaw Drop) and AU05 (Upper Lid Raiser) make the biggest positive contribution to the prediction. Then the blanks of the predefined text template were filled with the AU names that make most significant contribution. Finally, Text-to-Speech (TTS) generates voice explanations for robot speech.

4.3 Test on Pepper Robot

In this work, we have tested our multimodel explanation methods on the Pepper robot. During the experiments, a person interacts with the Pepper robot, who can simultaneously recognise their facial expressions and verbalise the human face emotion prediction as verbal feedback in HRI. The related speech is generated based on the emotion recognition results. For example, if the emotion recognition result is *happy*, the explainable speech sentence will be *As I noticed your Cheek Raiser and Lip Corner Puller, I think you are happy*. The speech voice is synthesized through the Text-To-Speech (TTS) tool of the Naoqi SDK of the Pepper robot. And based on the LRP model, the robot can extract the heatmap images as the pixel-level explanation for the interactor. The original face and the heatmap face are be shown in the Pepper chest screen as interpretable visual feedback. Through verbal and visual feedback, this explainable system has the benefit of supporting trustworthy human-robot interaction.

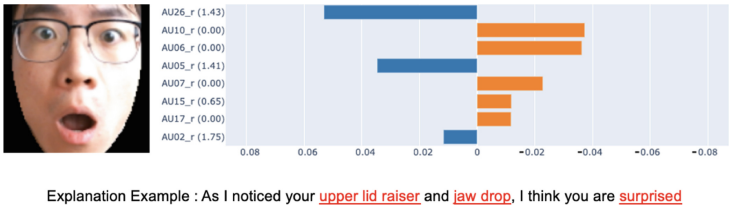

Fig. 7. An example of AUs-based explanation for surprise emotion

5 Conclusion and Future Work

Robotic systems may become more commonplace, but at the same time, more complex. When robots fail to express their intentions, people will feel not only discomfortable but also untrustworthy. It is necessary for people to know how the robot recognize human emotion to assess when such systems can be trusted, even if robots follow a reasonable decision-making process.

In conclusion, this paper integrates two explainable methods in emotion recognition for trustworthy HRI. Using the explainable method LRP, the robot can extract the facial heatmap that highlights significant parts of the facial pixels most responsible for the emotion prediction task. The visualized attention heatmap and verbal feedback, can help the user understand the perceptual mechanism the robot uses to recognise emotions. Thus the explainable method provides essential insights into the natural features of the prediction model.

In this work, we just completed the essential human facial emotion recognition with related explanation, but have not conducted much work on trust validation with our explainable model in human-robot interaction scenes. As for future work, more human-joined tests will be taken for trust evaluation to explore the effectiveness of our multimodel explanation. And we will explore

how we can use the feedback for human-in-the-loop robot learning to improve the robot's emotional perception ability in dynamic HRI scenes.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Arrieta, A.B., et al.: Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
3. 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), e0130140 (2015)
4. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
5. Dubey, A.K., Jain, V.: Automatic facial recognition using vgg16 based transfer learning model. J. Inf. Optim. Sci. **41**(7), 1589–1596 (2020)
6. Ekman, P., Friesen, W.V.: Facial action coding system. Environ. Psychol. Nonverbal Behav. (1978)
7. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the human face: Guidelines for research and an integration of findings, vol. 11. Elsevier (2013)
8. Ivanovs, M., Kadikis, R., Ozols, K.: Perturbation-based methods for explaining deep neural networks: a survey. Pattern Recogn. Lett. **150**, 228–234 (2021)
9. Kavila, S.D., Bandaru, R., Gali, T.V.M.B., Shafi, J.: Analysis of cardiovascular disease prediction using model-agnostic explainable artificial intelligence techniques. In: Principles and Methods of Explainable Artificial Intelligence in Healthcare, pp. 27–54. IGI Global (2022)
10. Lien, J.J., Kanade, T., Cohn, J.F., Li, C.C.: Automated facial expression recognition based on facs action units. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 390–395. IEEE (1998)
11. Lundqvist, D., Flykt, A., Öhman, A.: Karolinska directed emotional faces. Cogn. Emot. (1998)
12. Malik, S., Kumar, P., Raman, B.: Towards interpretable facial emotion recognition. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1–9 (2021)
13. Martinez, M., et al.: Emotion detection deficits and decreased empathy in patients with alzheimer's disease and parkinson's disease affect caregiver mood and burden. Front. Aging Neurosci. **10**, 120 (2018)
14. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 193–209 (2019)
15. Nie, W., Zhang, Y., Patel, A.: A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: International Conference on Machine Learning, pp. 3809–3818. PMLR (2018)
16. Rathod, J., Joshi, C., Khochare, J., Kazi, F.: Interpreting a black-box model used for scada attack detection in gas pipelines control system. In: 2020 IEEE 17th India Council International Conference (INDICON), pp. 1–7. IEEE (2020)
17. 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)

18. Robnik-Šikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. In: Zhou, J., Chen, F. (eds.) Human and Machine Learning. HIS, pp. 159–175. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90403-0_9
19. Rosenberg, E.L., Ekman, P.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, Oxford (2020)
20. 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. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
21. Taheri, A., Meghdari, A., Alemi, M., Pouretemad, H.: Human-robot interaction in autism treatment: a case study on three pairs of autistic children as twins, siblings, and classmates. Int. J. Social Rob. **10**(1), 93–113 (2018)
22. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach, Intell. **23**(2), 97–115 (2001)
23. Yao, L., Wan, Y., Ni, H., Xu, B.: Action unit classification for facial expression recognition using active learning and svm. Multimedia Tools Appl. **80**(16), 24287–24301 (2021)
24. Yin, P., Huang, L., Lee, S., Qiao, M., Asthana, S., Nakamura, T.: Diagnosis of neural network via backward deduction. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 260–267. IEEE (2019)
25. Yu, C.: Robot Behavior Generation and Human Behavior Understanding in Natural Human-Robot Interaction. Ph.D. thesis, Institut polytechnique de Paris (2021)
26. Yu, C., Tapus, A.: Interactive robot learning for multimodal emotion recognition. In: Salichs, M.A., et al. (eds.) ICSR 2019. LNCS (LNAI), vol. 11876, pp. 633–642. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35888-4_59
27. Yu, C., Tapus, A.: Multimodal emotion recognition with thermal and rgb-d cameras for human-robot interaction. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 532–534 (2020)
28. Zhang, H., Yu, C., Tapus, A.: Why do you think this joke told by robot is funny? the humor style matters. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 572–577. IEEE (2022)