Received: 9 March 2023 | Accepted: 19 June 2023

DOI: 10.1111/srt.13402

ORIGINAL ARTICLE

WILEY

Exploring artificial intelligence from a clinical perspective: A comparison and application analysis of two facial age predictors trained on a large-scale Chinese cosmetic patient database

Meng M. Zhang | Wen J. Di | Tao Song | Ning B. Yin | Yong Q. Wang

Center for Cleft Lip and Palate Treatment, Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Correspondence
Yong Q. Wang, Center for Cleft Lip and Palate Treatment, Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Ba-da-chu, Beijing 100144, China.
Email: surgeonfranks@163.com

Funding information
Teaching Reform Project of Peking Union Medical College, Grant/Award Number: 2021YXX006; Teaching Reform Project of Peking Union Medical College, Grant/Award Number: 2023zlgl011; Beijing Municipal Science & Technology Commission, Grant/Award Number: Z221100007422039

Abstract

Background: Age prediction powered by artificial intelligence (AI) can be used as an objective technique to assess the cosmetic effect of rejuvenation surgery. Existing age-estimation models are trained on public datasets with the Caucasian race as the main reference, thus they are impractical for clinical application in Chinese patients.

Methods: To develop and select an age-estimation model appropriate for Chinese patients receiving rejuvenation treatment, we obtained a face database of 10 529 images from 1821 patients from the author's hospital and selected two representative age-estimation algorithms for the model training. The prediction accuracies and the interpretability of calculation logic of these two facial age predictors were compared and analyzed.

Results: The mean absolute error (MAE) of a traditional support vector machine-learning model was 10.185 years; the proportion of absolute error ≤6 years was 35.90% and 68.50% ≤12 years. The MAE of a deep-learning model based on the VGG-16 framework was 3.011 years; the proportion of absolute error ≤6 years was 90.20% and 100% ≤12 years. Compared with deep learning, traditional machine-learning models have clearer computational logic, which allows them to give clinicians more specific treatment recommendations.

Conclusion: Experimental results show that deep-learning exceeds traditional machine learning in the prediction of Chinese cosmetic patients' age. Although traditional machine learning model has better interpretability than deep-learning model, deep-learning is more accurate for clinical quantitative evaluation. Knowing the decision-making logic behind the accurate prediction of deep-learning is crucial for deeper clinical application, and requires further exploration.

KEYWORDS
age estimation, application analysis, artificial intelligence, interpretability, plastic surgery

1 | INTRODUCTION

With the aging of human society, aging and anti-aging have emerged as research hotspots in plastic surgery and aesthetic medicine.[1] Facial aging—a significant component of overall body aging—is a fundamental biological rule that has significant psychological and sociological implications.[2,3] The distinctive changes of facial aging are frequently utilized as a crucial reference when attempting to estimate an individual's age. In social encounters, young features frequently convey an impression of vitality, health, and attractiveness, whereas aging faces frequently convey a sense of fatigue and lack of energy. With the rapid development of plastic surgery and the diversification of methods, facial rejuvenation surgery and operations have become more widely accepted. Local signs of aging, such as wrinkles, eye bags, and soft-tissue sagging, can be alleviated through a series of operations. Post-surgery faces suggest a reduced biological age; that is, people appear younger and more energetic. At present, preoperative and postoperative evaluations of rejuvenation treatment are mainly based on subjective indicators and methods, and there are few objective and quantifiable indicators. In traditional clinical evaluation, surgeons make subjective assessments of postoperative anti-aging effects based on clinical experience or semi-quantitative photograph evaluation scales or by using patients' postoperative satisfaction to evaluate clinical effects.[4] The results are often poorly reproducible owing to reliance on individual physician experience and a lack of uniformity among individuals. With regard to objectivity and repeatability, instrumental evaluation is preferable to conventional clinical evaluation. It can quantitatively and objectively measure various skin indicators, including wrinkles, pigmentation, and pores. However, the measurement results frequently only pertain to a limited area and thus are of little value to plastic surgeons who are concerned with the overall impact of rejuvenation on the face. In comparison, using facial biological age as an evaluation index is better aligned with patients' original motivation for receiving rejuvenation treatment, which is to make themselves look younger generally rather than only locally. Additionally, facial biological age is a reliable quantitative and objective indicator that is useful for analyzing and comparing the beneficial effects of treatment among various people.[5]

In recent years, facial age assessment has been extensively researched by computer scientists in China and other countries. However, because the face datasets used for developing and training models, such as FG-NET, MORPH, CACD, and LAP, mainly come from European and American countries, the Asian population is poorly represented, and the data specifications and clarity are uneven,[6] making it impossible to guarantee the applicability of these models to the Chinese population. The two primary categories of age-estimation models currently in use are deep-learning models and traditional machine-learning models.[7] Traditional model algorithms are usually based on statistical rules and assumptions that are defined and understood by humans. They have strong interpretability; the decision-making basis of the machine can be understood and obtained, and the reliability of the decision-making can be evaluated. The deep-learning model represented by the deep neural network (DNN), which uses a black-box algorithm of multiple neural-network structures, often has a high prediction accuracy, at the cost of interpretability. The calculation and decision basis behind it cannot be precisely understood.[7] The model's predictive effect and interpretability are both crucial for clinical applications. The former contributes to accurate assessment, and the latter helps doctors better understand which part of the face should be focused on and improved. Research efforts are mainly concentrated in the field of computer science because the application of artificial intelligence (AI) is limited by the demands for algorithm development expertise and model training resources.[1] To date, no comparative analytical work has been published on these two types of age-prediction models from the standpoint of clinical applicability.

Two age-estimation algorithms that were representative and effective on open datasets were selected for this study. A conventional large-scale face database of Chinese patients from a plastic-surgery hospital was used to train the two models, which were then clinically validated and compared.

2 | METHODS

2.1 | Face dataset creation

The study was approved by the ethics committee of the Plastic Surgery Hospital affiliated with Peking Union Medical College and was in accordance with the principles outlined in the Declaration of Helsinki. Preoperative images of patients receiving cosmetic surgery at our Department of Plastic Surgery from January 2012 to December 2021 were collected. All patients provided consent for their images to be used in the study. Lighting conditions were uniform during the face image acquisition, and the patients were required to be in a standing position, without makeup on the face. None of the patients had a fever or stayed up late in the week before the photography. The entire face and neck were exposed during the shooting, and the facial muscles were in a relaxed state. High-definition photographs were taken of each patient in frontal, lateral, and oblique orientations. The exclusion criteria for patients were as follows: those with congenital or acquired deformities of the face and neck; those with scars after trauma or infection on the face and neck; those with serious diseases (such as severe hypertension, diabetes, hepatitis, kidney disease, autoimmune disease, cancer, and immunosuppression); those with moderate to severe acne on the face; those with underlying diseases or other skin diseases; those who had undergone cosmetic surgery or treatment; women who were pregnant (or planning to become pregnant); those whose images were unclear and did not satisfy the criteria. Other attributes matched to each patient's face image were collected from medical records, including gender, age, and cosmetic treatments desired. There were 1821 patients with ages ranging from 18 to 80 in the final face dataset who satisfied the inclusion and exclusion criteria. There were 10 529 face images in total. Table 1 presents the included patients' gender and age distributions. A Nikon D850 camera with a 24–70-mm f/2.8E ED

TABLE 1 Gender and age distributions of the patients before plastic surgery (unit: Number of patients) (Dataset 1).

	18–20	21–30	31–40	41–50	51–60	61–70	71–80	Total
Female	226	753	366	221	87	36	1	1690
Male	24	40	25	19	19	3	1	131
Total	250	793	391	240	106	39	2	1821

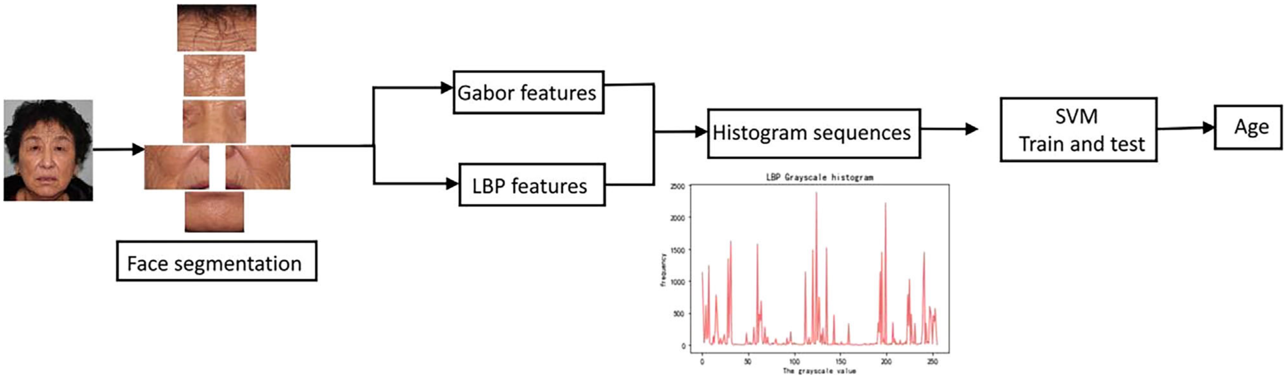

FIGURE 1 Flow diagram of support vector machine (SVM) age estimation.

VR lens, which had a resolution of 5184 × 3456 pixels, was used to capture the images.

2.2 | Facial age estimation based on traditional machine learning

Many academics in China and other countries conducted research on traditional machine learning for age estimates prior to the development of deep learning.[7] Among them, the Gabor wavelet combined with the local binary pattern (LBP) algorithm has attracted widespread attention because of its high accuracy and interpretability and has been applied to facial age estimation, quantitative wrinkle detection, and expression recognition.[8,9] The Gabor filter is a widely used descriptor based on local features. It is robust to changes in posture and lighting, can extract local texture features of images, and is useful for feature extraction of facial wrinkles—particularly fine wrinkles. As a description operator of image texture, the LBP can effectively extract the differences between each pixel and other points in its neighborhood as a global texture feature.[8] The theoretical basis for effective age estimation of this algorithm is to fuse and statistic local and global texture information to extract as comprehensive features as possible, including wrinkles, mottles, and sagging of soft tissue.

Figure 1 depicts the extraction of the face's texture using a Gabor wavelet and LBP histogram sequence. First, the Dlib library was used to detect the face and automatically mark 68 facial key points. According to the key points, the positioning and segmentation of the prominent areas of wrinkles were completed, and a total of six regional images of the forehead, the glabella, the nasion, both cheeks (nasolabial folds), and the chin were obtained. Subsequently, a Gabor filter with four scales (V = 4; 9 × 9, 11 × 11, 13 × 13, and 15 × 15 pixels) and four directions (U = 4; 0°, 45°, 90°, and 135°) was constructed to decompose the local face image and extract Gabor features of different scales and directions. The filtering results obtained for each image were 16 images, which were arranged in order of the Gabor angle and wavelength, called Gabor magnitude maps (GMMs). The texture features for different GMM orientations and scales were encoded using the eight-neighborhood LBP operator and stored as local Gabor binary pattern (LGBP) images.[8] The grayscale information in the LGBP image was converted into a histogram for description, and the histogram sequences of all sub-blocks of the LGBP for different face regions, directions, and scales were sequentially connected to obtain the texture feature vector of the face. Finally, the traditional support vector machine (SVM) algorithm was used to train the age classification model, that is, to establish the functional relationship between the texture feature vector and age. Owing to the SVM algorithm's suitability for classification problems and small sample sets,[8] 205 frontal images from the face dataset were arbitrarily selected according to age groups and included in the SVM model training. Details are presented in Table 2. Pre-experimental findings led to the age labels being changed to age groups with a 6-year age range (18–23, 24–29, 30–35, 36–41, 42–47, 48–53, 54–59, 60–65, 66–71, and 72–80). To train the model, 70% of the training set and 30% of the verification set were randomly selected. This process was repeated until the loss function stopped decreasing.

2.3 | Facial age estimation based on deep learning

In recent years, deep learning has been widely used in many computer-vision tasks, including face attribute recognition, and has achieved

TABLE 2 Age distribution of the frontal face dataset (unit: Number of images) (Dataset 2).

Age	18–23	24–29	30–35	36–41	42–47	48–53	54–59	60–65	66–71	72–80	Total
Total	20	22	24	24	24	24	24	23	13	7	205

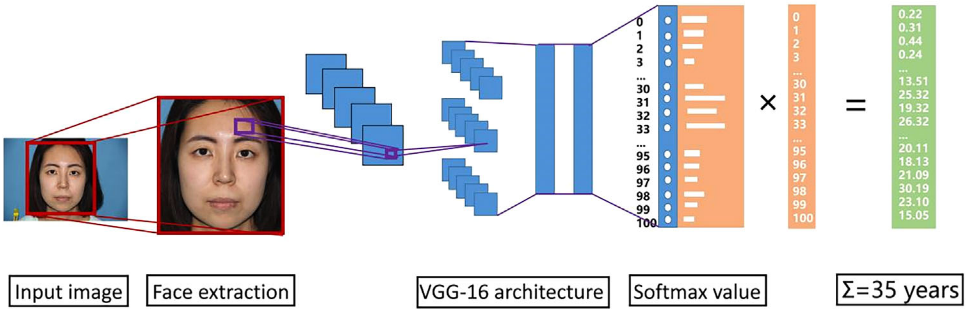

FIGURE 2 Flow diagram of VGG-16 age estimation.

remarkable success. Regarding the deep-learning model, we adopted the age–gender estimation algorithm. This algorithm is based on the VGG-16 architecture, which is well-known for its outstanding performance in the ImageNet challenge and is used for identifying age and gender from a face image.[6,10]

The age-estimation process of the algorithm is shown in Figure 2. First, the Dlib library was used to detect and located faces, extract facial regions, and ensure that faces were always in the same position and occupied the same proportions of the image. The resulting image was compressed to 224 × 224 pixels and used as input to a deep convolutional network. The normalized face image was input to the VGG-16 architecture for complex convolution operations, and the output was set to 101 values, corresponding to the numbers 0–100. The softmax function was used to convert these 101 values into a probability sequence. Each number (0–100) was multiplied by its corresponding probability value, and the products were summed to obtain the final prediction result, which was expressed in years. In view of the strong adaptability of the DNN to the shooting light, angle, and background information, we included 10529 face images in the face dataset, and the program randomly divided the images into a training set (80%, 8423 images) and a verification set (20%, 2106 images) for model training. The parameters were adaptively adjusted in the direction of the loss function decreasing until the loss function no longer decreased, at which point the final model was obtained.

2.4 | Comparison of two age-estimation models

To compare the age-estimation accuracies of the two models fairly, 92 frontal photos of patients were added, which were collected from January 2022 to June 2022. The inclusion and exclusion criteria were identical to those for the aforementioned face dataset, and it was ensured that the patients did not appear in this dataset. Details are presented in Table 3. The supplementary frontal face dataset was used as a test set to test the effects of the traditional machine-learning model and the deep-learning model, and comparative analyses were conducted.

3 | RESULTS

3.1 | Performance testing of traditional machine-learning model

As shown in Figure 3, when a 68-year-old woman's face image was supplied, the algorithm automatically segmented the image into six different parts: the forehead, the glabella, the nasion, both cheeks (nasolabial folds), and the chin. The GMM feature image was obtained by applying a Gabor filter to the original image, and it effectively displayed wrinkle features in different thicknesses and directions. The LBP operator encoded the texture features of different directions and scales of GMMs to obtain LGBP features. As shown in the figure, the LGBP more clearly reflected the start and end, depth, and direction of textures than the GMMs. Histogram sequences are used to describe LGBP images to obtain high-dimensional statistical features of facial textures.

The face texture data were trained and tested using the SVM—a traditional machine-learning technique—to create the age-estimation classification model. Figure 4 shows the test outcomes. The SVM confusion matrix is shown on the left, which is the specific classification result of 92 frontal face images in the test set. The vertical axis indicates the real age classification, and the horizontal axis indicates the predicted age classification. A darker color corresponds to a larger number of patients. A darker color of the squares on the diagonal from the upper left to the lower right corresponds to better performance of the SVM classifier, that is, more accurate

TABLE 3 Examination dataset containing 92 frontal face images (unit: Number of images) (Dataset 3).

Age	18–23	24–29	30–35	36–41	42–47	48–53	54–59	60–65	66–71	72–80	Total
Total	10	10	10	10	10	10	10	10	6	6	92

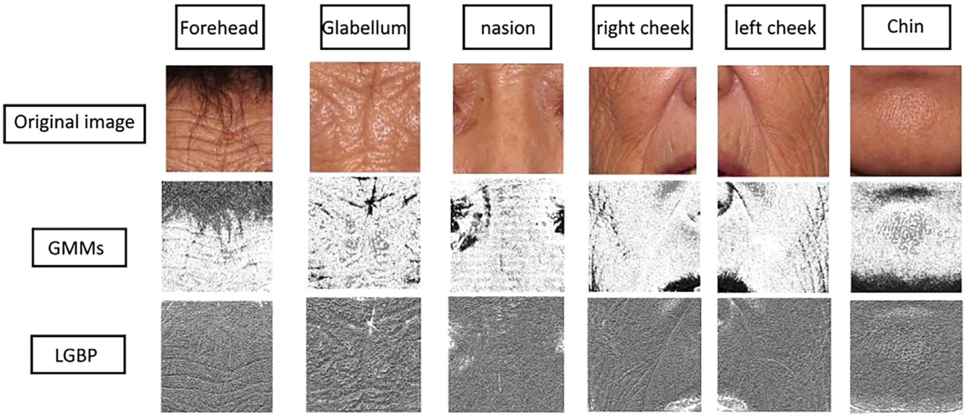

FIGURE 3 Gabor magnitude maps (GMMs) and local Gabor binary pattern (LGBP) features of 6 wrinkle-concentrated facial parts.

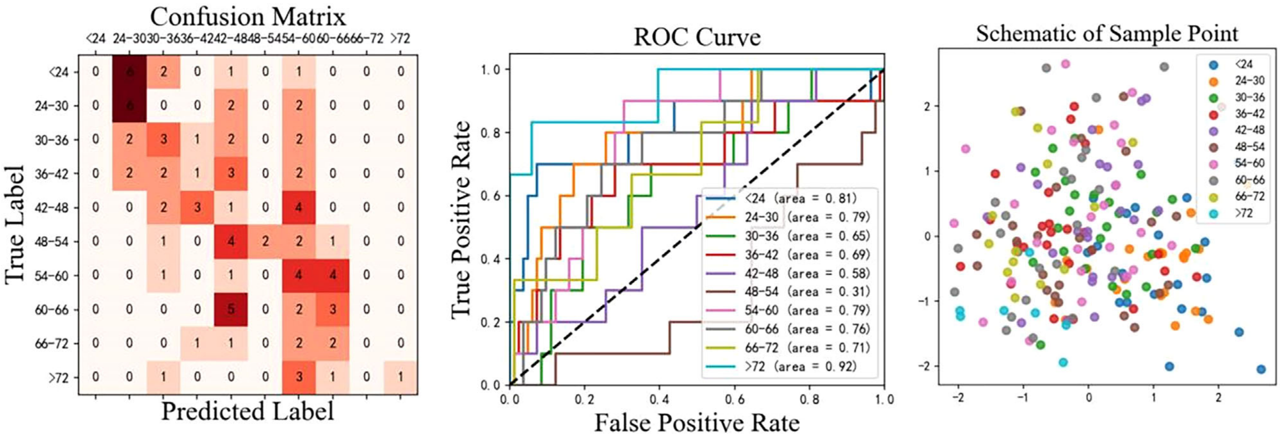

FIGURE 4 Test results of the traditional algorithm (SVM): Confusion matrix (left), ROC curve (middle), and schematic of sample point (right).

classification. The dark squares in the figure are largely distributed close to the diagonal, indicating that the classifier had a good predictive performance. The receiver operating characteristic (ROC) curve in the middle figure reflects the classification prediction effect of the classifier for different age groups. A larger area under the ROC curve (AUC) corresponds to a better classifier effect. The prediction effect for the different age groups was ranked as follows, from best to worst: > 72 years old (AUC = 0.92), < 24 years old (AUC = 0.81), 24–30 years old (AUC = 0.79), 54–60 years old (AUC = 0.79), 60–66 years old (AUC = 0.76), 66–72 years old (AUC = 0.71), 36–42 years old (AUC = 0.69), 30–36 years old (AUC = 0.65), 42–48 years old (AUC = 0.58), and 48–54 years old (AUC = 0.31). The results indicated that the prediction accuracies of the model for the elderly population (≥54 years old) and young population (< 30 years old) were better than that for the middle-aged population (30–54 years old). The right graph is a sample scatter diagram, which aggregates high-dimensional face features into two principal components and presents them on a plane to indicate the distribution of samples. Different colors in the figure represent individuals of different age groups. There was a clear boundary between individuals under 24 and those over 72. Samples in the middle age groups were scattered, with only a small number clustered together. Samples with similar ages were close to each other, and the overlap was significant. This illustrated that the texture features extracted by the model are obviously distinguishable for young and elderly groups, while those for the middle-aged lack good discrimination.

TABLE 4 Comparison between the two age-estimation algorithms.

Algorithm	VGG-16	SVM
Prediction type	Regression	Classification
MAE/year	3.011	10.185
SD/year	2.429	8.643
Minimum/year	0	1
25th percentile	1	4.25
Median	2.5	7
75th percentile	4.75	13
Maximum/year	10	40
Lower 95% CI of mean	2.508	8.395
Upper 95% CI of mean	3.514	11.97
Cumulative score (0)/accuracy	13.00%	0
Cumulative score (6)	90.20%	35.90%
Cumulative score (12)	100%	68.50%
Cumulative score (18)	100%	81.50%
Cumulative score (24)	100%	91.30%
Cumulative score (30)	100%	95.70%
Cumulative score (36)	100%	97.80%
Cumulative score (42)	100%	100%
Cumulative score (48)	100%	100%
p-Value	<0.0001	

3.2 | Performance testing and comparison of deep-learning model

The DNN model based on the VGG-16 algorithm framework was applied to the test set, and the results were compared with those of the SVM model, as shown in Table 4. The mean absolute error (MAE) was defined as the absolute error value between the predicted age and the true age. The MAE of the VGG-16 model was 3.011 ± 2.429 years, and the 95% confidence interval (CI) was 2.508–3.514 years. These results were significantly better than those of the SVM model (MAE of 10.185 ± 8.643 years and 95% CI of 8.395–11.97 years). The results of the two groups were subjected to a paired sample t-test using SPSS statistical analysis software. A p-value of < 0.0001 indicated a statistically significant difference. Comparisons of the quartile boxplots (left) and histograms (middle) of the two algorithms are shown in Figure 5. The cumulative score (CS, %) curve of the absolute error is on the right side of Figure 5, reflecting the age-estimation accuracies of the two algorithms. The point (x, y) on the curve indicates that the samples within the absolute error range of x accounted for y% of the population. As shown in Figure 5, the CS curve of VGG-16 was always higher than that of the SVM, and it reached a plateau (100%) at 12, indicating that the absolute error of the model for the test set was ≤ 12 years, whereas the absolute error of the SVM was within 24 years (91.30%). According to the experimental findings, VGG-16 significantly outperformed the SVM with regard to prediction stability and accuracy.

4 | DISCUSSION

Age is an important facial attribute. It is a key demographic and biological feature for human identification and plays an important role in social life.[7] The generation and update iterations of Automatic Age Estimation (AAE) have emerged in recent years as a result of the rapid development of the AI industry—particularly in response to the requirements for human–computer interaction, face aging simulation, and market information collection.[8] They are indispensable primarily because humans are incapable of estimating ages accurately. Experts in the field of machine learning have developed an AAE system that outperforms humans. They have successfully applied the findings of theoretical research to real-world applications, such as entertainment, security monitoring, and background data retrieval, and have achieved favorable results.[7]

With the continuous improvement of the reliability and accuracy of AI age estimation, some plastic surgeons are trying to apply this advanced method in clinical work to provide an objective reference for the assessment of facial aging before and after rejuvenation surgery.[5] However, clinicians frequently employ pre-trained models, as their understanding and application of algorithmic logic are constrained by their professional knowledge. These models are created by experts in computer vision, and the majority of them were developed using open-source Internet datasets, which are certainly not the optimal option for clinical applications. In addition, previous studies lacked analysis and interpretation of different algorithms from a clinical perspective, and most of them focused on specific algorithm descriptions and computer-vision effects and were obscure, which was unfavorable to plastic surgeons. To help plastic surgeons better understand and utilize facial age predictors, we trained two classical age predictors using a large-scale Chinese cosmetic patient face database and then analyzed and compared the algorithmic logic and results of the two predictors from a clinical perspective.

There are two important steps in training a successful age-estimation model: the selection of the face dataset and the determination of the algorithm logic. Numerous face datasets have been employed for AAE training, including Morph, FG-NET, AFAD, Lifespan, Faces, and the ICCV-2015 challenge dataset.[7] Taking the classical Morph dataset as an example, European and American populations account for the vast majority of the data, whereas Asian populations only account for 0.28%. A similar situation is observed for other datasets. The quality and specifications of the datasets cannot be guaranteed, owing to the various sources of these public datasets, including scans of individual photographs, face images from web pages, and movie clips. This affects the training of the model. Regarding the selection of AAE algorithms, they are mainly divided into traditional computer-vision algorithms and DNN architectures that have emerged in recent years. The former algorithms extract facial aging patterns through artificially designed operators and then feed them to traditional machine-learning algorithms to train an age classifier. However, deep learning relies on high-level semantic features extracted spontaneously by the learning network for recognition. Upon the emergence of deep learning, it quickly replaced traditional algorithms in the field

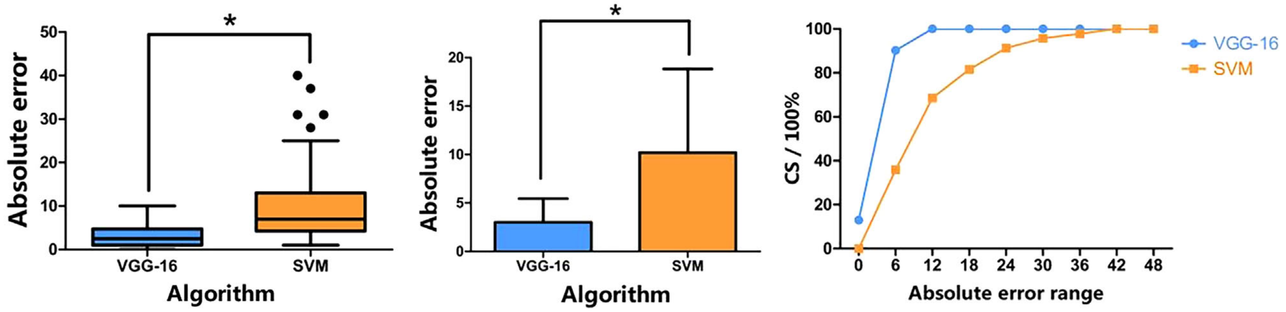

FIGURE 5 Comparison between two age estimation algorithms (VGG-16 vs SVM). Box-plot of absolute error(left), a bar graph of absolute error(middle), and cumulative score(CS) curve of absolute error(right).

of age estimation because of its excellent predictive performance. In contrast to entertainment and business applications, where prediction accuracy is the key concern, in clinical applications, it is also crucial to consider whether the machine-learning principles that underlie decision-making can be comprehended by humans. Such interpretability can help doctors better understand the aging characteristics of the face and can provide a treatment basis or direction beyond the traditional human experience.

In this study, as a traditional machine-learning algorithm, an SVM age classifier based on the Gabor wavelet and the LBP operator was used for training. The high-dimensional features of the face extracted during the process are shown in Figure 3. Different facial regions reflect different texture features, such as forehead lines, glabellar lines, nasolabial folds, and chin folds. Thus, the AI can show clinicians the basis of the age-estimating algorithm's decisions, helping them better understand the machine's thinking. This can help doctors to manually intervene and correct the algorithm and can suggest unknown individualized aging patterns to doctors. Zhou et al. used a similar method for the quantitative detection of wrinkles in face images and developed a face wrinkle evaluation model that breaks through the limitation of doctors' personal experience on the clinical rating of wrinkles and achieved good results.[9] As a classic DNN model, the deep learning VGG-16 algorithm creates a model with high prediction accuracy through a process of pure machine adaptation and ongoing parameter modification. The model's interpretability is significantly lower than that of traditional machine-learning algorithms because the learning process and operating mechanism of the automatically constructed model remain unknown.

Regarding the predictive performance, our results indicated that the deep-learning model VGG-16 (MAE = 3.011 ± 2.429 years old, 95%CI: 2.508–3.514 years) significantly outperformed the traditional SVM model (MAE = 10.185 ± 8.643 years, 95%CI: 8.395–11.97 years) for the MAE index, 95% CI, and CS. The results confirm the superiority of the deep-learning model for large-scale Chinese cosmetic patient data for the first time. Thus, we developed an accurate facial age-prediction model suitable for the Chinese population. Figure 6 shows the operation interface and examples for the VGG-16 age predictor. Regarding the SVM algorithm with slightly worse performance, we found that the algorithm had a higher prediction accuracy for young people (< 24 years old: AUC 0.81, 24–30 years old: AUC 0.79) and old people (> 72 years old: AUC 0.92, 66–72 years old: AUC 0.71, 60–66 years old: AUC 0.76, 54–60 years old: AUC 0.79) than for middle-aged people (30–36 years old: AUC 0.65, 36–42 years old: AUC 0.69, 42–48 years old: AUC 0.58, 48–54 years old: AUC 0.31). This may be because the texture features of young and old people are more significant than those of middle-aged people and have a better degree of discrimination; that is, young people have significantly less texture features, whereas old people have significantly more. In order to accurately predict the age of middle-aged individuals, it is not enough to rely solely on single facial texture features. More specific and subtle facial features must also be taken into consideration. This is a significant factor contributing to the superior performance of deep learning models compared to traditional machine learning models.

From a clinical perspective for Chinese individuals, facial aging primarily manifests in the changes of five clinical signs: increased wrinkles, skin tissue sagging, increased pigmentation, vascular disorders, and cheek skin pores.[11,12] Studies have shown that wrinkles/texture and ptosis/sagging criteria play a dominant role in the process of assessing one's perceived age of Chinese faces, accounting for 91% and 82% respectively in males and females. Meanwhile, the proportion of pigmentation disorders is similar in males and females, accounting for 9%. However, the weights of cheek skin pores and vascular disorders prove to be insignificant.[3,13–18] Therefore, although there is some correlation between pigmentation and perceived age among Chinese individuals, it is not as significant as texture and sagging in terms of appearance. In the field of traditional algorithms AAE, previous research has confirmed that facial pigmentation, large pores, and vascular disorders cannot act as reliable age descriptors when compared to wrinkles/texture and ptosis/sagging.[7] With the help of specific feature extraction algorithms, traditional age learning models can effectively extract a set of features to represent the age patterns of faces. These extraction algorithms mainly include Local Binary Patterns (LBP),[19] Biologically Inspired Features (BIF),[20] and Active Appearance Models.[21] In this study, Gabor filters, as a mature and efficient aging feature extraction algorithm, are used to extract local facial texture features in different regions and directions. Furthermore, the distribution shape of the LBP histogram well reflects the aging regularity of skin texture.[22] Past research has shown that the combination of these two algorithms can help better extract facial

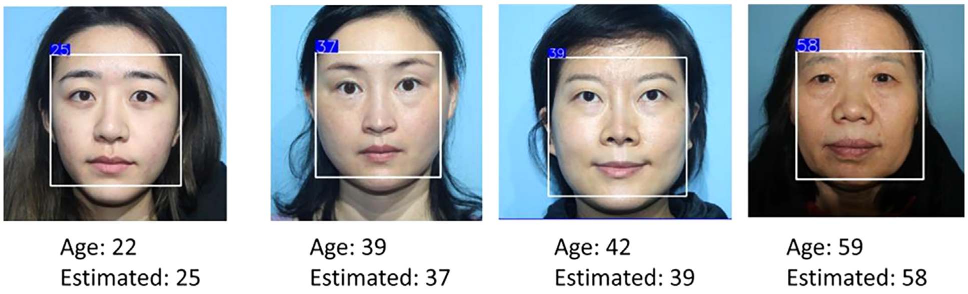

FIGURE 6 Age estimation examples via VGG-16 algorithm.

texture information, including local features, that is, wrinkles/texture, and global features, which are considered to reflect ptosis/sagging to a certain extent.[23,24] In summary, traditional machine learning AAE models have strong interpretability but are limited in available aging feature extraction algorithms. As a result, they are only accurate in extracting facial texture features. There is currently no good solution for accurately extracting facial pigmentation, vascular disorders, and large pores.

In order to more effectively and comprehensively establish the interpretation relationship between aging features and perceived age, there are two solutions. One is to design a more comprehensive aging pattern extraction operator and add the aging patterns confirmed by current clinical research to the existing operator, such as increased pigmentation, vascular disorders, and large pores.[4] Another approach is to employ deep learning AAE models, which can capture high-dimensional and abstract facial aging features, and explore methods to improve model interpretability. This involves making a connection between the facial digital features extracted by computers and clinical aging signs. Although several explainable deep learning techniques, such as backpropagation and attention mechanisms, have been proposed by computer-science scholars in recent years, they have not yielded satisfactory results. Therefore, reliable and effective model explanation methods remain a topic of exploration for relevant scholars.[25]

Similar to most age predictors, the two models we constructed were based on our established large-scale preoperative facial database with real age labels. Through training and calibration of the model, we predicted a relatively stable and accurate real age, which we call AI-perceived age.[1] However, there is currently insufficient evidence to prove that AI-perceived age is consistent with the apparent age given by human observers. After a patient undergoes some specific changes due to cosmetic procedures, the perceived age given by the age predictor is only related to the specific feature capture design and sensitivity of the algorithm. This may be consistent with the impression of the human on its age or may be inconsistent.[5] To our knowledge, unlike the relatively mature state of research for predicting the biological real age of a face, the research for predicting apparent age is still in its infancy. Building an apparent age database requires the collection of subjective opinions from a number of observers (from different genders, age groups, and social backgrounds) and data integration. Undeniably, Apparent age has great research value in the field of plastic surgery, both for patients and doctors, as it may better reflect social opinion before and after rejuvenation treatment. However, the requirement of a comprehensive and multi-dimensional collection of observers' opinions presents considerable challenges to the establishment of the apparent age database. Compared to real-age databases, the number and scale of apparent age databases are small, such as the LAP database provided by The organizers of ChaLearn Looking At People in 2015 and the APPA-REAL database—the first database containing both apparent age and real-age labels.[26,27] Furthermore, these databases are predominantly composed of Caucasians and have limited applicability to Chinese.[10,28] To address the above limitations, specialists in computer vision have developed transfer algorithms that make it possible to train an apparent-age predictor using a small apparent-age database on the foundation of a well-trained real-age predictor.[10] Above all, real age estimation by computer networks is not only an objective and non-dependent quantitative evaluation tool but also an algorithm basis for more in-depth research such as apparent age estimation[26]. Indeed, its acceptance, reliable results, and application verification in the field of plastic surgery as the concept of AI-perceived age have already been widely acknowledged.[1,5,29]

The limitations of this study are as follows. Owing to the uneven distribution of patients in the face database used for the model training, there were few males, and the age distribution was severely imbalanced. People under the age of 40 account for the majority of beauty seekers (79%). This situation is a reflection of the demographic characteristics of Chinese cosmetic patients, who consist mainly of women and young and middle-aged people. Because of traditional beliefs, the majority of males and the elderly do not accept cosmetic surgery. It is necessary to confirm whether the model's performance can be improved by using a gender- and age-balanced face dataset for training.

Although deep learning is a powerful technique, there have not been significant advancements in its interpretability, which has been a barrier to the continued growth of this field. AI should provide not only accurate prediction results but also insights into the algorithm's logic and thought process. We expect that in the future, more doctors will focus on algorithms' underlying logic and interpretability, closely collaborate with experts in AI, and employ advanced AI technology to alter current clinical evaluation systems. This will pave the way for the utilization of Big Data in plastic surgery, which is a significant challenge for plastic surgeons in the modern intelligent age.

ACKNOWLEDGMENTS

The authors thank Editage (www.editage.cn) for English language editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

FUNDING INFORMATION

This study is supported by the Teaching Reform Project of Peking Union Medical College (2021YXX006); Teaching Reform Project of Peking Union Medical College (2023zlgl011) and Beijing Municipal Science Technology Commission No Z221100007422039.

DATA AVAILABILITY STATEMENT

In order to comply with the relevant data privacy regulations, the participant information and photographic images captured in this study are processed for AI-based scoring and collected on dedicated, secure, and confidential servers.

ETHICS STATEMENT

The study was approved by the ethics committee of the Plastic Surgery Hospital affiliated with Peking Union Medical College and was in accordance with the principles outlined in the Declaration of Helsinki. All patients provided informed written consent for their images to be used in the study.

REFERENCES

1. Khetpal S, Peck C, Parsaei Y, et al. Perceived age and attractiveness using facial recognition software in rhinoplasty patients: A proof-of-concept study. *J Craniofac Surg*. 2022;33(5):1540-1544.
2. Hess U, Adams RJ, Simard A, Stevenson MT, Kleck RE. Smiling and sad wrinkles: Age-related changes in the face and the perception of emotions and intentions. *J Exp Soc Psychol*. 2012;48(6):1377-1380.
3. Flament F, Abric A, Amar D, et al. Changes in facial signs due to age and their respective weights on the perception of age, on a tired-look or a healthy glow among differently aged Chinese men. *Int J Cosmet Sci*. 2020;42(5):452-461.
4. Qiu HX, Long XH, Alice L, Roland B, Frederic F. Introduction to clinical tools for evaluating skin aging – Skin Aging Atlas (Asian). *J Clin Dermatol*. 2010;39(07):467-468.
5. Dorfman R, Chang I, Saadat S, Roostaeian J. Making the subjective objective: Machine learning and rhinoplasty. *Aesthet Surg J*. 2020;40(5):493-498.
6. Rothe R, Timofte R, Van Gool L. Deep expectation of real and apparent age from a single image without facial landmarks. *Int J Comput Vis*. 2018;126(2-4):144-157.
7. Othmani A, Taleb AR, Abdelkawy H, Hadid A. Age estimation from faces using deep learning: A comparative analysis. *Comput Vis Image Underst*. 2020;196:102961.
8. Huang B, Guo JC. Age estimation of facial images based on Gabor Wavelet and histogram sequence of LBP. *J Data Acquisit Proc*. 2012;27(03):340-345.
9. Zhou DL, Shuo Z. Detection of facial wrinkle based on improved maximum curvature points in image profiles. *Proceedings of the 3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)*. 2019:843-848.
10. Rothe R, Timofte R, Van Gool L. DEX: Deep EXpectation of apparent age from a single image. *IEEE*. 2015:252-257.
11. Bazin R, Flament F. *Skin Aging Atlas. Volume 2, Asian Type*. Editions Med'Com; 2010.
12. Flament F, Bazin R, Qiu H. *Skin Aging Atlas. Volume 5, Photo-aging Face & Body*. Editions Med'Com; 2017.
13. Flament F, Bourokba N, Nouveau S, Li J, Charbonneau A. A severe chronic outdoor urban pollution alters some facial aging signs in Chinese women. A tale of two cities. *Int J Cosmet Sci*. 2018;40(5):467-481.
14. Flament F, Ye C, Amar D. Assessing the impact of an aerial chronic urban pollution (UP) on some facial signs of differently-aged Chinese men. *Int J Cosmet Sci*. 2019;41(5):450-461.
15. Flament F, Amar D, Feltin C, Bazin R. Evaluating age-related changes of some facial signs among men of four different ethnic groups. *Int J Cosmet Sci*. 2018;40(5):502-515.
16. 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.
17. Flament F, Abric A, Amar D. Gender-related differences in the facial aging of Chinese subjects and their relations with perceived ages. *Skin Res Technol*. 2020;26(6):905-913.
18. Flament F, Bazin R, Qiu H, et al. Solar exposure(s) and facial clinical signs of aging in Chinese women: Impacts upon age perception. *Clin Cosm Investig Dermatol*. 2015;8:75.
19. Gunay ANVV. Automatic age classification with LBP. *IEEE International Symposium on Computer and Information Sciences*. 2008:1–4.
20. Han HOCJ. Age estimation from face images: Human vs. machine performance. *IEEE Int Conf Biometr*. 2013:1–8.
21. Lanitis A, Taylor CJ, Cootes TF. Toward automatic simulation of aging effects on face images. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(4):442-455.
22. Choi SE, Lee YJ, Lee SJ, Park KR, Kim J. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognit*. 2011;44(6):1262-1281.
23. Fu Y, Guo GD, Thomas S. Age Synthesis and Estimation via Faces: A survey. *IEEE Trans Pattern Anal Mach Intell*. 2010;32(11):1955-1976.
24. Markos G, Yannis P, Maja P. Modelling of facial aging and kinship: A survey. *Image Vision Comp*. 2018:1-69.
25. Ito H, Nakamura Y, Takanari K, et al. Development of a novel scar screening system with machine learning. *Plast Reconstr Surg*. 2022;150(2):465e-472e.
26. Agustsson E, Timofte R, Escalera S, et al. Apparent and real age estimation in still images with deep residual regressors on Appa-Real database. *IEEE 12th International Conference on Automatic Face and Gesture Recognition*. 2017:87-94.
27. Patcas R, Bernini DAJ, Volokitin A, et al. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. *Int J Oral Maxillofac Surg*. 2019;48(1):77-83.
28. Escalera S, Fabian J, Pardo P, et al. ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results. *IEEE International Conference on Computer Vision Workshops*. 2015: 243-251.
29. Chen SX, Zhang CJ, Dong M, et al. Using ranking-CNN for age estimation. *CVPR*. 2017:5183-5192.

How to cite this article: Zhang MM, Di WJ, Song T, Yin NB, Wang YQ. Exploring artificial intelligence from a clinical perspective: A comparison and application analysis of two facial age predictors trained on a large-scale Chinese cosmetic patient database. *Skin Res Technol*. 2023;29:e13402. https://doi.org/10.1111/srt.13402