Development of Diagnostic Algorithm Based on Individual Skin Properties, Lifestyle and Genetic Data for Personalized Solution

Eunbi Ko[1], Hyo Sil Kim[1], Kyoungmin Cho[4], Yoo Jin Song[4], Ji Hye Kim[1], Kiyoung Sung[1], Kyung-Won Hong[2], Seung Jae Baik[1], Young Ho Park[1], **Hye One Kim[3*]**, **Taeyoung Park[4*]**

[1]Department of R&I Center, Amorepacific Corp., South Korea, [2]Department of Bio Institute, Theragen Bio, South Korea, [3]Department of Dermatology, Hallym Univ., South Korea, [4]Department of Statistics and Data Science, Yonsei Univ., South Korea

* Hye One Kim, Department of Dermatology, Hallym univ. Seoul, Korea
Tel: 82-2-829-5221, E-mail: hyeonekim@hallym.or.kr
* Taeyoung Park, Department of Statistics and Data Science, Yonsei Univ., Seoul, Korea
Tel: 82-2-2123-2542, E-mail: tpark@yonsei.ac.kr

Abstract

Background

To select optimized skin-care solutions, it is essential to analyze and comprehend the skin. Both genetic and environmental factors act as important factors for skin aging and change. Understanding the effects of these factors not only explains why individuals differ in phenotypes, but also helps predict future phenotypes. This study aimed to quantitatively analyze the effects of environment, lifestyle, and innate genes on current skin conditions.

Methods

Six phenotypes (wrinkles, melanin, redness, dullness, hydration, and oiliness) were examined in the present study using data from 2526 women aged 20–60 years. Feature selection for the exposome and gene data was conducted using the XGBoost algorithm. Each phenotype was categorized into five classes: worst, bad, normal, good, and best. The CatBoost algorithm was used to predict the phenotype classes of the participants based on the selection of significant features, while SMOTE mitigated the corresponding class imbalance problem.

Results

Through feature selection, we identified 10 key genetic features per phenotype that were highly associated with the phenotype. The overall accuracy of the predictive model was calculated to be 46% for wrinkles, 54% for melanin, 37% for redness, 39% for dullness, 41% for hydration, and 33% for oiliness, showing an improvement of 31%, 17%, 12%, 5%, 24%, and 120%, respectively, compared with that calculated without including genetic features.

Conclusion

The ultimate goal of this study is to use big data and AI technology to predict future skin conditions based on the outcomes of our current lifestyle choices and environmental exposure as well as genetic factors. On the basis of this study, future longitudinal studies will be able to accurately predict the effect of current intrinsic and extrinsic factors on future skin conditions and develop it into a study that aids in personalized skincare solutions.

Keywords: skin diagnosis; skin type; genetic data; exposome; personalized solution

Introduction.

Skin aging and sensitivity are caused by a combination of intrinsic and extrinsic factors, including increased expression of genes, lifestyle (eg, sleep and stress), and environmental exposure (eg, change of temperature, ultrafine dust concentration and ultraviolet [UV] light).[1-5] Recently, the incidence of COVID-19 has increased people's awareness of the effects of unexpected environmental influences on the skin.[6-8] Therefore, the need for personalized skincare to keep the skin healthier is growing. To improve the current condition and prevent problems of the skin, it is necessary to identify both genetic and external factors that affect the skin.

In 2005, after American cancer epidemiologist Christopher Wilde coined the term “exposome”, Dr. Jean Krutmann proposed that environmental factors that are part of the skin aging exposome fall into the following major categories: (i) sun radiations, (ii) air pollution, (iii) tobacco smoke, (iv) nutrition, (v) under-researched, miscellaneous factors, and (vi) cosmetic products.[9-10] Data from the correlation analysis between the various skin-related variables and exposomic factors acquired by IOPE Lab (Seoul, Korea) in 2016 showed that the exposome does affect the skin. For instance, the effect of wearing a mask and fine dust on skin sensitivity were quantitatively analyzed. According to the decision tree analysis, use of sunscreen had the largest influence on aging, followed by life experiences such as pregnancy or childbirth.

Understanding the genetic characteristics of the skin will provide an optimal solution for personalized care and improvement of the skin. The genetic identification of skin-related phenotypes has become recently possible using genome-wide association studies (GWAS).

However, most GWAS have focused on skin properties that are unique to Caucasian populations. Hence, genomic studies are needed for Asian populations. In 2016, the *MC1R* gene was found to determine the age of Dutch women, and in 2021, P&G discovered genes related to “sensitive skin” through another GWAS study.[11-13] It has been reported that yet another GWAS study helped discover multiple genetic loci related to skin color in Korean women, and a machine learning algorithm that predicts the skin characteristics of Korean women was developed.[14-15]

We previously researched the prediction of the characteristics of the skin by investigating the phenotype-genotype interaction[16] and discovered markers that have a common correlation with wrinkles and melanin; additional research helped identify and report the underlying mechanism involved.[17] It has been confirmed that certain common markers related to redness, moisture content, ceramide concentration, and skin temperature are highly related to skin barrier function and are likely to be used as skin sensitivity prediction markers.[18]

In this study, we attempted to identify factors for predicting changes in skin condition through correlation analysis with lifestyle, climate/environment, and innate genes that affect the current skin condition. We used a machine learning algorithm to predict changes in the participants’ skin using variables generated during feature engineering. After converting standardized scores for each phenotype into 5-point scales, we developed a classification model for multiple classes. The most interesting aspect of this study is the role of the genotype–phenotype correlation analysis in providing insight into the effects of acquired lifestyle and environment, at large, and cosmetics, in particular, on current skin conditions. We present our initial model, which correlates the effect of each factor, including genes, on probable skin aging with good predictive power using a machine learning algorithm.

Materials and Methods.

1) Participants

The participants were recruited from the IOPE Lab (Seoul, Korea). A total of 2526 samples were gathered between 2016 and 2019. All participants provided written informed consent, and this study was approved by the institutional review board (2017-1EF-N022R).

2) Facial phenotype examination

Skin Touch® was used to measure the degree of hydration and oiliness. Antera 3D® (Miravex, Ltd, Dublin, Ireland) was used to measure not only the concentration and homogeneity of pigmentation and redness distributed over the skin but also the depth of periorbital wrinkles. Homogeneity of pigmentation is expressed as dullness in this paper.

3) Materials

This study involved 2526 Korean women aged 20 to 60 years, with 91% in their twenties and thirties. Information on 41 distinct genes and 33 lifestyle and environmental variables was gathered from the participants. The effectiveness of both lifestyle and environmental factors was demonstrated through correlation analysis using data from 2526 subjects in above sections. This research will also explain how those characteristics can be used to predict the degree of phenotypes along with genetic features from 452 subjects. In the prediction model, six phenotypic levels (wrinkle, melanin, redness, dullness, hydration, and oiliness) were measured and used as response variables. Each phenotypic level was standardized, and the minus sign was used to invert the standardized value only for wrinkles, melanin, redness, and dullness. By doing so, a larger value for each phenotypic level indicated a better skin condition. Each phenotypic level was then divided into five groups: < 10th, 10th to 30th, 30th to 70th, 70th to 90th, and ≥ 90th percentiles. Each phenotypic group was assigned one to five points, which were intuitively interpreted as worst, bad, normal, good, and best, respectively. Statistical analyses were performed using PLINK version 1.9 and SPSS program, as well as Python 3.9.

4) Genotyping and SNP quality control

Oral swab samples were obtained, and DNA was extracted using ExgeneTM Tissue SV (GeneAll, Seoul, Korea). All DNA samples were amplified and randomly portioned into 25–125 bp fragments, which were in turn purified, re-suspended, and hybridized in Axiom Genome-Wide Human Array Plates following hybridization; the bound targets were washed under stringent conditions to remove non-specific background and minimize noise resulting from random ligation events. The 902,527 single nucleotide polymorphisms (SNPs) were genotyped according to the manufacturer's instructions using an Axiom Precision Medicine Research Array (Affymetrix, Santa Clara, CA, USA), which provided genome-wide coverage in five major populations as well as imputation accuracy for GWAS markers 0.90 and 0.94 with minor allele frequencies (MAF)>1% and >5%, respectively, for the 7.4 million imputed markers in Asian population. To reduce potential concerns regarding batch effects and the possibility of false associations, we applied highly stringent quality control measures while selecting SNPs for use in the case and control datasets. Quality control procedures were performed for each of the 902 K SNPs before the association tests were conducted. The SNP set was filtered based on genotype call rates (≥ 0.98) and MAF (≥0.01). The Hardy–Weinberg equilibrium (HWE) was calculated for individual SNPs using an exact test. After filtering, 312,942 polymorphic SNPs were analyzed on chromosomes 1 to 22, and 6,416 and 816 SNPs were analyzed on chromosomes X and Y, respectively.

5) GWAS

Six phenotypes were tested by linear regression analysis with an additive model after adjustment for age. P-values were not adjusted for multiple tests. Statistical significance was determined at value of $P < 0.0001$ or functional associations.

6) Feature Selection

In the model for phenotype prediction, the covariates included 41 genetic variables, 16 lifestyle-related variables, and 17 environmental variables. Owing to the large number of features compared to the number of subjects, feature selection was required to improve performance. In this study, machine learning-based feature selection methods for selecting the genetic features were effective for removing insignificant variables. A machine learning-based feature selection method was used for selecting the genetic variables. XGBoost was used to filter out the most influential genetic predictors. The XGBoostClassifier was used to determine feature importance, and the 10 most important genetic variables per phenotype were identified.

7) Oversampling

When phenotypic traits were converted to 5-point scales, extreme skin conditions were captured using percentiles with unequal spacing. However, this often results in a class imbalance, which may lead to overfitting and poor predictive performance. Thus, SMOTE, one of the most effective oversampling strategies, was employed in this study to achieve a balance between the classes and compensate for data-related problems.

8) Machine-Learning model

CatBoost algorithm was employed for multi-class classification. This strategy works well with data containing many categorical variables. As the genetic traits and other covariates in our data were ordered categorical variables, CatBoost's encoding algorithm was useful. The data were randomly split into two sets: training (90 percent) and test (10 percent). The hyperparameters were adjusted for each response variable of the six phenotypes.

Results.

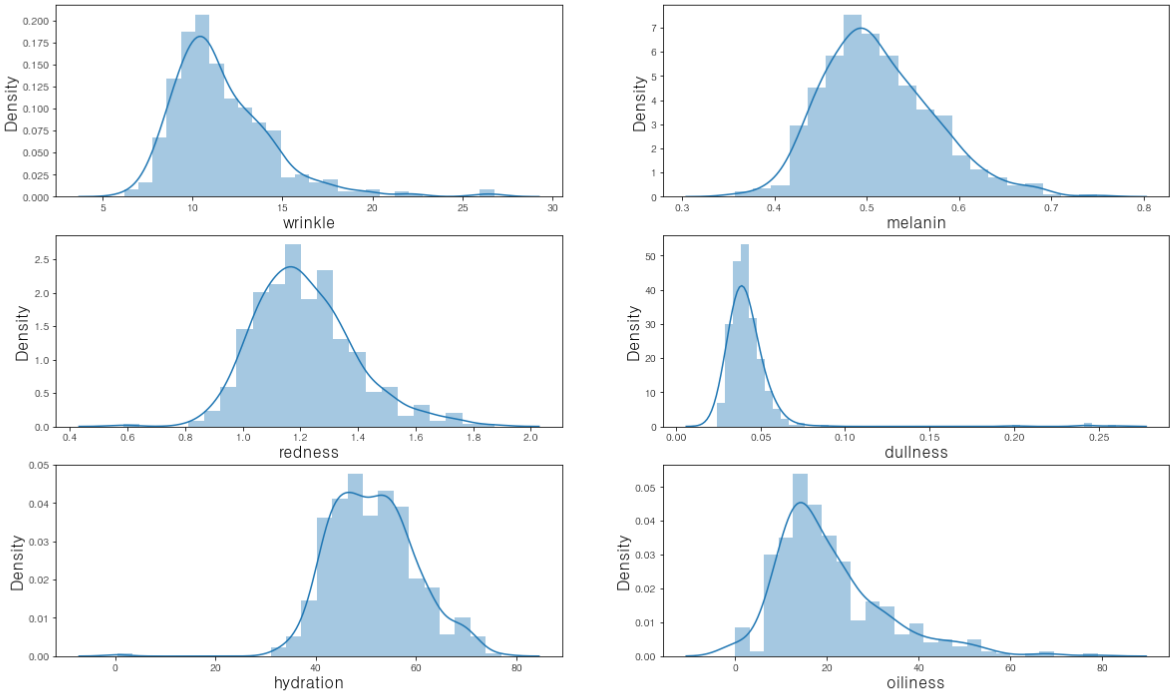
The mean and standard deviation for six unstandardized phenotypic levels are as follows: 11.65 ± 2.83 for wrinkle, 0.51 ± 0.06 for melanin, 1.22 ± 0.18 for redness, 0.04 ± 0.02 for dullness, 51.44 ± 8.56 for hydration, and 20.73 ± 12.03 for oiliness. (Table 1). Distribution plots of skin indicators showed in Figure 1.

1) Phenotype Measurements

Table 1: Parameters, Mean Values, and Standard Deviations (SDs) of the Subjects

Parameters	Site	Mean (± SD)
Wrinkles (A.U.)	Crow’s feet	11.65 (± 2.83)
Melanin (A.U.)	Cheek	0.51 (± 0.06)
Redness (A.U.)	Cheek	1.22 (± 0.18)
Dullness (A.U.)	Cheek	0.04 (± 0.02)
Hydration (A.U.)	Cheek	51.44 (± 8.56)
Oiliness (A.U.)	Cheek	20.73 (± 12.03)

Figure 1: Distribution plots of Skin Indicators



2) Machine Learning

First, an experiment was done to determine prediction performance utilizing a dataset lacking genetic information. The accuracy, precision, recall, and F1 score of the CatBoost algorithm for multi-class classification prediction of skin conditions in the test data without genetic characteristics are shown in Table 2.

Table 2. Predictive Performance of the CatBoost Algorithm without Genetic Information in Test Data

	wrinkle	melanin	redness	dullness	hydration	oiliness
Accuracy	0.35	0.46	0.33	0.37	0.33	0.15
Precision	0.18	0.48	0.29	0.35	0.45	0.16
Recall	0.32	0.46	0.29	0.36	0.45	0.24
F1 score	0.23	0.45	0.27	0.35	0.33	0.10

The model's performance without genetic information was low, according to the evaluation measures. As a result, we used feature selection to identify relevant genetic traits and paired them with lifestyle and environmental variables to predict six phenotypes. Table 3 lists certain genetic characteristics.

Table 3. Feature Importance of Genetic features

Phenotype	SNP	Gene	Feature Importance
Wrinkle	rs74718616	TLL2	0.483
	rs74650929	PRKCH	0.385
	rs142918295	FGFR1OP2	0.323
	rs142331737	SLC15A5	0.232
	rs75165433	CNKSR3	0.189
	rs140464409	ASB7	0.131
	rs181465583	SPOCK3	0.087
	rs35702263	LRRC6	0.061
Melanin	rs77310600	SIAE	0.899
	rs28641937	ADAM28	0.349
	rs16865318	CLDN1	0.157
Redness	rs74914748	KRT9	2.291
Dullness	rs438669	NAV3	0.423
	rs73353749	RCAN1	0.264
	rs3747250	CELSR1	0.150
Hydration	rs75165433	CNKSR3	0.258
	rs438669	NAV3	0.103
	rs117880177	PPP2R2B	0.004
Oiliness	rs142918295	FGFR1OP2	2.802
	rs1386821	IL6R	2.531

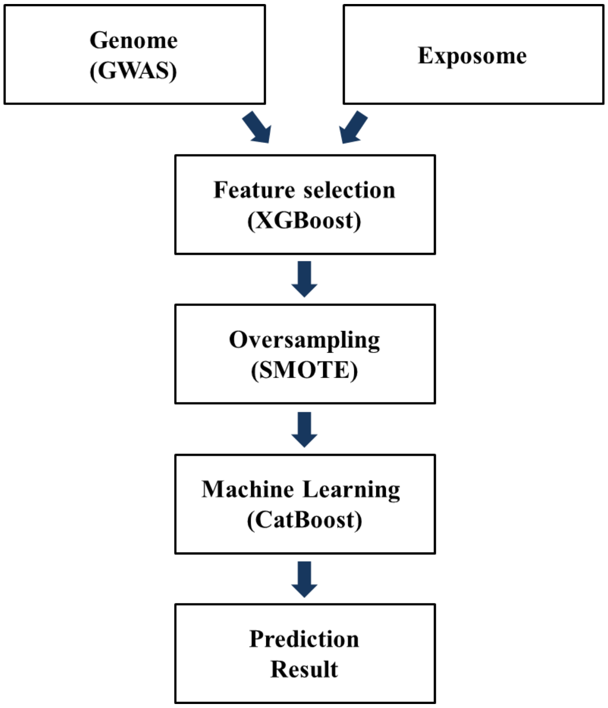
For each phenotype, we identified 10 significant genetic traits based on a machine learning feature-selection mechanism. Among the significant genetic traits, a total of 20 genetic features were chosen to have particularly good predictive power - 8 were related to wrinkles, 3 to melanin, 1 to redness, 3 to dullness, 3 to hydration, and 2 to oiliness (Table 3).

Predictive performance was greatly improved by adding genetic information to the exposome. Table 4 lists the accuracy, precision, recall, and F1 score of the CatBoost algorithm based only on the exposome. In particular, the accuracy, precision, recall, and F1 score of the predictive model for melanin were improved by 17%, 25%, 20%, and 7%, respectively, compared with the cases without genetic features.

Table 4. Predictive Performance of the CatBoost Algorithm in Test Data

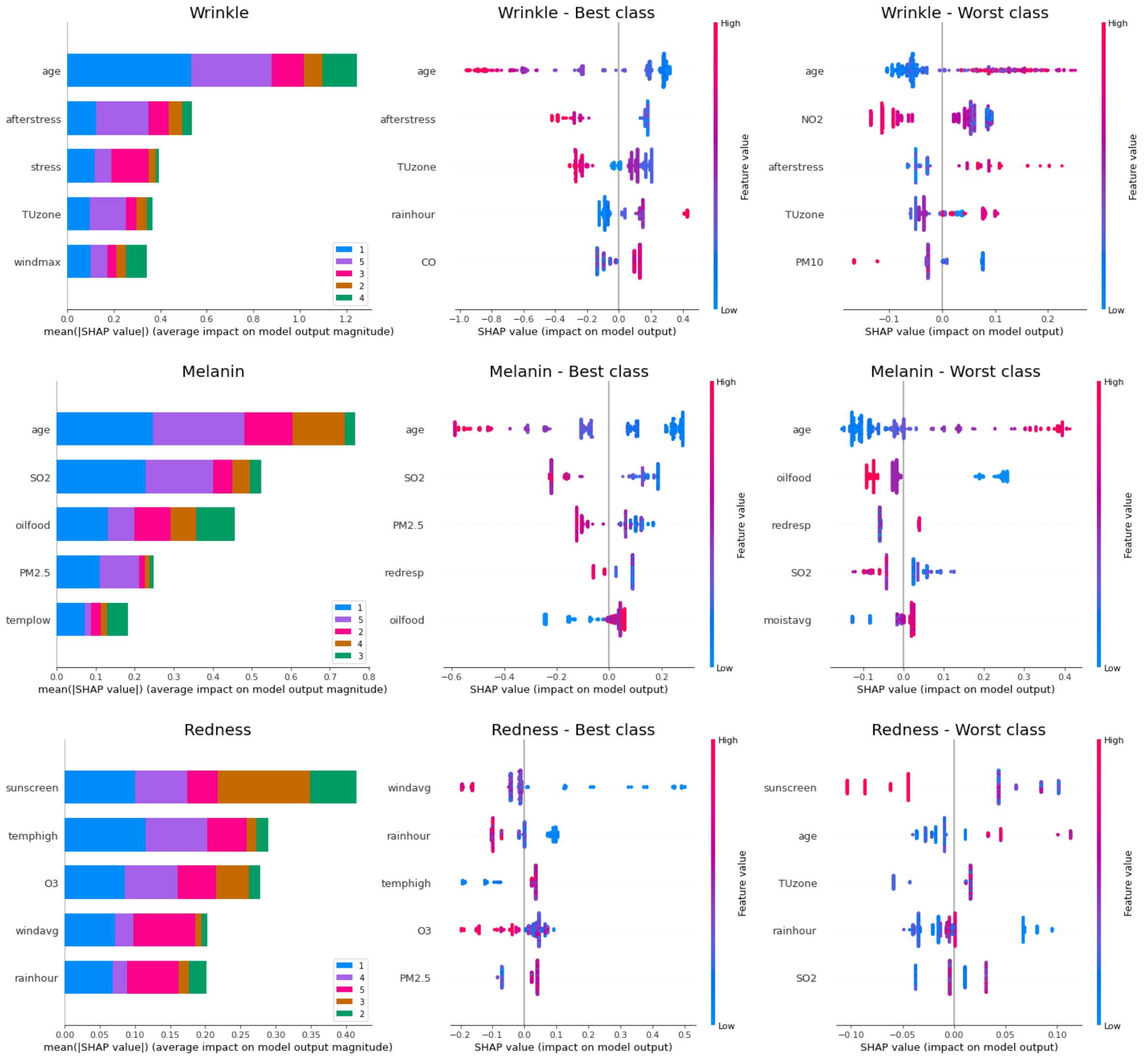
	wrinkle	melanin	redness	dullness	hydration	oiliness
Accuracy	0.46	0.54	0.37	0.39	0.41	0.33
Precision	0.43	0.60	0.38	0.46	0.38	0.23
Recall	0.44	0.55	0.39	0.45	0.44	0.25
F1 score	0.43	0.48	0.37	0.38	0.37	0.23

Figure 2. Flowchart for predicting six phenotypes



The main objective of the current study is to use big data and AI to anticipate future skin problems as the outcomes of various factors. The current study used genome and exposome data to investigate six phenotypes (wrinkles, melanin, redness, dullness, moisture, and oiliness). The XGBoost algorithm was used to extract important features from the exposome and gene data. Because each phenotype was divided into five groups (the worst, bad, normal, good, best conditions), the CatBoost algorithm was used to predict the participant's phenotypic classes based on the identification of relevant characteristics, and SMOTE was utilized to address the corresponding class imbalance issue (Figure 2).

Figure 3. SHAP values for each class of phenotype



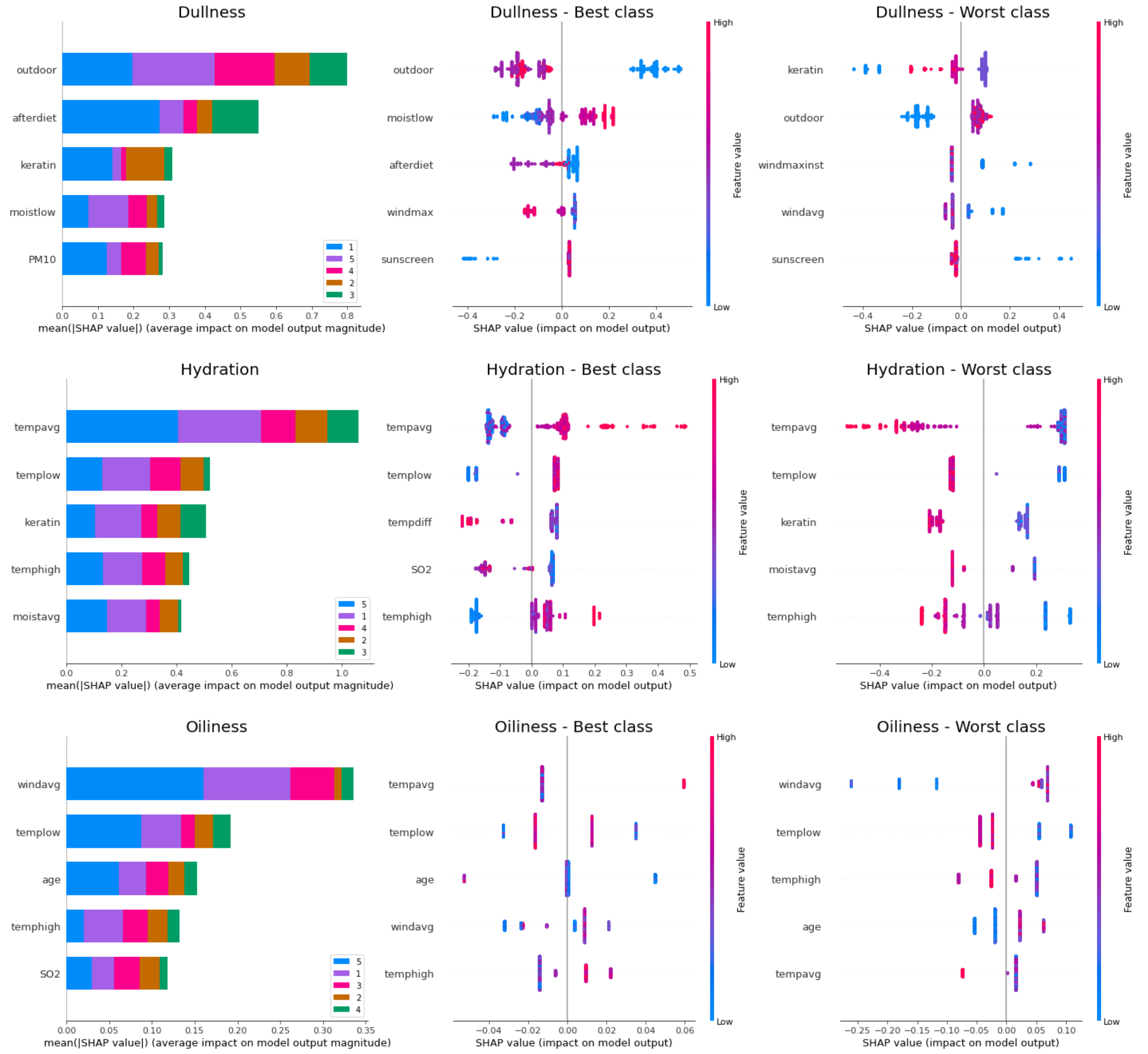


Figure 3 shows SHAP values (Lundberg and Lee, 2017) for each class of phenotype. In machine learning models, SHAP is used to understand the importance of features on model predictions. The left column of Figure 3 corresponds to a barplot of mean absolute SHAP values for features, showing the global importance of features on model predictions in a decreasing order. In a prediction model for wrinkle, top 3 features that have the greatest impact on the model predictions are age, skin deterioration after getting stress, and stress level. Furthermore, age, amount of sunscreen that subjects put on, outdoor activity level, average temperature, and average wind speed are the most influential features for melanin, redness, dullness, hydration, and oiliness, respectively.

The middle and right columns of Figure 3 correspond to beeswarm plots of SHAP when predicting the best and worst skin conditions, respectively, which show how each feature contributes on the model predictions. Each point in the figure is the SHAP value of the attribute and instance. The position of the Y-axis depends on the characteristics, and the position of the X-axis is determined by the SHAP value. The color represents the characteristic value, which means the higher the SHAP value goes to red. In a beeswarm plot, features are listed in the order of importance of impact on each of the best and worst skin conditions. For example, as a subject gets older, the likelihood that corresponding subject will be assigned to the class with the worst wrinkles increases. Conversely, it is less likely the subject will be allocated to the best class regarding wrinkles as age increases. As a result, the best wrinkle condition is positively associated with lower age and lower skin deterioration level after getting stressed while the worst wrinkle condition is positively associated with higher age and lower NO_2. Trouble level in T zone and U zone is negatively correlated with wrinkle, which means the subjects with much trouble is more likely to have many wrinkles. High values of age have negative contribution on melanin as they do on wrinkle. In the best class of melanin, environmental factors like SO_2 and PM2.5 have negative correlation with melanin. On the other hand, lifestyle related features such as skin redenss reaction level are more relevant than environmental features in the worst class. The more frequent skin redness response people have, the higher level of melanin they get.

In the case of redness, dullness, hydration, and oiliness, many other variables besides age are highly influential compared to wrinkle and melanin. Average wind speed and precipitation level are negatively correlated with the probability of belonging to the best class regarding redness, while the amount of sunscreen people put on is negatively associated with the worst class. Outdoor activity level strongly affects dullness, that is, the more outdoor activities people do, the darker their skin becomes. Hydration and Oiliness have particularly strong relationship with temperature. Temperature related features like daily average temperature and daily lowest temperature have positive impact on hydration as they become higher. Similar pattern occurs in oiliness while the association is not as evident as it is in hydration. Consequently, we can determine what a person needs to improve their skin condition by examining SHAP values from the prediction model. For instance, it is clear that a subject should put on more sunscreen to get out of the worst class of redness. In addition, as highly ranked features frequently fall under the categories of environmental and lifestyle-related variables, we can determine their significance.

Figure 4. Skin Prediction Schematic Diagram

Getting older (22 to 40) & Putting on less Sunscreen

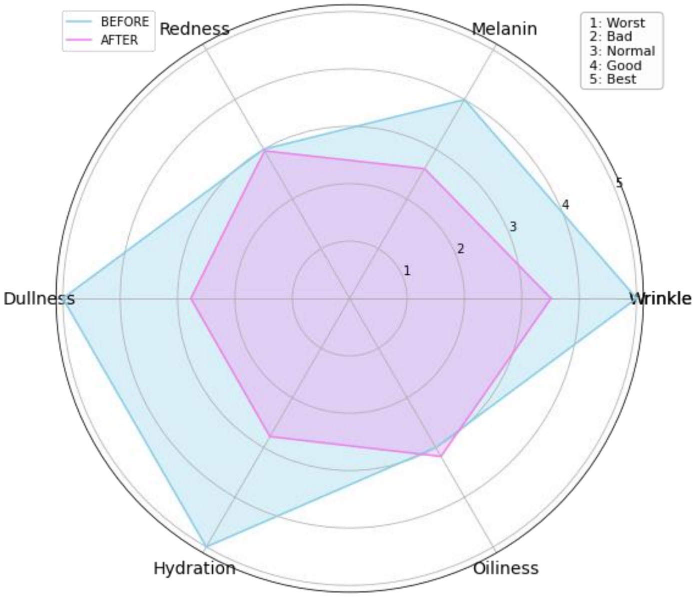


Figure 4 displays the skin prediction value based on Before (22 years old) and After (40 years old + putting on less screen) as a Schematic Diagram. It is clear that melanin moved from normal-good to bad, wrinkle changed from normal-bad to worst, hydration changed from normal-good to worst, and dullness adversely changed from normal-good to worst, despite the fact that there was no discernible change in redness or oiliness.

Discussion.

Customers use various skincare and functional cosmetics to prevent skin aging and improve the current condition of the skin. It is important to analyze and understand the skin to select optimized skin care solutions. Many cosmetics companies are introducing various customized services to provide tailor-made solutions for each customer's personal skin conditions and needs, because every individual is born with different skin type and lives in different environments. The types of customizations being attempted while selecting cosmetics vary from “on-site mixed type,” where the product is manufactured and provided based on skin diagnosis and counseling done in the field, to “device type,” where the product is provided based on data obtained using internet of things (IoT) technology. Recently, customized solutions have been attempted to reflect genetic characteristics through direct-to-consumer (DTC) genetic testing.

In this study, we identified factors for predicting skin changes through a correlative analysis of lifestyle, climate/environment, and innate genes that affect current skin conditions. The result obtained through hierarchical and simple correlation analyses is intended to be used as a variable of the correlation function in the future. For example, the lifestyle factor that affects wrinkles the most was found to be skin deterioration level after getting stressed, which means the participants whose skin deteriorates less after stressed out showed a significant difference in wrinkles depending on other lifestyle-related features such asthe amount of stress they got and trouble level on T/U zone.. As the influence of the variable may vary depending on the result of the lifestyle response, it was intended to be reflected in the skin prediction logic.

Next, a cross-analysis of the skin indicators and external environmental data was conducted. Consequently, the correlation between the environmental factors such as temperature, humidity, wind speed, amount of sunlight, fine dust, and air quality and the current skin condition was confirmed. However, this pattern changed slightly after customers started wearing masks owing to COVID-19. For example, we observed a significant increase in redness levels of the skin during the years 2019–2021 despite there being no change in the average temperature or ultrafine dust concentration compared with data from previous years. The reason could be attributed to the elevated temperature and humidity inside the mask and continuous physical stimulation.

Thus, the current skin condition is influenced by an individual’s lifestyle, changes in the climate, and environment. Genetic data does seem to act as an important variable in the diagnosis and prediction of skin changes. Therefore, among the genetic variables screened, those with correlations to lifestyle and age predicted the skin index. For example, a person who smoked and did not use sunscreen during outdoor activities presented a high concentration of melanin in addition to mutations of *MC1R*, *OCA2*, *AGER*, and *ASIP* genes; the gene mutations were found to have an effect on the skin condition.

Taken together, this knowledge will enable us to provide more proactive and personalized solutions in the field of cosmetics and in life care, such as lifestyle choices, eating habits, and environmental responses. We aim to enhance the diagnostic algorithm with further studies by collecting and analyzing data on various races and sexes.

Conclusion.

Current skin conditions are affected not only by genomic factors but also by the individual's lifestyle, climate changes, and the environment. Recently, various service models have been developed to provide updated solutions that reflect the data collected directly from the customers. In this study, we developed an algorithm for diagnosing and predicting the "skin type" to provide a hyper personalized solution based on various data. We identified factors for predicting changes in skin condition through correlation analysis of variables that affect the current skin condition such as lifestyle, climate, environment, and innate genes. Prediction algorithms that reflect these variables can then propose customized solutions based on lifestyle and genetic information.

Acknowledgments.

We would like to thank the IOPE Lab of AmorePacific Co. for collecting all the data used in the analysis. We also thank Theragen Bio for their support.

Conflict of Interest Statement. NONE.

References.

1. Guinot C, Malvy DJ-M, Ambroisine L, et al. (2002) Relative contribution of intrinsic vs extrinsic factors to skin aging as determined by a validated skin age score. Arch Dermatol. 138(11):1454–1460.
2. M A Farage et al. (2008) Intrinsic and extrinsic factors in skin ageing: a review. Int J Cosmet Sci. 30(2):87-95.
3. Kim M, Park T, Yun JI, Lim HW, Han NR, Lee ST. (2020) Investigation of age-related changes in the skin microbiota of Korean women. Microorganisms.

8(10):1581.

4. Jang SI, Lee M, Han J, et al. (2020) A study of skin characteristics with long-term sleep restriction in Korean women in their 40s. Skin Res Technol. 26(2):193–199.
5. Fussell JC, Kelly FJ. (2019) Oxidative contribution of air pollution to extrinsic skin ageing. Free Radic Biol Med. 151:111–122.
6. Yan Y, Chen H, Chen L, et al.(2020) Consensus of Chinese experts on protection of skin and mucous membrane barrier for health-care workers fighting against coronavirus disease 2019. Dermatol Ther. 33(4):e13310.
7. Park SR, Han J, Yeon YM, Kang NY, Kim E. (2021) Effect of face mask on skin characteristics changes during the COVID-19 pandemic. Skin Res Technol. 27(4):554–559.
8. Seok Young Kang (2021) Clinical manifestations and patch test results for facial dermatitis associated with disposable face mask use during the COVID-19 outbreak: A case-control study. J Am Acad Dermatol. 85(3):719-721.
9. Christopher Paul Wild, et al. (2005) Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 14(8):1847-50.
10. Jean Krutmann, M.D, et al. (2017) The skin aging exposome. Journal of Dermatological Science. 152–161
11. Manfei Zhang et al.(2017) A Genome-Wide Association Study of Basal Transepidermal Water Loss Finds that Variants at 9q34.3 Are Associated with Skin Barrier Function. J Invest Dermatol. 137(4):979-982.
12. Fan Liu et al. (2016) The MC1R Gene and Youthful Looks. Curr Biol. 26, 1213–1220.
13. Miranda A Farage et al. (2020) Genome-Wide Association Study Identifies Loci Associated with Sensitive Skin. Cosmetics, 7(2), 49
14. Jung Yeon Seo et al. (2021) GWAS Identifies Multiple Genetic Loci for Skin Color in Korean Women. Journal of Investigative Dermatology
15. Hye-Young Yoo et al. (2022) A Genome-Wide Association Study and Machine-Learning Algorithm Analysis on the Prediction of Facial Phenotypes by Genotypes in Korean Women. Clinical, Cosmetic and Investigational Dermatology. 15 433–445

16. Eunbi Ko et al. (2017) Genotype-Phenotype interation analysis of skin properties via genome-wide assiciation studies in 411 Korean females. International Federation of Societies of Cosmetic Chemists

17. A young Kim et al. (2019) Malonyl Co-A Decarboxylase (MLYCD), Korean Specific Genetic Marker Associated With Skin Aging, Regulates Melanogenesis. Journal of the Society of Cosmetic Scientists of Korea

18. Eunbi Ko et al. (2019) Genome-wide association study in Korean females identifies genetic variants associated with skin barrier. World congress of Dermatology

19. Prokhorenkova, L., Gusev, G., Vorobev, A., et al (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

20. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining:785-794.

21. Chawla, N. V., Bowyer, K. W., Hall, L. O., et al (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321-357.

22. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.