iFSCC 35TH
CONGRESS
FRANCE_CANNES 15-18 SEPT 25
THE FUTURE IS SCIENCE
ifscc2025.com
IFSCC
SOCIÉTÉ FRANÇAISE DE COSMÉTOLOGIE

IFSCC 2025 full paper (IFSCC2025-1120)

“Real World Makeup Study Exploration：Lip Fading Testing Methods and Analysis Based on Remote Image Collecting”

Hua Sun [1] , Linghan Liu [1], Yanwen Jiang [1*] and Gang Jin [1,]

[1] Shanghai China-norm Quality Technical Service Co., Ltd., Shanghai, China

1. Introduction

In recent years, consumer demand for makeup products has exhibited a dual deepening trend: shifting from basic decorative functions to the synergistic enhancement of precise efficacy and scenario adaptability. Consumers are not only concerned with the core parameters of products, such as long-lasting performance and anti-friction properties, but also demand consistent performance across various life scenarios. For instance, they require all-day long-lasting makeup in commuting scenarios.This evolving demand presents multidimensional challenges to existing cosmetic efficacy evaluation systems. Although current Chinese regulations impose no mandatory requirements for assessing the "beautifying and decorative" efficacy of color cosmetics, the deepening scientific literacy among consumers and intensifying market competition necessitate brand expansion into multidimensional efficacy evaluation methods [1]. These methods must provide robust evidence for product claims while addressing rapidly changing market demands and consumer expectations.

The current efficacy evaluation system for color cosmetics primarily relies on two methodologies: laboratory testing and human trial studies[2]. While laboratory evaluations achieve parameter control through in vitro models or expert evaluations, they still face challenges in standardization, potential standard drift in subjective human assessments, and inherent limitations. Firstly, laboratory environments struggle to simulate dynamic variables inherent in real-world usage scenarios, including temperature/humidity fluctuations, frictional forces, and variations in consumer behavior patterns, leading to discrepancies between evaluation results and actual user experiences. Secondly, the operational costs of long-cycle laboratory tests (e.g., wear resistance assessments) constrain sample sizes and data diversity, hindering the identification of differentiated needs across heterogeneous user groups. Most crucially, traditional laboratory methods focus on rapid product efficacy verification while neglecting correlation analysis between consumer behavior patterns and usage scenarios, thereby failing to provide insights beyond basic efficacy KPIs.

To address these limitations, this study introduces Real-World Study (RWS) methodology. Originating from medical research[3,4], RWS evaluates intervention effects in naturalistic settings through non-interventional observation, demonstrating core advantages in ecological validity, dynamic tracking capability, and population diversity. RWS has been increasingly introduced into the consumer goods field in recent years and has also been applied in the efficacy evaluation of cosmetics[5]. We believe that by mining the data on consumer product usage behaviors in real-life scenarios, it will be more helpful to reconstruct a multi-dimensional evaluation system for color cosmetics.It enables not only the characterization of performance attenuation patterns in real-world conditions but also reveals user experience drivers through behavior-formula correlation analysis. Notably, the synergistic development of smartphone imaging and artificial intelligence (AI) technologies provides technical feasibility for RWS implementation[6,7]. Mobile data acquisition transcends spatial-temporal constraints of laboratory settings, enabling large-scale, continuous efficacy monitoring while significantly reducing research costs.

This study focuses on lip color attenuation as the research entry point. Through deep learning techniques, we develop an automated analysis algorithm based on smartphone imaging and clinical-grade assessment results, aiming to explore the feasibility of remote cosmetic efficacy evaluation within RWS frameworks. This methodology may provide innovative solutions for makeup products efficacy assessment that balance cost-effectiveness, objectivity, and ecological validity.

2. Materials and Methods

2.1 Participants

A total of 157 Chinese female participants aged 22–55 years were recruited. All participants exhibited no apparent large-area facial imperfections (e.g., pigmentation, enlarged pores), scars, or raised textures (including acne-induced protrusions) that might interfere with assessments. Participants voluntarily joined the study and provided signed informed consent forms.

2.2 Image Acquisition Protocol

Participants underwent facial imaging after cleansing with water and a 0.5-hour equilibration period. Image acquisition was conducted in two phases. Phase 1: 107 participants were imaged using a VISIA-CR (Canfield) facial imaging system, yielding 320 valid images. Of these, 222 images were allocated to the training set and 98 to the test set. Phase 2: 50 participants were photographed using smartphone cameras, generating 300 valid mobile phone images (MBI) for remote image-based algorithm development.

Imaging timepoints were consistent across both phases:T0 (bare lips), Baseline (BL) immediate (15 minutes post-application), and T1-T6 (1–6 hours) post-application. Professional makeup artists standardized cosmetic application, while trained technicians executed data collection using standardized protocols. Image acquisition workflow illustrated in Figure 1.

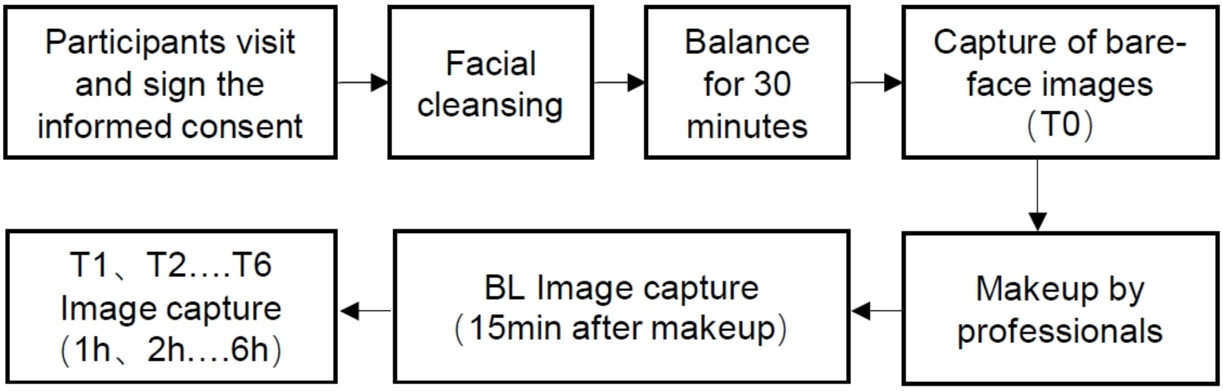
Participants visit and sign the informed consent
Facial cleansing
Balance for 30 minutes
Capture of bare-face images（T0）
Makeup by professionals
BL Image capture（15min after makeup）
T1、T2….T6 Image capture（1h、2h….6h）

Figure1. Image acquisition workflow

2.3 Clinical Grading

Clinical grading and subsequent algorithm development utilized identical stimulus materials to ensure evaluation fairness. A double-blind method was employed to score lipstick attenuation on a 0–9 scale (10 levels), with higher scores indicating greater fading severity. All images were randomized and assessed independently by two dermatologists, who referenced a standardized lip makeup grading atlas (Figure 2). Their scores were averaged for subsequent clinical evaluation data.

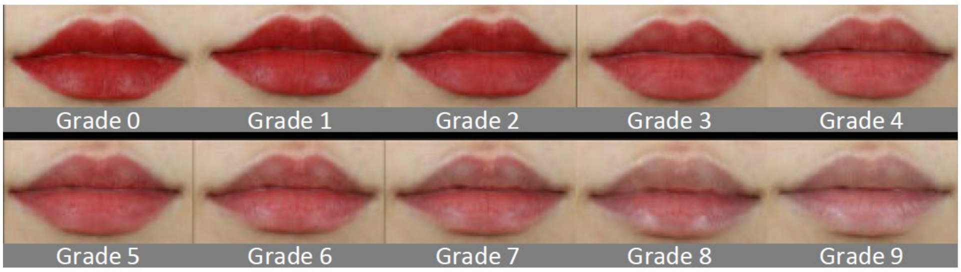

Figure 2. Atlas of Wearability (Fading/Peeling off) with 10-point Scale*

** The lip fading Atlas (developed in-house, not yet publicly released) has been validated through clinical applications and internal reliability assessments. Full methodological details will be published separately. Grade 0 indicates no fading/peeling off, grade 9 indicates severe fading/peeling off, no residue of lipstick on lips*

2.4 Image Processing and Algorithm Development

To establish a reliable analytical benchmark, a lip ROI (Region of Interest) extraction pipeline was first implemented: a face detection model was applied to preliminarily localize facial regions, followed by precise lip contour localization using Dlib's 68-point facial landmark detector. The ROI boundary was smoothed via B-spline curve approximation, and dynamic threshold segmentation was employed to isolate the lip region from surrounding skin, ensuring spatial consistency in feature extraction.

During the feature engineering phase, multidimensional feature vectors were extracted from preprocessed ROI images. Through recursive feature elimination (RFE) and XGBoost-based feature importance evaluation, three critical predictors were identified: 1) chroma shift (Δa) in the Lab color space's a-channel, reflecting the intensity of lipstick color variation over time; 2) percentage of bare skin exposure area, calculated by analyzing the dominant color value in bare lip images within a defined delta tolerance and its proportional coverage in post-application images; and 3) color difference ΔE between T0 and Timm timepoints, quantified using the CIEDE2000 standard formula. These features collectively constructed a multidimensional index system for characterizing lip makeup longevity.

The model architecture employed a gradient-boosted decision tree (GBDT) regression model within the XGBoost framework, leveraging technical advantages in regularization enhancement and computational efficiency. Specifically, L1/L2 regularization terms were introduced into the objective function to control overfitting, second-order Taylor expansion was utilized to approximate the loss function for accelerated convergence, and feature binning preprocessing enabled parallel computation acceleration. With decision trees as base learners and linear regression as the objective function, the model adopted mean absolute error (MAE) as the loss function, and hyperparameters were optimized via grid search.

The training process utilized a stratified 5-fold cross-validation strategy, where the dataset was randomly partitioned into five mutually exclusive subsets for iterative training and validation. Final metrics were averaged across five experiments to mitigate overfitting and ensure robustness on test sets. Model performance evaluation extended beyond conventional machine learning metrics, encompassing three dimensions: 1) MAE for absolute prediction error assessment; 2) intraclass correlation coefficient (ICC(2,1)) under a two-way random model to quantify consistency between machine predictions and expert ratings; 3) Shapley Additive ex-Planations (SHAP)-based feature contribution analysis to validate clinical rationality in decision-making. This validation framework effectively balanced model complexity and generalization capability, achieving high agreement with expert evaluations while maintaining low error margins, thereby delivering an interpretable machine learning solution for cosmetic efficacy assessment.

3. Results

3.1 Feature Extraction of Lip Makeup Images

To investigate the role of image features in model decisions and evaluate clinical interpretability, we quantified feature contributions through Recursive Feature Elimination (RFE) and XGBoost's intrinsic feature importance assessment. As shown in Figure 3, three key predictors were identified: percentage area of base color exposure in lip makeup degradation regions, chromatic difference in the a*-channel of L*a*b* color space, and Delta E color difference between target regions vs baseline. This finding provides an interpretable computational foundation for clinical parameter analysis.

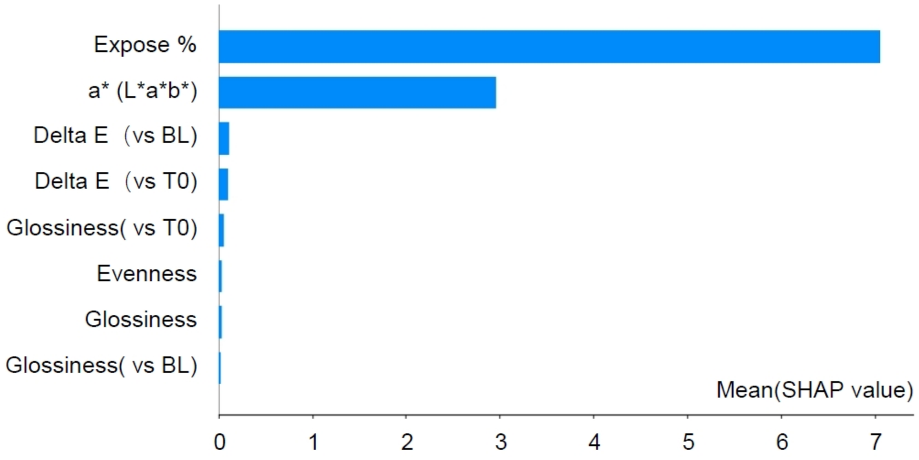

Figure 3. Contribution of image feature metrics to model predictions

In clinical scoring, the percentage of base color exposure area was assigned a higher weight. Furthermore, employing the SHAP interpretation method for individual-level analysis of feature contributions across test set samples, we found that the model exhibited heightened sensitivity to this metric when processing individuals with pronounced makeup degradation.

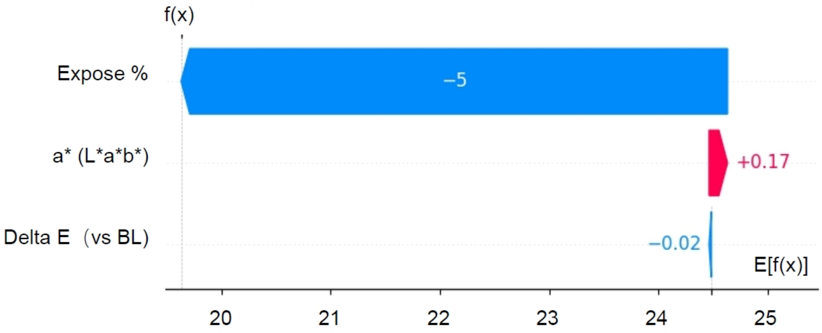

Figure 4. Per-Image Feature Contribution Analysis Across Predictive Metrics

3.2. Algorithm Establishment Based on Visca-CR Image Training and Test Sets

The algorithm performance on Visia-CR images is summarized in Table 1. The model achieved a MAE of 0.51 and an ICC of 0.93 with expert clinical assessments in the training set, while maintaining MAE=0.91 and ICC=0.77 in the test set.

Table 1. Algorithm prediction results for the training set and test set

Dataset	N	Image Metrics and Clinical Grading Data Analysis	
		ICC	MAE
Training set	222	0.93	0.51
Test set	98	0.77	0.91

3.3 Algorithm Analysis Establishment Based on Mobile Phone Images

To enhance cross-device compatibility, an enhanced algorithm optimized for mobile-captured images was developed based on the established framework. As shown in Table 2, the refined model demonstrated MAE=2.08 and ICC=0.78, indicating preserved measurement consistency across imaging devices。

Table 2. Algorithm prediction results for the mobile phone images

Dataset	N	Image Metrics and Clinical Grading Data Analysis	
		ICC	MAE
MBI validation	300	0.78	2.08

4. Discussion

This study has achieved objective quantitative assessment of lip makeup fading degree through machine learning models, with its core value lying in transforming clinical evaluation standards into reproducible technical metrics. Although the MAE values for the test set increased compared to the training set (MAE = 0.51), the model demonstrated clinical translation potential by attaining an ICC value of 0.77 in the test set, thereby validating the feasibility of constructing quantitative indicators based on image features [8].Notably, although the ICC in the MBI external validation set remains within the clinically acceptable range (0.78), the MAE increased to 2.08. This phenomenon may stem from gamut differences between the Visia system and smartphone cameras (Visia covers 98% Adobe RGB, while smartphone cameras average 92% sRGB coverage), leading to device-dependent biases in ΔE calculations [9,10]. These findings highlight the need to establish device-agnostic color calibration protocols in the future to improve algorithmic precision. This implies that implementing aesthetic detection on mobile devices will require stricter control over image acquisition and analysis, with additional technical challenges to address, such as light source standardization and cross-device imaging variations.

SHAP interpretability analysis through individualized feature contribution analysis of test set samples demonstrated that the model exhibited stronger dependency on the bare area percentage metric in cases with evident makeup fading. This pattern aligns closely with the mechanism described in previous multispectral imaging-based studies of longlasting foundation, wherein "makeup fading predominantly manifests as pigment shedding and base exposure" [11], thereby robustly validating the clinical rationality of the model's decision-making.

The mobilephone image analysis algorithm developed in this study effectively bridges the technical demands between laboratory precision evaluation and dynamic monitoring in real-world scenarios. By integrating multi-source heterogeneous information from VISIA professional instrument data, clinical evaluation standards, and smartphone images, we have constructed a "laboratory-consumer grade" dual-modal analytical framework. This technical approach constitutes a methodological complement to previous cosmetic evaluation studies based on multispectral imaging [12,13], where the former focuses on laboratory-level precision validation, while this research prioritizes resolving real-environment adaptability challenges. We believe this methodology will also generate broader innovative momentum for the industry, such as optimizing R&D workflows through objective quantification of image features, which could shorten testing cycles and reduce human errors compared to subjective scoring or clinical studies. Crucially, the versatility of mobilephone imaging algorithms pioneers new pathways for acquiring multidimensional consumer data, breaking through the limitations of previous clinical studies confined to high-precision fixed settings. Compared to past real-world cosmetic studies[14,15], this approach also provides a scientifically grounded, objective, convenient, and user-friendly evaluation methodology. Thereby advancing methodological applications in real-world contexts and enabling multidimensional dynamic consumer insights.

The limitations of this study primarily manifest in two dimensions: Firstly, the current training data scale fails to meet the requirements for modeling nonlinear interactions in high-dimensional feature spaces, potentially limiting the model's representational capacity for extreme makeup fading cases. Secondly, while algorithm development was based on standardized acquisition protocols under controlled environments (including device models, illumination conditions, and imaging parameters), significant variations exist in real-world user-captured images regarding illumination uniformity, focus accuracy, and shooting angles. This necessitates future technological breakthroughs in cross-device robustness enhancement and user-friendly image acquisition guidance system development.

5. Conclusion

This study achieved objective quantitative evaluation of lip makeup fading degree based on smartphone images through machine learning models. Experimental results demonstrated that the model attained an intraclass correlation coefficient (ICC) $\geqslant 0.7$ with clinical assessments in smartphone image verification, meeting consistency requirements for clinical testing. This marks a critical step forward in cosmetic efficacy evaluation transitioning from controlled laboratory environments to real-world consumer scenarios: The lip fading image analysis algorithm combined with the RWS method provides brands a scientific approach to deeply explore consumer needs and product feedback in real-world scenarios. This study not only lays the foundation for future exploration of more areas or differentiated efficacy in RWS, but also offers us the opportunity to construct larger-scale, long term tracking and more convincing efficacy evidence. By establishing an analytical framework that integrates professional evaluation systems with consumer-grade imaging data, this work overcomes the limitations of traditional research relying on single data sources, offering the industry a replicable technical paradigm for implementing RWS. The digital transformation in cosmetic and wellness sectors must be rooted in authentic scenarios - only when technical models become deeply embedded in users' actual usage environments can they unlock multidimensional data value. Building upon this, we anticipate this research will stimulate collaborative innovation with industry partners to advance the paradigm shift in cosmetic efficacy evaluation from "laboratory gold standards" to "real-world silver standards".

References

[1] Anon. , Practical Human Efficacy Evaluation of Color Cosmetics Products [J]. China Quality Certification, 2023(9):82-83.

[2] Ying W, Rui L, Zongqi H. Research on In Vitro Evaluation Methods for Long-Lasting Performance of Color Cosmetics Products [J]. Shanghai Light Industry, 2023(6):123-127.

[3] Qi Y , Zhao K , Chen N ,et al.Understanding the landscape and promoting the use of guides for real-world study: a scoping review[J].Journal of Clinical Epidemiology, 2024, 176.DOI:10.1016/j.jclinepi.2024.111551.

[4] Kim S , Min W K .Toward High-Quality Real-World Laboratory Data in the Era of Healthcare Big Data[J].Annals of Laboratory Medicine, 2025, 45(1).DOI:10.3343/alm.2024.0258.

[5] Suh D H , Kim T E , Lee S J ,et al.Enhanced Tolerability and Improved Outcomes in Acne Management: A Real-World Study of Dermocosmetic Adjunctive Therapy[J].Journal of Cosmetic Dermatology, 2025, 24(1).DOI:10.1111/jocd.16772.

[6] Ye D, Dengfeng Y, Zhihang L, et al. Efficacy evaluation of foundation make-up products by image analysis method [J]. Detergent & Cosmetics, 2019, 42(7):5.

[7] Rachel L. D, Chelsea E. K, Katarina R. K. Artificial Intelligence Predicts Fitzpatrick Skin Type, Pigmentation, Redness, and Wrinkle Severity From Color Photographs of the Face [J]. Journal of Cosmetic Dermatology, 2025, 24: e70050. https://doi.org/10.1111/jocd.70050

[8] Zineb N , Rachid B , Fatine E .Enhancing Cosmetic Supply Chain Efficiency Through Demand Forecasting Using Machine Learning[C]//International TRIZ Future Conference.Springer, Cham, 2025.DOI:10.1007/978-3-031-75923-9_13.

[9] Goldsberry A , Hanke C W , Hanke K E .VISIA system: A possible tool in the cosmetic practice[J].Journal of drugs in dermatology: JDD, 2014, 13(11):1312-1314.

[10] Bao C, Wenyu W, Ao F. et al, Relevant Evaluation and Comparative Study on Sunscreen and Sunburn Detection Effect Based on the Mirror Application of Huawei Mobile Phone [J]. China Illuminating Engineering Journal, 2023, 34(6):95-101.

[11] Nagaoka T , Kimura Y .Quantitative cosmetic evaluation of long-lasting foundation using multispectral imaging[J].Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI), 2019, 25(3):318-324.DOI:10.1111/srt.12651.

[12] Chengtong L, Hua Z, Min W. Efficacy evaluation of cosmetics (Ⅸ)——Application of image analysis method in the evaluation of cosmetic efficacy [J]. China Surfactant Detergent & Cosmetics, 2018, 048(010):551-557.DOI:10.13218/j.cnki.csdc.2018.10.002.

[13] Xiaomin Z, Yunshan Z, Xin Q. Application of quantitative image analysis on clinical efficacy evaluation of cosmetics [J]. Detergent & Cosmetics, 2016(1):5.

[14] Faad M G , Faad C F , Chilukuri S ,et al.Real-world experience with a treatment with a skincare regimen of products containing the Macrocystis pyrifera ferment for optimizing facial skin rejuvenation[J].Journal of Cosmetic Dermatology, 2024, 23(Sup2):11.DOI:10.1111/jocd.16420.】

[15] Fluhr J W ,Agnès Voisard, Nikolaeva D G ,et al.Stratum Corneum Hydration Measurements with a Bluetooth Wireless Probe: A Real-Life Study at Home Compared to Measurements under Laboratory Conditions[J].Skin Pharmacology and Physiology, 2024, 37(1-3):9.DOI:10.1159/000539411.