Differences in the annotation between facial images and videos for training an artificial intelligence for skin type determination

Gabriele Maria Lehner, Laura Gockeln, Bettina Marie Naber, Janis Raphael Thamm, Sandra Schuh, Gabriel Duttler, Anna Rottenkolber, Dennis Hartmann, Frank Kramer, Julia Welzel

Angaben zur Veröffentlichung / Publication details:

Lehner, Gabriele Maria, Laura Gockeln, Bettina Marie Naber, Janis Raphael Thamm, Sandra Schuh, Gabriel Duttler, Anna Rottenkolber, Dennis Hartmann, Frank Kramer, and Julia Welzel. 2024. "Differences in the annotation between facial images and videos for training an artificial intelligence for skin type determination." *Skin Research and Technology* 30 (3): e13632. https://doi.org/10.1111/srt.13632.

Received: 6 September 2023 | Accepted: 4 January 2024

DOI: 10.1111/srt.13632

ORIGINAL ARTICLE

WILEY

Differences in the annotation between facial images and videos for training an artificial intelligence for skin type determination

Gabriele Maria Lehner[1] | Laura Gockeln[1] | Bettina Marie Naber[1] | Janis Raphael Thamm[1] | Sandra Schuh[1] | Gabriel Duttler[2] | Anna Rottenkolber[2] | Dennis Hartmann[3] | Frank Kramer[3] | Julia Welzel[1]

[1]Department of Dermatology and Allergology, University Hospital Augsburg, Augsburg, Germany

[2]GRANDEL-The Beautyness Company, Augsburg, Germany

[3]IT Infrastructure for Translational Medical Research, University of Augsburg, Augsburg, Germany

Correspondence
Gabriele Marie Lehner, MD, Department of Dermatology and Allergology, University Hospital Augsburg, Sauerbruchstr. 6, 86179 Augsburg, Germany.
Email: MarieGabriele.Lehner@uk-augsburg.de

Abstract

Background: The Grand-AID research project, consisting of GRANDEL-The Beautyness Company, the dermatology department of Augsburg University Hospital and the Chair of IT Infrastructure for Translational Medical Research at Augsburg University, is currently researching the development of a digital skin consultation tool that uses artificial intelligence (AI) to analyze the user's skin and ultimately perform a personalized skin analysis and a customized skin care routine. Training the AI requires annotation of various skin features on facial images. The central question is whether videos are better suited than static images for assessing dynamic parameters such as wrinkles and elasticity. For this purpose, a pilot study was carried out in which the annotations on images and videos were compared.

Materials and Methods: Standardized image sequences as well as a video with facial expressions were taken from 25 healthy volunteers. Four raters with dermatological expertise annotated eight features (wrinkles, redness, shine, pores, pigmentation spots, dark circles, skin sagging, and blemished skin) with a semi-quantitative and a linear scale in a cross-over design to evaluate differences between the image modalities and between the raters.

Results: In the videos, most parameters tended to be assessed with higher scores than in the images, and in some cases significantly. Furthermore, there were significant differences between the raters.

Conclusion: The present study shows significant differences between the two evaluation methods using image or video analysis. In addition, the evaluation of the skin analysis depends on subjective criteria. Therefore, when training the AI, we recommend regular training of the annotating individuals and cross-validation of the annotation.

Gabriele Marie Lehner and Laura Gockeln contributed equally to this work.

KEYWORDS
annotation, artificial intelligence, cosmetics, imaging, skin analysis, skin care

1 | INTRODUCTION

The appearance of the human face changes throughout life. Skin physiology and exposome (UV exposure, nicotine, and alcohol consumption, stress, etc.) play an important role in this process. These two factors influence facial structure, skin coloration, and skin sagging, which leads to wrinkles, pigmentation, and slack skin. Depending on the extent of the two factors, external appearance does not always correlate with biological age.[1–3] Artificial intelligence (AI) has made enormous progress in recent years and has also become indispensable in the pharmaceutical industry. The use of AI in pharmaceutical research and development offers numerous benefits, such as speeding up the development process of new drugs and improving the accuracy of disease diagnosis, especially in the diagnosis of skin cancer.[4]

So far, there are few publications that use AI to assess skin characteristics (e.g., skin type) as well as extrinsic skin ageing.[5,6]

The interest in dermatological cosmetics, measures, and interventions for skin rejuvenation is increasingly growing. The development of digital tools for skin analysis detection is helpful in this regard. In this process, the human face becomes the subject of correction, care, or beautification procedures. It is also possible to obtain a personalized skin analysis and a customized skin care routine through this.

The aim of the Grand-AID research project is to develop an image-driven tool that analyzes skin characteristics and provides individual skin care recommendations. Artificial intelligence (AI) will be used to automate the analysis. For this purpose, static facial photographs are usually taken for annotation and training of the algorithm as well as for evaluation of the skin characteristics. Dynamic changes such as mimic wrinkles, elasticity, and slackening of the skin can only be judged to a limited extent on static images, especially since self-photographs are often beautified.

For the development of the algorithm, the question arose whether sequential individual recordings with different facial expressions or short video sequences are more suitable for training the AI, in particular whether there are differences in the evaluation of the properties depending on the recording mode. Videos are significantly more complex than images, especially in terms of capturing, annotation, and training the algorithm. Since the annotation of the images has a major impact on the validity and reliability of the AI results, we also examined whether there are inter-individual differences in the annotation of the images.

Furthermore, the practicability of two different scoring systems 0 to 3 and 1 to 10 was investigated. The grading can generally be done on a semi-quantitative or a continuous scale. When the results are presented to customers, a scale that allows grading from 0 to 10 might be more understandable than a semi-quantitative scale that only shows mild, medium, or severe expression of parameters. Changes over time could be more visible with a more detailed scale. On the other hand, a detailed grading from 0 to 10 could pose a problem for annotation and then also for training the algorithm if the extremes (0 = absent and 10 = maximally present) are underrepresented in the data set. Therefore, both scales were used for annotation to check which scale was more appropriate and whether there were differences.

2 | MATERIALS AND METHODS

A preliminary study was conducted to test the parameters for the image capture as well as the annotation process.

First, it was discussed and determined which features of the facial skin should be assessed. Then, two different scores were applied to evaluate their practicability for annotation: a semi-quantitative score ranging from 0 = not present to 3 = maximally pronounced and a linear score on a scale of 10 cm from 0 to 10 in terms of a visual analog score. Static images of the face of skin-healthy subjects were taken from the front and side, first with neutral facial expression, then with different facial expressions (Figure 1). In addition, a short video sequence was recorded from the front with the same facial expressions.

Four raters with dermatological expertise (physicians in further training to become specialists in dermatology) at Augsburg University Hospital were selected: LG, BN, GL (all female), and JT (male). The training of the evaluators was carried out by using already annotated example images based on standardized photographic scales [20] and subsequently a direct learning success control. The annotated images from the study will then serve as a base data set to train other graders, who will then annotate the entire test and evaluation data set of several thousand images in a standardized manner.

Ethical issued were discussed between the partners. This is a non-interventional pilot study on healthy subjects, not involving patients, drugs, or devices. The image evaluation was carried out retrospectively on anonymized data sets. A data protection agreement was concluded, and the subjects gave their written informed consent to participate in the study. The general data protection aspects of the final project, when customers upload their recordings to a platform for AI-based classification, were discussed intensively between the cooperation partners and sealed with a data exchange agreement. Customers explicitly agree to the use of their images for assessment by artificial intelligence. Other uses are excluded.

The evaluation was based on two different scores. The score 0 to 3 was divided as follows: "0" stood for the non-existence of the mentioned characteristic, with "1" a mild expression, with "2" a moderate as well as with "3" a severe expression was present. The linear score "0 to 10" is supposed to represent the gradual presence of the mentioned characteristic on a scale of from 0 to 10. "0" stands for non-existence,

FIGURE 1 Images of a test person with neutral facial expression (A), grim facial expression (B), smiling (C), frowning (D). Wrinkles and sagging skin can be better assessed and easier annotated in the facial expressions than in the static image.

"10" for massive expression. A total of 25 subjects with healthy skin were recruited (23 female and 2 male). The age range was 19 to 65 years with a mean value of 36 years and a standard deviation of 12 years. All test subjects were of Caucasian origin.

Pictures and videos of all subjects were evaluated for the study. The resolutions of the pictures were 72 DPI and of the videos: 3840 × 2160 pixels. To capture the pictures and videos, an Iphone 12 was mounted on a tripod and the subjects' images were taken by a third person. A ring light was used for even lighting and care was taken to ensure that the subjects were at an even distance from the camera. The face and the upper part of the body up to the chest can be seen in the image details. Exact standardization of lighting and image capture was not achieved, but this also reflects the situation when customers later take their own images.

When taking the photos and videos, care was taken to ensure that the eight features to be evaluated were clearly visible: Wrinkles, redness, shine, pores, pigmentation spots, dark circles, sagging skin, blemished skin. When taking the photos, the subjects were photographed in both frontal and side views. For the frontal photos, subjects were asked to show the following facial expressions/grimaces to clearly depict the features being evaluated: neutral facial expression, smiling, frowning, grim facial expression. When recording the videos, the head was dynamically rotated to the left and right, resulting in a 180° view.

The evaluation of 25 subjects was carried out by each rater. For each rating, the two available scores ("score 0 to 3″ and linear score 0 to 10) were assessed in parallel. Image annotations were performed in a cross-over study design: Two raters initially annotated half of the subjects to be rated in video format, while the other two raters annotated them in image format. The assignment was done randomly. Subsequently, this assignment was changed for the second half of the subjects to be scored, so that the evaluators now had to evaluate images that had previously been scored in video format and vice versa. In each case, both scores were applied, and all characteristics were evaluated.

Statistically, mean values and frequency distributions were calculated. In addition, comparisons between video and image recordings as well as comparisons between the raters LG, BN, JT, and GL were carried out. Methodologically, the Wilcoxon test for paired differences in connected samples or the Friedman rank test with significance at $p < 0.05$ was used for this purpose.

3 | RESULTS

The evaluators with dermatological expertise stated that they could not notice any difference in labeling of images or videos. Both rating methods could be applied without any problems. For pictures it seemed sufficient to use only the frontal view, a side view, and the facial expressions of laughing, frowning, and grimacing to score the criteria. The raters also indicated that both scores, the semi-quantitative score of 0 to 3 and the linear rating of 10 cm, were easy to assign.

Statistical analysis showed that the two different ratings were very similar:

In the semi-quantitative score from 0 to 3, wrinkles ($p = 0.025$) and dark circles (0.010) were rated significantly higher in the videos than in the images (Figure 2).

For skin sagging, the scores were significantly higher in the videos than in the pictures ($p = 0.007$), too. There were no significant differences between the two imaging methods for the other parameters.

Using the linear score of 10 cm, wrinkles ($p = 0.023$), pores ($p = 0.028$), and dark circles ($p = 0.00$) had significantly higher scores in the videos than in the images (Figure 3). There were no significant differences between the two imaging methods for the other parameters.

Furthermore, a statistical evaluation was performed with respect to the four raters (LG, GL, BN, JT):

Significant differences were found between the raters, especially between rater JT and the other three raters (Figure 4). In the semi-quantitative evaluation, a significant difference was found between the raters only for the skin laxity parameter. JT rated skin sagging significantly more pronounced than the other three raters ($p = 0.028$). There was no significant difference in scores for the other characteristics.

When the linear score was applied to the parameters of wrinkles, dark circles, and skin sagging, there was a significant difference between the raters (Figure 5), too. JT rated wrinkles ($p = 0.022$) significantly less pronounced and skin sagging ($p = 0.018$) significantly more

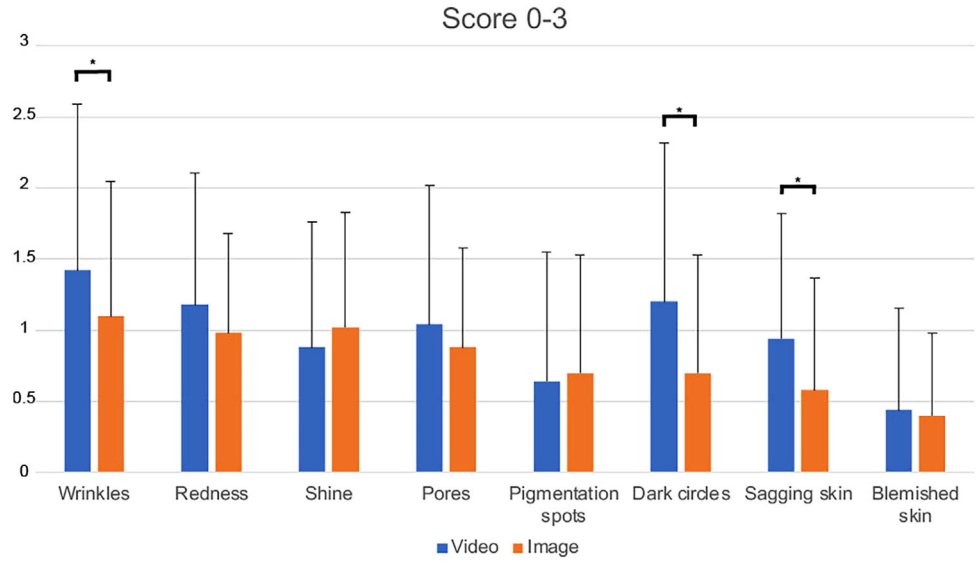

FIGURE 2 In the semiquantitative score of 0 to 3, there is a significant difference between wrinkles and dark circles in the videos and the images.

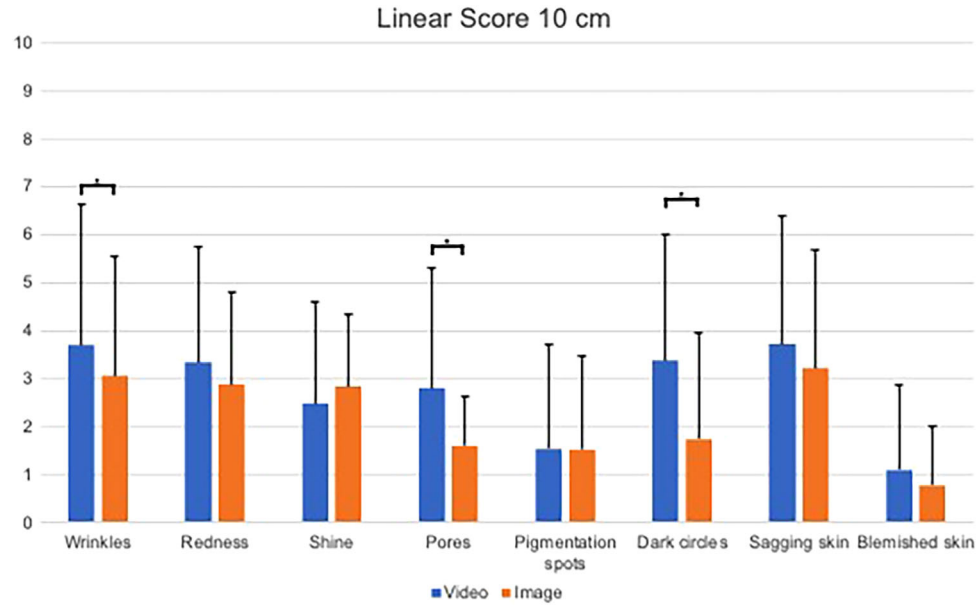

FIGURE 3 Using the linear score of 10 cm, wrinkles, pores, and dark circles showed significantly higher scores in the videos than in the images.

pronounced than the other three raters. For dark circles ($p = 7.08$) JT chose significantly lower scores than LG and GL. There were no significant differences for the other features.

4 | DISCUSSION

The goal of the research project is to evaluate facial features in a selfie image using AI. Before training the AI, we performed a preliminary study to determine if there are differences in the annotation with respect to image and video analysis, if the features can be assessed in static images or if mimics are more useful and finally if there are interindividual differences in annotation. The use of AI-based technologies is a relatively new approach that is predominantly used in the medical field.[7–10] In cosmetic sciences, AI-based tools for automated quantification of skin characteristics are a growing market. Supervised training of the algorithms requires an annotation of the features. The reliability and validity of the annotation, which is based on subjective criteria, has so far not been examined in detail.

The automatic evaluation system used for this purpose does not replace a diagnostic tool. Rather, it is intended to provide advice or guidance to customers. In previous studies, the automated assessment

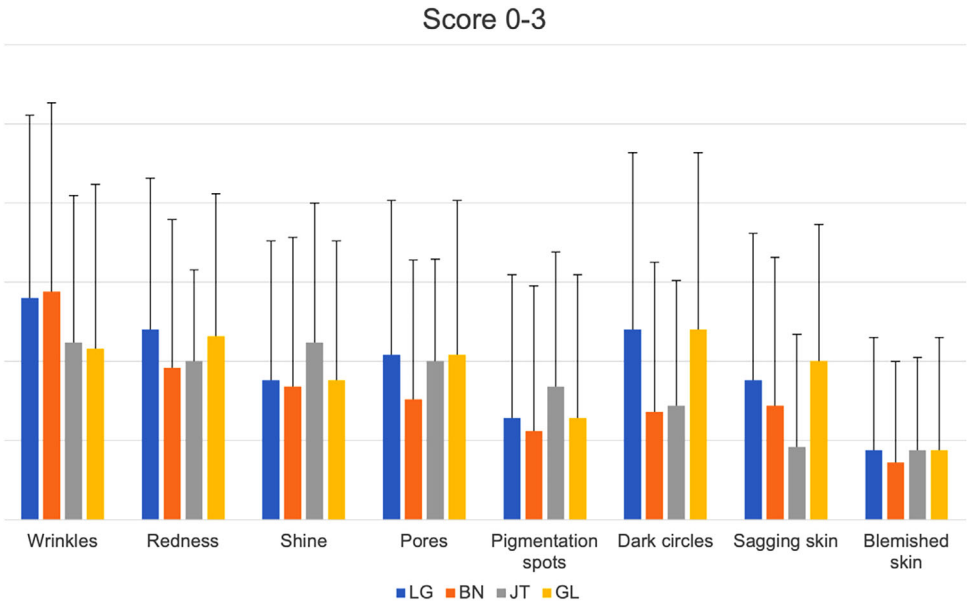

FIGURE 4 In the semi-quantitative evaluation, a significant difference was found between the raters only for the parameter skin sagging. Rater JT rated skin sagging significantly more pronounced than the other three raters. There was no significant difference in the ratings for the other characteristics.

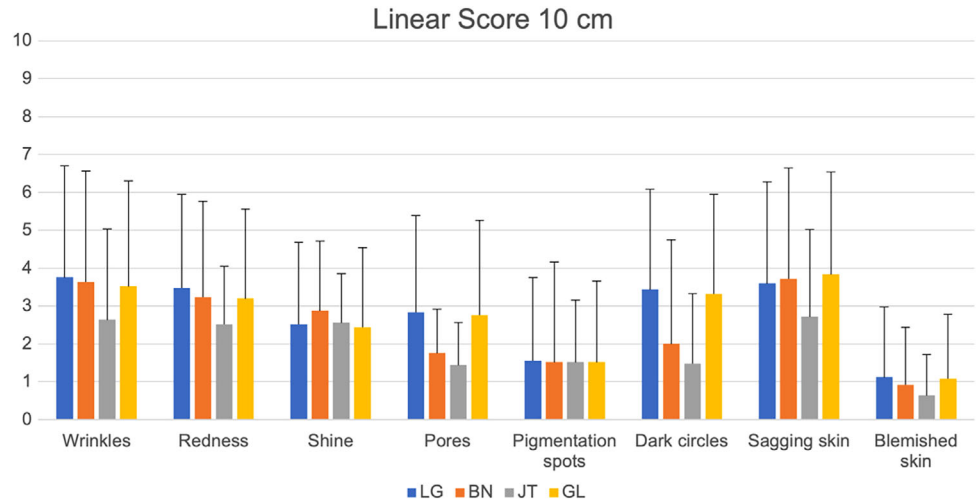

FIGURE 5 When applying the linear score for the parameters such as wrinkles, dark circles and skin sagging, there was a significant difference between the raters. Rater JT rated wrinkles significantly less pronounced and skin sagging significantly more pronounced than the other three raters. For dark circles, JT chose significantly lower scores compared to LG and GL.

format has been shown to be robust, feasible, and in close agreement with the clinical assessment of experts and dermatologists who have made assessments based on images.[11]

In addition to the use of classic imaging techniques (standardized photographs, videos, optical microscopy images, etc.), a few papers describe the use of smartphone images (no selfies).[12,13]

Static facial photographs are usually used for annotation and training of the algorithm as well as for evaluation of the skin characteristics. Dynamic changes such as mimic wrinkles, elasticity, and slackening of the skin can only be judged to a limited extent on static images, especially since self-photographs are often beautified.

Most apps use these static "beautiful" images from image collections for the training of the algorithm. However, characteristics such as elasticity and wrinkles are much easier to detect and evaluate in dynamic shots. Therefore, we compared video sequences with static facial expression sequences to assess which technique is better suited for depicting skin characteristics in a natural way. The results of our study confirm this assumption, with wrinkles, dark circles, pore formation, and skin sagging being assessed more distinctly using video analysis than images.

Furthermore, eight facial features were assessed in the present study, whereas other studies have examined many more features. For

example, Flament et al. included 23 facial features including hair to recommend make-up treatments. Certainly, the appearance of the face is a composite of different facial features. However, for the product recommendation, the eight facial features specified were sufficient. An extension of the facial features is thus dependent on the question.

Cosmetic products are often marketed worldwide. In order to be able to offer adequate skin care products to the individual population groups, the specific characteristics of the aging process in the various geographical regions must be known. Thus, an app or platform with integrated AI-based assessment of skin characteristics should take this diversity of skin types into account. A fundamental problem when using AI is that the training data does not take into account the diversity of the real data. This can result in misjudgment and discrimination against individual groups. To avoid this, the training data set should be as diverse as possible and a representative section of the real data. It is important for future studies to also consider data on the exposome, that is, on the factors known to influence the skin aging process.

The influence of the annotation on the quality of the AI has so far been little examined and considered. In our study, there were significant differences between rater JT and the other three raters. Specifically, in the semiquantitative score, the feature skin laxity was rated significantly more pronounced. Semiquantitative scores allow few gradations in the annotation, so that subjective fluctuations must be considered here. This will naturally affect the result of the AI. This bias should be overcome by regular review of the annotators and follow-up training.

A possible more objective system for skin type classification was applied by Seo et al.[14] In this study, noninvasive biotechnological devices were used to obtain measurement results on skin parameters, which were used to train the AI. In the study by Malihi et al.,[15] neural networks were used to classify wound types, and in the Gibstein et al. study,[16] neural networks were applied to assess surgical outcomes after face lift surgery. In our study, the AI was trained using clinical images and facial expressions, which may be subject to significantly more subjective variation. A more objective skin type classification system offers the possibility of producing reproducible results to a greater degree and could have been a useful addition to our study.

However, studies by Flament et al.[17] and Zhang et al.[18] confirmed that it is possible to have facial features accurately assessed by an AI using selfie images. Here, a strong correlation of the AI-based evaluations with the evaluations of experts or dermatologists was shown. To ensure reproducible results with consistently high quality, regular training of the evaluators is of great relevance. Intensive training of the raters was also a core component of the preparation and evaluation phase in our study.

Furthermore, in the studies by Flament et al.[19] and Zhang et al.,[18] the evaluation of the AI was even slightly superior to the evaluations of the experts in some characteristics. This again emphasizes the immense opportunities of this technology, which will be ubiquitous and useful for many issues in the future. Certainly, our preliminary study also offers an important contribution to the further development of AI.

Whatsoever, the present study owns further limitations. On the one hand, the work, which was declared as a preliminary study, had a rather short elaboration phase, which meant that further questions could not be addressed. For example, it would be interesting to find out whether a significantly improved precision of the AI assessment can be ensured by more deficit-oriented training. Another problem is the small sample size of this study, which is not representative of larger populations. In addition, our population consisted of predominantly young subjects, who often offered less pronounced facial features such as wrinkles and pigment spots for training the AI. Another limitation in this context is the rather homogeneous subject cohort. Studies by Flament et al.[17] can score here with much larger subject populations and heterogeneous factors such as large age differences.

In general, another limitation of our study is the small number of only four raters. Accordingly, the differences between raters in the analyses were large. Here, the study by Flament et al.[17] can contrast 50 US dermatologists of various profiles and provide a much more precise data set.

However, for all other features, there were no significant differences between raters in the semiquantitative score. It can also be assumed that the feature skin sagging is a more complex feature to evaluate, as the individual facial anatomy of the subjects can only be approximated on a photo sequence or video sequence. The face is one of the most complex regions of the body and is subject to a wide variety of influences, especially in the aging process, from which many morphologic changes result.[19]

In the linear ratings, there were significant differences in several characteristics such as wrinkles, dark circles, and skin sagging. JT rated wrinkles significantly less pronounced and skin sagging significantly more pronounced than the other three raters. For eye circles, JT chose significantly lower scores compared to LG and GL. One could assume that this could be a gender-specific factor and that men rate differently than women, but there is currently no scientific data on this. Alternatively, this may be due to insufficient training of the raters. In a study by Flament et al.,[19] the following solution to this problem was found: "If a dermatologist grading differed from the trained dermatologists' reference value by ± 0.4 grading units, they were requested to regrade the image three times, and then another set of previously graded images was sent to them to check their accuracy (…)". Special retraining will be useful in the future to obtain replicable annotations. There were no significant differences in scores for the other features. The conclusion should be that individuals should be trained on standard images whenever they mount. Another possibility of correction is that the scores of persons who consistently annotate lower or higher values than the average are computationally corrected and adjusted to the mean. In addition, the linear score has a finer gradation, so that greater variability must be expected.

This pilot study served to evaluate possible influencing factors on the annotation of large image data sets by numerous people. We learned from this that the image data should be as diverse as possible and reflect the entire variety of the target group. A regular review of the annotation procedure and, if necessary, retraining in the event of significant deviations is of importance to guarantee reliable results of the algorithm.

5 | CONCLUSION

In summary, the application of AI is a useful tool not only in the medical field but also in the pharmaceutical industry and cosmetic sciences. Our study shows that there are significant differences whether annotation of skin parameters is performed on moving or static images. Regular training of the annotating individuals and cross-validation of the annotation allows for accurate and meaningful assessment.

ACKNOWLEDGMENTS

The project is part of the "FuE-Kooperationsprojekt Grand-AID", which received funding from the funding program "Zentrales Innovationsprogramm Mittelstand" of the Bundesministerium für Wirtschaft und Klimaschutz (BMWK)-funding code 16KN093231. The manuscript reflects only the author's views and the BMWK is not liable for any use that might be made of information contained herein.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

GD and AR are employees of GRANDEL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Gabriele Maria Lehner https://orcid.org/0009-0001-4890-9374
Laura Gockeln https://orcid.org/0009-0002-3596-7601
Bettina Marie Naber https://orcid.org/0009-0005-3488-1349
Janis Raphael Thamm https://orcid.org/0000-0001-6205-7959
Sandra Schuh https://orcid.org/0000-0002-1470-7619
Dennis Hartmann https://orcid.org/0000-0002-8751-2505
Frank Kramer https://orcid.org/0000-0002-2857-7122
Julia Welzel https://orcid.org/0000-0002-6099-7418

REFERENCES

1. Flament F, Abric A, Adam AS. Evaluating the respective weights of some facial signs on perceived ages in differently aged women of five ethnic origins. *J Cosmet Dermatol*. 2021;20(3):842-853. doi:10.1111/jocd.13612
2. Batres C, Porcheron A, Latreille J, Roche M, Morizot F, Russell R. Cosmetics increase skin evenness: evidence from perceptual and physical measures. *Skin Res Technol*. 2019;25(5):672-676. doi:10.1111/srt.12700
3. Jiang R, Kezele I, Levinshtein A, et al. A new procedure, free from human assessment that automatically grades some facial skin structural signs. Comparison with assessments by experts, using referential atlases of skin ageing. *Int J Cosmet Sci*. 2019;41(1):67–78. https://doi-org.ezproxy.bibliothek.uni-augsburg.de/10.1111/ics.12512
4. Beltrami EJ, Brown AC, Salmon PJM, Leffell DJ, Ko JM, Grant-Kels JM. Artificial intelligence in the detection of skin cancer. *J Am Acad Dermatol*. 2022;87(6):1336-1342.
5. Flament F, Velleman D, Yamashita E, et al. Japanese experiment of a complete and objective automatic grading system of facial signs from selfie pictures: validation with dermatologists and characterization of changes due to age and sun exposures. *Skin Res Technol*. 2021;27(4):544–553. https://doi-org.ezproxy.bibliothek.uni-augsburg.de/10.1111/srt.12982
6. Flament F, Jacquet L, Ye C, et al. Artificial intelligence analysis of over half a million European and Chinese women reveals striking differences in the facial skin ageing process. *J Eur Acad Dermatol Venereol*. 2022;36(7):1136-1142.
7. Connor CW. Artificial intelligence and machine learning in anesthesiology. *Anesthesiology*. 2019;131(6):1346-1359. doi:10.1097/ALN.0000000000002694
8. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. *Comput Biol Med*. 2020;127:104065. doi:10.1016/j.compbiomed.2020.104065
9. Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. *Br J Dermatol*. 2020;183(3):423-430. doi:10.1111/bjd.18880
10. Sengupta S, Mittal N, Modi M. Improved skin lesions detection using color space and artificial intelligence techniques. *J Dermatolog Treat*. 2020;31(5):511-518. doi:10.1080/09546634.2019.1708239
11. Flament F, Hofmann M, Roo E, et al. An automatic procedure that grades some facial skin structural signs: agreements and validation with clinical assessments made by dermatologists. *Int J Cosmet Sci*. 2019;41(5):472–478. https://doi-org.ezproxy.bibliothek.uni-augsburg.de/10.1111/ics.12563
12. Ngoo A, Finnane A, McMeniman E, Tan JM, Janda M, Soyer HP. Efficacy of smartphone applications in high-risk pigmented lesions. *Australas J Dermatol*. 2018;59(3):e175-e182. doi:10.1111/ajd.12599
13. Chuchu N, Takwoingi Y, Dinnes J, et al. Cochrane Skin Cancer Diagnostic Test Accuracy Group. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. *Cochrane Database Syst Rev*. 2018;12(12):CD013192. doi:10.1002/14651858.CD013192
14. Seo JI, Ham HI, Baek JH, Shin MK. An objective skin-type classification based on non-invasive biophysical parameters. *J Eur Acad Dermatol Venereol*. 2022;36(3):444-452. doi:10.1111/jdv.17793
15. Malihi L, Hüsers J, Richter ML, et al. Automatic wound type classification with convolutional neural networks. *Stud Health Technol Inform*. 2022;295:281-284. doi:10.3233/SHTI220717
16. Gibstein AR, Chen K, Nakfoor B, et al. Facelift surgery turns back the clock: artificial intelligence and patient satisfaction quantitate value of procedure type and specific techniques. *Aesthet Surg J*. 2021;41(9):987-999. doi:10.1093/asj/sjaa238
17. Flament F, Jiang R, Houghton J, et al. Accuracy and clinical relevance of an automated, algorithm-based analysis of facial signs from selfie images of women in the United States of various ages, ancestries and phototypes: a cross-sectional observational study. *J Eur Acad Dermatol Venereol*. 2023;37(1):176–183.
18. Zhang Y, Jiang R, Kezele I, et al. A new procedure, free from human assessment, that automatically grades some facial skin signs in men from selfie pictures. Application to changes induced by a severe aerial chronic urban pollution. *Int J Cosmet Sci*. 2020;42(2):185–197. https://doi-org.ezproxy.bibliothek.uni-augsburg.de/10.1111/ics.12602
19. Bazin R, Doublet E. Skin aging atlas. Volume 1, *Caucasian type*. Editions Med'Com; 2007.

How to cite this article: Lehner GM, Gockeln L, Naber BM, et al. Differences in the annotation between facial images and videos for training an artificial intelligence for skin type determination. *Skin Res Technol*. 2024;30:e13632. https://doi.org/10.1111/srt.13632