International Journal of Scientific Engineering and Science
Volume 10, Issue 3, pp. 59-62, 2026. ISSN (Online): 2456-7361

Multimodal Fusion Skin Disease Classification System Based on Improved EfficientNet Network

Xinyue Fang[1], Yi Zhao[1], Shukuan Sun[1], Yuting Liu[1*]
[1]School of Wannan Medical University, Wuhu, Anhui, China
Email address: liuyuting@wnmc.edu.cn

Abstract*—The incidence and mortality of skin cancer are on the rise. Traditional dermoscopy diagnosis is highly dependent on physicians' experience, time-consuming, and costly; meanwhile, high similarity of skin lesions and sample scarcity result in poor generalization of conventional models. To address these issues, this study proposes a skin disease classification algorithm based on multimodal deep feature fusion and constructs a corresponding classification and recognition system. The algorithm improves the EfficientNet architecture by integrating multi-scale spatial pyramid pooling (SPP) and efficient channel attention (ECA) mechanisms to boost feature extraction. A dual-path fusion structure is designed, which dynamically correlates clinical text and image features via gated attention units, and combines feature concatenation to retain cross-modal complementary information. This work provides a reliable basis for personalized and precise diagnosis and treatment of skin diseases, and offers effective assistance for clinical diagnosis.*

***Keywords*—** *Dermatoscope; multimodal; EfficientNet; skin disease classification.*

I. Introduction

Skin cancer is one of the most prevalent malignant tumors globally, with its incidence and mortality on a steady rise. Statistics show ~3 million new non-melanoma skin cancer cases and 130,000 melanoma cases worldwide annually [1]. Though accounting for only 5% of all skin cancers, melanoma—characterized by high malignancy and metastatic risk—causes 75% of skin cancer deaths, making it the leading cause of mortality from such cancers [2]. Early-stage skin cancers (e.g., basal cell carcinoma, squamous cell carcinoma) are curable via local excision, with a 5-year survival rate over 95% [3]; by contrast, advanced melanoma has a 5-year survival rate below 20% [4]. These stark disparities highlight that early detection and treatment can drastically lower skin cancer mortality, underscoring the critical value of skin disease identification research.

In the field of dermatological classification research, deep learning technology is leading groundbreaking explorations aimed at overcoming the limitations of traditional methods and improving diagnostic performance. Early research frameworks were primarily built upon manual feature engineering and shallow machine learning models: the Celebi [5] team utilized Euclidean distance transformation to segment dermoscopy images, thereby extracting color and texture features to accomplish the classification task; Shoieb [6] and colleagues adopted a multi-class linear Support Vector Machine (SVM) architecture, effectively improving the sensitivity and accuracy of the classification system. However, these methods are generally constrained by technical bottlenecks such as insufficient feature salience and subtle differences between categories.

With the iterative upgrades of computing infrastructure, Convolutional Neural Networks (CNNs) have gradually established their central role in this field. In 2017, Esteva [7] and his team were the first to integrate transfer learning strategies into the GoogLeNet and Inception-v3 models, achieving accuracy rates of 72.1% and 55.4% in 3-class and 9-class skin disease classification tasks, respectively, and surpassing the diagnostic performance of professional dermatologists for the first time. This landmark achievement sparked a surge in multi-classification research: Matsunaga [8] achieved precise differentiation among three disease categories—melanoma, pigmented nevi, and others—through a multi-network ensemble architecture.

Innovations in model architecture and ensemble learning paradigms have garnered significant attention in recent years: the block attention mechanism proposed by Gessert [9] significantly improved the classification performance of high-resolution dermoscopy images; the multi-classifier network constructed by Dai et al. [10] achieved an accuracy of 93.77% in classification tasks for common skin diseases such as vitiligo and acne. Ensemble learning further improves classification accuracy by combining the strengths of different network architectures, but faces practical challenges such as high computational costs and limited generalization capabilities. In the realm of interpretability research, Li et al.'s study, which combines a multi-task learning framework to reveal model decision-making mechanisms, has laid a methodological foundation for building trust in deep learning technologies within clinical settings.

II. Data Set Preparation and Preprocessing

2.1 Selection of Datasets

This paper adopts the PAD-UFES-20 dataset downloaded from Kaggle, a public dataset released in 2020 by the Dermatology and Surgical Assistance Program of the Federal University of Espírito Santo for skin disease classification. It comprises 2,298 dermatoscopic images from 1,373 patients, covering 6 categories: 3 skin cancers (Basal Cell Carcinoma [BCC], Squamous Cell Carcinoma [SCC], Melanoma [MEL]) and 3 benign conditions (Actinic Keratosis [ACK], Seborrheic Keratosis [SEK], Nevus [NEV]). Each image is annotated with category label, patient age, lesion location and lesion diameter.

International Journal of Scientific Engineering and Science
Volume 10, Issue 3, pp. 59-62, 2026. ISSN (Online): 2456-7361

The detailed distribution of disease types is presented in Table I.

TABLE I. PAD-UFES-20: Disease Names and Number of Images

Disease Name	Abbreviation	Number of images
Actinic keratosis	ACK	730
Basal cell carcinoma	BCC	845
Melanoma	MEL	52
Nevus	NEV	244
Squamous cell carcinoma	SCC	192
Seborrheic keratosis	SEK	235

2.2 Data Preprocessing

The PAD-UFES-20 dataset's 21 label dimensions are converted into 81 distinct features for network training in skin disease classification. Data preprocessing adopts one-hot encoding, with missing values denoted as -5 and tailored mappings for different data types.

Dermoscopy images are prone to device noise, surface reflections and uneven lighting, making image preprocessing critical for model efficiency and accuracy. Specific steps are as follows: non-local average filtering is used to remove noise, glare and fine hair while retaining lesion texture details, with parameters set as: 21×21 search window (to capture similar blocks), 7×7 similar block size (for local feature matching), and filter strength h=0.1×σ (where σ is image noise standard deviation, dynamically adjusted per noise level). Beyond standard rotation and cropping, a color correction algorithm is applied. Ambient light-induced color deviations affect image color and classification results, as capture devices only record incoming light signals and cannot restore the lesion's intrinsic color.

III. Selection of Model Algorithms

A hybrid gated attention unit approach is adopted, which integrates patients' clinical information to provide image decision support and addresses key dermatological classification challenges (high inter-class similarity, limited sample size, and underutilized information). This study employs a Meta-EfficientNet multimodal model fusing text and image data: first, D-F-EfficientNet serves as the image backbone to strengthen skin lesion feature extraction; second, a dual-path fusion architecture is designed, which leverages an improved gated attention unit (MetaBlock) for dynamic correlation between patients' clinical text features and image regions, while incorporating global feature concatenation (Concat) to retain cross-modal complementary information.

3.1 Image Feature Extraction

This project employs D-F-EfficientNet for image feature extraction. Based on the EfficientNet-B5 architecture, it incorporates an improved spatial pyramid pooling structure to address the issue of limited receptive fields. By utilizing multi-layer feature fusion, it enhances the distinctiveness of shallow-layer features. The D-F-EfficientNet architecture is shown in Fig.1.

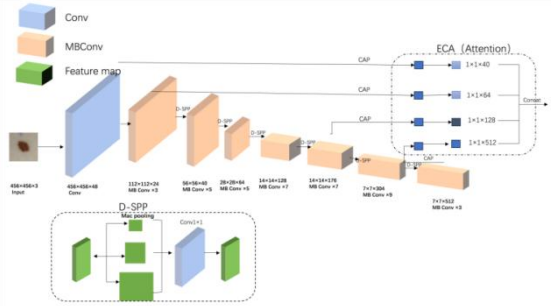

Fig. 1. D-F-EfficientNet Network Architecture Diagram.

3.2Feature Fusion

The feature fusion section proposes a dual-path classifier consisting of an improved MetaBlock network and a concatenation network. Metadata is fed into the network model during training, and two distinct small networks are used to update the weights of the feature maps. This approach provides additional information about lesion regions from multiple levels, effectively leveraging the complementary information from different modalities to improve the accuracy of clinical diagnosis.

(1) Meta-Efficient Networks

It comprises an improved MetaBlock network and a concatenation network. The MetaBlock network serves as a gated attention transformation unit that converts patient metadata into feature vectors for inclusion in network training. By combining this with image feature information, it updates the weights of the feature maps and selects those with strong correlations to the metadata for output; meanwhile, the concatenation network directly integrates the metadata processed through fully connected layers.

Compared to common deep learning models, the Meta-EfficientNet model combines text and image information to construct the network, incorporating preprocessed metadata as part of the network architecture.

(2) MetaBlock Network

MetaBlock is a gated attention mechanism conversion unit that directly maps metadata (such as patient age and gender) to attention weights via fully connected layers. While it features static weight allocation, it suffers from limitations such as the absence of cross-channel interaction. In contrast, D-MetaBlock introduces an adaptive cross-channel interaction mechanism that uses dynamic convolutional kernels to learn fine-grained associations between metadata and image features, thereby significantly improving multimodal fusion performance.

(3) Concatenation Network

A concatenation network is a fundamental method for multimodal feature fusion. It forms a high-dimensional joint feature representation by directly concatenating feature vectors from different modalities (such as image features and text features), followed by classification via fully connected layers.

Meta-EfficientNet adopts a dual-path fusion architecture, which combines the concatenation network and the improved MetaBlock via weighted averaging. The concatenation network features a serial image feature path-based classification structure, forming a complementary image-text framework. For metadata processing, two architecture-adaptive fully connected

International Journal of Scientific Engineering and Science
Volume 10, Issue 3, pp. 59-62, 2026. ISSN (Online): 2456-7361

layers are used for nonlinear mapping; after feature correlation learning, the output is concatenated directly with the image feature map. The layers are followed by batch normalization, ReLU activation, and a Dropout rate of 0.3. The model structure is illustrated in Fig.2.

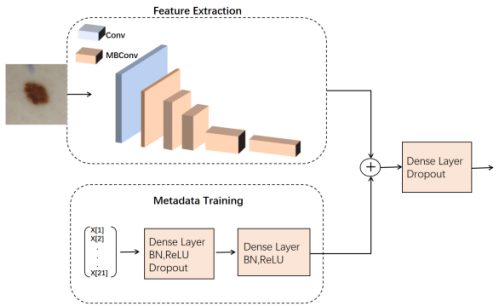

Fig. 2. Feature Fusion Based on Stitching

IV. Analysis of Results

Meta-EfficientNet was trained for 50 epochs. During training, high accuracy on the training set coupled with low loss on the validation set indicates that the model is performing well. As shown in Fig.3, the accuracy curve rises gradually with each iteration, and the final accuracy on the validation set reaches 90.80%; The loss curve continued to decrease with each iteration, with the validation set loss ultimately dropping to 0.190. After approximately 20 iterations (marked as the "Cross Point Epoch" in the figure), the model gradually converged, and the curve demonstrated good training performance.

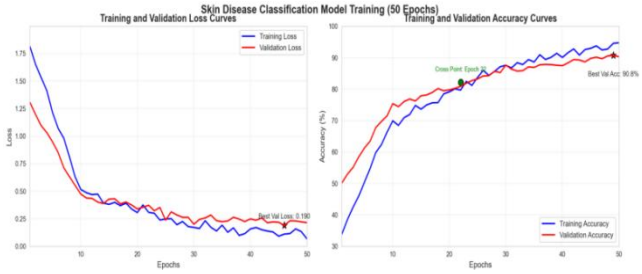

Fig. 3. Network training curve (left) and loss curve (right) Accuracy curve

To comprehensively compare the performance of algorithms in predicting dermatological classifications, this project designed a comparative experiment. This paper employs ensemble learning to compare three different multimodal networks: Concat, MetaBlock, and Meta-EfficientNet. The experimental results for these three networks are shown in Table II, with BACC used as the evaluation metric.

TABLE II. BACC Values for Each Category in the Feature Fusion Network

Types of Skin Diseases	Concat	MetaBlock	Meta-EfficientNet
ACK	0.830	0.851	**0.865**
BCC	0.710	0.877	**0.903**
MEL	0.683	0.882	**0.889**
NEV	0.772	0.920	**0.921**
SCC	0.385	0.252	**0.412**
SEK	0.830	0.788	**0.875**

As shown in Table II, the Meta-EfficientNet network achieves varying degrees of improvement in balanced accuracy across all disease categories compared to other networks. Notably, in the SCC category, Meta-EfficientNet outperforms the MetaBlock network by 15.8% and the Concat network by 2.7%, effectively addressing the MetaBlock network's low recognition rate for SCC and enhancing the network's stability. The Meta-EfficientNet network achieves higher BACC values than the other two networks, demonstrating greater effectiveness. The experimental results indicate that the multimodal feature fusion network proposed in this paper, which integrates image and text data, is both feasible and effective.

V. Conclusion

To address the drawbacks of traditional dermatological diagnosis and single-modal deep learning models, this study proposes an improved multimodal skin disease classification framework based on EfficientNet. Multi-scale SPP and ECA attention are integrated to boost image feature extraction, and a dual-path fusion architecture—combining an enhanced gating attention unit and feature concatenation—is designed to fuse dermatoscopic images with clinical metadata. Experiments on the PAD-UFES-20 dataset show that the proposed Meta-EfficientNet model attains 90.80% validation accuracy, as well as superior BACC, sensitivity and AUC. It effectively mitigates overfitting induced by data imbalance and enhances the recognition of similar lesions, thus enabling reliable AI-assisted diagnosis for primary care, reducing misdiagnosis and supporting the AI+primary healthcare strategy. Future research will focus on dataset expansion and model lightweighting for clinical deployment.

Acknowledgment

We sincerely thank the Undergraduate Innovation and Entrepreneurship Training Program for their support (Grant Nos. 202510368002 and 202510368064), which provided essential resources and a supportive research environment. We also appreciate the financial and technical assistance from the Horizontal Research Project of Wannan Medical University(H202530). We are deeply grateful to our supervisor for professional guidance and continuous support, and to our laboratory and college for providing excellent experimental conditions.

References

[1] Fuduli A, Veltri P, Vocaturo E, et al. Melanoma detection using color and texture features in computer vision systems[J]. Advances in Science, Technology and Engineering Systems Journal, 2019, 4(5): 16-22.

[2] FORSCHNER A，EICHNER F，AMARAL T，et al. Improvementof overall survival in stage IV melanoma patients during 2011-2014：analysis of real-world data in 441 patients of the German CentralMalignant Melanoma Registry(CMMR)[J]. J Cancer Res ClinOncol，2017，143(3)：533-540. DOI：10.1007/s00432-016-2309-y.

[3] Di WANG, Xiaoqi LÜ, Jing LI. Dermoscopic image classification based on multi-scale and three-dimensional interaction feature optimization[J]. Optics and Precision Engineering, 2024, 32(24): 3644.

[4] Wang, L. X., Zhang, L. Inner-Capsule Network for Dermoscopic Image Recognition[J]. Pattern Recognition and Artificial Intelligence, 2020, 37(11): 986–998.

[5] Celebi M E, Kingravi H A, Uddin B, et al. A methodological approach to the classification of dermoscopy images [J]. Computerized Medical imaging and graphics, 2007, 31(6): 362-373.
[6] Shoieb D A, Youssef S M, Aly W M. Computer-aided model for skin diagnosis using deep learning [J]. Journal of Image and Graphics, 2016, 4(2): 122-129.J. L. Author, "Title of paper," to be published.
[7] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. nature, 2017, 542(7639): 115-118.
[8] Matsunaga Y, Ogura Y, Ehama R, et al. Establishment of a mouse skin model of the lichenification in human chronic eczematous dermatitis[J]. British Journal of Dermatology, 2007, 156(5): 884-891.
[9] Gessert N, Sentker T, Madesta F, et al. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting[J].IEEE Transactions on Biomedical Engineering, 2019, 67(2): 495-503.
[10] Dai, C. C., Luan, H. J., Yang, X. Y., et al. Research on Multi-Binary Classifiers for Skin Disease Diagnosis Based on Convolutional Neural Network[J]. High Technology Letters, 2023, 32(10): 1025–1035.A. Harrison, private communication, 1995.