The Power of Properties: Uncovering the Influential Factors in Emotion Classification

Tim Büchner*[1] [0000-0002-6879-552X], Niklas Penzel*[1] [0000-0001-8002-4130], Orlando Guntinas-Lichius[2] [0000-0001-9671-0784], and Joachim Denzler[1] [0000-0002-3193-3300]

[1] Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany
{firstname.surname}@uni-jena.de
[2] Dept. of Otorhinolaryngology, Jena University Hospital, 07747 Jena, Germany
orlando.guntinas@uni-jena.de

Abstract. Facial expression-based human emotion recognition is a critical research area in psychology and medicine. State-of-the-art classification performance is only reached by end-to-end trained neural networks. Nevertheless, such black-box models lack transparency in their decision-making processes, prompting efforts to ascertain the rules that underlie classifiers' decisions. Analyzing single inputs alone fails to expose systematic learned biases. These biases can be characterized as facial *properties* summarizing abstract information like age or medical conditions. Therefore, understanding a model's prediction behavior requires an analysis rooted in causality along such selected *properties*. We demonstrate that up to 91.25% of classifier output behavior changes are statistically significant concerning basic *properties*. Among those are age, gender, and facial symmetry. Furthermore, the medical usage of surface electromyography significantly influences emotion prediction. We introduce a workflow to evaluate explicit *properties* and their impact. These insights might help medical professionals select and apply classifiers regarding their specialized data and *properties*.

Keywords: Facial Emotion Recognition · Property Analysis · Model Behavior · Facial Asymmetry · Medical Facial Analysis · Facial Palsy

1 Introduction

Human emotion recognition via facial expressions is essential in psychology and medical research. High classification performance is achieved by neural networks trained end-to-end and tailored for broad usage on large datasets. The application of these black box models to unseen data often leads to irregular behavior. Previous research, for example, [22], finds a loss in predictive performance when observing various so-called subpopulation shifts. While we follow a similar approach, the performance decline attributed to subpopulation shifts is insufficient. To uncover potential causes, we model these shifts as changes in what we call

* These authors contributed equally to this work.

properties. Consequently, by investigating different *property* manifestations, we can uncover deviations in the model behavior. Hence, we go beyond simple predictive performance and uncover influential factors in emotion classification.

Some of these potential factors (*properties*) such as age or gender, have a broad impact on the visual appearance of a face. Regarding emotion recognition, this is even more evident in the case of medical conditions influencing mimicry. Particularly in facial palsy cases, where there is unilateral paralysis of the facial nerve. The pronounced facial asymmetry may impact a model's prediction or muscle activation studies with joint surface electromyography (sEMG) [2, 3, 9].

We assess HSEmotion-7 [18] and ResidualMaskNet [13], specifically their application on medically acquired facial data. Both models can predict the six basic emotions after Ekman [7], with an additional class for neutral expressions. We record 36 healthy probands to investigate the presence of sEMG electrodes and additionally their artificial removal [2, 3]. Further, we capture 36 patients with facial palsy to evaluate the influence of facial asymmetry [4]. This setup enables us to capture a multitude of relevant *properties*. Allowing us to extend a general performance analysis for subpopulation shifts [22] by utilizing additional explanations [16]. These explanations are based on causal principles [16], permitting us to test for statistically significant deviations in model behavior.

Our experimental results reveal two principal insights. First, both models exhibit a variation in predictive performance; Second, and more critically, there is a significant ($p < 0.01$) behavior change concerning the manifestations of varying *properties*. While for some, the shift in model behavior only significantly occurs for some emotions, other *properties* are important irrespective of the predicted class. Examples include visibly attached sEMG electrodes and whether a person suffers from unilateral facial palsy. We find statistically significant changes in the model behavior in up to 91.25% of the analyzed *properties*.

2 Facial Emotion Classification

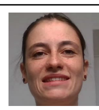
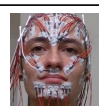
Facial emotion classification is a broad research field with many possible applications, especially in psychology and medicine. State-of-the-art performance is currently attainable exclusively through end-to-end trained convolutional neural networks. Our study focuses on two such models: HSEmotion-7 (HSE-7) [18] and ResidualMaskNet (RMN) [13]. Both models predict the six basic emotions defined by Ekman [7], with an added class for neutral expressions.

A general reduction in prediction accuracy is anticipated [22]. However, our interests are the subpopulation shifts resulting from different *property* manifestations. Therefore, we do not fine-tune to prevent distorting the interpretability of the general behavior analysis. Toward this goal, we require a custom evaluation dataset that captures *properties* typically unaccounted for in large datasets.

2.1 Emotion Evaluation Dataset

Our facial emotion images were captured using a standardized procedure where participants had to mimic the six basic emotions four times in a random se-

Table 1: We list the accuracies by emotion, sEMG attachment, and model (without fine-tuning) for probands and patients. A reference image is shown for each set. The patients are recorded only for the *happy* expression without sEMG.

	Probands						Patients	
sEMG	Without		Attached		Removed		Without	
Emotion	RMN	HSE-7	RMN	HSE-7	RMN	HSE-7	RMN	HSE-7
angry	82.29	66.32	36.25	36.07	82.86	54.82	-	-
disgusted	60.42	82.99	0.00	10.89	44.11	66.07	-	-
fearful	33.33	60.76	0.71	58.39	20.36	45.89	-	-
happy	95.83	87.85	66.96	45.89	73.04	46.07	60.64	65.81
sad	13.19	81.25	0.18	75.18	8.57	76.96	-	-
surprised	65.62	55.90	98.21	45.00	47.50	49.46	-	-
Mean Acc.	58.45	72.51	33.72	45.24	46.07	56.54	-	-
Reference Image								

quence. Therefore, we eliminate the risk of human-annotation bias, depending exclusively on the participants' capacity for facial mimicry. Using a frontal camera, we collected data from 36 healthy probands (18-67 years, 17 male, 19 female) four times with and twice without attached high-resolution surface electromyography (HR-sEMG) [9]. We follow the work of [2,3] to remove sEMG electrodes artificially. For each of these groups, we list the accuracy per model and emotion in Table 1. Furthermore, 36 patients (25-72 years, 8 male, 28 female) with unilateral chronic synkinetic facial palsy, which is presumed to be a significant factor affecting classification, were recorded three times using the 3dMD face system (3dMD LLC, Georgia, USA). This system creates 3D scans of the patients. We simulate a frontal view by calculating the camera position based on facial landmarks [4], ensuring consistency among participants; see Table 1. The first set of recordings only includes *happy* expressions. Combined, our dataset consists of 8,952 annotated images for evaluation.

2.2 Facial Properties: Selection and Manifestations

To evaluate the model's behavior regarding subpopulation shifts, we must select *properties*. The selection criteria are discussed herein, while Table 2 provides an overview comprising the manifestation types: scalar [`S`] or binary [`B`]. The medical setup grants access to `bodily` *properties* typically unaccounted for in large datasets. These include age, weight, gender, and presence of facial palsy.

Additionally, we utilize the experiment repetitions to evaluate model behavior across `recordings` for the same participant. Individual recording sessions are also scrutinized to detect any learned expressions. We indicate sEMG attachment and its subsequent artificial removal for the probands. We selected

Table 2: We list our selected *properties* used to examine model behavior. Their manifestations are scalars [S] or a binary [B] value.

	Property	Abbr.	Description
Bodily	Age [S]	$\mathfrak{B}_A$	The participant's age at the moment of recording
	Weight [S]	$\mathfrak{B}_W$	The measured weight at the moment of recording
	Gender [B]	$\mathfrak{B}_G$	The self-declared gender during questionnaire
	Facial Palsy [B]	$\mathfrak{B}_F$	The participant suffers from unilateral facial palsy
Recording	Participant ID [S]	$\mathfrak{R}_P$	Unique participant ID, to check individuum bias
	Session ID [S]	$\mathfrak{R}_S$	Recording session, to check for task memorization
	Attached sEMG [B]	$\mathfrak{R}_A$	If sEMG electrodes are attached [9]
	Removed sEMG [B]	$\mathfrak{R}_R$	If sEMG electrodes are removed artifically [2,3]
Symmetry	Facial Volume [S]	$\mathfrak{S}_V$	The lateral facial volume difference [4]
	eye-level dev. [S]	$\mathfrak{S}_E$	The deviation angle for the horizontal eye line [21]
	midline dev. [S]	$\mathfrak{S}_M$	The deviation angle for nose bridge line [21]
	LPIPS [S]	$\mathfrak{S}_L$	The image similarity between facial halves [23]

four computed metrics to assess facial `symmetry`, as facial palsy alone is insufficient for impact evaluation. We compute the lateral volume difference using the 3D patient scans [4]. Further, two *properties* leverage facial landmark symmetry [21]. LPISP calculates the image similarity between facial halves, with images aligned and center-cropped along the eye-line to reduce rotation artifacts [2,23].

It should be noted that a *property* like `age` encompasses a complex mixture of features, including wrinkles, hair color, or age spots. This type of subdivision can be accomplished for nearly all chosen *properties*. While a more comprehensive analysis would give more insight, it is not feasible within the scope of this study.

3 Methodology

Interested in emotion recognition model behavior relative to subpopulation shifts, we base the analysis on selected *properties* and related manifestations. Following the approach of [22], we assess accuracy change with respect to these shifts, see Table 1. We build our main analysis upon a causally-grounded method detailed in [16] to detect statistically significant behavioral shifts in the model in response to property manifestation. In [16], the authors develop a structural causal model (SCM) encompassing supervised learning building on the causality framework of Pearl [10]. Using this SCM together with Reichenbach's common cause principle [14], the question of whether a trained classifier uses a *property* becomes a statistical conditional independence (CI) test [16]. The selection of suitable CI tests is a vital hyperparameter choice. However, in [19], the authors prove that there cannot be a non-parametric CI test that controls for Type-I errors (false positives) in all cases. Following the analysis in [11,12,15], we form a committee of nonlinear tests to assess a model's feature usage [12], specifically conditional HSIC [8], RCoT [20], and CMIknn [17]. We report the consensus results with a significance level of $p < 0.01$ in Table 3.

Table 3: Significant changes ($p < 0.01$) per model and emotion are denoted with ✓. Additionally, we denote the average usage ($\varnothing$) by *property* and model.

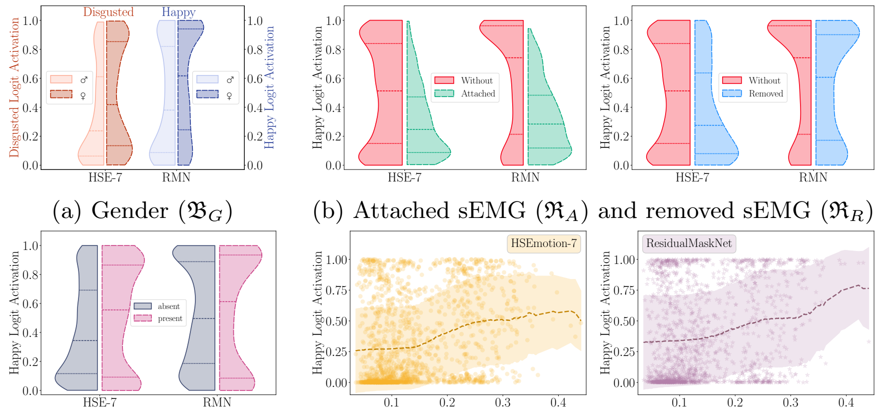
		Bodily				Recording				Symmetry				
		$\mathfrak{B}_A$	$\mathfrak{B}_W$	$\mathfrak{B}_G$	$\mathfrak{B}_F$	$\mathfrak{R}_P$	$\mathfrak{R}_S$	$\mathfrak{R}_A$	$\mathfrak{R}_R$	$\mathfrak{S}_V$	$\mathfrak{S}_E$	$\mathfrak{S}_M$	$\mathfrak{S}_L$	
HSEmotion-7 [18]	angry	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	
	disgusted	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	
	fearful	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	
	happy	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✓	
	sad	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	
	surprised	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	
	neutral	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	$\varnothing$	$6/7$	$7/7$	$6/7$	$7/7$	$6/7$	$3/7$	$7/7$	$7/7$	$6/7$	$5/7$	$3/7$	$7/7$	$\sum$ $70/84$
ResMaskNet [13]	angry	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	
	disgusted	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	
	fearful	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	
	happy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	sad	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	
	surprised	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	
	neutral	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	
	$\varnothing$	$7/7$	$7/7$	$5/7$	$7/7$	$6/7$	$6/7$	$7/7$	$7/7$	$5/7$	$3/7$	$6/7$	$7/7$	$\sum$ $73/84$

Since both models output logits use softmax [13, 18], observable changes in one logit may be counterbalanced by opposing logit groups. Hence, we verify each emotion to address this, as displayed in Table 3. This approach enables more profound insight into *property* utilization beyond simple summary statistics.

4 Results

We investigate subpopulation shifts resulting from different *property* manifestations by applying black-box neural networks. Specifically, we focus on ResidualMaskNet (RMN) [13] and HSEmotion-7 (HSE-7) [18] for facial emotion recognition. Each image of our dataset has all *properties*, as referenced in Table 2. Using these images, we capture the logit activations of the pre-trained RMN and HSE-7. Hence, using the reference annotations, we test for significant changes in behavior [16]. Statistically significant results ($p < 0.01$) are denoted per model and emotion in Table 3. Of the 84 possible combinations, RMN utilizes information in 91.25% ($73/84$) and for HSE-7 in 87.50% ($70/84$) of cases. The findings show that the models use *properties* in their decision-making processes. We now examine the models' behavior concerning several *properties* in detail.

First, we focus on self-declared gender. Other work [6] demonstrated that gender ($\mathfrak{B}_G$) affects emotion classification, a claim we corroborate for 11 of 14 cases as seen in Table 3. The visualization, see Fig. 1a, displays the logit activation distribution regarding men and women. We choose *disgusted* for HSE-7 due to its significance and difficulty in execution [2, 5, 9]; see Table 3. The remaining



(a) Gender ($\mathfrak{B}_G$) (b) Attached sEMG ($\mathfrak{R}_A$) and removed sEMG ($\mathfrak{R}_R$)

(c) Facial palsy ($\mathfrak{B}_F$) (d) LPIPS ($\mathfrak{S}_L$) symmetry value for lateral face halves

Fig. 1: Model behavior visualization: In the binary [B] case, we use violin plots for logit distributions (a, b, c). For the scalar [S] *property* $\mathfrak{S}_L$, we plot a regression analysis by estimating Gaussians with a sliding window [11]. For all visualizations, we only use images of the corresponding emotion [16].

investigations focus on *happy* expressions due to the observed prediction stability for probands and patients; see Table 3. The displayed distributions indicate HSE-7 struggles to classify *disgusted* expressions for men compared to women. Moreover, the RMN more frequently attributes *happy* expressions to women.

Secondly, we investigate the attached ($\mathfrak{R}_E$) and artificially removed ($\mathfrak{R}_R$) sEMG electrodes, as they are influential for both models and all seven emotions; see Table 3. As argued above, we visualize the *happy* logit distribution concerning $\mathfrak{R}_E$ and $\mathfrak{R}_R$ in Fig. 1b. The overall lower activations are anticipated for the attached case [2,3]. Yet we see in Table 1 that the quality of the removal highly depends on the model and emotion when recovering predicted performance. For instance, regarding *happy* expressions, the RMN recovers a higher predictive performance and logit distribution (see Fig. 1b). We assume that the recovery introduced artifacts or the sEMG influences the participants' facial mimicry.

Lastly, we assess whether facial palsy-induced asymmetry ($\mathfrak{B}_F$) or symmetry in general ($\mathfrak{S}$) impacts model behavior. In Table 3, we observe in 77.14% ($^{54}/_{70}$) of combinations a significant influence of facial symmetry. Especially, facial palsy ($\mathfrak{B}_F$) and different LPIPS ($\mathfrak{S}_L$) manifestations influence the behavior of both models and all seven emotions. We follow [11] and regress the *property* manifestation trend by estimating Gaussians with a sliding window approach for LPIPS. Again, focusing on *happy* expressions, Fig. 1d show that higher facial symmetry leads to higher logit activation on average. However, facial palsy specifically, see Fig. 1c, shows increased uncertainty. The indicated quartiles suggest a strong bimodal distribution with many high and low activations. This suggests that the models are influenced by facial symmetry. Hence, their application for unilateral facial palsy patients should be approached with caution.

5 Conclusion

This work studied emotion classifiers applied in a medical context. To go beyond performance metrics, we used the causal-based framework from [16]. We demonstrated that up to 91.25% of classifier output behavior changes are statistically significant concerning varying *properties*, including age, gender, and facial symmetry. To obtain such *properties* unaccounted in other datasets, we recorded 36 probands and 36 patients with facial palsy, a disease affecting facial expressions.

To summarize, we observe differences in model behavior regarding gender and facial symmetry. Hence, their application on medical conditions should be approached with care. Additionally, the obstruction of facial features during medical studies can significantly impact the model behavior, as we observe for attached sEMG electrodes. However, our observations do not necessarily indicate harmful biases but open the discussion beyond simple predictive performance.

Finally, our selected *properties* are not exhaustive and do not cover all possible biases. Many other properties related to different downstream tasks could be conceived and studied in the future. Additionally, different models and *property* aware training are promising [1, 6]. We hope to prompt more extensive analysis and inspire researchers to analyze facial recognition models beyond prediction performance.

Acknowledgments. Partially supported by Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) project 427899908 BRIDGING THE GAP: MIMICS AND MUSCLES (DE 735/15-1 and GU 463/12-1).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Blunk, J., Penzel, N., Bodesheim, P., Denzler, J.: Beyond debiasing: Actively steering feature selection via loss regularization. In: DAGM German Conference on Pattern Recognition (DAGM-GCPR) (2023)
2. Büchner, T., Guntinas-Lichius, O., Denzler, J.: Improved obstructed facial feature reconstruction for emotion recognition with minimal change cyclegans. In: Advanced Concepts for Intelligent Vision Systems (Acivs). pp. 262–274. Springer-Nature (august 2023). https://doi.org/10.1007/978-3-031-45382-3_22
3. Büchner, T., Sickert, S., Volk, G.F., Anders, C., Guntinas-Lichius, O., Denzler, J.: Let's get the facs straight - reconstructing obstructed facial features. In: International Conference on Computer Vision Theory and Applications (VISAPP). SciTePress (march 2023). https://doi.org/10.5220/0011619900003417
4. Büchner, T., Sickert, S., Volk, G.F., Guntinas-Lichius, O., Denzler, J.: From Faces to Volumes - Measuring Volumetric Asymmetry in 3D Facial Palsy Scans. In: Advances in Visual Computing. Lecture Notes in Computer Science, Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-47969-4_10
5. Büchner, T., Sickert, S., Graßme, R., Anders, C., Guntinas-Lichius, O., Denzler, J.: Using 2d and 3d face representations to generate comprehensive facial electromyography intensity maps. In: International Symposium on Visual Computing (ISVC). pp. 136–147 (2023). https://doi.org/10.1007/978-3-031-47966-3_11, https://link.springer.com/chapter/10.1007/978-3-031-47966-3_11

6. Chen, Y., Joo, J.: Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14960–14971. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01471
7. Ekman, P.: An argument for basic emotions. Cognition and Emotion **6**(3-4), 169–200 (1992). https://doi.org/10.1080/02699939208411068
8. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. Advances in neural information processing systems **20** (2007)
9. Guntinas-Lichius, O., Trentzsch, V., Mueller, N., Heinrich, M., Kuttenreich, A.M., Dobel, C., et al.: High-resolution surface electromyographic activities of facial muscles during the six basic emotional expressions in healthy adults: a prospective observational study. Scientific Reports **13**(1), 19214 (2023)
10. Pearl, J.: Causality. Cambridge university press (2009)
11. Penzel, N., Kierdorf, J., Roscher, R., Denzler, J.: Analyzing the behavior of cauliflower harvest-readiness models by investigating feature relevances. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 572–581. IEEE (2023)
12. Penzel, N., Reimers, C., Bodesheim, P., Denzler, J.: Investigating neural network training on a feature level using conditional independence. In: European Conference on Computer Vision. pp. 383–399. Springer (2022)
13. Pham, L., Vu, T.H., Tran, T.A.: Facial expression recognition using residual masking network. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4513–4519 (2021). https://doi.org/10.1109/ICPR48806.2021.9411919
14. Reichenbach, H.: The direction of time, vol. 65. Univ of California Press (1956)
15. Reimers, C., Penzel, N., Bodesheim, P., Runge, J., Denzler, J.: Conditional dependence tests reveal the usage of abcd rule features and bias variables in automatic skin lesion classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1810–1819 (2021)
16. Reimers, C., Runge, J., Denzler, J.: Determining the relevance of features for deep neural networks. In: European Conference on Computer Vision. Springer (2020)
17. Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: International Conference on Artificial Intelligence and Statistics. PMLR (2018)
18. Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: International Conference on Machine Learning. vol. 202. PMLR (2023), https://proceedings.mlr.press/v202/savchenko23a.html
19. Shah, R.D., Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics **48**(3), 1514–1538 (2020)
20. Strobl, E.V., Zhang, K., Visweswaran, S.: Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. Journal of Causal Inference (2019)
21. Wei, W., Ho, E.S.L., McCay, K.D., Damaševičius, R., Maskeliūnas, R., Esposito, A.: Assessing Facial Symmetry and Attractiveness using Augmented Reality. Pattern Analysis and Applications **25**(3) (2022). https://doi.org/10.1007/s10044-021-00975-z
22. Yang, Y., Zhang, H., Katabi, D., Ghassemi, M.: Change is hard: A closer look at subpopulation shift. arXiv preprint arXiv:2302.12254 (2023)
23. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Apr 2018). https://doi.org/10.48550/arXiv.1801.03924