IFSCC 2025 full paper (IFSCC2025-1194)

"CNN based data-driven study for "suits the skin" makeup recommendation."

Ken Nishino[1,*], Midori Takazawa[1], Asami Kato[1], Akito Hattori[1], Noriko Nakamura[1], Kahori Han'yama[1], Natsuko Okiyama[1]

[1] Cosmetic Product Research Laboratory, Kao Corporation, Odawara, Japan

1. Introduction

Human skin textures are strongly influence impressions [1-4]. Base makeup such as cosmetic foundations are designed to correct skin texture to enhance the impression of the skin. However, base makeup can make skin problems such as pores and wrinkles more noticeable, or can agglomerate into pores and wrinkles, resulting in an undesirable finish. Such a poor finish is a major source of dissatisfaction for customers. One of the reasons for this is the relationship between skin condition and makeup formulation, i.e., the base makeup formulation was not suitable for the skin. In order to provide base makeup that suits the skin of all people, it is necessary to clarify the relationship between skin condition and makeup finish. In this study, we developed deep learning models which can analyze both the state of bare skin and the finish of makeup, and applied them to discover "suit the skin" logic and develop solutions.

How can we find a relationship between skin condition and makeup finish from a wide variety of skin conditions? A data science approach, in which multifaceted analysis of skin conditions and makeup finishes is performed on a vast amount of data, is considered indispensable for this purpose. We have attempted to develop and apply a deep layer convolutional neural network (CNN) model for detailed analysis of bare skin and makeup finish in order to elucidate the makeup finish in such a data-driven manner.

There have already been several reports on the application of CNNs for skin evaluation. The most popular application is medical, especially in the identification and detection of skin cancer [5-10]. Regarding makeup, for example, facial attractiveness evaluation, automatic makeup generation, makeup pattern suggestion, and makeup finish evaluation method have been reported already [11-15]. CNN is effective for both skin condition and makeup finish analysis, and it is likely that the approach of combining these two methods to approach the relationship between skin condition and makeup finish is reasonable.

In this study, we first developed a CNN model that can evaluate various skin conditions and makeup finishes. Next, we applied the CNN model to clarify the relationship between skin conditions and visually evaluated "undesirable makeup finish". Finally, focusing on "dry skin," we evaluated both changes in skin conditions and makeup finishes during a day using the CNN model to investigate the possibility of a multi-perspective study of "foundations that suit the skin" by evaluating skin and makeup finishes using the CNN model. All procedures followed the Helsinki protocols and were approved by the Ethics Committee of Kao

Corporation. The research volunteers were given oral explanations and gave consent in advance regarding the purpose, method, safety considerations, and risks of the experiment.

2. Materials and Methods

2.1. CNN based models for skin and makeup finish evaluation

In this study, two types of skin image evaluation algorithms, "patch-based" and "part-based," which include CNN models, were used (Figure 1). The advantage of the patch-based algorithm is that the point-by-point evaluation results can be visualized.

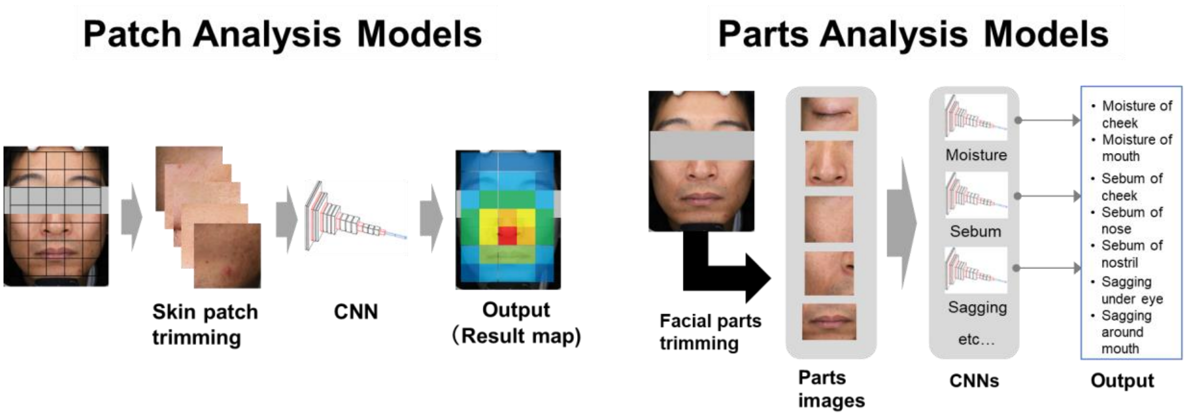

Figure 1. Two skin evaluation algorithms including CNN models.

However, this algorithm can only be used for appearance features that do not depend on facial regions and cannot be used to estimate features that occur in specific regions, such as "wrinkles at corner of the eyes". In addition, a dataset requires annotation information for each patch. On the other hand, the part-based algorithm can be used for datasets with clear site-specific annotation information, such as "wrinkles around the eyes" and "moisture content in the cheeks". We trained over 60 assessment metrics using these two algorithms. However, in this report, we only report on the CNN models of the assessment metrics and algorithms shown in Table 1, which are used in our experiments.

CNN models were trained using Japanese bare/makeup skin images, visual assessment scores and measurement data. Face images were captured using the face image capture devices with built-in single-lens reflex cameras, VISIA CR and VISIA Evolution (Canfield). The data set size varied depending on the evaluation metric, but overall, images for a total of 565 Japanese men and women were used. Although the machine learning model trained with Japanese images, the performance of the models on diverse ethnic groups (Asian: n=40, Caucasian: n=10, and African: n=10) was evaluated.

Table 1. Assessment metrics and measured positions

Assessment metrics	Method	Position
Moisture	Skicon 200 EX	Cheek Mouth
TEWL	Tewameter	Cheek
Ceramide	Tape stripping	Cheek
Skin hardness	Cutometer R0	Cheek
Skin viscoelasticity	Cutometer R7	Cheek
Dermal collagen density	DermaLab	Cheek
Roughness	Primos CR	Cheek
Sebum	SEBU meter	Forehead Cheek Nose
Sagging	Visual assessment	Under eye Mouth
Pore prominence	Visual assessment	Nose Forehead Cheek
Wrinkle	Visual assessment	Nasolabial Corner of eye Under eye Forehead Between eyebrows
Wrinkling under eyes	Visual assessment	Under eye
Scale prominence	Visual assessment	Skin Patch

2.1. CNN-models-based skin conditions / makeup finish relationship analysis

To clarify the relationship between bare skin and makeup finish, makeup finish of Japanese women facial images was visually assessed by experts and their bare skin images were analyzed with the developed models. The images to be visually evaluated were the five areas shown in Figure 2, which were extracted from the VISIA evolution standard images (267 Japanese women in their 20s to 70s). Evaluators responded to the items shown in Table 2 on a 2AFC scale. The evaluators were 37 skilled makeup evaluation in their 20s to 50s who belonged to Kao Corporation. For all images, the selection rate for all items was calculated and the correlation coefficient with the skin measurements predicted using a CNN model was evaluated.

Figure 2. Five facial parts for the visual assessment

Table 2. Visual assessment items

	Forehead	Eye	Nostril	Cheek	Smile line
Comparison with bare skin	Vertical wrinkles were noticeable	Wrinkles under eyes were noticeable	Pores were noticeable	Pores were noticeable	Pores were noticeable
	Horizontal wrinkles were noticeable	Increased dryness in appearance	Scales were noticeable	Scales were noticeable	Scales were noticeable
	Pores were noticeable		Pore plugs were noticeable	Skin texture roughness was noticeable	Wrinkles were noticeable
	Scales were noticeable			Increased dryness in appearance	
Monadic judgment	Accumulation in pores	Accumulation in double eyelid	Accumulation in nostril	Accumulation in pores	Accumulation in pores
	Accumulation in wrinkles	Uneven finish	Accumulation in pores	Uneven finish	Uneven finish
	Uneven redness		Uneven finish		

2.3. Evaluation of the actual state of makeup finish related to dry skin

Focusing on "scale prominence," a characteristic of dry skin and one of the undesirable features of makeup finish, we conducted a test to observe changes in skin condition over the course of a day in order to clarify the relationship between the daily changes in skin condition and the impression of the skin's appearance. In this test, facial photographs were taken using VISIA Evolution at five time points: bare skin before makeup, immediately after makeup application, three hours after makeup application, seven hours after makeup removal, and after makeup removal. This process was performed twice, once in summer and once in winter, each time using different makeup products. The study participants were 20 Japanese women in their 20s.

The facial photographs were analyzed using the developed CNN model. The scores of good makeup finish evaluated by a visual evaluation skilled person at three timings of the winter test, immediately after makeup application, 3 hours after makeup application, and 7 hours after makeup application, were used to evaluate the correlation between the CNN model's evaluation metric "scale prominence" and the appearance of the makeup finish. Seasonal differences in scale prominence changes caused by makeup breakdown were also evaluated by comparing scale prominence immediately after makeup and 7 hours after makeup. In addition, skin changes due to daytime activities were evaluated by comparing the scale prominence of bare skin before makeup (at morning) and after makeup removal (at night). The area where the test site is located tends to be drier in winter, with an average humidity of around 70% in summer versus less than 50% in winter. The assessment of seasonal differences may reflect differences in the dryness of the outside air.

3. Results and Discussion

3.1. CNN based models for skin and makeup finish evaluation

The prediction accuracy of the trained machine learning models averaged 0.691 in terms of correlation coefficients. Visual assessment metrics tended to have better accuracy, with an average correlation coefficient of 0.816, while the other metrics had an average of 0.555. Since the algorithm for predicting skin conditions from facial images involves finding clues for various evaluation metrics from skin surface images and making predictions, visual evaluation metrics, which are the skin appearance features themselves, are considered to easily obtain high accuracy. On the other hand, the average correlation of 0.555 is a significant accuracy that cannot be considered coincidental, so the predicted value obtained from it certainly reflects the skin image characteristics associated with the metric and is considered effective in data-driven analysis.

When the developed machine learning model was applied to facial images from individuals of various ethnic backgrounds, the model's performance showed an average correlation coefficient of 0.706 for specific visual evaluation metrics. However, for other metrics, the average correlation coefficient was 0.275, indicating lower accuracy. If the evaluation metric directly correlates with image features, small-scale validation and fine-tuning could potentially be performed for diverse ethnic groups. However, to accurately assess other skin properties and similar characteristics across different ethnic groups, further data collection and retraining of the model are recommended.

3.2. CNN-models-based skin conditions / makeup finish relationship analysis

Using the developed CNN model, correlation coefficients between predicted skin characteristics and visually assessed undesirable makeup finishes were calculated. All calculated results are described as heat maps in Figure 3.

Correlation	Undesirable changes (be noticeable) due to makeup														Unfavorable finish									
	Wrinkles				Pores				Scales				Other		Accumulation in pores				Other accumulations		Uneven makeup finish			
	Vertical forehead	Horizontal forehead	Under eye	Mouth	Forehead	Nostril	Cheek	Mouth	Forehead	Nostril	Cheek	Mouth	Roughness of cheek	Pore plugs	Forehead	Nostril	Cheek	Mouth	Forehead wrinkle	Nostril	Eye	Nostril	Cheek	Mouth
Moisture (Cheek)	0.124	0.095	0.201	0.224	0.158	0.201	0.197	0.203	-0.039	-0.187	-0.071	-0.247	0.247	-0.039	0.165	0.122	0.121	0.064	0.078	-0.066	0.06	-0.001	0.076	-0.082
Moisture (Mouth)	0.218	0.291	0.154	0.263	0.304	0.166	0.28	0.286	-0.141	-0.345	-0.228	-0.421	0.271	-0.066	0.192	0.092	0.179	0.122	0.212	-0.202	0.002	-0.141	0.062	-0.276
TEWL (Cheek)	-0.151	-0.227	-0.208	-0.276	-0.087	-0.098	-0.118	0.01	0.116	0.249	0.066	0.176	-0.338	0.197	0.071	0.048	0.004	0.075	-0.154	0.116	-0.05	0.011	-0.135	0.077
Ceramide (Cheek)	0.055	0.159	0.083	0.11	0.057	0.028	-0.009	-0.104	-0.087	-0.214	-0.001	-0.06	0.248	-0.157	-0.133	-0.085	-0.095	-0.116	0.069	-0.174	0.006	-0.123	0.115	-0.079
Skin hardness (Cheek)	-0.093	0	-0.001	-0.13	0.076	-0.128	-0.073	0.071	0.061	-0.001	-0.105	0.05	0.075	0.055	-0.013	-0.038	-0.061	0.071	-0.107	-0.065	0.078	-0.08	-0.081	0.008
Prominence of pores (Nose)	0.252	0.326	0.076	0.144	0.242	0.336	0.102	0.016	-0.166	-0.217	-0.162	-0.231	0.231	0.128	0.159	0.209	0.062	-0.03	0.317	-0.256	-0.12	0.026	0.11	-0.169
Prominence of pores (Forehead)	0.102	0.172	0.012	0.002	0.326	0.179	0.119	0.154	-0.163	-0.06	-0.117	-0.203	0.002	0.17	0.333	0.256	0.097	0.072	0.187	-0.131	-0.1	0.076	-0.036	-0.138
Prominence of pores (Cheek)	0.214	0.237	0.058	0.143	0.241	0.231	0.295	0.258	-0.121	-0.114	-0.169	-0.331	0.073	0.141	0.332	0.248	0.247	0.115	0.308	-0.165	-0.064	0.027	-0.006	-0.242
Dermal collagen density (Cheek)	-0.278	-0.282	-0.121	-0.285	-0.206	-0.263	-0.334	-0.256	0.075	0.087	0.049	0.327	-0.166	-0.017	-0.282	-0.21	-0.277	-0.111	-0.361	0.108	0.009	-0.073	-0.145	0.188
Viscoelasticity (Cheek)	-0.406	-0.489	-0.22	-0.435	-0.241	-0.372	-0.283	-0.106	0.174	0.219	0.084	0.409	-0.37	0.084	-0.161	-0.178	-0.213	0.029	-0.535	0.227	-0.036	-0.095	-0.266	0.196
Sagging (Under eye)	0.431	0.606	0.269	0.417	0.211	0.373	0.179	-0.012	-0.144	-0.241	-0.108	-0.346	0.445	-0.095	0.101	0.168	0.104	-0.101	0.663	-0.28	0.041	0.145	0.303	-0.143
Sagging (Corner of mouth)	0.433	0.594	0.241	0.425	0.197	0.351	0.196	-0.038	-0.115	-0.206	-0.048	-0.293	0.425	-0.094	0.078	0.163	0.112	-0.127	0.645	-0.275	0.048	0.142	0.332	-0.124
Roughness (Cheek)	0.265	0.395	0.191	0.197	0.322	0.228	0.246	0.34	-0.081	-0.185	-0.216	-0.397	0.261	0.077	0.293	0.245	0.165	0.149	0.378	-0.24	0.066	-0.011	0.018	-0.257
Wrinkling (Under eye)	0.472	0.611	0.313	0.446	0.209	0.366	0.2	0.038	-0.114	-0.223	-0.101	-0.331	0.466	-0.068	0.11	0.177	0.127	-0.054	0.674	-0.29	0.067	0.146	0.314	-0.122
Sebum (Forehead)	-0.369	-0.459	-0.206	-0.279	-0.171	-0.273	-0.104	-0.003	0.045	0.255	0.105	0.298	-0.38	0.111	-0.049	-0.089	-0.045	0.084	-0.521	0.196	-0.057	-0.09	-0.189	0.137
Sebum (Cheek)	-0.324	-0.483	-0.248	-0.416	-0.127	-0.296	-0.21	0.027	0.115	0.219	0.062	0.302	-0.413	0.187	0.015	-0.105	-0.103	0.148	-0.537	0.196	-0.086	-0.125	-0.297	0.081
Sebum (Nostril)	-0.219	-0.36	-0.285	-0.333	-0.037	-0.159	-0.135	0.021	0.116	0.212	-0.032	0.207	-0.328	0.296	0.103	0.031	-0.059	0.144	-0.363	0.181	-0.116	-0.037	-0.271	0.087
Wrinkle (Forehead)	0.359	0.533	0.177	0.305	0.127	0.309	0.124	-0.081	-0.12	-0.208	-0.109	-0.275	0.361	-0.018	0.033	0.136	0.013	-0.15	0.618	-0.282	-0.037	0.085	0.182	-0.202
Wrinkle (Between eyebrows)	0.364	0.444	0.139	0.266	0.06	0.249	0.055	-0.107	-0.057	-0.117	-0.085	-0.225	0.303	-0.028	0.017	0.124	0.034	-0.165	0.622	-0.21	-0.002	0.104	0.255	-0.114
Wrinkle (Under eye)	0.449	0.598	0.266	0.457	0.202	0.377	0.204	0.014	-0.123	-0.222	-0.101	-0.34	0.46	-0.049	0.101	0.186	0.119	-0.064	0.66	-0.266	0.039	0.134	0.317	-0.165
Wrinkle (Corner of eye)	0.445	0.566	0.239	0.388	0.201	0.35	0.129	-0.063	-0.137	-0.207	-0.069	-0.29	0.404	-0.079	0.075	0.148	0.074	-0.142	0.643	-0.268	0.03	0.147	0.306	-0.135
Wrinkle (Nasolabial)	0.417	0.569	0.215	0.437	0.191	0.362	0.212	-0.019	-0.107	-0.194	-0.065	-0.33	0.383	-0.05	0.129	0.203	0.18	-0.078	0.641	-0.257	0.07	0.139	0.323	-0.146

Figure 3. Relation map between skin properties and makeup finish

Here, positive correlations are colored red and negative correlations are colored blue, and the shading corresponds to the magnitude of the correlation coefficient. Cells with p-values less than 0.05 corrected by Bonferroni's correction are shown in bold. In the visual evaluation, items evaluated by comparing bare skin with skin after makeup are described on the left side of the figure as "Undesirable changes due to makeup," and items evaluated by looking only at the image after makeup are described on the right side of the figure as "Unfavorable finish."

The figure shows that many makeup finish features correlate with various skin conditions. However, these correlations are highly collinear and show a very simple relationship. The dominant factor is age. Younger skin without wrinkles and sagging is more prone to the phenomena of "scales are noticeable" and "pore plugs are noticeable," but the incidence of these phenomena related to morphological characteristics of the skin, such as "wrinkles and pores after makeup are noticeable" and "accumulation in wrinkles and pores," increases with age.

Even if age is the dominant factor in the data, it is certainly not the only factor that contributes to the perception of "suit the skin," and further analysis is needed as more data is accumulated. In addition to the "undesirable makeup finish" shown here, there are a wide variety of factors that contribute to the perception of "suit the skin," including the color, texture, and feel of the finished product, as well as its change over time. However, as data becomes more multidimensional and the number of data increases, data-driven analysis methods are expected to become more effective. In this analysis, the correlation with the visual evaluation was evaluated because the machine learning model developed did not cover the "undesirable makeup finish" of the visual evaluation item. However, we believe that the development of a machine learning model that can accurately and comprehensively evaluate this cosmetic finish will greatly accelerate data accumulation and the development of data-driven solutions.

3.3. Evaluation of the actual state of makeup finish related to dry skin

First, we evaluated the correlation between Scale prominence predicted by the CNN model and visual evaluation results. The results are shown in Figure 4. As shown in the figure, the predicted scale prominence in the makeup finish is found to have an unfavorable trend in appearance. Compared with other visual evaluation results, it was also found that the skin flake and base makeup agglomerates were similar in appearance, giving the impression that the base makeup was adhering unevenly to the skin.

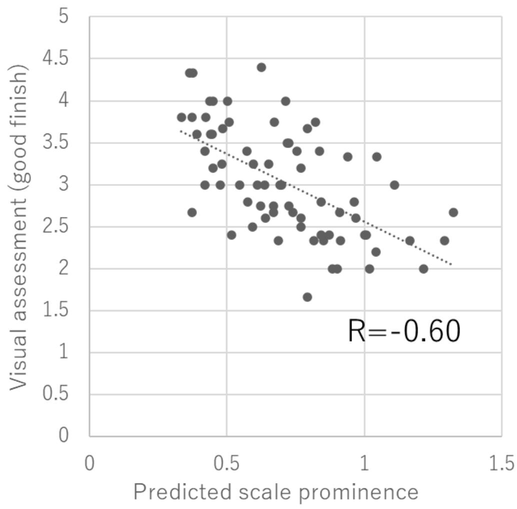

Figure 4. Comparison of the predicted scale prominence by the CNN model and visual assessment (good finish).

Figure 5 shows the results of comparing Scale prominence immediately after makeup application and after 7 hours, for bare skin before makeup (morning) and bare skin after cleansing (night). The error bars represent the Standard Error. For the bare skin condition, both the change over time and the difference between seasons were tested with a paired t-test. For the makeup condition, on the other hand, only the change over time was tested with a paired t-test, since the cosmetics used during the summer and winter months differed, and therefore, no correspondence could be established. (†: $p<0.1$, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$)

As shown in the figure, "scale prominence" increased in both comparisons during the winter season. Figure 5(a) shows makeup breakdown over time, closing the fact that scale prominence can occur as a makeup breakdown specific to dry climates. Figure 5(b) is a comparison of skin conditions in the morning and at night, so the scales are more noticeable at night than in the morning, and this tendency is more pronounced in the dry winter months, which may reflect the skin burden caused by dryness throughout the day.

Figure 6 shows one example of photographs and analysis results in winter. The figure shows that the distribution of scale prominence increased with changes in makeup application, time, and cleansing. n particular, the forehead and mouth area showed a monotonous increasing trend. On the other hand, there were some areas, such as the nose area, that showed an increase after 7 hours of makeup application but a decrease after cleansing. In these areas, it is possible that the makeup foundation had a significant effect on the appearance of scale, which was eliminated by cleansing. It is desirable to explore the differences in the phenomenon between the different areas and its effect on user satisfaction, and to develop further "suit the skin" solutions.

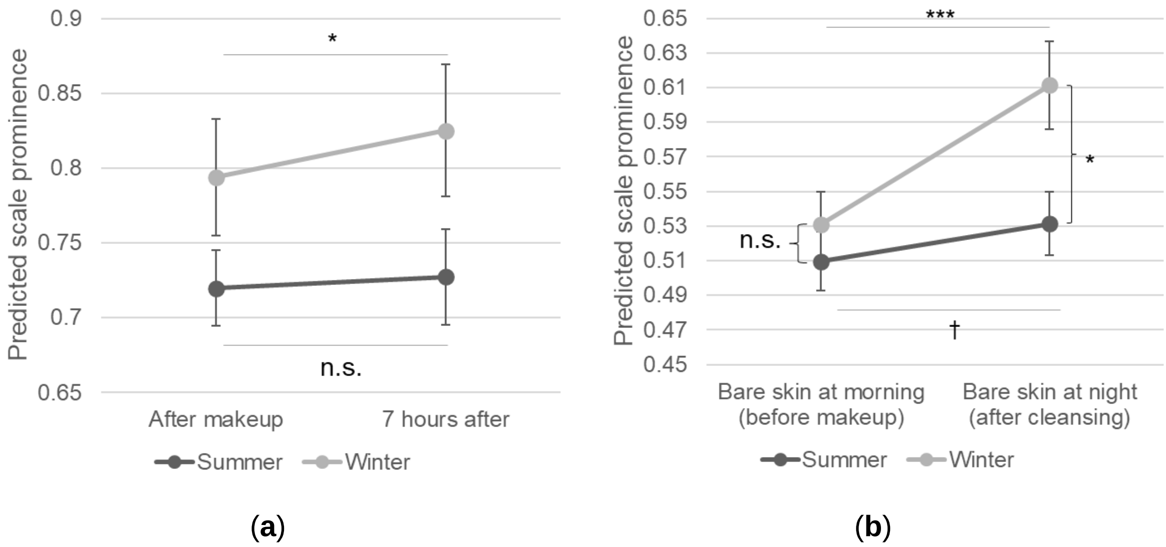

Figure 5. Scale prominence changes within a day in different seasons: (a) After makeup and 7 hours after; (b) Bare skin at morning and at night. Mean values were tested by paired t-test. (†: p<0.1, *: p<0.05, **: p<0.01, ***: p<0.001)

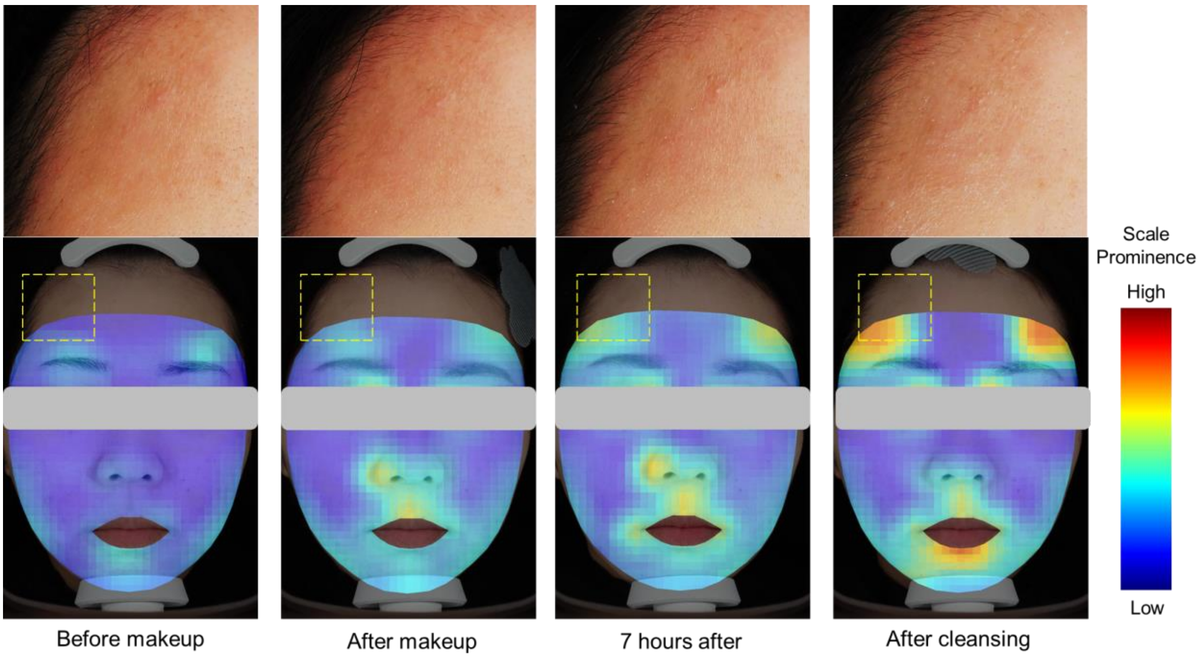

Figure 6. Example of typical daily scale prominence changes. The heatmap in the lower row shows the point-by-point scale prominence predicted by the developed CNN model. The upper panel shows the change in appearance of the skin surface on the forehead, indicated by the yellow dashed line in the lower heatmap.

4. Conclusion

Data-driven analysis of bare / makeup skin using machine learning models was shown to be effective in obtaining an overall picture of the actual state of makeup finish. In addition, the study using a machine learning model focusing on a specific cosmetic finish feature, the "scale prominence," provided new insights into skin conditions and makeup finishes. We believe that combining an understanding of the overall picture based on large-scale data with detailed analysis focusing on specific conditions and specific metric will enable us to gain a more accurate understanding of the actual situation and to develop solutions. In the future, we will further deepen our understanding of the relationship between skin condition and makeup finish, and work to develop solutions for a world in which everyone can achieve their own personal beauty.

References

1. Samson, N., et al. (2010). “Visible skin condition and perception of human facial appearance.” International Journal of Cosmetic Science 32(3): 167-184.
2. Fink, B., et al. (2006). “Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces.” Evolution and Human Behavior 27(6): 433-442.
3. Samson, N., et al. (2010). “Visible changes of female facial skin surface topography in relation to age and attractiveness perception.” Journal of Cosmetic Dermatology 9(2): 79-88.
4. Matts, P. J., et al. (2007). “Color Homogeneity and Visual Perception of Age, Health, and Attractiveness of Female Facial Skin.” Journal of the American Academy of Dermatology 57(6): 977-984.
5. Esteva, A., et al. (2017). “Dermatologist-level classification of skin cancer with deep neural networks.” nature 542(7639): 115-118.
6. Dalila, F., et al. (2017). “Segmentation and classification of melanoma and benign skin lesions.” Optik 140: 749-761.
7. Oliveira, R. B., et al. (2016). “Computational methods for the image segmentation of pigmented skin lesions: A review.” Comput Methods Programs Biomed 131: 127-141
8. Reboucas Filho, P. P., et al. (2018). “Automatic histologically-closer classification of skin lesions.” Comput Med Imaging Graph 68: 40-54.
9. Hosseinzadeh Kassani, S. and P. Hosseinzadeh Kassani (2019). “A comparative study of deep learning architectures on melanoma detection.” Tissue Cell 58: 76-83.
10. Pathan, S., et al. (2019). “Automated detection of melanocytes related pigmented skin lesions: A clinical framework.” Biomedical Signal Processing and Control 51: 59-72.
11. Alashkar, T., et al. (2017). Examples-rules guided deep neural network for makeup recommendation. AAAI.
12. Shen, S., et al. (2017). Fooling neural networks in face attractiveness evaluation: Adversarial examples with high attractiveness score but low subjective score. 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), IEEE.
13. Nishino, K. (2024). Skin patch based makeup finish assessment technique by deep neural network. Skin Research and Technology, 30(2), e13561.
14. Chang, H., et al. (2018). Paired-cyclegan: Asymmetric style transfer for applying and removing makeup. Proceedings of the IEEE conference on computer vision and pattern recognition.

15. Jiang, W., et al. (2020). Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.