Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis

Article *in* IEEE Transactions on Cybernetics · September 2021

DOI: 10.1109/TCYB.2021.3069920

5 authors, including:

Wang Sutong
Sichuan University
18 PUBLICATIONS **271** CITATIONS
SEE PROFILE

Yunqiang Yin
University of Electronic Science and Technology of China
230 PUBLICATIONS **4,775** CITATIONS
SEE PROFILE

Dujuan Wang
Sichuan University
94 PUBLICATIONS **2,117** CITATIONS
SEE PROFILE

Yanzhang Wang
Dalian University of Technology
87 PUBLICATIONS **1,262** CITATIONS
SEE PROFILE

Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis

Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin, *Fellow, IEEE*

***Abstract*—Skin lesion diagnosis is a key step for skin cancer screening, which requires high accuracy and interpretability. Though many computer-aided methods, especially deep learning methods, have made remarkable achievements in skin lesion diagnosis, their generalization and interpretability are still a challenge. To solve this issue, we propose an interpretability-based multimodal convolutional neural network (IM-CNN), which is a multiclass classification model with skin lesion images and metadata of patients as input for skin lesion diagnosis. The structure of IM-CNN consists of three main paths to deal with metadata, features extracted from segmented skin lesion with domain knowledge, and skin lesion images, respectively. We add interpretable visual modules to provide explanations for both images and metadata. In addition to area under the ROC curve (AUC), sensitivity, and specificity, we introduce a new indicator, an AUC curve with a sensitivity larger than 80% (AUC_SEN_80) for performance evaluation. Extensive experimental studies are conducted on the popular HAM10000 dataset, and the results indicate that the proposed model has overwhelming advantages compared with popular deep learning models, such as DenseNet, ResNet, and other state-of-the-art models for melanoma diagnosis. The proposed multimodal model also achieves on average 72% and 21% improvement in terms of sensitivity and AUC_SEN_80, respectively, compared with the single-modal model. The visual explanations can also help gain trust from dermatologists and realize man–machine collaborations, effectively reducing the limitation of black-box models in supporting medical decision making.**

***Index Terms*—Deep learning, interpretability, multimodal convolutional neural network, skin lesion diagnosis.**

I. Introduction

SKIN cancer is a major public health problem with 104 350 estimated new cases and 11 650 estimated deaths in the U.S. in 2019, which belongs to five leading cancer types for the estimated new cancer cases [1]. It costs the national health system around 8 billion dollars for this disease, which continues to rise each year [2]. The early diagnosis of skin lesion provides more treatment time to improve the survival rate and reduce the negative impact.

In order to curb rising skin cancer cases, medical experts need to race against time to identify the skin lesion types of patients as early as possible. The data show that patients possess more than a 99% chance of surviving within five years if they are treated properly at the early stage [3]. Once the cancer spreads to a patient's lymphatic system, the chance of 5-year survival drops to less than 15%. Thus, the early diagnosis of skin cancer plays an essential role, which can be detected by professional dermatologists through dermoscopy. Dermoscopy is an imaging technique that eliminates the reflections of skin surface. It visualizes the deeper levels of skin by removing surface reflection and provides opportunity for the diagnosis of various skin lesions with unaided eyes. However, the accurate diagnosis of skin lesion with dermoscopy requires rich experience and training, which takes years of time and plenty of money to train and cultivate an outstanding dermatologist [4]. With the increase of the world's population, the shortage of medical resources and the serious uneven distribution of medical resources have become a very prominent social contradiction, which aggravates the shortage of medical resources in underdeveloped countries. Besides, misdiagnosis is common in diagnosis due to human mistakes.

Computer-aided methods have attracted increasing research interest and have provided opportunities to solve the abovementioned problems. At the early stage, some simple computeaided diagnosis (CAD) methods, such as edge detection and line fitting, are used to assist diagnosis. With the development of machine learning and pattern recognition, skin lesion classifiers based on features extracted by humans are used. It helps to identify the skin lesion but is highly time consuming to collect features and hard to achieve satisfactory performance. Machine learning, especially deep learning, has become popular in recent years as it can automatically extract features from images and achieve great success [5]–[8]. Through the end-to-end artificial intelligence technology, the efficiency of skin lesion diagnosis has been greatly improved and the

Manuscript received September 29, 2020; revised January 21, 2021; accepted March 25, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 71871148, Grant 71971041, and Grant 71533001; in part by the Outstanding Young Scientific and Technological Talents Foundation of Sichuan Province under Grant 2020JDJQ0035; in part by the Major Program of National Social Science Foundation of China under Grant 20&ZD084; and in part by National Key Research and Development Program of China under Grant 2019YFB1404702. This article was recommended by Associate Editor J. Han. *(Corresponding authors: Dujuan Wang; Yaochu Jin.)*

Sutong Wang and Yanzhang Wang are with the Institute of Information and Decision Technology, Dalian University of Technology, Dalian 116000, China (e-mail: wangsutong@mail.dlut.edu.cn; yzwang@dlut.edu.cn).

Yunqiang Yin is with the School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: yinyunqiang@126.com).

Dujuan Wang is with the Business School, Sichuan University, Chengdu 610064, China (e-mail: djwang@scu.edu.cn).

Yaochu Jin is with the Department of Computer Science, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: yaochu.jin@surrey.ac.uk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2021.3069920.

Digital Object Identifier 10.1109/TCYB.2021.3069920

cost has been considerably reduced. With the increasing popularity of mobile phones and the enhancement of their pictures, AI-based skin lesion diagnosis applications show great promises, allowing people to conduct low-cost skin lesion screening without going to hospitals. The extensive medical resources allow more families to enjoy the convenience that technologies bring to life. Therefore, it is of great practical significance to develop skin lesion diagnosis systems based on AI.

Despite the advantages of the end-to-end deep learning methods in terms of accuracy, their black-box nature in operation has been criticized by the medical community. In the presence of noise, the diagnosis result may become inaccurate, giving rise to hidden risks of medical security [9]. The data might be maliciously manipulated, leading to misdiagnosis and other more serious consequences. Thus, it is highly desirable to take into account of interpretability in addition to the accuracy for medical applications of AI systems [10]. However, popular deep learning methods can achieve high accuracy but lack medical interpretability, while traditional methods with domain knowledge such as ABCD rules, namely, asymmetry, border irregularity, color variegation, and diameter, can provide clinical interpretation for dermatologists. Thus, it makes sense to introduce domain knowledge into deep learning methods [11]. It makes improvement in interpretability while maintaining high accuracy. Apart from image data, it is more human-like to integrate more models of information into final diagnosis.

The main contributions of the article are summarized as follows.

1) We propose a novel interpretability-based multimodal convolutional neural network (IM-CNN) for skin lesion diagnosis, which is comprised of three paths to deal with metadata of the patient, features extracted from lesion segmentation, and skin lesion images.
2) We design an interpretability visual module to give explanations for the three branches, consisting of gradient-weighted class activation mapping (Grad-CAM) for image learning and Shapley additive explanation (SHAP) for metadata learning. This enables us to achieve high accuracy and good interpretability, including multimodal visualization analysis.
3) We adopt various strategies to improve the accuracy of our proposed model, including a loss function handling imbalanced data, data augmentation, regularization, optimizer selection, evaluation index selection, learning rate adjustment, and early-stop strategy.

The proposed model has not only better interpretability but also achieved the state-of-the-art performance for skin lesion diagnosis on the HAM10000 dataset, which opens up a new avenue for deep learning application in skin lesion diagnosis.

The remainder of this article is organized as follows. Section II reviews the related work on skin lesion diagnosis with traditional, deep learning, and multimodal methods. The proposed IM-CNN model for skin lesion diagnosis is described in Section III. Section IV provides the experimental results and discussion. Conclusions and future work are discussed in Section V.

II. Related Work

A. Skin Lesion Diagnosis With Traditional Methods

The traditional methods focus mainly on dermoscopy image features and patient information, and often diagnose the skin lesion with domain knowledge. The 7-point checklist is an efficient way for skin lesion diagnosis, including pigment network, blue whitish veil, vascular structures, pigmentation, streaks, dots and globules, and regression structures [12]. In addition to the 7-point checklist, the ABCD rule is also a common method for dermatologists to know about the physical conditions and progression of skin lesions [13]. However, these two methods place a high demand on dermatologists.

Many scholars devote themselves to the extraction of image features to reduce the reliance on dermatologists, such as line detection and color detection. Popular classifiers, such as random forest, support vector machine (SVM), and *k*-nearest neighbors (KNNs), are often utilized for extracting image features. Mirzaalian *et al.* [14] proposed a graph-based approach to matching moles among images and tested it on a mass of synthetic images and some real images. It provides a convenient way for finding moles in different images of the same person and greatly improves the diagnosis efficiency. Barata *et al.* [15] proposed CAD systems for the skin lesion diagnosis. They used the correspondence latent Dirichlet allocation algorithm to obtain a probabilistic model for the detection of relevant colors and classified the skin lesion with color information and achieved the best performance using random forests. Madooei *et al.* [16] proposed a new approach to identifying the blue-white structure (BWS), which was an important criterion for skin lesion diagnosis. They labeled the image and localized the BWS regions with the multiple instance learning (MIL) framework and achieved state of the art in comparative experiments. Chen *et al.* [17] designed a surgical wounds assessment system for self-care after surgeries, consisting of superpixel segmentation, skin area detection, wound area detection, and wound assessment. Mirzaalian *et al.* [18] used dual-channel quaternion tubularness filters and multilabel optimization based on the Markov random field (MRF) to enhance hairs in dermoscopic images, overcoming a big challenge in skin lesion segmentation and diagnosis. They validated the proposed approach on 40 real clinical images and 94 synthetic images, and their results showed that the method outperformed SVD and DCT.

The traditional methods rely heavily on image preprocessing, leading to poor generalization toward images from different data sources. The handcrafted features extracted by traditional methods are not able to effectively represent the skin lesion images, and consequently, the classification accuracy based solely on hand-crafted features is low. However, the domain knowledge in traditional methods is easy for dermatologists to understand. So, we extract the domain knowledge into our framework to retain its interpretability.

B. Skin Lesion Diagnosis With Deep Learning Methods

Deep learning has shown great success in various fields, and recent studies have confirmed the overwhelming superiority of deep learning to dermatologists in terms of classification of skin lesions.

Brinker *et al.* [19] compared the skin diagnosis performance of convolution neural networks (CNNs) with that of dermatologists tested on 100 images, including melanomas and nevi. The experimental results showed that CNNs achieve better performance in terms of the mean sensitivity and specificity. Haenssle *et al.* [20] devoted a considerable amount of effort to diagnostic performance comparisons between 58 dermatologists and the convolutional neural network. In study level-I, dermatologists were significantly surpassed by the CNN when provided with dermoscopic images only. In study level-II, dermatologists were provided with more clinical information besides images, which was more in line with the actual situation of the dermatologists' diagnosis. The results showed much improvement in dermatologists' diagnostic performance while their performance was still inferior to the CNN in terms of specificity. This reflected the value of additional information for diagnosis. Some dermatologists were reluctant to follow the recommendation of CNN when they did not trust it, which will hinder the application of deep learning in the medical field [21]. Fujisawa *et al.* [22] developed a skin tumor diagnosis system with a deep CNN. It was trained on a relatively small dataset comprised of 4867 images and achieved higher average sensitivity and specificity than the dermatologists and dermatology trainees. Xia *et al.* [23] designed a transfer learning model to collect skin lesion-related images from Web source and presented a comprehensive visual method for skin lesion analysis.

Some researchers tried to improve the CNN for better performance. They used various methods to improve its accuracy and robustness. Kassani and Kassani [24] conducted a comparative study of deep learning architectures for skin lesion detection and found ResNet50 with augmentations and preprocessing performed the best for melanoma detection. Barata *et al.* [25] introduced color compensation techniques to reduce the influence of the discrepant color features from multiple sources. The shades of the gray method were used to apply color constancy to both RGB and HSV images. The results showed a significant improvement in the lesion classification. Pollastri *et al.* [26] proposed a framework for augmenting images with generative adversarial networks (GANs) in the skin lesion segmentation. The comparative experiments showed that the augmented images significantly increase the accuracy of melanoma skin lesion segmentation. Nida *et al.* [27] used a deep region-based convolutional neural networks (RCNNs) to detect skin lesion regions. The method was evaluated on dataset ISIC-2016, and the superiority of the proposed method was confirmed. Yu *et al.* [28] proposed a two-stage method consisting of segmentation and classification. They incorporated a multiscale contextual information integration to make a fully convolutional residual network (FCRN) for segmentation and then classified the skin lesions with very deep residual network (DRN). The experiments were conducted on the ISBI 2016 dataset, and the results proved the effectiveness of the proposed method. Therefore, we introduce the effective deep learning methods to our framework to improve the model accuracy.

C. Multimodal Methods for Skin Lesion Diagnosis

Multimodal machine learning is a new trend that builds models and relates information from multiple modalities. This is a dynamic multidisciplinary field that is increasingly important and has extraordinary potential [29], [30]. Zhang *et al.* [31] proposed a multimodal feature integration framework to capture the high-order correlation among multimodal features. Besides, multimodal approaches allow the absence of modalities and to perform more reliably when some attributes are missing [32], [33].

Multimodal fusion has found wide applications in medical analysis, including brain diagnosis, breast cancer tumor detection, and brain tumor segmentation. Yang *et al.* [34] designed a robust multimodal data integration method called SMSPL to identify significant multiomics signatures and to predict cancer subtypes of cancers. Zheng *et al.* [33] proposed a multimodal emotion recognition framework called EmotionMeter to combine electroencephalography (EEG) signals and eye movements. The experimental results showed the significant enhancement in accuracy compared with the single modality. The multimodal medical image fusion has been proven to be an effective method for improving diagnosis accuracy and advancing medical reliability [35].

The multimodality concept has been introduced into skin lesion diagnosis and achieved notable results. Mahbod *et al.* [22] proposed a fully automatic computerized method based on a novel ensemble scheme for CNNs that combined intra-architecture and interarchitecture networks for the skin lesion classification. Zhang *et al.* [36] presented a synergic deep learning (SDL) model for melanoma detection, where different deep convolutional neural networks learned from each other. This was an effective mechanism for reducing errors, which achieves the state of the art on ISIC-2016 and ISIC-2017 datasets. Chai *et al.* [37] designed a multibranch neural network (MB-NN) model to take advantage of both hidden features and domain knowledge for glaucoma diagnosis. Kawahara *et al.* [12] developed multitask multimodal neural nets for skin lesion diagnosis. They used various loss functions for the robustness of the model in different conditions. In the comparable results, their model showed superiority in sensitivity, specificity, and AUROC, which opened up a new direction for multimodal fusion in skin lesion diagnosis. Hagerty *et al.* [4] combined the handcrafting process with deep learning image processing for melanoma detection. The comparative results showed that fusion of handcrafted features and deep learning features achieved higher accuracy in melanoma diagnosis.

Most previous studies are related to decision making about whether to make a biopsy according to the skin lesion types. It is more valuable to give a reasonable explanation of the decision-making process with high accurate classification, which is more human-like. Note, however, there exists a tradeoff between accuracy and interpretability. The rule-based methods are highly interpretable but not very accurate, while deep learning methods are usually accurate with good generalization ability but unexplainable. Some researchers tried to improve the interpretability.

Li *et al.* [38] proposed an interpretable deep neural network called WaveletKernelNet (WKN) for industrial intelligent diagnosis. They explored the interpretability of WKN through the visualization of the feature map of the continuous wavelet convolutional layer and its inner product matching. Zhang *et al.* [39] proposed an attention residual learning convolutional neural network (ARL-CNN) model, which gives a full play to the advantages of residual learning in terms of deep network and attention learning for important regions detection. The model was evaluated on the ISIC-skin 2017 dataset and compared to a few state-of-the-art models, including ResNet, residual attention network, and SEnet in terms of accuracy, sensitivity, specificity, and AUC. As a complementary work to the research from the multimodal diagnosis perspective, we propose a novel IM-CNN for skin lesion diagnosis.

III. Methodology

The specific domain knowledge, including age, location, asymmetry, and border irregularity, are the main judgement criteria of skin lesions for dermatologists. To make full use of the domain knowledge in dermatology and to utilize the entire skin lesion images in the meantime, we propose an IM-CNN for skin lesion diagnosis. The metadata and skin lesion images of patients are optional as input of IM-CNN, which allows for the absence of one modal data. The diagnosis results of skin lesions and visual analysis are obtained as the output of IM-CNN.

A flowchart of IM-CNN is presented in Fig. 1. IM-CNN consists of three paths (branches). The first branch deals with the basic information of the patient, such as age and location of the skin lesion. The second branch introduces the image features to the metadata. It segments the skin lesion areas with a segmentation neural network, and extracts image features, such as shapes and colors, from the view of a dermatologist using an automatic algorithm. The specific domain knowledge is added into the framework construction through the first two branches. The third branch focuses on the skin lesion image and uses an ensemble of state-of-the-art deep learning models as the backbone to capture the global and local details. In order to improve the interpretability of the model and make it easier for dermatologists to use, we add the interpretability modules, which consist of SHAP for metadata and Grad-CAM for skin lesion images. The dermatologists obtain not only the prediction results but also the judgement reason based on the domain knowledge of Branch 1 and Branch 2, which enhance the dermatologists' trust to models.

In the sequel, feature extraction based on lesion segmentation, backbone model, interpretability visual module, and multimodality fusion will be explained in detail.

A. Feature Extraction Based on Lesion Segmentation

In traditional diagnostic methods, visual domain knowledge including the ABCD rule is often obtained manually, which requires plenty of time to deal with and is a common valuable method for doctors to know about the physical conditions and progression of the skin lesion. In branch 2, we enrich ABCD rules by adding patterns and texture to capture more comprehensive features and call it the ABCDPT rules. ABCDPT stands for asymmetry, border irregularity, color variegation, diameter, patterns, and texture of skin lesion images. We extract features according to ABCDPT rules with deep learning methods based on lesion segmentation automatically, and combine them with branch 1.

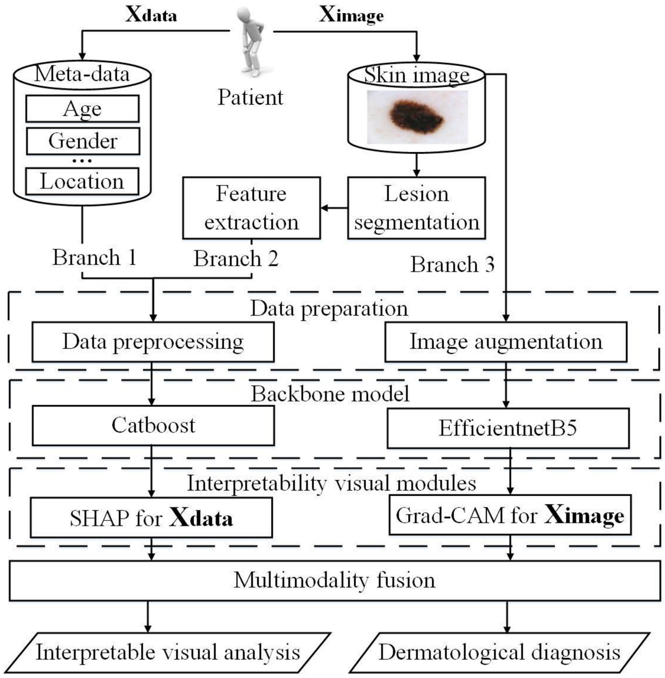

Fig. 1. Flowchart of IM-CNN.

In the HAM10000 dataset, there exists a normal skin area unrelated to skin lesion diagnosis in images. Skin lesion segmentation is the first step to extract features from images. A popular deep learning model, called U-Net, a well-known model with a U-shaped structure based on FCN [40], is then used for skin lesion segmentation. It consists of downsampling and upsampling structures similar to the capital English letter U, and uses skip connection in the same stage, which ensures that more high-level and low-level features are integrated into the final feature map. U-Net fits for the medical image segmentation because medical images are usually simple with a fixed structure, requiring both high-level and low-level features. The refined segmented masks are easily derived through it. We describe patterns with a pigment network, negative pigment network, and globules, which are of great significance for skin lesion classification in clinical diagnosis. The examples of morphological patterns are shown in Fig. 2.

The pigment network is comprised of cross brown lines with a grid-like reticular pattern. The negative pigment network, also known as the reverse or inverse network, consists of relatively bright or dark areas filled with obvious holes. The shallower area shows serrated and the darker area resembles elongated tubes or curved spheres. This feature is highly specific for melanoma, particularly for a melanoma arising in a nevus. Globules are well-defined structures with a shape of circle or oval, which are often located at the

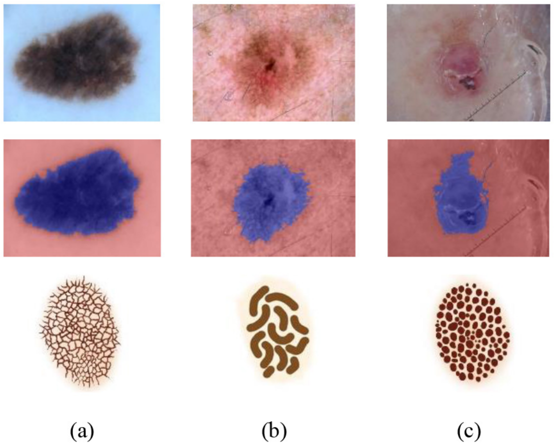

Fig. 2. Examples of morphological patterns, segmented area, and topology of skin lesion images. (a) Pigment network. (b) Negative network. (c) Globules.

edge of melanocytic lesions. The distribution of the globules is important: the globules around the melanocytic lesion show a horizontal growth phase, which can occur in growing melanocytic nevus or superficial melanoma. In melanoma, the globules vary in size, shape, and color, and are often found locally and irregularly around the lesion [41], [42]. U-Net is used for extraction of the pigment network, negative network, and globules. We notice that not all skin images have the skin features mentioned above, so we transfer the output results to the Boolean form to judge if they have the corresponding pattern and then blend them into the metadata.

The gray-level co-occurrence matrix (GLCM) [43] measures consistency of patterns and colors in an image, and is able to derive the main texture properties, such as correlation, homogeneity, energy, and contrast from segmented skin lesion images. Suppose that a skin lesion image consists of $M \times N$ resolution cells. The image I can be expressed as a function $I(M \times N \rightarrow \text{GLCM})$ that assigns grayscale to resolution cells. Given a random position (h, k) in I and offset $(\Delta x, \Delta y)$, $\text{GLCM}(i, j)$ can be represented as follows:

$$\text{GLCM}(i,j) = \sum_{h=1}^{N}\sum_{k=1}^{M}\begin{cases} 1, & I(h,k) = i, I(h+\Delta x, k+\Delta y) = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\text{GLCM}(i, j)$ means the probability that a pixel at grayscale i has a grayscale value of j with another pixel at its corresponding position. Correlation, homogeneity, energy, and contrast are calculated based on $\text{GLCM}(i, j)$. The descriptions and computation formulas of features above are shown in Table I, where P and A denote the image perimeter and the image area, respectively. σ_R, σ_G, and σ_B are the standard deviations of red, green, and blue components of the lesion area, respectively, while M_R, M_G, and M_B are the maximum values of red, green, and blue components in the lesion region.

B. Backbone Model

Branch 1 and branch 2 are blended together as the metadata and processed with the Catboost [44] classifier. Catboost is a popular ensemble boosting method that introduces an ordered target statistic strategy to deal with categorical features effectively. It replaces the category features ranked before the sample with the expectation of the feature value, and adds the priority and its weight. So, it can convert the category features into numerical features, effectively reducing the noise of low-frequency categorical feature and enhancing the robustness of the algorithm. The metadata $\boldsymbol{X}_{\text{data}}$ consist of original descriptive information of the patient such as age, and extracted features from images, such as pigment network and asymmetry. These features are well defined and composed of categorical features and numerical features. Data preprocessing, such as one-hot coding for categorical features and min–max normalization for numerical features, is necessary before inputting into the model. Suppose that $\boldsymbol{X}_{\text{data}} = (\boldsymbol{X}_{\text{cat}} + \boldsymbol{X}_{\text{num}}) = (\boldsymbol{x}_{\text{cat,i}} + \boldsymbol{x}_{\text{num,i}}), i = 1, 2, \ldots, n$, where n denotes the number of samples, $\boldsymbol{X}_{\text{cat}}$ represents the categorical feature vector, while $\boldsymbol{X}_{\text{num}}$ represents the numerical feature vector.

Suppose the random permutation of samples is $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$. The sample $x^j_{\rho_U}$ at the jth feature in sequence ρ_U can be expressed as follows:

$$x^j_{\rho_U} = \frac{\sum_{k=1}^{U-1} I\left(x^j_{\rho_k} = x^j_{\rho_U}\right) \cdot y_k + a \cdot U}{\sum_{k=1}^{U-1} I\left(x^j_{\rho_k} = x^j_{\rho_U}\right) + a} \tag{2}$$

where U denotes the added prior term, and $a > 0$ is the weight coefficient of the prior term.

In branch 3, we adopt deep learning models to deal with skin lesion images. A recent study [45] has shown that better models in the ImageNet challenge also have better performance on transferability. To achieve better performance on skin lesion diagnosis, it is necessary to choose an accurate and efficient network on ImageNet as our backbone model.

EfficientNet is proposed by Tan and Le [46]. It achieves state-of-the-art performance with fewer parameters compared with other models. It improves the prediction accuracy and efficiency significantly through the compound scaling method, which scales the depth, width, and resolution of convolutional neural networks. The compound scaling method is a key step in EfficientNet, which pays more attention to the relevant area and obtains more details on classification targets. The layer ℓ of input image $\boldsymbol{X}_{\text{image}}$ can be expressed as tensors $\langle H_\ell, W_\ell, C_\ell \rangle$, where H_ℓ, W_ℓ, and C_ℓ denote the dimension of height, width, and the number of channels, respectively. Thus, the EfficientNet can be defined as follows:

$$h_{\text{image}} = \odot_{\ell=1,\ldots,s} f_\ell^{R_\ell}\left(\boldsymbol{X}_{\langle H_\ell, W_\ell, C_\ell \rangle}\right) \tag{3}$$

where $f_\ell^{R_\ell}$ means that the function operation f_ℓ is repeated R_ℓ times at layer ℓ, and $\boldsymbol{X}_{\langle H_\ell, W_\ell, C_\ell \rangle}$ denotes the input image $\boldsymbol{X}$ with tensor $(H \times W \times C)$. In this work, we adopt EfficientNetB5 pretrained on ImageNet as network weights.

C. Interpretable Visual Module

It is hard for the human beings to distinguish similar classes, for example, melanoma from melanocytic nevus, and basal cell carcinoma from vascular skin lesion, with the naked eye. The interpretability visual module provides effective support for

TABLE I
DESCRIPTION AND FORMULAS OF SKIN LESION FEATURES

Feature	Formulas	Description
Asymmetry	$\frac{1}{2}\left(\frac{A-h_A}{A}+\frac{A-v_A}{A}\right)$	Asymmetric index and eccentricity
Border irregularity	$\frac{P^2}{4\pi A}$	Compact index
Color variegation	$\mathrm{C}_R=\frac{\sigma_R}{M_R},\mathrm{C}_G=\frac{\sigma_G}{M_G},\mathrm{C}_B=\frac{\sigma_B}{M_B}$	Normalized standard deviation
Diameter	$\frac{P}{2\pi}$	Lesion region diameter
Correlation	$\sum_{i,j}\frac{GLCM(i,j)(i-\mu_i)(j-\mu_j)}{\sigma_i\sigma_j}$	Linear dependency measurement of image
Homogeneity	$\sum_{i,j}\frac{GLCM(i,j)}{1+(i-j)^2}$	Distribution closeness of elements in GLCM
Energy	$\sqrt{\sum_{i,j}GLCM(i,j)^2}$	Summation of squared elements for texture uniformity measurement
Contrast	$\sum_{i,j}(i-j)^2GLCM(i,j)$	Comparison of the intensity of a pixel and its neighbors over the entire image

finding the differences between similar skin lesions. It highlights the related regions in skin lesion image and analyzes the features of metadata that promote or suppress the classification results.

For image $\boldsymbol{X}_{\text{image}}$, Grad-CAM is conducted, and abnormal regions are highlighted through visualization. The visual heat map is adopted as the display form, and regions of interest (ROIs) show a high clinical relevance of skin lesion, to which more attention should be paid. Deeper layers in the convolutional neural network are expected to have rich semantic information, which is used as local information guidance of the shallow layers. Grad-CAM can be regarded as a generalized form of CAM, which is not limited by the model structure. CAM replaces the fully connected layer with a GAP layer and retrains to obtain weights while Grad-CAM uses the global average of the gradient to calculate the weights.

The Grad-CAM heat-map is a weighted combination of feature maps and can be expressed as follows:

$$L_{ij}^{c}=\mathrm{relu}\left(\sum_{k}\omega_{k}^{c}A_{ij}^{k}\right)\tag{4}$$

where relu is the rectified linear unit activation function and ω_k^c can be calculated as follows:

$$\omega_{k}^{c}=\sum_{i}\sum_{j}\alpha_{ij}^{kc}\mathrm{relu}\left(\frac{\partial Y^{c}}{\partial A_{ij}^{k}}\right)\tag{5}$$

where α_{ij}^{kc} is the weight coefficient for the pixelwise gradients for class c and convolutional feature map A^k, which can be calculated through backpropagation. Compared to the original version Grad-CAM, relu and weight gradient α_{ij}^{kc} are added to the weight representation of the feature map corresponding to each class.

For metadata $\boldsymbol{X}_{\text{data}}$, SHAP [47] is introduced to analyze the feature importance in the experiment and obtain the relations between features and classification results with Shapley values. Inspired by the cooperative game theory Shapley values, SHAP constructed an additive interpretation model and all features are considered to be contributors. For each prediction sample, the model generates a prediction value, and the SHAP value is the value assigned to each feature in the sample. It calculates the marginal contribution of a feature when it is added to the model, and then considers the different marginal contributions of that feature in the case of all feature sequences.

Let x_{data}^{ij} denote the jth feature of the ith sample in a set of skin lesion patients, y_{data}^{i} denotes the predicted value of the model for the sample, and y_{data}^{base} denotes the baseline of the entire metadata model. The SHAP value of y_{data}^{i} can be formulated as follows:

$$\begin{aligned}y_{data}^{i}&=y_{data}^{\text{base}}+f\left(x_{data}^{i1}\right)+f\left(x_{data}^{i2}\right)+\cdots+f\left(x_{data}^{iM}\right)\\&=y_{data}^{\text{base}}+\sum_{m=1}^{M}f\left(x_{data}^{im}\right)\end{aligned}\tag{6}$$

Algorithm 1: SHAP Value Computing

```
Input: Data matrix $X_{data}$, iterations T, selected sample
       $x^{i}_{data}$, feature $j$, machine learning model $h$
Output: Shapley value $f(x^{i}_{data})$
Begin
For t = 1 to T:
    Randomly choose another sample $z_i$ from $X_{data}$
    Choose a random order $\tau$ of the feature values
    Sequence samples $x_\tau = (x_1, \ldots, x_j, \ldots, x_m)$ and
      $z_\tau = (z_1, \ldots, z_j, \ldots, z_m)$
    Construct two new samples based on $x_\tau$ and $z_\tau$ with
      and without the $jth$ feature
    Sequence samples including feature $j$:
      $x_{+j} = (x_1, \ldots, x_{j-1}, x_j \ldots, z_m)$
    Sequence samples excluding feature $j$:
      $x_{-j} = (x_1, \ldots, x_{j-1}, z_j, \ldots, z_m)$
    $f(x^{it}_{data}) = f(x_{+j}) - f(x_{-j})$
$f(x^{i}_{data}) = \frac{1}{T}\sum_{t=1}^{T} f(x^{it}_{data})$
End
```

where $f(x^{ij}_{data})$ denotes the SHAP values of x^{ij}_{data}, M denotes the number of the features, and y^{base}_{data} is the mean value of all $f(x^{ij}_{data})$. When $f(x^{ij}_{data}) > 0$, the jth feature of the ith sample has the positive effect on the prediction results y^{i}_{data}, while when $f(x^{ij}_{data}) < 0$, the jth feature imposes a negative effect on y^{i}_{data}. It reflects the positive and negative effects on the prediction result.

In terms of feature importance, SHAP considers the impact and synergistic effect of all the features compared with the mean decrease impurity, permutation importance, and drop column importance, which can be easily computed by

$$I_j = \sum_{i=1}^{N} \left| f\left(x^{ij}_{data}\right) \right| \tag{7}$$

where $|f(x^{ij}_{data})|$ means the absolute Shapley value of the jth feature across the data. The feature importance can be sorted according to the values of I_j. The pseudocode of the SHAP value computing is presented in Algorithm 1.

This process is repeated until the Shapley values of all features are computed. Tree SHAP is able to compute the marginal contribution ratio of tree-based models such as Catboost. So, we use tree SHAP to reflect the relations between features of metadata and the diagnostic results.

D. Multimodality Fusion

IM-CNN takes both the skin lesion image and metadata as the input, and outputs the probability of diagnostic result and the corresponding visual explanations from the view of image and metadata. It takes full advantages of each modal to obtain an accurate and interpretable diagnosis. Some fundamental notations are given as follows.

Particularly, bold capital letters and bold lowercase letters are used to express matrices and vectors, respectively. Given an input feature space $\boldsymbol{X}$ of a skin lesion patient, we divide it into image feature $\boldsymbol{X}_{\text{image}}$ and metadata feature $\boldsymbol{X}_{\text{data}}$. $\boldsymbol{X}_{\text{image}}$ indicates the pixel matrix of the skin lesion and $\boldsymbol{X}_{\text{data}}$ contains metadata information of patients. Thus, $D = (\boldsymbol{x_i}, y_i) = ([\boldsymbol{x}_{\text{image}}, \boldsymbol{x}_{\text{data}}], y_i), i = 1, 2, \ldots, n$ is a collection of n labeled instances, where $\boldsymbol{x_i} \in \boldsymbol{X}$, $\boldsymbol{x}_{\text{image}} \in \boldsymbol{X}_{\text{image}}$, and $\boldsymbol{x}_{\text{data}} \in \boldsymbol{X}_{\text{data}}$. We denote a set of skin lesion diagnostic results as $y_i = \{0, 1\}^d$, where y_i is the one-hot encoding of y, and d corresponds to the type of diagnosed skin lesions. Given d classes, the set of class labels $= \{c_1, c_2, c_3, \ldots, c_d\}$. If the class label of the ith sample is k, it is denoted as $y_{i,k} = 1$. One-hot encoding is used and the true label of each sample is a one-hot vector, with only one position being 1.

Imbalanced data are common in the medical field, which makes the classifier focus more on the major classes but neglect the minor classes. It results in a low sensitivity to the minor classes and a low specificity to the major classes, which can be addressed to a certain degree by revising the loss function.

The categorical cross-entropy loss is a popular loss function in multiclass classification learning. It assigns the same weight to each class, which leads to little attention to the minor classes and results in a low sensitivity for underrepresented classes. To overcome the effect of imbalanced data, IM-CNN introduces focal loss as the loss function. Focal loss is a variant of the categorical cross-entropy loss, which has been proposed for handling imbalance data. It introduces a class weight α to balance the class proportion and modulating factor $(1 - p_t)^\gamma$ to balance the learning difficulty. Each class loss is reweighted by its inverse frequency. The focal loss is defined as

$$\text{Focal\_Loss} = -\sum_{i=0}^{N}\sum_{c=0}^{C} \alpha_{i,c} y_{i,c} \left(1 - p_{i,c}\right)^\gamma \log p_{i,c} \tag{8}$$

where $p_{i,c}$ and $y_{i,c}$ denote the predictions and the ground truth for c class of the ith sample of given dataset. i and c denote the data samples and the class, respectively. $\alpha_{i,c}$ reflects the weight of c class, and γ regulates the rate of weight reduction of simple samples. Thus, the model puts more focus on difficult and misclassified samples. In the HAM10000 dataset, over half of the entire samples belong to melanocytic nevus (NV), and the visual features of NV are clearly different from other categories. As a consequence, it is necessary to use the focal loss to regulate the NV class.

We add a merging layer h_{merge} at the end of the outputs of two classifiers h_{image} and $h_{\text{meta-data}}$, which is used to fuse multimodal features, and can be expressed as follows:

$$h_{\text{merge}} = \sum_{i=1}^{H} \text{Softmax}\left(h_{\text{output}}\right) \tag{9}$$

where H and h_{output} indicate the number of models and the output of each modal, $h_{\text{meta-data}}$ and h_{image}, in this work. We use softmax as the activation function to obtain the probability of multiclassification. The softmax function compresses the multimodal vectors into 0–1 vector. The final classification results belong to the classes with the highest probabilities.

The results have been proven to have remarkable performance in comparison to the original one such that

$$\mathrm{Softmax}\left(h_{\mathrm{output}}\right)_c = \frac{e^c}{\sum_{c=1}^{C} e^c}. \tag{10}$$

IV. Skin Lesion Diagnosis Experiment

IM-CNN is implemented based on Keras and Tensorflow. All the experiments were run on a tower workstation with a NVIDIA GeForce RTX 2080 Ti GPU, a 2.10 GHz Intel Xeon E5-2620 v2 CPU and 4*8G DDR3 RAM. To evaluate the performance of the proposed model, we set three baseline sets for comparison. The first baseline set is composed of the traditional method based on extracted features from images. The second is made up of the popular deep learning method, VGG-16/19, ResNet50, DenseNet-121/169, Inception_v3, Inception Resnet v2, and Xception. We also compare the proposed model with its single modality to validate the effectiveness of introducing multimodality.

A. Data Preparation

The comparative studies are conducted on the HAM10000 dataset introduced by Tschandl *et al.* [48] and Codella *et al.* [49], which is so far the largest publicly available collection of quality controlled dermoscopic images of skin lesions and screened for both privacy and quality assurance.

The HAM10000 dataset consists of 10 015 images, each of which has 600×450 pixels. It includes representative samples of seven classes, including actinic keratoses and intraepithelial carcinoma/Bowen's disease (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus, such as keratoses and BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular skin lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, and VASC). Fig. 3 shows some typical samples of seven skin lesion dermoscopy images, and Table II gives a summary of the seven skin lesions. More than 95% of all lesions encountered in clinical practice will fall into one of the seven diagnostic categories. In practice, the clinician's task is to distinguish different skin lesions, and make specific diagnoses because different lesions, such as melanoma and basal cell carcinoma, can be treated in different ways and time frames.

MEL is the most dangerous form and is a malignant tumor derived from melanocytes. If removed early, it can be cured with a simple surgical excision. Melanoma can be invasive or noninvasive and is also often confused with NV. NV is a benign neoplasm of melanocytes and appears in numerous variants that are included in the dataset. In contrast to melanoma, they are usually symmetric in color and structure.

There are many common features among BCC, VASC, and DF. BCC is a common variant of epithelial skin cancer and rarely metastases. It grows destructively and accompanied by morphological variations, such as flattening and pigmentation if untreated timely. DF is considered benign hyperplasia or a minimal inflammatory response to trauma. The most common dermoscopy appearance is the surrounding reticular line with a central white spot indicating fibrosis. VASC is red or purple, with well-defined structures called red clods or loose solids. It incorporates cherry hemangiomas and pyogenic granuloma, among others.

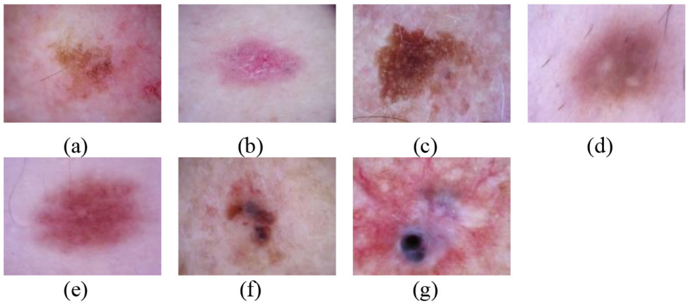

Fig. 3. Some typical samples of seven skin lesion dermoscopy images. (a) AKIEC. (b) BCC. (c) BKL. (d) DF. (e) NV. (f) MEL. (g) VASC.

TABLE II
Summary of Seven Skin Lesions

Abbr.	Class	Count	Distribution
AKIEC	Actinic keratosis / Bowen's disease (intraepithelial carcinoma)	327	3.27%
BCC	Basal cell carcinoma	514	5.13%
BKL	Benign keratosis lesion	1099	10.97%
DF	Dermatofibroma	115	1.15%
NV	Melanocytic nevus	6705	66.95%
MEL	Melanoma	1113	11.11%
VASC	Vascular skin lesions	142	1.42%

AKIEC, a mixture of actinic keratosis and Bowen's disease, is a noninvasive variant of squamous cell carcinomas that may develop into the invasive type, usually manifested by surface scales and a lack of pigmentation. Actinic keratosis is more common on the face, and Bowen's disease is more common on other parts of the body.

BKL is a generic type of keratosis and can be confused with the other classes, including seborrheic keratosis, solar lentigo, and lichen-planus-like keratosis (LPLK). They are biologically similar and are often reported in the same general terms in histopathology. The sensitivity of BKL is less important than other classes.

B. Data Preprocessing

Data preprocessing is carried out for images and metadata. The purpose of preprocessing for images and metadata is to reduce the effect of noise and imbalanced classes in the dataset so as to increase the ability of models to learn important features hidden in metadata and images.

Metadata consist of two parts: the first part comes from the basic information of patients, such as age and sex; and the second part derives from the feature extracted from images, including pigment network, globules, asymmetry, and border irregularity.

Data preprocessing for metadata is necessary for model performance. The missing values are first processed with the mean insertion method for numerical values and the mode insertion method for categorical values. Then, features of metadata can be divided into numerical and categorical types. The statistics summarizing the numerical and categorical features are shown in Tables III and IV, respectively. Min–max normalization is used for numerical features, which scales and translates each feature into the interval [0, 1]. Normalization is conducted as follows:

$$x^{*}_{ij,\text{data}} = \frac{x_{ij,\text{data}} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)} \tag{11}$$

where $x_{ij,\text{data}}$ denotes the jth feature value of the ith sample in metadata, and $\min(x_j)$ and $\max(x_j)$, respectively, denote the minimum and maximum value of the jth feature for all the samples.

The one-hot encoding method is applied for producing vectors and converting categorical features into dummy features, which can effectively prevent transformed categorical features from being assigned ordinal meaning.

Fig. S1 in the supplementary material shows some basic statistics plots of skin lesion data. In Fig. S1 in the supplementary material, we observe that the age may be helpful to differentiate NV, MEL, and BKL in Fig. S1(a) in the supplementary material. The incidence rate of male is slightly higher than that of female in Fig. S1(b) in the supplementary material. In Fig. S1(c) in the supplementary material, the location of skin lesion includes extremity, scalp, back, ear, foot, and so on. MEL and NV share similar locations, and AK is more frequent in head or neck.

Skin images may be disturbed by noises, such as hairs, light, angle, and black frames, which greatly affects the feature extraction of CNN. Image data augmentation is a popular method for image preprocessing. Hair removal is a useful way to mitigate the effects of hairs. We first convert the color image to a grayscale, and then apply morphological transformation on the grayscale image. Masks for hairs are created and removed from original images with an inpainting algorithm. Comparison before and after hair removal is shown in Fig. S2 in the supplementary material.

Data imbalance is a serious problem in classification task, which has a significant impact on classification accuracy. If the model is trained on imbalanced data, it usually classifies the new samples into the majority class. From Table II, we can find that data are imbalanced among seven classes. For instance, images of Melanocytic Nevus are approximately 58 times as many as those of Dermatofibroma. The adequate data imbalance processing method is necessary, and image augmentation is an effective way to deal with it. We extend the number of images according to the frequency of each class, that is, the less the number of images, the more the number of extensions. We first employ rotations with random horizontal and vertical flips between −180° and 180°. Then, we apply randomly width or height shifting with 0.2 range, and zooming and shearing with 0.15 range. Finally, we introduce the color constancy method to augment color, brightness, contrast and eliminate the variance of luminance with the random factor from 0.8 to 1.2. All normalized images for training and testing are resized to 224 × 224 pixels. The ultimate image count of each class is similar.

TABLE III
STATISTICS SUMMARY OF NUMERICAL FEATURES

Numerical feature	mean	std	median	min	max
Age	51.57	17.36	50.00	0.00	85.00
AsymIdx	0.22	0.13	0.18	0.00	0.93
Eccentricity	0.69	0.16	0.71	0.11	1.00
CI	2.54	1.71	1.95	1.08	26.37
StdR	0.16	0.06	0.16	0.01	0.40
StdG	0.20	0.06	0.21	0.02	0.37
StdB	0.20	0.05	0.21	0.02	0.37
Diameter	215.33	96.65	207.96	32.11	586.32
Correlation	1.00	0.00	1.00	0.99	1.00
Homogeneity	0.70	0.06	0.71	0.44	0.89
Energy	0.08	0.02	0.07	0.03	0.23
Contrast	2.39	1.60	1.94	0.30	30.03

TABLE IV
STATISTICS SUMMARY OF CATEGORICAL FEATURES

Categorical feature	unique	top	frequency
Sex	3	male	5406
Location	15	back	2192
Diagnosis type	4	histopathology	5340
Pigment network	2	true	6203
Negative network	2	true	9823
Globules	2	true	9634

C. Parameter Settings

The dataset is split into 90% training set and 10% test set with a stratified train-test split method to ensure the same proportion of each class. Each image and its corresponding metadata in the training set and test set are unique and irrelevant.

Referring to the parameter settings in the comparative literature [12], [19], [28], [36], [39], [50] and considering our conditions of experimental equipment, we set the key parameters as follows. All models in the experiments are fine-tuned for 50 epochs using the Adam optimization with a starting learning rate of 0.0001. The batch size is equal to 32. Dropout is an efficient method that temporarily drops the network units according to a certain probability during the training process. It is used with a probability of 0.3 for alleviating overfitting. The learning rate is set to be reduced by a factor of 0.25, and early stopping strategy is adopted to prevent overfitting if the

TABLE V
DIAGNOSTIC CONFUSION MATRIX

Confusion Matrix		Ground truth	
		Condition positive	Condition negative
Predict results	Predicted positive	True positive (TP)	False positive (FP) Type I error
	Predicted negative	False negative (FN) Type II error	True negative (TN)
Indicator		$SEN = \frac{\text{TP}}{\text{TP} + \text{FN}}$	$SPE = \frac{\text{TN}}{\text{FP} + \text{TN}}$
		$ACC = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	

models fail to improve the validation loss within ten epochs. In terms of network training, transfer learning is employed to take advantages of similar features from ImageNet [51] and obtain more robust performance.

D. Evaluation Metrics

The skin lesion diagnosis of seven categories is a multiclass classification problem. To evaluate the model comprehensively and compare the model objectively, we use multidimensional indicators, including accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristics (AUCs) of all diagnosis categories. The calculation of the evaluation metrics for each category adopts the one versus rest strategy. For example, the sensitivity of MEL is calculated through

$$\text{SEN}_{\text{MEL}} = \frac{\text{TP}_{\text{MEL}}}{\text{TP}_{\text{MEL}} + \text{FN}_{\text{MEL}}}. \tag{12}$$

All the indicators are computed based on the confusion matrix as Table V. The following metrics are computed between prediction results and ground truth.

In Table V, true or false refers that the classification is right or not, while positive or negative refers that ground truth belongs to the category or not. Thus, TP and TN are measures of correctly classifying positive or negative cases for skin lesion category, respectively. FP is the measure of incorrect rejection as correct ones, also called Type I error, while FN is the measure of correct rejection as incorrect ones called Type II error. They are the main components of a confusion matrix, which lay the foundation for calculating the indicators for measuring the performance of classification models. The formulas of sensitivity (SEN), specificity (SPE), and accuracy (ACC) are listed in the Table V.

There exists a tradeoff between sensitivity and specificity. It is of great significance of detecting skin cancer with a reasonable sensitivity and specificity on skin lesion diagnosis in clinical application. The receiver operating characteristic (ROC) curve is an efficient way to balance them and model classification errors. It is typically created by plotting TPR against FPR at different threshold values. Based on the ROC curve, the area under the ROC curve can be calculated, denoted as AUC, which is a common index summarizing the information contained in the curve. It represents the probability assuming positive ranks higher than negative, which is equivalent to the ranks of the Wilcoxon test. To be precise, AUC can be expressed as follows:

$$\begin{aligned} \text{AUC} &= \int_{x=0}^{1} \text{TPR}\Big(\text{FPR}^{-1}(x)\Big)dx \\ &\approx \frac{\sum_{i \in P} \text{rank}_i - \frac{1}{2}\big(1 + N^{+}\big)N^{+}}{N^{+} \times N^{-}} \end{aligned} \tag{13}$$

where P, N^{+}, and N^{-} denote the positive class, the number of the positive class and the negative class, respectively, and $N = N^{+} + N^{-}$. The probability scores are first sorted from the largest to the smallest, and then sum the ranks of all the positive samples. The number of samples with positive scores greater than negative scores can be calculated by getting rid of $N^{+} - 1$ pairs of two positive samples.

In the medical diagnosis area, sensitivity plays a more important role compared with specificity. Thus, we introduce a new indictor, AUC_SEN_80, which only calculates the area for the region with sensitivity larger than 80%, and is defined as follows:

$$\text{AUC\_SEN\_80} = \int_{x=0.8}^{1} \text{TPR}\Big(\text{FPR}^{-1}(x)\Big)dx. \tag{14}$$

AUC_SEN_80 puts more focus on sensitivity and filters the samples with sensitivity larger than 80%. The AUC_SEN_80 is usually smaller than AUC but more representative for model evaluation especially for medical area. All the equations above will be ill defined if the denominator is 0. So we set the final output as 1 when this happens. The indicators above for each category and macroaverage measures of all categories are computed, and weighted by the category proportion. The macroaverage measure shows the average performance per category, which can efficiently avoid inflated performance estimates on imbalanced datasets. It can be expressed as

$$\text{Macro} - \text{average} = \frac{1}{c}\sum_{c=1}^{C} \text{indicator}_c \tag{15}$$

where indicator denotes accuracy, sensitivity, specificity, AUC, and AUC_SEN_80, and C denotes the corresponding category.

E. Comparison With Deep Learning Models

To verify the performance of our proposed model, IM-CNN, we select some state-of-the-art deep learning models, such as DenseNet121, DenseNet169, Inception_Resnet_v2, Resnet50, Inception_v3, Vgg16, Vgg19, and Xception for comparison. We use the training set to train the model and the test set to evaluate the models in terms of the evaluation metrics. All the deep learning models are constructed based on pretrained ImageNet. Each experiment is conducted five times to reduce the impact of randomness, and outputs the mean value and the standard deviation. The comparative results of skin lesion diagnosis for different models are listed in Table VI. The values in bracket denote the standard deviation and the best performance is highlighted in bold.

TABLE VI
COMPARISON RESULTS WITH POPULAR MODELS

Model	Predict Time/Patient	ACC	SEN	SPE	AUC	AUC_SEN_80
DenseNet121	0.011 (0.002)	0.920 (0.015)	0.708 (0.090)	0.909 (0.014)	0.926 (0.030)	0.855 (0.048)
DenseNet169	0.014 (0.003)	0.919 (0.016)	0.724 (0.079)	0.909 (0.017)	0.919 (0.019)	0.838 (0.048)
Inception_Resnet_v2	0.011 (0.001)	0.929 (0.009)	0.663 (0.059)	0.911 (0.007)	0.911 (0.016)	0.833 (0.042)
Resnet50	**0.006** (0.001)	0.923 (0.017)	0.676 (0.087)	0.911 (0.021)	0.914 (0.010)	0.819 (0.033)
Inception_v3	0.007 (0.002)	0.921 (0.020)	0.765 (**0.051**)	0.901 (0.025)	0.914 (0.014)	0.813 (0.058)
Vgg16	0.012 (0.003)	0.910 (0.022)	0.675 (0.095)	0.898 (0.023)	0.927 (0.020)	0.870 (0.035)
Vgg19	0.012 (0.003)	0.900 (0.025)	0.621 (0.094)	0.889 (0.033)	0.912 (0.025)	0.838 (0.049)
Xception	0.009 (0.002)	0.935 (0.012)	0.721 (0.089)	0.919 (0.012)	0.910 (0.042)	0.818 (0.086)
EfficientnetB5	0.012 (0.001)	0.933 (0.012)	0.724 (0.065)	0.928 (0.012)	0.876 (0.049)	0.781 (0.079)
Proposed model	0.012 (**0.001**)	**0.951** (**0.008**)	**0.835** (0.067)	**0.932** (**0.010**)	**0.978** (**0.006**)	**0.960** (**0.009**)

In our experiments, an NVIDIA RTX 2080 Ti GPU is used to train the proposed IM-CNN model. The prediction time on a single patient is relatively fast and has no significant difference with the fastest Resnet50, taking less than 0.02 s on average. The quick prediction time suggests that our model can be used in a routine clinical workflow.

From the comparison results, we can find that the proposed model achieves the best performance among all the models with an accuracy of 0.951, sensitivity of 0.835, specificity of 0.932 AUC of 0.978, and AUC_SEN_80 of 0.960. It also achieves the best performance in the standard deviation of accuracy, specificity, AUC, and AUC_SEN_80. EfficientnetB5 is one of the base backbones of the proposed model. The great improvement in sensitivity and AUC shows the effectiveness of introducing multimodality to the model and taking full advantages of base backbone models. In addition, AUC_SEN_80 is an effective indicator considering both sensitivity and AUC. Our proposed model achieves a 22.92% improvement compared with EfficientnetB5 in terms of AUC_SEN_80. Fig. S3 in the supplementary material plots the ROC curve of the proposed model, which achieves a great and balanced performance on imbalanced category.

Melanoma diagnosis is the most important in skin lesion diagnosis. Numerous studies are related to melanoma diagnosis. We compare our proposed models with other skin diagnosis models on melanoma in terms of accuracy, sensitivity, specificity, and AUC. The results are summarized in Table VII. The missing values of indicators are replaced by "–."

TABLE VII
COMPARISON RESULTS WITH OTHER DIAGNOSIS MODELS ON MELANOMA

Model	ACC	SEN	SPE	AUC
ARL-CNN50[50]	0.85	0.658	0.896	0.875
DRN-50[28]	0.855	0.547	**0.931**	0.783
x-combine[12]	0.643	-	0.806	0.822
CNN-PA[19]	0.721	-	-	-
SDL[36]	0.872	0.715	-	0.874
Fusion of fine-tuned network[51]	-	-	-	0.873
dermatologists[19]	-	0.741	0.6	0.671
Multimodal CNN[53]	0.72	0.729	-	0.861
Proposed model	**0.906**	**0.838**	0.915	**0.951**

The compared skin diagnosis models in Table VII only process the melanoma diagnosis, which is a binary classification problem and easier than multiclass classification. We extract melanoma class individually and convert it into a binary classification problem. From Table VII, we can see that our proposed IM-CNN performs the best in terms of accuracy, sensitivity, and AUC. It achieves 0.906 in accuracy, 0.838 in sensitivity, and 0.951 in AUC. What is more, our proposed model shows great superiority to dermatologists, which has been shown to be superior to deep learning methods in skin lesion diagnosis [11], in terms of AUC. Although our model performs worse than DRN-50 in terms of specificity, however, our model puts more focus on sensitivity, which plays a more vital role than specificity in melanoma diagnosis.

F. Comparison With Single-Modal Model

In order to validate the effectiveness of introducing multimodality, we compare our proposed multimodal model with a single-modal model. The architecture of the proposed model IM-CNN is listed in Table VIII.

We evaluate three branches to measure the contribution of each branch. In Table VIII, branch 1 means the metadata only including essential information of patients, such as age and location. Branch 2 deals with the feature information extracted from the image such as pigment network. Branch 3 corresponds to the deep learning backbone EfficientNetB5. It is expected to act directly on the skin lesion image. The final

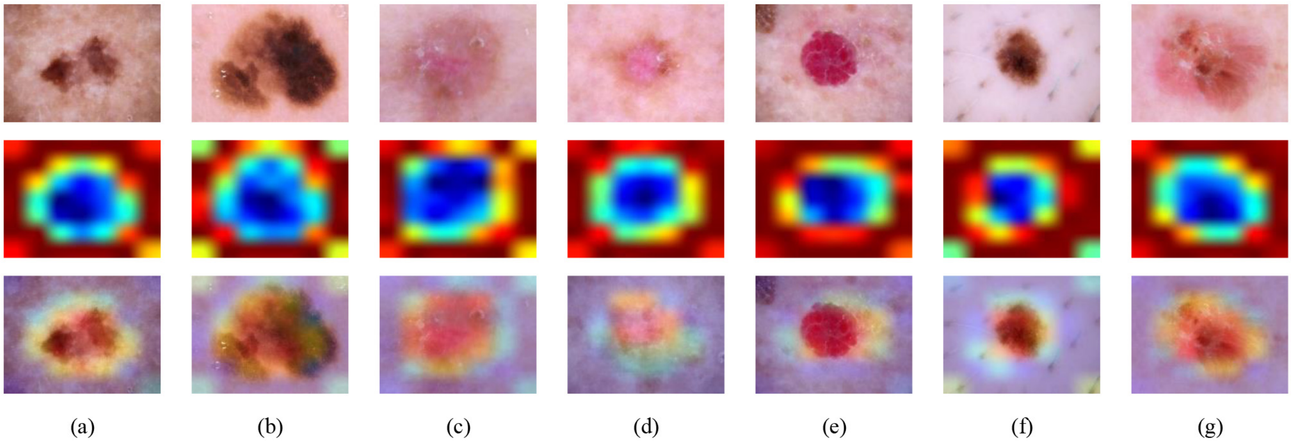

Fig. 4. (a)–(g) indicate the visual heat map images of MEL, NV, BCC, DF, VASC, BKL, and AKIEC samples, respectively. The original skin lesion images are listed in first row, and the heat map and their combination with the original image are listed in second and third rows.

TABLE VIII
COMPARISON WITH SINGLE-MODAL MODELS

Branch	Class	ACC	SEN	SPE	AUC	AUC_ SEN_80
Branch 1	MEL	0.820	0.486	0.862	0.831	0.714
	Avg.	0.905	0.402	0.883	0.889	0.775
Branch 2	MEL	0.798	0.559	0.828	0.807	0.629
	Avg.	0.891	0.426	0.868	0.855	0.717
Branch 3	MEL	0.849	0.622	0.878	0.827	0.616
	Avg.	0.933	0.739	0.929	0.891	0.763
Final output	MEL	**0.906**	**0.838**	**0.915**	**0.951**	**0.915**
	Avg.	**0.951**	**0.835**	**0.932**	**0.978**	**0.960**

output fuses the multimodal branches and outputs the accurate classification results. The single modality model can be processed separately when the data of another modal are missing. From Table VIII, we can see that the performance of the multimodal model is superior to any single-modal ones in terms of average accuracy, sensitivity, specificity, AUC, and AUC_SEN_80. Especially, the final output achieves the best result in melanoma diagnosis, which confirms the benefit of multimodality. The improvement in sensitivity is the most significant, which has 108%, 96%, and 13% improvement compared with branch 1, branch 2, and branch 3 on average, respectively. It also has 5% improvement in accuracy, 4% in specificity, 11% in AUC, and 28% in AUC_SEN_80 on average compared with single modal methods. The complete version is shown in Table SI in the supplementary material.

V. DISCUSSION

While it is very important to achieve accurate results of dermatological diagnosis, it is equally important to explain the results. As we have adopted multimodal data to help diagnose the skin lesion, we can use various ways to show the interpretation through the proposed framework.

After deriving the classification results from IM-CNN, we can also output the interpretable visual analysis with Grad-CAM and SHAP from multimodality. We randomly select seven different diagnostic categories of samples and show their corresponding visual explanation of image and metadata. The visual explanations of seven different kinds of skin lesions are shown in Fig. 4. The heat map of the image highlights the related ROIs that give prominence to the abnormal region in skin lesion images. The highlighted areas can effectively reduce the impact of noise and help dermatologists pay more attention to the discriminative positions. The SHAP value visualization of a MEL patient sample is shown in Fig. 5. It shows the contribution of each feature in metadata. The base value is the mean value of the predicted value of the dataset. Red and blue arrows, respectively, show the positive and negative contributions of each feature to the final prediction result, and push the prediction result from the base value to the final output. Arrows with a longer interval play a more important role than shorter ones. It clearly shows the relationship between various medical indicators and diagnosis results in a quantitative form. The contrast of the patient is 4.526, which is much higher than the mean value 2.39 with a standard deviation of 1.6 and median value 1.94 of contrast. A large contrast value increases the risk of MEL, which also holds for other features listed in Fig. 5. The samples of other categories are shown in Fig. S4 in the supplementary material.

MEL is the most dangerous form in skin lesion diagnosis and is often confused with NV, which is a benign neoplasm of melanocytes. From the view of ROIs, we can find that the shapes and colors of MEL and NV are similar. It is hard to distinguish the MEL from NV by just looking at the image, and additional attributes are needed for distinction. BCC, VASC, and DF share some common features such as pink colors and are often subject to confusion.

With the aid of visual classification, dermatologists can understand the operating mechanism of models and generate trust to the model gradually. Besides, the interpretability visual module shows the deviation between models and dermatologist, which can further improve the performance of models and

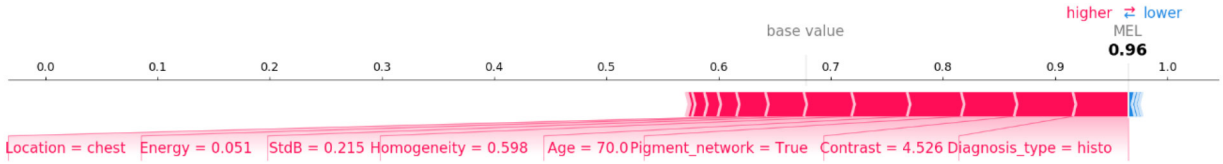

Fig. 5. Indication of the SHAP value visualization of metadata of a MEL sample.

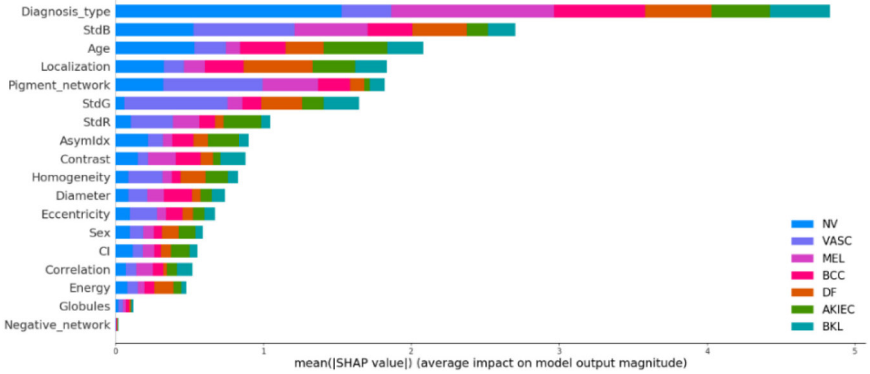

Fig. 6. Feature importance with mean SHAP values.

help dermatologist discover new patterns. Thus, the virtuous circle of man–machine collaboration will emerge.

From Fig. 6, we can discover that age, color (StdR, StdG, and StdB), location, and pigment network play a crucial part in classification of seven skin lesions. Different features have different effects on different skin lesion categories.

VI. Conclusion

In this article, we proposed a novel IM-CNN model for skin lesion diagnosis based on images and meta-data. We embedded a three-branch structure and introduced domain knowledge into the deep learning model, which addresses the limitation of black-box models and improves the interpretability. As far as we know, it is the first time that both multimodal and interpretability have been considered simultaneously. Our experimental results show the promising performance of the proposed model compared to popular models, while its time complexity remains similar. In addition, IM-CNN achieves 72%, 9%, and 21% improvement on average in terms of sensitivity, AUC, and AUC_SEN_80 compared with single modal.

The visual explanations make it easier for dermatologists to understand the operation mechanism of the model, which is helpful for gaining trust of dermatologists and man–machine collaboration. It also facilitates the discovery of new domain knowledge in terms of skin lesion diagnosis from the view of image and metadata.

Future work can be divided into two parts. First, we are going to consider more modalities to improve the data quality and the model performance, such as electronic medical record (EMR) with text mode. Second, it is valuable to compare the proposed model with dermatologists with and without auxiliary of IM-CNN.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, Jan. 2019.

[2] H. W. Lim *et al.*, “The burden of skin disease in the United States,” *J. Amer. Acad. Dermatol.*, vol. 76, no. 5, p. 958–972, May 2017.

[3] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[4] J. R. Hagerty *et al.*, “Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1385–1391, Jul. 2019.

[5] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, “P-CNN: Part-based convolutional neural networks for fine-grained visual categorization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2019, doi: 10.1109/TPAMI.2019.2933510.

[6] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[7] S. Xu, H. K. Chan, E. Ch’ng, and K. H. Tan, “A comparison of forecasting methods for medical device demand using trend-based clustering scheme,” *J. Data Inf. Manage.*, vol. 2, no. 2, pp. 85–94, 2020.

[8] J. Liu, L. Ding, X. Guan, J. Gui, and J. Xu, “Comparative analysis of forecasting for air cargo volume: Statistical techniques vs. machine learning,” *J. Data Inf. Manage.*, vol. 2, no. 4, pp. 243–255, 2020.

[9] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019.

[10] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, “An improved random forest-based rule extraction method for breast cancer diagnosis,” *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105941.

[11] D. Zhang, J. Han, Y. Zhang, and D. Xu, “Synthesizing supervision for learning deep saliency network without human annotation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[12] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, “Seven-point checklist and skin lesion classification using multitask multimodal neural nets,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 538–546, Mar. 2019.

[13] F. Nachbar *et al.*, “The ABCD rule of dermatoscopy,” *J. Amer. Acad. Dermatol.*, vol. 30, no. 4, pp. 551–559, Apr. 1994.

[14] H. Mirzaalian, G. Hamarneh, and T. K. Lee, “A graph-based approach to skin mole matching incorporating template-normalized coordinates,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2152–2159.

[15] C. Barata, M. E. Celebi, J. S. Marques, and J. Rozeira, “Clinically inspired analysis of dermoscopy images using a generative model,” *Comput. Vis. Image Understand.*, vol. 151, pp. 124–137, Oct. 2016.

[16] A. Madooei, M. S. Drew, and H. Hajimirsadeghi, "Learning to detect blue–white structures in dermoscopy images with weak supervision," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 779–786, Mar. 2019.

[17] Y. W. Chen, J. T. Hsu, C. C. Hung, J. M. Wu, F. Lai, and S. Y. Kuo, "Surgical wounds assessment system for self-care," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 12, pp. 5076–5091, Dec. 2020.

[18] H. Mirzaalian, T. K. Lee, and G. Hamarneh, "Hair enhancement in dermoscopic images using dual-channel quaternion tubularness filters and MRF-based multilabel optimization," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5486–5496, Dec. 2014.

[19] T. J. Brinker *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, May 2019.

[20] H. A. Haenssle *et al.*, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018.

[21] A. Hauschild *et al.*, "To excise or not: Impact of MelaFind on German dermatologists' decisions to biopsy atypical lesions," *J. Deutschen Dermatologischen Gesellschaft*, vol. 12, no. 7, pp. 606–614, Jul. 2014.

[22] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, Jan. 2019.

[23] Y. Xia, L. Zhang, L. Meng, Y. Yan, L. Nie, and X. Li, "Exploring Web images to enhance skin disease analysis under a computer vision framework," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3080–3091, Nov. 2018.

[24] S. H. Kassani and P. H. Kassani, "A comparative study of deep learning architectures on melanoma detection," *Tissue Cell*, vol. 58, pp. 76–83, Jun. 2019.

[25] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 1146–1152, May 2015.

[26] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting data with GANs to segment melanoma skin lesions," *Multimedia Tools Appl.*, vol. 79, nos. 21–22, pp. 15575–15592, Jun. 2020.

[27] N. Nida, A. Irtaza, A. Javed, M. H. Yousaf, and M. T. Mahmood, "Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering," *Int. J. Med. Inform.*, vol. 124, pp. 37–48, Apr. 2019.

[28] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[29] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[30] W. Wu *et al.*, "Multimodal vigilance estimation using deep learning," *IEEE Trans. Cybern.*, early access, Oct. 7, 2020, doi: 10.1109/TCYB.2020.3022647.

[31] L. Zhang, Y. Gao, C. Hong, Y. Feng, J. Zhu, and D. Cai, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

[32] E. Debie *et al.*, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, Mar. 2021.

[33] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[34] Z. Yang, N. Wu, Y. Liang, H. Zhang, and Y. Ren, "SMSPL: Robust multimodal approach to integrative analysis of multiomics data," *IEEE Trans. Cybern.*, early access, Jul. 22, 2020, doi: 10.1109/TCYB.2020.3006240.

[35] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, Sep. 2014.

[36] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[37] Y. Chai, H. Liu, and J. Xu, "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models," *Knowl. Based Syst.*, vol. 161, pp. 147–156, Dec. 2018.

[38] T. Li *et al.*, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Jan. 20, 2021, doi: 10.1109/TSMC.2020.3048950.

[39] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[41] T. Russo *et al.*, "Dermoscopy pathology correlation in melanoma," *J. Dermatol.*, vol. 44, no. 5, pp. 507–514, May 2017.

[42] H. Kittler *et al.*, "Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International society of dermoscopy," *J. Amer Acad. Dermatol.*, vol. 74, no. 6, pp. 1093–1106, Jun. 2016.

[43] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[44] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY, USA: Curran Assoc., Dec. 2018.

[45] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2656–2666.

[46] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.

[47] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., May 2017.

[48] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161.

[49] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, 2018, pp. 168–172.

[50] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imag. Graph.*, vol. 71, pp. 19–29, Jan. 2019.

[51] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[52] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, Nov. 2018.

Sutong Wang received the B.S. degree in electronic commerce from the China University of Mining and Technology, Xuzhou, China, in 2017. He is currently pursuing the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China.

He has published several papers in journals, including *Information Sciences* and *Applied Soft Computing*. His current research interests include interpretable machine learning, data mining, and computer-aided medical diagnosis and prognosis, and he is also interested in deep learning and evolutionary computation.

Yunqiang Yin received the Ph.D. degree from Beijing Normal University, Beijing, China, in 2009.

He is currently a Professor with the School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China. He has published more than 80 papers in various international journals, including *Naval Research Logistics*, *Omega*, *European Journal of Operational Research*, *International Journal of Production Research*, *Computers & Operations Research*, *Journal of Scheduling*, and IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS. His research interests are in production and operations management, medical operations management, and machine learning.

Dujuan Wang received the B.S. and M.S. degrees in computer science and technology, and the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China, in 2004, 2007, and 2016, respectively.

She is currently a Professor with the Business School, Sichuan University, Chengdu, China. She has published more than 50 papers in various international journals, including *Naval Research Logistics*, *Omega*, *European Journal of Operational Research*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, *International Journal of Production Economics*, *Computers & Operations Research*, and *Knowledge-Based Systems*. Her research interests are in medical machine learning, service operation management and optimization, and artificial intelligence.

Yanzhang Wang received the M.S. and Ph.D. degrees in management science and engineering from the Dalian University of Technology (DUT), Dalian, China, in 1983 and 1989, respectively.

He was promoted to an Associate Professor and a Professor by breaking a rule in 1990 and 1992, and was authorized by the Degree Council of the State Council as a Doctoral Advisor in 1993. He is the Director of Information and Decision Technology Institute and a Distinguished Professor with the School of Management, DUT. He has (co)authored over 200 peer-reviewed journal and conference papers and books. His current research interests include knowledge management, decision support, and systems engineering.

Yaochu Jin (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in automatic control from the Electrical Engineering Department, Zhejiang University, Hangzhou, China, in 1988, 1990, and 1996, respectively, and the Dr.-Ing. degree in computer science and engineering from Ruhr University Bochum, Bochum, Germany, in 2001.

He is currently a Distinguished Chair and a Professor of Computational Intelligence with the Department of Computer Science, University of Surrey, Guildford, U.K., where he heads the Nature Inspired Computing and Engineering Group. He was a “Finland Distinguished Professor” with the University of Jyväskylä, Finland, a “Changjiang Distinguished Visiting Professor” with Northeastern University, Shenyang, China, and a “Distinguished Visiting Scholar” with the University of Technology Sydney, Ultimo, NSW, Australia. His main research interests include data-driven evolutionary optimization, evolutionary learning, trustworthy machine learning, and morphogenetic self-organizing systems.

Prof. Jin is the recipient of the 2018 and 2021 IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION Outstanding Paper Award, the 2015, 2017, and 2020 *IEEE Computational Intelligence Magazine* Outstanding Paper Award, and the Best Paper Award of the 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. He was the General Co-Chair of the 2016 IEEE Symposium Series on Computational Intelligence and the Chair of the 2020 IEEE Congress on Evolutionary Computation. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS and *Complex & Intelligent Systems*. He was an IEEE Distinguished Lecturer and a Vice President for Technical Activities of the IEEE Computational Intelligence Society. He was named by the Web of Science as “a Highly Cited Researcher” in 2019 and 2020.