Explainable AI Survey

Research description:

This research is done in the context of my Bachelor thesis Artificial Intelligence at Utrecht University.

In this research, I want to analyze and evaluate new ways of explaining uninterpretable machine learning models.
The purpose of this survey is to quantify the quality of automatic explanations (e.g. in terms of clarity and plausibility) generated for two types of Deep Neural Network models trained to predict facial expressions.

Your task is to evaluate and compare different explanation methods for two machine learning models. This will be done using closed questions. No personal data is required or being collected. The survey takes 6-8 minutes to complete.

---

* Required

Consent Form

The participant states:
- I voluntarily agree to participate in the research project.
- I agree that I will not be paid for my participation.
- I have been informed of the nature of the research project.
- I understand that statistical data gathered from this survey can be used in a scientific publication.
- I understand that my participation will remain anonymous.
- I agree that my data can be shared with other researchers to answer possible other research questions.

1. Consent *

Check all that apply.

- [ ] I consent

Contact

If you have any questions about this research, you can contact:
- Kaya ter Burg: k.terburg@students.uu.nl
- My supervisor dr. H. Kaya: h.kaya@uu.nl

If you have any issues regarding your privacy during this research, you can contact Utrecht University's privacy officer: privacy@uu.nl

General questions

2. What is the highest degree or level of school you have completed? *

Mark only one oval.

- No degree
- Elementary school
- High school
- MBO
- HBO
- Bachelor's degree
- Master's degree
- Doctorate degree

3. I am very knowledgeable on the subject of Artificial Intelligence (AI). *

Mark only one oval.

- 1 - Strongly disagree
- 2 - Disagree
- 3 - Neutral
- 4 - Agree
- 5 - Strongly agree

Image set 1 / 3

This is the first batch of example images. You will be shown the input image together with the emotion probability distribution of the classification model.

Image 1 - The model classifies this image as AFRAID. This is correct.

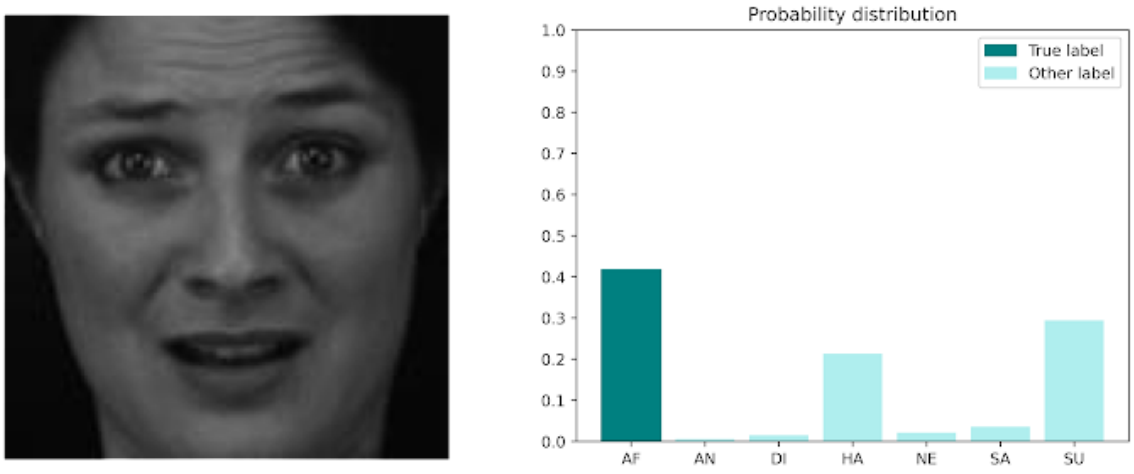


Image 2 - The model classifies this image as ANGER. This is correct.

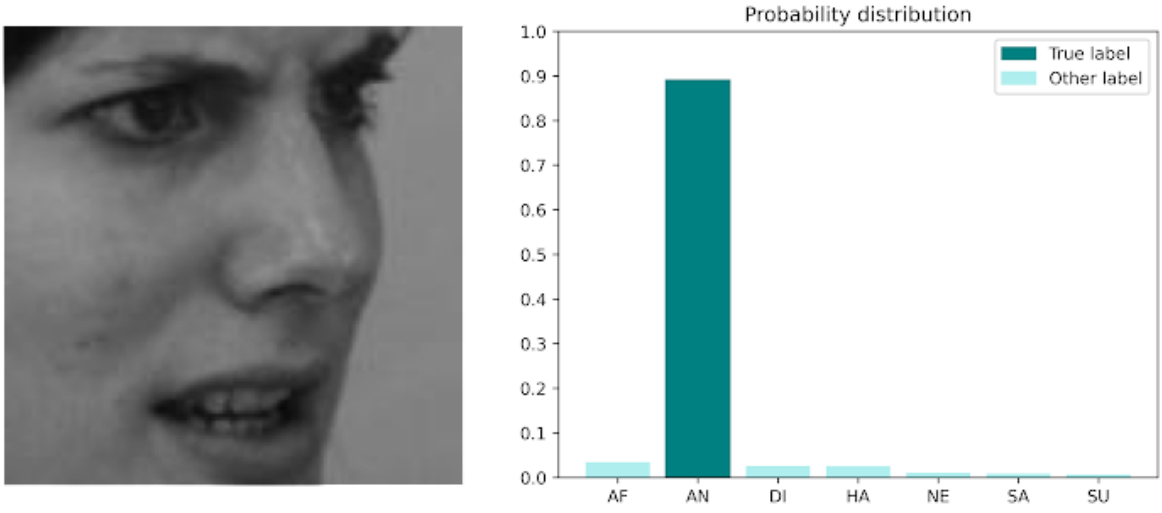


Image 3 - The model classifies this image as DISGUST. This is incorrect. The correct label is ANGER.

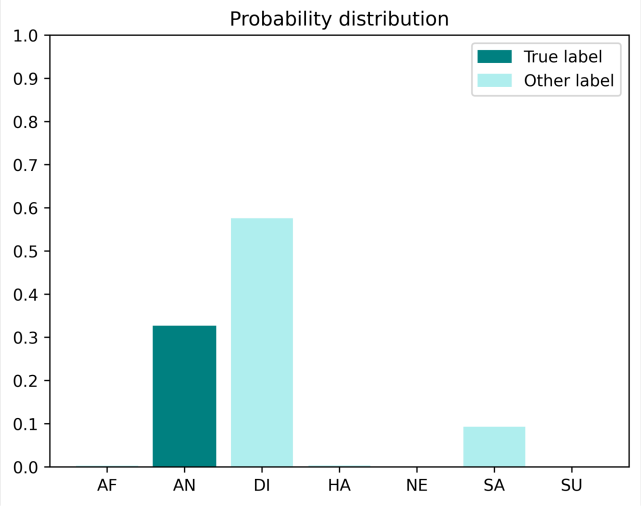


Image 4 - The model classifies this image as HAPPY. This is correct.

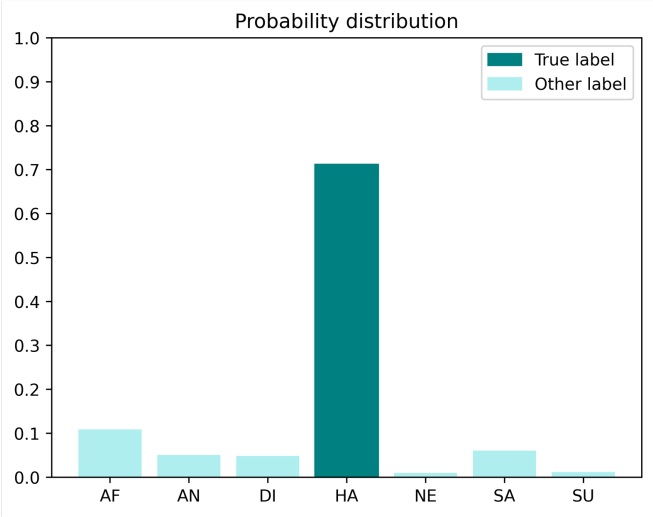
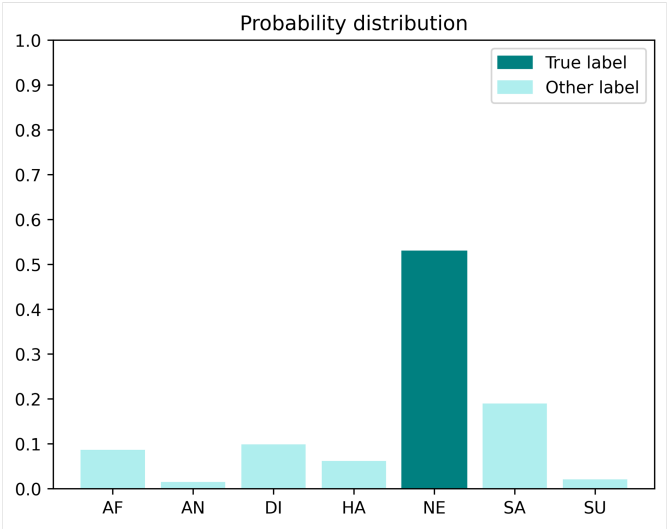


Image 5 - The model classifies this image as NEUTRAL. This is correct.



Evaluation questions

The following questions regard the example images shown above.

The questions will be answered through a scaling system:

1 = Strongly disagree
2 = Disagree
3 = Neutral / Neither agree nor disagree
4 = Agree
5 = Strongly agree

4. 1) The output representations help me understand how the model works.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

5. 2) The output representations of how the model works are satisfying.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

6. 3) The output representations are sufficiently detailed.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

7. 4) The output representations let me know how confident the model is for individual predictions.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

8. 5) The output representations let me know how trustworthy the model is.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

9. 6) I found the output representations unnecessarily complex.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

10. 7) I think I would need an expert to give me additional explanations.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

11. 8) The outputs of the model are very predictable.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

12. 9) The model can perform the task better than a novice human.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

13. 10) I am confident in the model. I believe it works well.

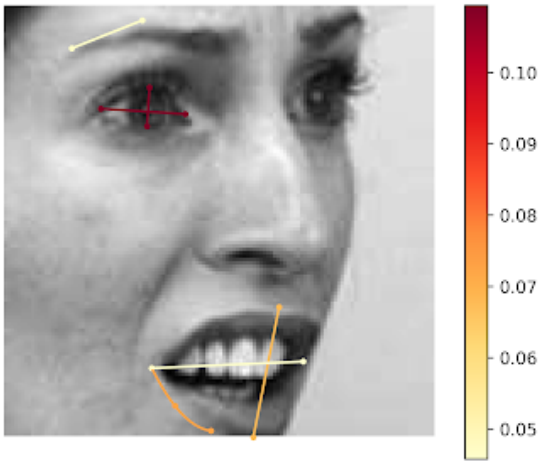
Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

Image set 2 / 3

Now you will see a few predictions from the same model, but this time with visual and textual explanations.

Image 1

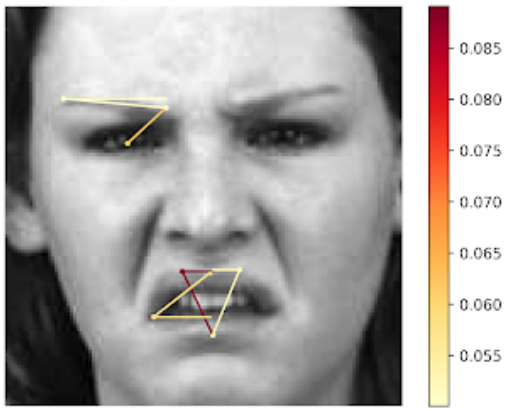


This person's emotion is classified as AFRAID. This classification is CORRECT.

The following 5 features, listed from most important to less important, contributed for 42.0% to the decision:

1. Left eye aspect ratio (ratio between eye width and eye height)
2. Curve of the left lower-outer lip
3. Height of the mouth
4. Slope of the outer to centre part of the left eyebrow
5. Width of the mouth

Image 2

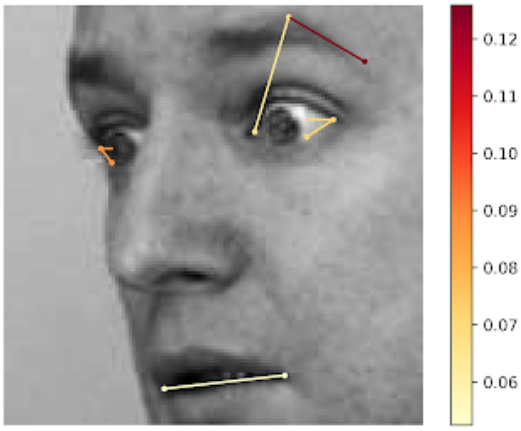


This person's emotion is classified as DISGUST. This classification is CORRECT.

The following 5 features, listed from most important to less important, contributed for 42.9% to the decision:

1. Angle from left mouth corner to left upper mouth
2. Distance between the centre of the left eye and the left inner eyebrow
3. Angle from top of left mouth side to centre of bottom mouth
4. Angle from top of right mouth side to centre of bottom mouth
5. Left eyebrow angle

Image 3

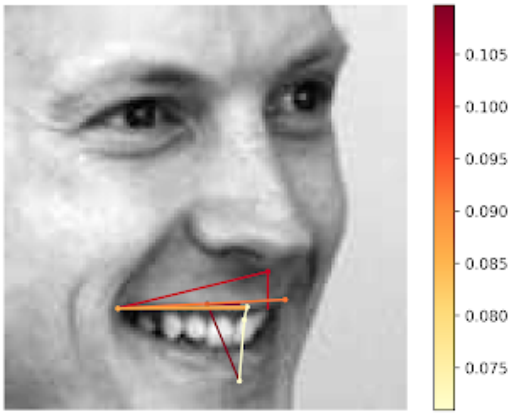


This person's emotion is classified as SURPRISE. This classification is INCORRECT; the correct label is AFRAID.

The following 5 features, listed from most important to less important, contributed for 46.1% to the decision:

1. Slope of the outer to centre part of the right eyebrow
2. Left lower eye outer angle
3. Right lower eye outer angle
4. Distance between the right inner eye and the centre of the right eyebrow
5. Width of the mouth

Image 4

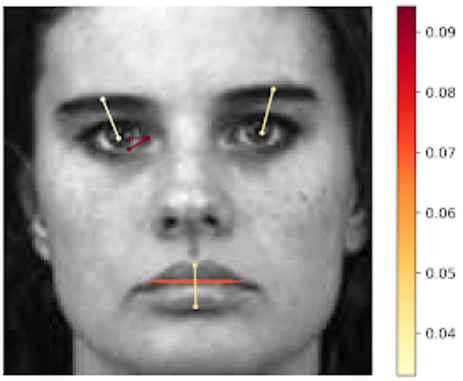


This person's emotion is classified as HAPPY. This classification is CORRECT.

The following 5 features, listed from most important to less important, contributed for 47.2% to the decision:

1. Angle from left mouth corner to left upper mouth
2. Angle from nose tip to left mouth corner
3. Width of the mouth
4. Angle from left mouth corner to top of the mouth
5. Ratio between the lower mouth height and the upper mouth height

Image 5



This person's emotion is classified as NEUTRAL. This classification is CORRECT.

The following 5 features, listed from most important to less important, contributed for 42.8% to the decision:

1. Left lower eye inner angle
2. Opening of the mouth
3. Ratio between the lower mouth height and the upper mouth height
4. Distance between the centre of the right eye and the centre of the right eyebrow
5. Distance between the centre of the left eye and the centre of the left eyebrow

Evaluation questions

The following questions regard the example images shown above. They are the same questions as with the previous batch of images.

The questions will be answered through the same scaling system as before:

1 = Strongly disagree
2 = Disagree
3 = Neutral / Neither agree nor disagree
4 = Agree
5 = Strongly agree

14. 1) The explanations help me understand how the model works.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

15. 2) The explanations of how the model works are satisfying.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

16. 3) The explanations are sufficiently detailed.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

17. 4) The explanations let me know how confident the model is for individual predictions.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

18. 5) The explanations let me know how trustworthy the model is.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

19. 6) I found the explanations unnecessarily complex.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

20. 7) I think I would need an expert to give me additional explanations.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

21. 8) The outputs of the model are very predictable.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

22. 9) The model can perform the task better than a novice human.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

23. 10) I am confident in the model. I believe it works well.

Mark only one oval.

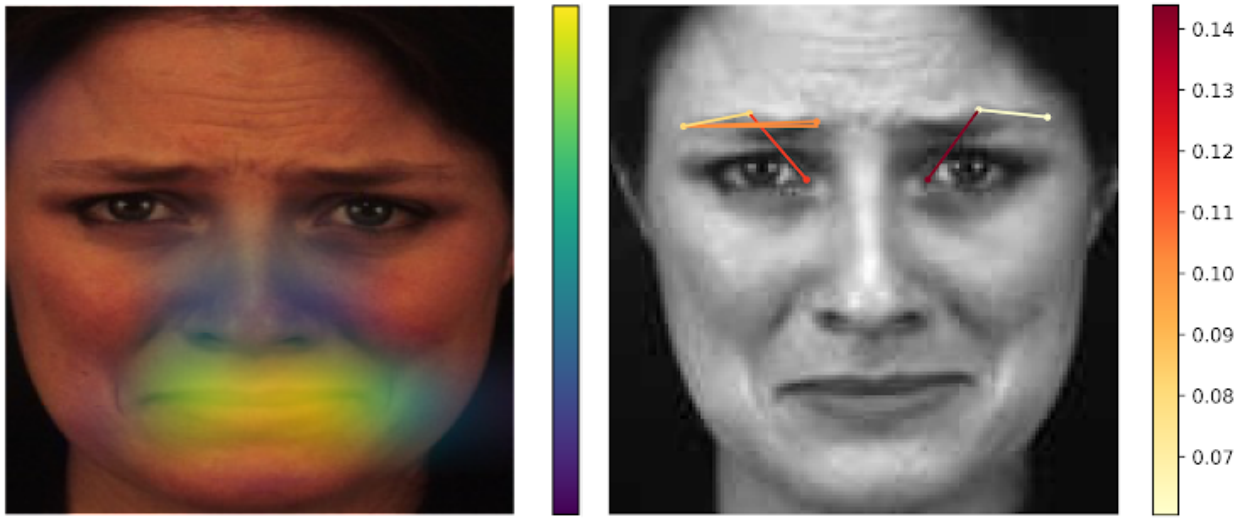
	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

Image set 3 / 3

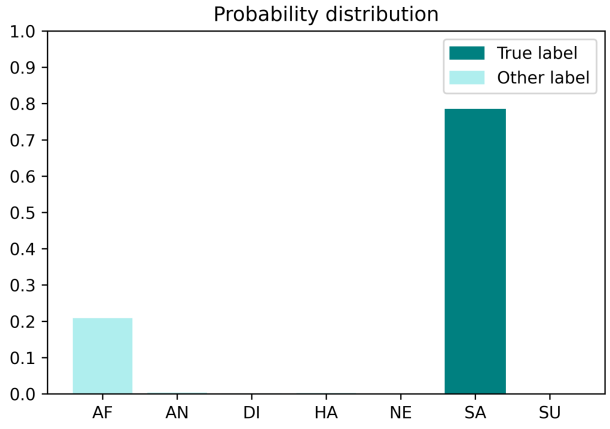
For the last set of examples, you will be shown two explanation side-by-side. Both explanations are from different models, but the images are the same.

Model 1 is always the left explanation and model 2 the right one.

Image 1. Model 1 - Model 2



Probability distributions for image 1. Model 1 - Model 2



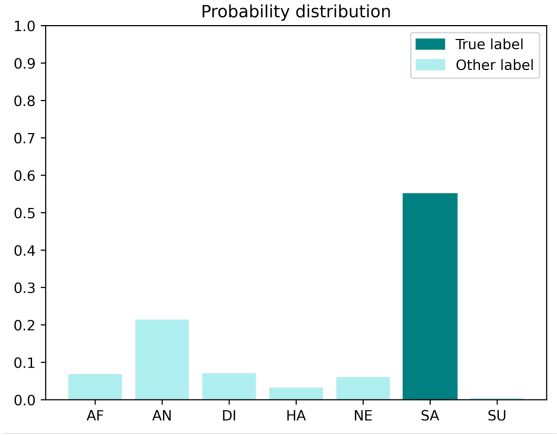
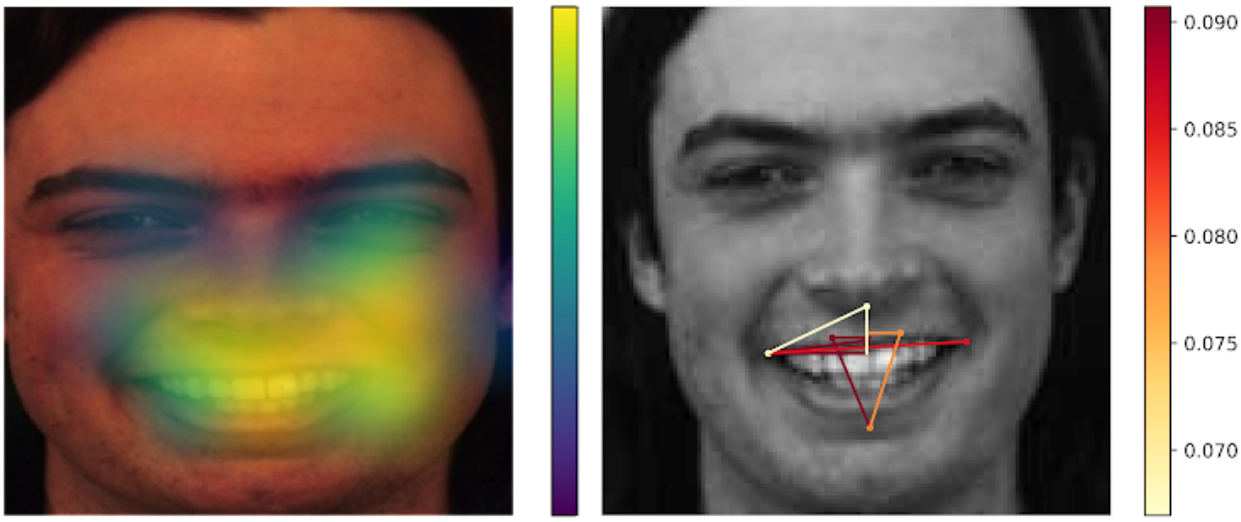
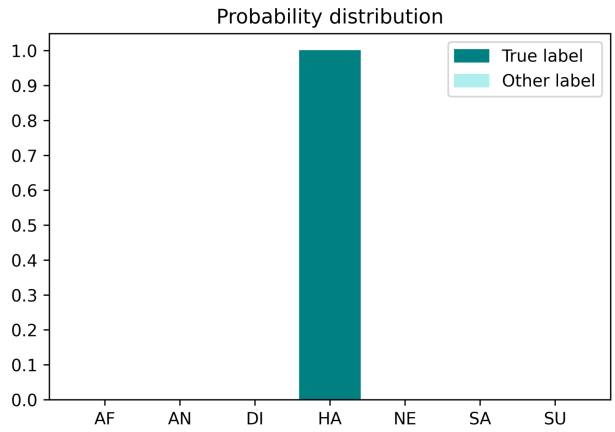


Image 2. Model 1 - Model 2



Probability distributions for image 2. Model 1 - Model 2



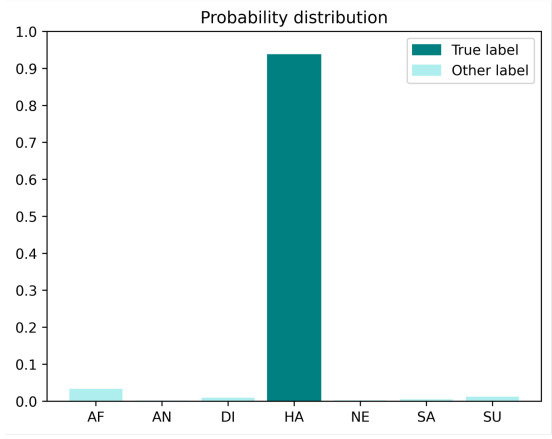
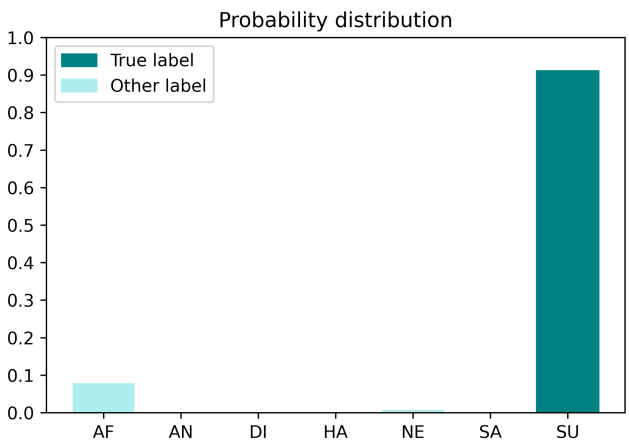


Image 3. Model 1 - Model 2

Probability distributions for image 3. Model 1 - Model 2



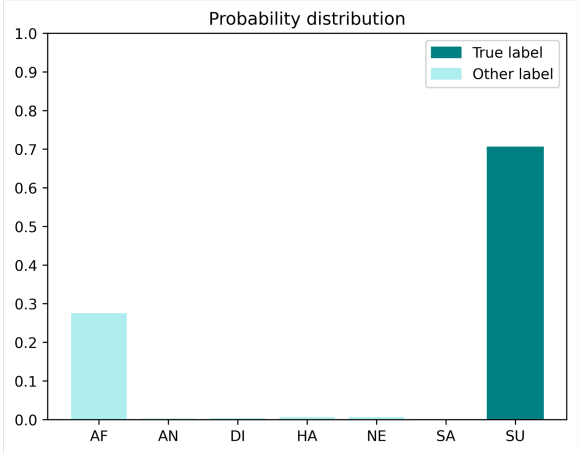
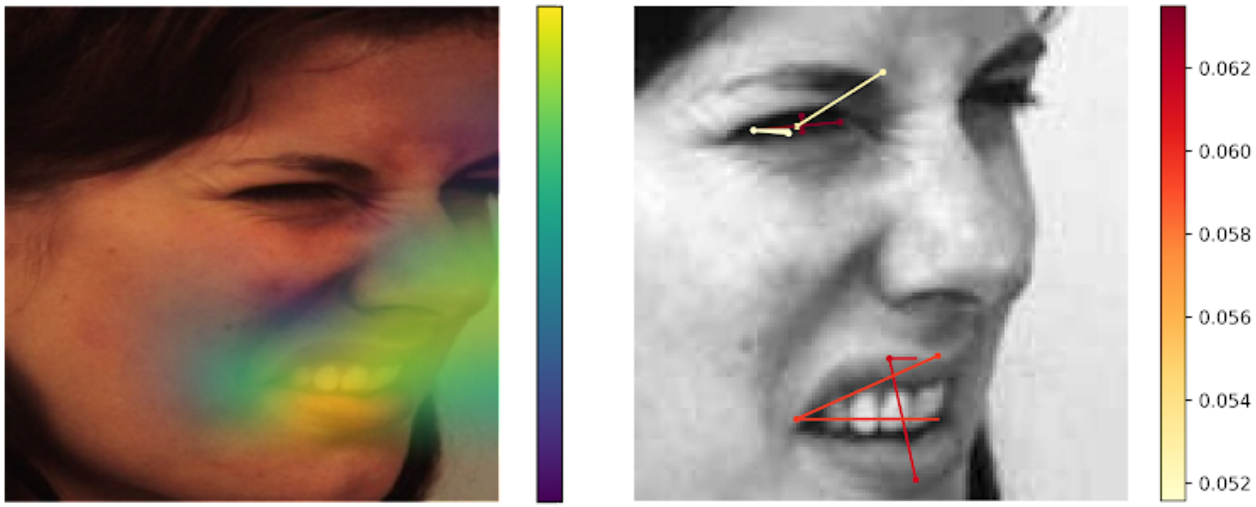


Image 4. Model 1 - Model 2



Probability distributions for image 4. Model 1 - Model 2

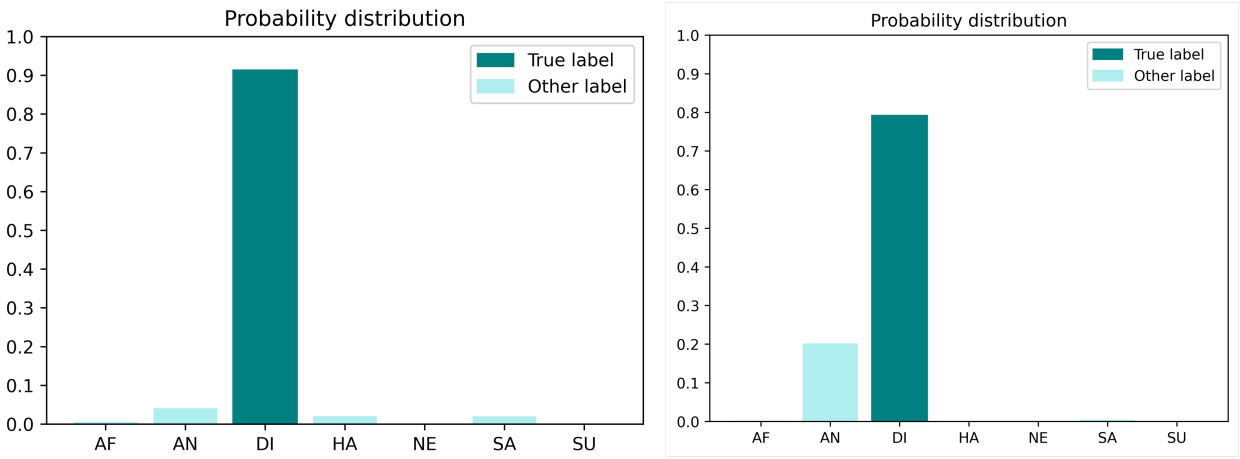
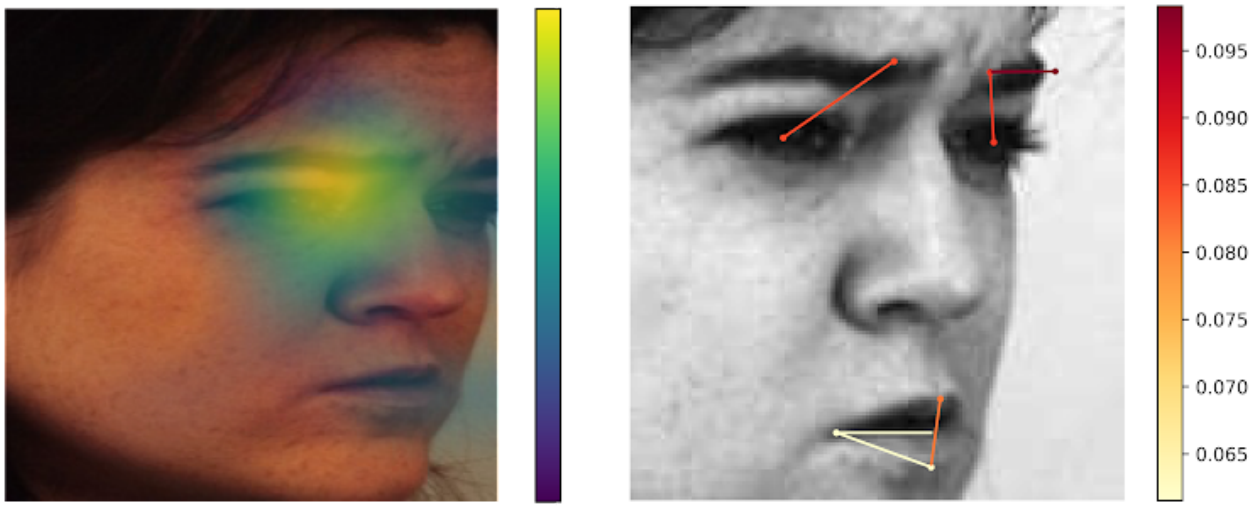


Image 5. Model 1 - Model 2



Probability distributions. Model 1 - Model 2

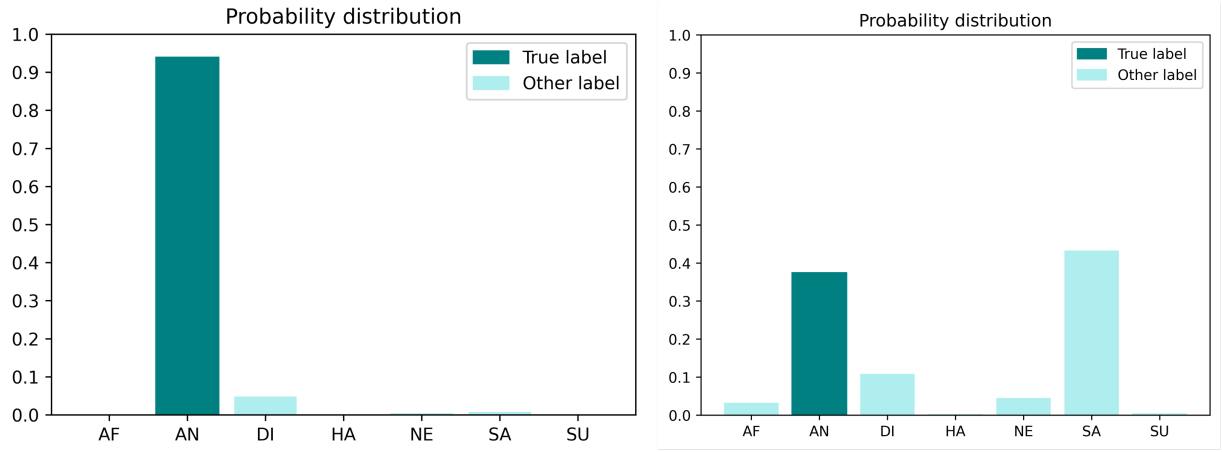
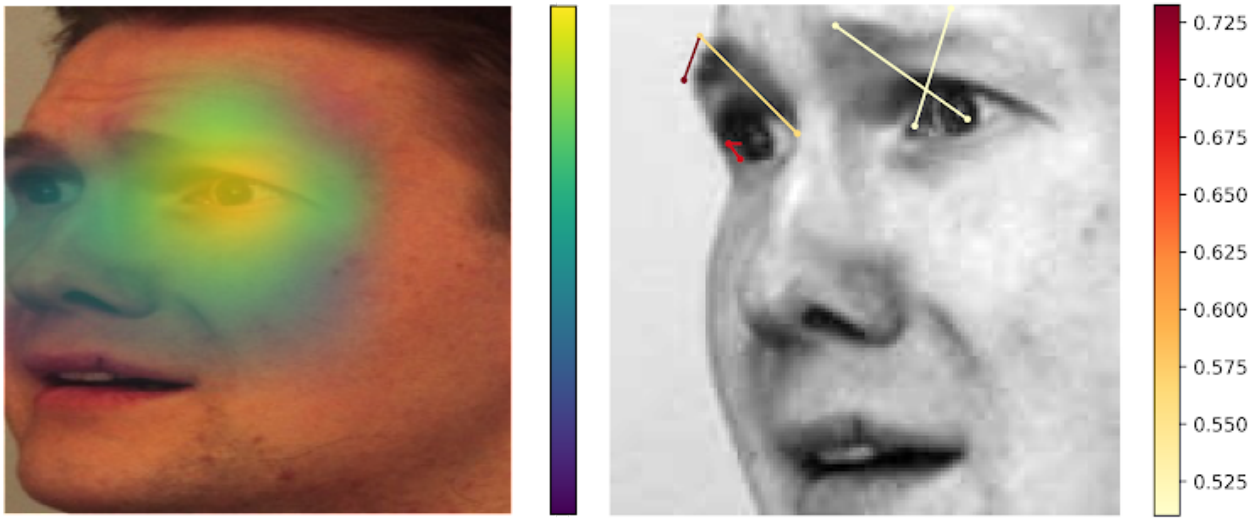
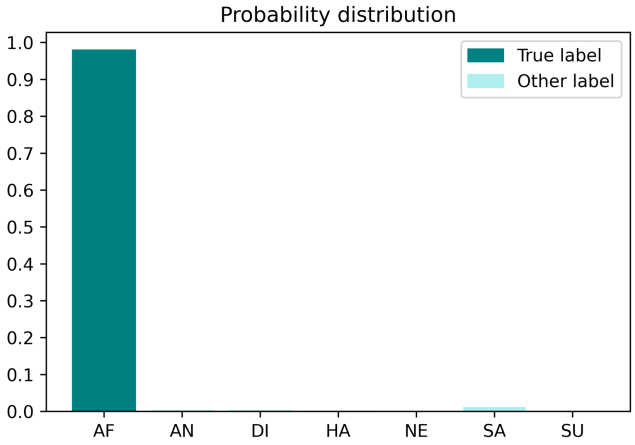


Image 6. Model 1 - Model 2



Probability distributions for image 6. Model 1 - Model 2



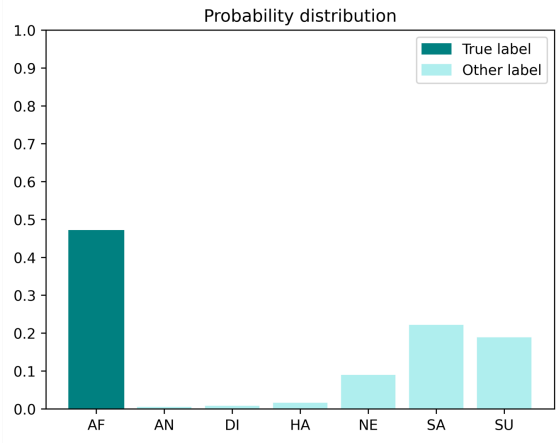
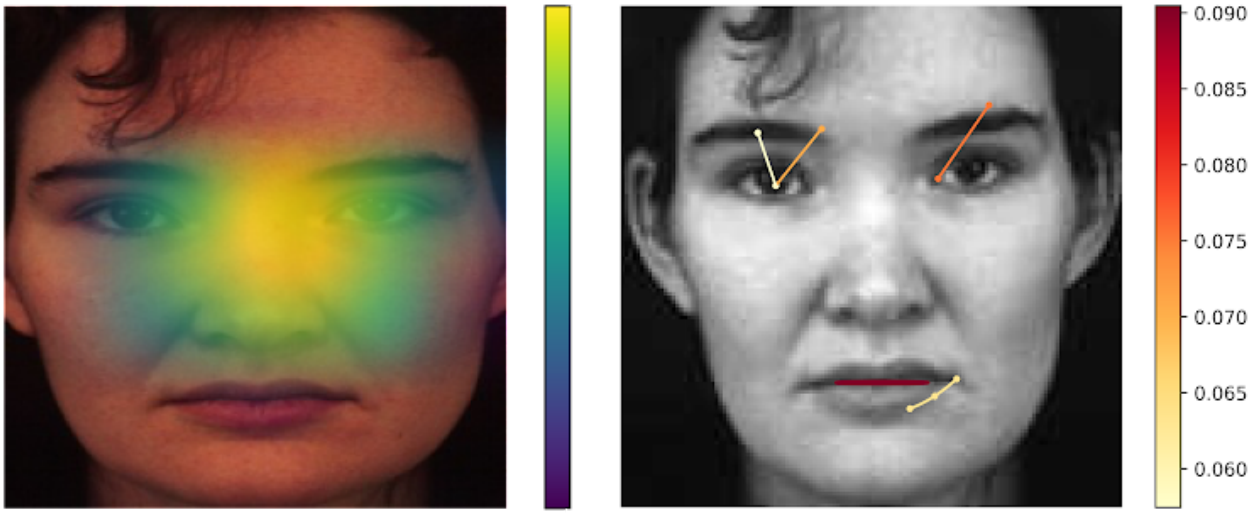
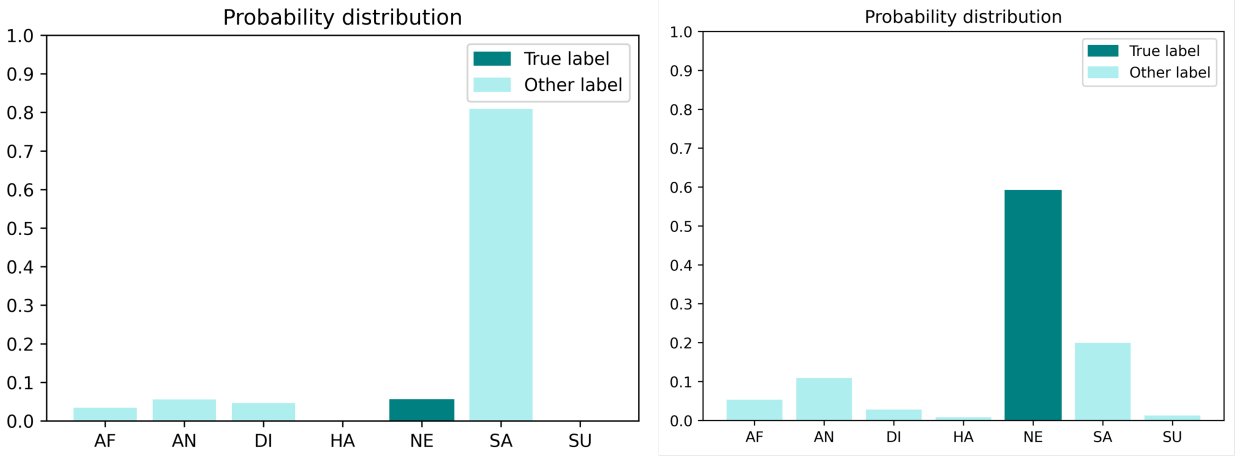


Image 7. Model 1 - Model 2



Probability distributions for image 7. Model 1 - Model 2



Evaluation questions.

Below are questions for the comparison between the explanations of the two models.

The same scaling system as before is used:

1 = Strongly disagree
2 = Disagree
3 = Neutral / Neither agree nor disagree
4 = Agree
5 = Strongly agree

24. 1) The explanations for model 1 are more understandable than those for model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

25. 2) I trust model 1 more than model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

26. 3) I would prefer the explanations of model 1 over those for model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

27. 4) The explanations for model 1 are more detailed than those for model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

28. 5) The explanations for model 1 are clearer on the model's accuracy than those for model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

29. 6) The explanations for model 1 reflect the model's confidence on each prediction better than those of model 2

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

30. 7) Model 1's explanations are more unnecessarily complex than those of model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

31. 8) The explanations for model 1 were more precise than those for model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

32. 9) I would follow model 1's advice over that of model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

33. 10) The outputs of model 1 were more predictable than those of model 2.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

Thank you!

You've reached the end of the questionnaire. Thank you very much for participating! I really appreciate you took the time to answer the questions.