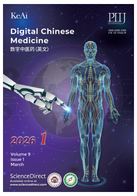
Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed



Multi-label fundus disease classification using dual-branch deep learning: an intelligent diagnosis framework inspired by traditional Chinese medicine Five Wheels theory

Xin He[a], Xiaohui Li[a], Jun Peng[b], Lei Sun[a], Dan Shu[a], Li Xiao[c], Qinghua Peng[c*], Xiaoxia Xiao[a*]

a. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

b. Ophthalmology and Otolaryngology, The First Hospital of Hunan University of Chinese Medicine, Changsha, Hunan 410007, China

c. School of Traditional Chinese Medicine, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

ARTICLE INFO

Article history
Received 04 September 2025
Accepted 14 January 2026
Available online 25 March 2026

Keywords
Multi-label classification
Fundus images
Deep learning
Dual-branch network
Traditional Chinese medicine ophthalmology
Five Wheels theory

ABSTRACT

Objective To develop a dual-branch deep learning framework for accurate multi-label classification of fundus diseases, addressing the key limitations of insufficient complementary feature extraction and inadequate cross-modal feature fusion in existing automated diagnostic methods.

Methods The fundus multi-label classification dataset with 12 disease categories (FMLC-12) dataset was constructed by integrating complementary samples from Ocular Disease Intelligent Recognition (ODIR) and Retinal Fundus Multi-Disease Image Dataset (RFMiD), yielding 6 936 fundus images across 12 retinal pathology categories, and the framework was validated on both FMLC-12 and ODIR. Inspired by the holistic multi-regional assessment principle of the Five Wheels theory in traditional Chinese medicine (TCM) ophthalmology, the dual-branch multi-label network (DBMNet) was developed as a novel framework integrating complementary visual feature extraction with pathological correlation modeling. The architecture employed a TransNeXt backbone within a dual-branch design: one branch processed red-green-blue (RGB) images to capture color-dependent features, such as vascular patterns and lesion morphology, while the other processed grayscale-converted images to enhance subtle textural details and contrast variations. A feature interaction module (FIM) effectively integrated the multi-scale features from both branches. Comprehensive ablation studies were conducted to evaluate the contributions of the dual-branch architecture and the FIM. The performance of DBMNet was compared against four state-of-the-art methods, including EfficientNet Ensemble, transfer learning-based convolutional neural network (CNN), BFENet, and EyeDeep-Net, using mean average precision (mAP), F1-score, and Cohen's kappa coefficient.

Results The dual-branch architecture improved mAP by 15.44 percentage points over the single-branch TransNeXt baseline, increasing from 34.41% to 44.24%, and the addition of FIM further boosted mAP to 49.85%. On FMLC-12, DBMNet achieved an mAP of 49.85%, a Cohen's kappa coefficient of 62.14%, and an F1-score of 70.21%. Compared with BFENet (mAP: 45.42%, kappa: 46.64%, F1-score: 71.34%), DBMNet outperformed it by 4.43 percentage points in mAP and 15.50 percentage points in kappa, while BFENet achieved a marginally higher

*Corresponding author: Xiaoxia Xiao, E-mail: amily_x@hnucm.edu.cn. Qinghua Peng, E-mail: pqh410007@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2026.02.007

Citation: HE X, LI XH, PENG J, et al. Multi-label fundus disease classification using dual-branch deep learning: an intelligent diagnosis framework inspired by traditional Chinese medicine Five Wheels theory. Digital Chinese Medicine, 2026, 9(1): 80-90.

F1-score. On ODIR, DBMNet achieved an F1-score of 85.50%, comparable to state-of-the-art methods.

Conclusion DBMNet effectively integrates RGB and grayscale visual modalities through a dual-branch architecture, significantly improving multi-label fundus disease classification. The framework not only addresses the issue of insufficient feature fusion in existing methods but also demonstrates outstanding performance in balancing detection across both common and rare diseases, providing a promising and clinically applicable pathway for standardized, intelligent fundus disease classification.

1 Introduction

Traditional Chinese medicine (TCM) ophthalmology is founded upon a holistic diagnostic paradigm for fundus examination. Central to this is the classical Five Wheels theory, which systematically maps specific fundus regions to corresponding Zang-organ systems (liver, heart, spleen, lung, and kidney), indicating the comprehensive assessment of multiple visual manifestations across the entire fundus for syndrome differentiation [1]. In clinical practice, TCM practitioners simultaneously evaluate vascular morphology, retinal pigmentation, lesion distribution, and other subtle signs to identify complex syndrome patterns such as blood stasis, Qi deficiency, or Yin deficiency. However, the inherent complexity and subjective nature of this framework poses significant challenges to its standardization and objectification, thereby limiting its integration into contemporary clinical workflows and quality control systems.

Advancing the objectification of TCM ophthalmology constitutes a critical step toward overcoming these limitations. Computational approaches, particularly deep learning-based image analysis, offer a promising avenue for the objective identification and quantification of fundus manifestations. Automated diagnostic systems capable of providing standardized, reproducible assessments of multiple coexisting pathological features can facilitate both the preservation of TCM diagnostic principles and their integration with evidence-based medical practice. Nevertheless, the application of artificial intelligence (AI) to TCM ophthalmology demands technical frameworks that can encapsulate the holistic, multi-regional assessment emphasized by classical theory—a requirement that aligns closely with multi-label classification paradigms in modern medical imaging [2].

In contemporary ophthalmology, fundus diseases increasingly present as complex conditions with multiple coexisting pathologies, a trend exacerbated by aging populations and growing prevalence of chronic systemic diseases, such as diabetes and hypertension [3]. Single-disease classification models are thus becoming inadequate for clinical needs, as patients frequently present with concurrent pathologies, such as diabetic retinopathy combined with macular edema or glaucoma combined with cataract. This underscores the necessity for multi-label classification approaches that can simultaneously detect and differentiate multiple pathological features within a single fundus image.

Current research in automated multi-label fundus diagnosis is primarily dominated by three technical paradigms. Convolutional neural networks (CNNs), including EfficientNet and ResNet architectures, extract hierarchical features through deep layers; representative works have employed multi-label classification frameworks [4], automatic lesion detection systems [5], deep neural networks for multi-class diagnosis [6, 7], transfer learning-based approaches [8], knowledge distillation strateg ies [9], dual-branch designs with bilateral fundus images [10], attention mechanisms and feature fusion [11, 12], discriminative convolution networks for imbalanced datasets [13], and attention-guided image enhancement [14], as well as multi-modal data fusion [15]. Transformer-based models leverage self-attention to capture global dependencies across the fundus image [16, 17], proving particularly beneficial for diseases with spatially distributed manifestations. Recent work has further extended this paradigm through multi-modality learning with semantic dictionary [18]. Graph convolutional networks (GCNs) explicitly model label co-occurrence patterns and disease correla tions [19, 20], achieving robust performance by integrating learned relationships with visual features. Despite these advancements, each paradigm exhibits distinct limitations: CNNs struggle with modeling global context information, Transformers require substantial computational resources and large-scale datasets, and GCNs depend on reliable co-occurrence statistics that are often unavailable for rare diseases.

Current multi-label fundus diagnosis methods are confronted with five critical limitations. First, most approaches exclusively process red-green-blue (RGB) images, neglecting the complementary diagnostic information contained within several specific spectral channels, particularly the green channel, which indicates superior contrast for vascular structures and hemorrhages, both of which are critical for detecting diabetic retinopathy and blood stasis patterns in TCM ophthal-mology. Second, existing architectures inadequately integrate local and global features: microaneurysms require fine-grained detail, while glaucomatous changes demand global morphometric assessment. Third, dual-branch methods employ simplistic late fusion strategies without facilitating deep feature interaction, precluding the acquisition of complementary feature representations. Fourth, severe class imbalance leads to suboptimal model performance for rare diseases, a challenge further exacerbated when attempting to detect subtle TCM syndrome manifestations. Fifth, limited cross-dataset generalizability restricts clinical deployment across diverse imaging protocols and patient populations.

These technical challenges directly impede the application of computational methods to TCM ophthalmology. The Five Wheels theory emphasizes on holistic, multiregional fundus assessment, which requires the simultaneous evaluation of vascular patterns (optimally captured in the green channel), lesion pigmentation (requiring RGB information), and spatial feature distribution—capabilities that are absent in current single-modality architectures. TCM syndrome differentiation further demands the integration of subtle, spatially distributed signs; for instance, distinguishing blood stasis from Qi deficiency requires recognizing the specific combinations of hemorrhages, vascular tortuosity, and retinal discoloration across different fundus regions. Without systematic integration of complementary visual modalities and effective feature interaction mechanisms, computational tools cannot adequately support the standardization and objectification of TCM diagnostic frameworks.

To address these limitations, this study developed a novel framework, namely the dual-branch multi-label network (DBMNet), which integrated TCM's holistic diagnostic principles with modern deep learning for multilabel fundus pathology detection. Inspired by the emphasis of Five Wheels theory on comprehensive multi-regional assessment, DBMNet employs a dual-branch architecture that processes RGB and grayscale (derived from green channel) images to capture complementary chromatic and morphological information, which is analogous to TCM's integrated observation of vascular patterns and retinal coloration for syndrome differentiation. We further designed a feature interaction module (FIM) to enable deep cross-modal feature fusion at multiple scales, ensuring effective integration of local lesion details and global structural patterns.

A TransNeXt-based dual-branch architecture that integrates complementary visual modalities through parallel feature extraction was constructed for addressing the information insufficiency of single-modality approaches. The FIM was designed to facilitate effective local-global feature interaction between the color and grayscale branches, enhancing the network's representational capacity for spatially diverse pathological manifestations. Through extensive experiments on the FMLC-12 and ODIR datasets, we demonstrated superior performance in multi-label classification while providing interpretable feature visualizations relevant to both modern clinical screening and TCM ophthalmology assessment.

2 Data and methods

2.1 Multi-label fundus datasets and preprocessing

Two primary publicly available datasets (ODIR [21] and RFMiD [22]) for multi-label classification of fundus images are utilized. The ODIR dataset comprised fundus images from a large-scale, multi-center clinical study involving 487 hospitals across 26 provinces in China, and was derived from a source pool of over 1.6 million images maintained in a private clinical database, with a final set of 10 000 high-quality fundus images (from both eyes of 5 000 patients) selected. These images were annotated for eight categories: Normal (N), Diabetic Retinopathy (D), Glaucoma (G), Cataract (C), Age-related Macular Degeneration (A), Hypertension (H), Myopia (M), and Others (O) (Table 1). Annotations were performed by six ophthalmologists, ensuring high-quality labels through a rigorous adjudication process. The RFMiD dataset consisted of 3 200 fundus images with expert annotations, covering 45 distinct disease categories, including N, D, A, Drusen (DN), M, Branch Retinal Vein Occlusion (BRVO), Tessellation (TSLN), Optic Disc Cupping (ODC), and Optic Disc Edema (ODE) as representative examples (Table 2). The dataset was split into three subsets for model development: 60% (1 920 images) for training, 20% (640 images) for validation, and 20% (640 images) for testing.

Table 1 The number of images for each category label on ODIR dataset

Label	Number of images
N	2 876
D	1 800
G	326
C	313
A	280
H	193
M	267
O	1 188

Table 2 The number of images for representative category labels on RFMiD dataset

Label	Number of images
N	669
D	632
A	169
DN	230
M	167
BRVO	119
TSLN	304
ODC	445

Some public multi-label fundus datasets had significant shortcomings. Both ODIR and RFMiD predominantly contained single-label images, with limited examples of co-occurring diseases. Class imbalance in each dataset constrained their utility for complex lesion patterns. To better utilize available data, we constructed FMLC-12 by integrating complementary samples from the ODIR and RFMiD datasets.

We integrated the datasets in four steps. First, all disease categories common to both datasets were retained for label consistency, including N, D, A, and M. Second, categories from RFMiD with at least 100 samples were included for ensuring sufficient training data and mitigating class imbalance. Categories such as N (230 samples), BRVO (119 samples), TSLN (304 samples), and ODC (445 samples) were added. Although ODE had only 96 samples, it was included due to its high clinical relevance and frequent co-occurrence with other pathologies. Third, exclusive categories from ODIR, including G, C, and H, were retained due to sufficient sample sizes (over 190 images each). Fourth, additional images from the O category of ODIR were reclassified based on diagnostic keywords in the metadata, adding 347 images to five newly introduced categories. Specifically, D (319 images), BRVO (150 images), TSLN (431 images), ODC (521 images), and ODE (120 images) were updated, while 841 images with insufficient samples were excluded.

All images were screened for quality (requiring a minimum 512 × 512 pixel resolution) and adequate contrast for visualizing the key structures, such as the optic disc, vessels, macula, and the absence of severe artifacts. For shared categories, we adopted existing labels from both datasets. RFMiD provided pathology descriptions in its annotation files, which we used to map images to corresponding categories. ODIR images in the O category were reassigned based on diagnostic keywords in ODIR's metadata.

The final FMLC-12 dataset comprised 6 936 unique fundus images across 12 disease categories constructed by merging and de-duplicating the selected and reclassified images from both ODIR and RFMiD sources. As shown in Table 3, class distribution ranged from 2 272 images for D to 120 images for ODE with a ratio of 19 : 1.

In multi-label datasets, the statistics in Table 1 – 3 represent per-category image counts, where individual images may be assigned to multiple disease categories simultaneously. Consequently, the sum of counts across all categories exceeds the total number of unique images in each dataset. Additionally, the label of disease categories employ their respective official labeling conventions of datasets, which are preserved in this study to maintain consistency with the source datasets. The publicly available datasets used in this study, ODIR and RFMiD, can be obtained through the official ODIR website at https://odir2019.grand-challenge.org/ and via the doi link: 10.3390/data6020014, respectively.

Table 3 The number of images for each category label on FMLC-12 dataset

Label	Number of images
N	1 592
D	2 272
G	326
C	313
A	449
H	193
M	435
DN	408
BRVO	151
TSLN	333
ODC	445
ODE	106

2.2 TransNeXt-based heterogeneous branch network

To address the challenges of feature insufficiency and cross-branch information limitation in fundus image classifications, a TransNeXt-based dual-branch model, called DBMNet, was proposed. The architecture was conceptually inspired by the Five Wheels theory of TCM ophthalmology, which emphasized the assessment of fundus regions and visual manifestations for holistic diagnosis. Similar to how TCM practitioners simultaneously evaluate the vascular patterns, tissue coloration, and structural morphology across different fundus regions, the DBMNet systematically integrated the complementary visual information from RGB and grayscale images to capture both chromatic features and morphological characteristics, which were essential for accurate multi-label pathology classification. This architecture implemented cross-branch information fusion incorporating an FIM, with its overall structure illustrated in Figure 1.

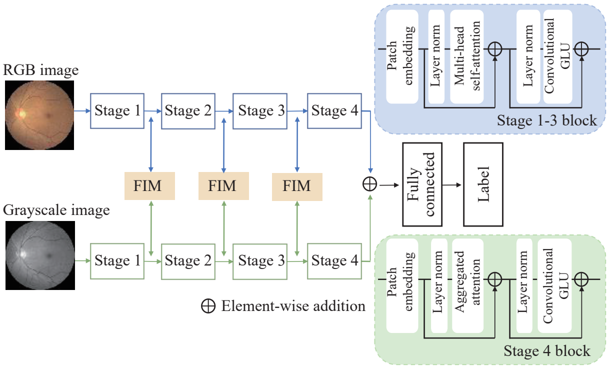


Figure 1 DBMNet model architecture diagram

GLU, gated linear unit.

TransNeXt [23], as an advanced Transformer variant, demonstrated significant advantages over CNNs in global information modeling. Its core innovation lies in combining pixel-level attention with aggregated attention mechanisms to effectively mitigate the local information fragmentation caused by patch embedding in Vision Transformers (ViT). The structure of TransNeXt is shown in Figure 2.

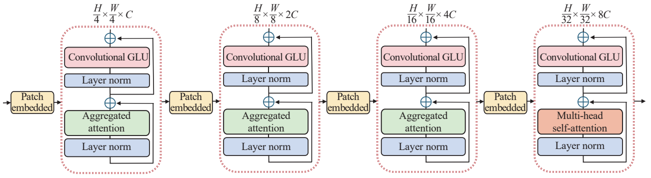


Figure 2 TransNeXt model architecture diagram

TransNeXt processed an input feature map through four stages. The first three stages employed an aggregated attention mechanism through three steps. First, the depthwise separable convolution (DSC) is applied to the input features X to obtain local representations X_1'. Second, global average pooling captures channel-wise context, which is normalized by a sigmoid activation function σ to obtain X_g. Finally, this global context is combined with local features via element-wise multiplication, creating a representation that balances both local and global information. The aggregated attention computation can be mathematically expressed as:

$$X_1' = \mathrm{DSC}(X) \in \mathbb{R}^{H\times W\times C} \tag{1}$$

$$X_g = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot GAP(X))) \in \mathbb{R}^{1\times 1\times C} \tag{2}$$

$$Y_{\mathrm{agg}} = X_1' \odot X_g + X \tag{3}$$

Here, H, W, and C denote the height, width, and number of channels of the feature map, respectively. GAP(X) represents the global average pooling. $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$, which are weight matrices with a reduction ratio $r = 16$. The symbol $\odot$ denotes element-wise multiplication.

The features processed through the first three stages were then fed into the fourth stage, where TransNeXt employs the multi-head self-attention mechanism to capture global contextual relationships, culminating in a comprehensive feature representation.

The dual-branch architecture processed two types of input images to capture complementary diagnostic information. This design aligned with the holistic assessment principle in TCM ophthalmology, where practitioners integrated observations of both color-related manifestations and structural patterns. The color information and texture features in fundus images were preserved in RGB images, which is important for classifying the pigmentation-related lesions, such as macular degeneration and hemorrhagic lesions. In TCM diagnostic terminology, color assessment is crucial for identifying patterns, such as blood stasis (characterized by dark red discoloration and hemorrhages) or Qi deficiency (manifesting as retinal pallor). The color distribution and contrast in RGB images provided essential complementary information for lesion characterization, enabling the model to identify and differentiate these regions more accurately.

In contrast, grayscale images demonstrated clearer structural features, including the optic disc, vascular network, and macular region. By removing redundant color information, grayscale images highlighted local texture and morphological details, which made them especially suitable for classifying structural abnormalities, such as optic disc edema and retinal detachment. In TCM practice, this focus on morphology is crucial for assessing vascular patterns to identify specific conditions, such as liver Yang rising (associated with vascular tortuosity), or to assess the structural integrity related to kidney essence deficiency. The integration of these two complementary information sources via DBMNet enabled comprehensive fundus assessment analogous to TCM's multifaceted observational approach.

2.3 Feature interaction module

To fully leverage the complementary nature of dual-branch features and enhance the model's ability to detect complex disease patterns, an FIM was designed for promote information sharing and fusion between the two branches. FIM served as a critical function analogous to TCM's integrative diagnostic process, where observations from different modalities (color, structure, and distribution) are synthesized into comprehensive pattern assessment. Rather than processing RGB and grayscale features independently, FIM enables explicit cross-modal feature interaction, allowing the network to learn complementary fused representations that capture both chromatic and morphological characteristics simultaneously. The structure of FIM is shown in Figure 3. At its core, the FIM employed a large selection kernel (LSK), which is particularly effective at modeling local details and global contextual information.

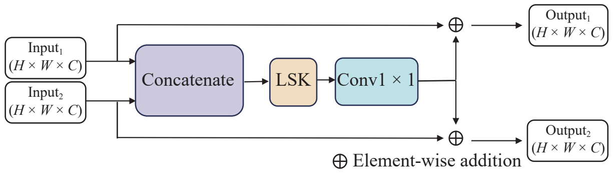

Input$_1$ $(H \times W \times C)$
Input$_2$ $(H \times W \times C)$
Concatenate
LSK
Conv1 × 1
Output$_1$ $(H \times W \times C)$
Output$_2$ $(H \times W \times C)$
⊕ Element-wise addition

Figure 3 Feature interaction module structure

As illustrated in Figure 4, LSK adopted a dual-pathway architecture for multi-scale feature extraction through parallel processing. The design rationale was to capture both fine-grained local details and broader contextual information simultaneously, which is crucial for accurate fundus disease classification. This multi-scale approach reflected the TCM's diagnostic principle for assessing the fine lesion details (e.g., microaneurysms) and global structural patterns (e.g., vascular distribution). Given the concatenated input feature map $X_{\mathrm{in}} = \mathrm{Cat}\,(F_{\mathrm{RGB}}, F_{\mathrm{Gray}}) \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 2C}$, the LSK performed dual-pathway feature extraction. Path A applied a 5 × 5 DSC (stride = 1, padding = 2) to extract detailed local information. The 5 × 5 kernel size offered a sufficient local receptive field while maintained computational efficiency. This operation is formulated as:

$$X_{\mathrm{A}} = \mathrm{DSC}_{5\times5}(X_{\mathrm{in}}) \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 2C} \tag{4}$$

$$X_{\mathrm{A}}(i,j,c) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} X_{\mathrm{in}}(i+m, j+n, c)\cdot K_{\mathrm{A}}(m+2, n+2, c) \tag{5}$$

where i and j are spatial coordinates, m and n are kernel offsets, and c is the channel index.

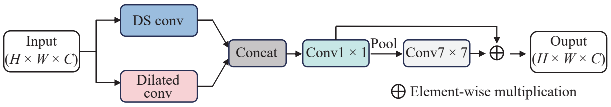


Figure 4 LSK structure

DS conv, depthwise separable convolution.

Path B incorporated a dilated convolution (DC) with a kernel size of 7 × 7 (stride = 1, padding = 9) and dilation rate of 3 to expand the receptive field. This configuration effectively captured long-range dependencies while avoiding excessive parameter increase. The operation is defined as:

$$X_{\mathrm{B}} = \mathrm{DC}_{7\times7,d=3}(X_{\mathrm{in}}) \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 2C} \tag{6}$$

$$\mathrm{X}_{\mathrm{B}}(i,j,c) = \sum_{m=-3}^{3}\sum_{n=-3}^{3} X_{\mathrm{in}}(i+3m, j+3n, c)\cdot K_{\mathrm{B}}(m+3, n+3, c) \tag{7}$$

Feature fusion was first performed by concatenating the outputs from two pathways (A and B) along with the channel dimension, thereby constructing a comprehensive feature representation. To reduce computational overhead while preserving critical feature information, we achieved channel dimensionality reduction via two consecutive 1 × 1 convolutions (Conv1: $4C \rightarrow 2C$, stride = 1; Conv2: $2C \rightarrow C$, stride = 1). The module incorporated a global feature enhancement mechanism, applying both global average pooling and global maximum pooling to the dimensionally-reduced features for capturing contextual information from perspectives of different statistical properties. These were subsequently concatenated to form a comprehensive global descriptor. The entire FIM computation process can take the RGB images and their corresponding grayscale images as inputs, with details described as follows.

$$F_{\mathrm{Cat}} = \mathrm{Concat}(F_{\mathrm{RGB}}, F_{\mathrm{Gray}}) \in \mathbb{R}^{H\times W\times 2C} \tag{8}$$

$$F_{\mathrm{LSK}} = \mathrm{LSK}(F_{\mathrm{Cat}}) \in \mathbb{R}^{H\times \mathrm{W}\times 2C} \tag{9}$$

$$F_{\mathrm{Conv}} = \mathrm{Conv}_{1\times1}(F_{\mathrm{LSK}}) \in \mathbb{R}^{H\times W\times C} \tag{10}$$

$$\mathrm{Output}_1 = F_{\mathrm{RGB}} + F_{\mathrm{Conv}} \tag{11}$$

$$\mathrm{Output}_2 = F_{\mathrm{Gray}} + F_{\mathrm{Conv}} \tag{12}$$

Here, $\mathrm{Conv}_{1\times1}$ denotes a convolutional layer in the dimensionality reduction stack, with a kernel size of 1×1, a stride of 1, and an output channel dimension of C. This operation facilitated channel-wise feature fusion while preserving the spatial resolution. The subsequent residual connection adds the original branch features to the convolved output, thereby enriching each modality with cross-modal information while retaining their distinct original characteristics, which also aids in gradient flow during training.

In summary, the FIM processed the concatenated dual-branch features through the LSK module (which performed multi-scale extraction and global enhancement) and then distributed the refined, shared information back to both branches via residual connections. By enabling deep, explicit interaction between color-based and structure-based features, the FIM achieved a comprehensive feature integration that mirrors TCM's holistic approach of synthesizing diverse visual observations. Within the DBMNet, FIM modules can be deployed at multiple hierarchical levels to facilitate multi-scale feature interaction.

2.4 Experimental setup

All experiments were implemented in PyTorch 2.3.0 and conducted on an NVIDIA GeForce RTX 3090 GPU (24 GB). Models were trained for 100 epochs with a batch size of 16. The AdamW optimizer was applied with an initial learning rate of 1 × 10^{-4}. The learning rate was scheduled using a cosine annealing warm restarts scheduler. For data augmentation, we applied random horizontal flipping, random rotation (± 15 °), and color jittering. All input images were resized to 512 × 512 pixels and normalized with the ImageNet mean and standard deviation (SD).

2.5 Ablation study and comparison study design

The systematic ablation experiments were conducted to evaluate the contribution of each component. First, different backbone architectures including DenseNet121, DenseNet201, ResNet101, and TransNeXt were compared for multi-label classification on the ODIR dataset. These architectures represented different design paradigms: DenseNet for dense connectivity, ResNet for residual learning, and TransNeXt for modern Transformer-based approaches. Second, on the FMLC-12 dataset, the impact of the dual-branch structure and FIM module was evaluated by progressively adding components to the baseline TransNeXt model. Input modality ablation was also performed to assess different input configurations: using RGB paired with red, blue, or green channel, as well as using a standalone grayscale branch.

The methods used in this study were compared against state-of-the-art methods, including EfficientNet Ensemble [4], EyeDeep-Net [6], Transfer learning-based CNN [8], and BFENet [10], using a consistent evaluation framework on both the FMLC-12 and ODIR datasets. For the FMLC-12 dataset, all comparison methods were retrained from scratch using their publicly available implementations with default hyperparameters, under identical experimental settings, including the same data splits, preprocessing procedures (512 × 512 pixels resizing and ImageNet normalization), and evaluation protocols. For the ODIR dataset, the performance results of the compared methods were directly taken from their original publications, as retraining under fully identical experimental conditions was not feasible due to differences in data preprocessing, label definitions, or the unavailability of complete training details.

2.6 Evaluation metrics

For multi-label classification, we adopted mean average precision (mAP), F1-score, and Cohen's kappa coefficient as the evaluation metrics. To handle class imbalance in these datasets, precision, recall, and F1-score were computed via micro-averaging, each instance-label pair was treated as an independent binary decision, with true positives (TP), false positives (FP), and false negatives (FN) aggregated globally across all samples and categories before calculating the final metrics. For Cohen's kappa, we used macro-averaging: first computing the per-label binary kappa score for each category, then averaging across all categories to reflect per-class agreement beyond chance. These micro-averaged metrics were robust to class frequency by weighting each prediction equally, while macro-averaged kappa highlights class-level consistency. The specific calculation formulas are as follows:

$$\text{Precision} = \frac{\sum_{i=1}^{C} \text{TP}_i}{\sum_{i=1}^{C} (\text{TP}_i + \text{FP}_i)} \tag{13}$$

$$\text{Recall} = \frac{\sum_{i=1}^{C} \text{TP}_i}{\sum_{i=1}^{C} (\text{TP}_i + \text{FN}_i)} \tag{14}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}(i) \tag{16}$$

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{17}$$

3 Results

3.1 Backbone network selection

The different backbone architectures on the ODIR dataset were compared (Table 4). TransNeXt significantly outperformed other backbones across all metrics, achieving an mAP of 66.29%, a kappa of 74.06%, and an F1-score of 77.55%. Based on these results, TransNeXt functioned as the backbone network.

Table 4 Performance comparison of different baseline models on the ODIR dataset

Backbone	mAP (%)	Kappa (%)	F1-score (%)
DenseNet121	42.79	48.88	64.10
DenseNet201	45.65	53.51	65.33
ResNet101	43.80	47.26	56.37
TransNeXt	66.29	74.06	77.55

3.2 Ablation study results

The ablation study results on the FMLC-12 dataset were summarized in Table 5, indicating the contributions of each component. The baseline TransNeXt achieved an mAP of 34.41%, a kappa of 40.77%, and an F1-score of 60.51%. Adding the dual-branch structure improved performance to 44.24% mAP, 55.10% kappa, and 62.45% F1-score. Finally, the complete architecture with FIM further improved to 49.85% mAP, 62.14% kappa, and 70.21% F1-score. Compared with the baseline TransNeXt, the complete model achieved a 15.44 percentage points increase in mAP and a 21.37 percentage points increase in the Cohen's kappa coefficient.

The input modality ablation results were compared across different channel combinations (Table 6). Among individual color channels, the green channel performed best with 45.44% mAP. In contrast, the grayscale modality achieved the highest performance across all metrics with 49.85% mAP, 62.14% kappa, and 70.21% F1-score.

Table 5 Ablation studies of DBMNet components on the FMLC-12 dataset

Model configuration	mAP (%)	Kappa (%)	F1-score (%)	Params (M)	Inference time (ms)
TransNeXt	34.41	40.77	60.51	89.72	69.70
TransNeXt + dual-branch	44.24	55.10	62.45	184.33	158.36
TransNeXt + dual-branch + FIM	49.85	62.14	70.21	190.82	161.14

Table 6 Input modality ablation on the FMLC-12 dataset

Input configuration	mAP (%)	Kappa (%)	F1 (%)
RGB + red channel	42.92	53.14	65.56
RGB + blue channel	44.28	55.03	66.23
RGB + green channel	45.44	56.64	70.91
RGB + grayscale	49.85	62.14	70.21

3.3 Comparison with state-of-the-art methods

The performance of the DBMNet was compared with state-of-the-art methods on the FMLC-12 and ODIR datasets (Table 7 and 8). For FMLC-12, all comparison results were obtained by re-implementing the corresponding methods under a consistent experimental setting. For ODIR, although the official evaluation platform provides multiple metrics, the binary F1-score was the only metric that could be reliably compared across methods due to the limitations in reproducible outputs. Therefore, Table 8 reports only the binary F1-score and includes fewer methods, as only a limited number of published approaches provide results that are verifiable under the same evaluation protocol. On FMLC-12, DBMNet achieved the highest mAP of 49.85% and kappa of 62.14%. BFENet achieved a slightly higher F1-score of 71.34% than DBMNet (70.21%), but considerably lower mAP and kappa. On ODIR, DBMNet achieved an F1-score of 85.50%, slightly higher than that of the Transfer learning-based CNN (85.30%), but lower than that of BFENet (89.20%).

Table 7 Performance comparison of different methods on the FMLC-12 dataset

Method	mAP (%)	Kappa (%)	F1-score (%)
EfficientNet Ensemble	43.44	56.60	65.58
Transfer learning-based CNN	42.48	50.59	64.67
BFENet	45.42	46.64	71.34
EyeDeep-Net	41.15	45.02	66.20
DBMNet	49.85	62.14	70.21

Table 8 Performance comparison of different methods on the ODIR dataset

Method	F1-score (%)
Transfer learning-based CNN	85.3
BFENet	89.2
DBMNet	85.5

3.4 Visualization analysis of model attention on fundus lesion regions

Figure 5 shows the heatmap comparisons between DBMNet and other approaches. DBMNet demonstrated more concentrated attention on lesion regions, with particularly clear hotspots in tessellation regions and diabetic retinopathy detection. In contrast, the heatmaps of EfficientNet Ensemble showed more dispersed attention patterns, and BFENet showed weaker attention on critical features.

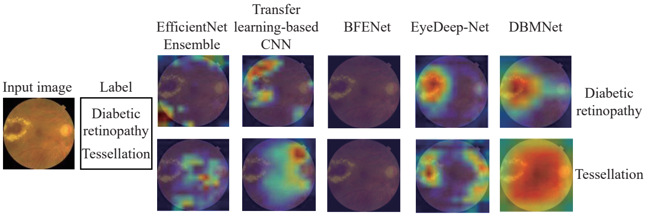


Figure 5 Visualization of attention heatmaps for diabetic retinopathy and tessellation lesion detection across different methods

4 Discussion

4.1 Balanced multi-class performance and clinical significance

This study addresses a key challenge in automated fundus diagnosis: achieving balanced performance across coexisting pathologies under severe class imbalance. On the complex FMLC-12 dataset, the proposed DBMNet achieved an mAP of 49.85% and a Cohen's kappa coefficient of 62.14%, demonstrating competitive overall accuracy while maintaining superior agreement across disease categories.

A notable finding is that DBMNet substantially outperformed BFENet in terms of Cohen's kappa coefficient (62.14% vs. 46.64%), despite a slightly lower F1-score (70.21% vs. 71.34%). This divergence highlights an important limitation of relying solely on aggregate metrics such as F1-score in multi-label medical diagnosis [24]. BFENet's high F1-score but low kappa suggests strong performance on dominant classes but insufficient sensitivity to rare diseases. In contrast, the higher kappa achieved by DBMNet indicates more equitable performance across both common and low-frequency pathologies, including clinically critical conditions such as branch retinal vein occlusion and optic disc edema.

From a clinical perspective, this balanced sensitivity is particularly important. Rare diseases, although less frequent, often carry severe consequences if missed. Diagnostic systems optimized only for dominant classes may therefore introduce unacceptable clinical risk. By prioritizing balanced multi-class performance, DBMNet is better aligned with real-world screening requirements, where reliable detection across the full disease spectrum is essential.

4.2 Architectural contributions to robust multi-label classification

The balanced performance of DBMNet can be attributed to its architectural design. The dual-branch structure processes complementary visual information: the RGB branch captures chromatic cues relevant to lesions such as exudates and drusen, while the grayscale branch emphasizes structural and vascular patterns important for hemorrhages and vascular abnormalities [25]. Ablation experiments confirmed that grayscale processing outperformed individual color channels, indicating its effectiveness in enhancing structural contrast across diverse lesion types.

Furthermore, the FIM enables explicit cross-branch feature fusion, allowing the network to learn synergistic representations that are difficult to obtain through single-branch or late-fusion strategies. In addition, the global attention mechanism of the TransNeXt backbone facilitates modeling long-range spatial dependencies across the entire fundus image. This capability is particularly beneficial for multi-label scenarios, where multiple pathologies may appear in spatially distant regions yet reflect related pathological processes.

4.3 Cross-dataset generalization and robustness

Cross-dataset evaluation revealed important insights into model robustness. On ODIR (8 categories, 18.6% multi-label rate), DBMNet achieved competitive F1-score of 85.50%. However, its advantage became more pronounced on the more complex FMLC-12 dataset (12 categories, 23.40% multi-label rate). When transferring from ODIR to FMLC-12, DBMNet exhibited smaller performance degradation (15.29%) compared with BFENet (17.86%), representing 2.57 percentage points less performance loss. More critically, DBMNet maintained substantially higher Cohen's kappa coefficient of 62.14% versus BFENet's 46.64% on FMLC-12, indicating superior preservation of balanced multi-class performance.

This divergent performance pattern reveals a key advantage of the dual-branch architecture: by prioritizing feature diversity through complementary RGB and grayscale processing, DBMNet generates more robust representations that generalize beyond training-specific label correlations. In contrast, methods optimized for specific label structures may overfit to dataset-specific disease co-occurrence patterns, leading to performance collapse when encountering expanded label spaces or novel disease combinations. Such robustness is critical for clinical deployment, where disease prevalence, imaging protocols, and coexisting pathologies vary substantially across populations and healthcare settings [26]. The ability to maintain equitable detection across disease categories under diverse conditions makes DBMNet particularly suitable for real-world screening applications where systematic under-detection of rare diseases poses unacceptable clinical risk.

4.4 Limitations and future work

Several limitations of this study should be acknowledged. First, although DBMNet achieved superior balanced performance, its F1-score remains slightly lower than that of BFENet, indicating room for further optimization. Future work may explore advanced loss functions or optimization strategies to improve overall accuracy without compromising class balance. Second, the dual-branch architecture incurs a relatively high computational cost, which may limit its deployment in resource-constrained environments. Model compression techniques, such as pruning or knowledge distillation, should be investigated to reduce inference time and parameter size while preserving diagnostic performance. Third, this study lacks validation on completely independent external cohorts. Although the FMLC-12 and ODIR datasets incorporate multi-source data, prospective multi-center validation is necessary to confirm generalizability across diverse clinical settings and imaging devices. Future research should focus on large-scale external validation, improving computational efficiency, and enhancing model interpretability to facilitate clinical integration.

5 Conclusion

This study presents DBMNet, a novel dual-branch TransNeXt-based framework for multi-label retinal disease classification, which systematically integrates complementary visual modalities (RGB and grayscale) through a dedicated feature integration module. Experimental results on the FMLC-12 and ODIR datasets demonstrate that DBMNet achieves balanced performance across both common and rare diseases, addressing key challenges in automated fundus diagnosis and supporting equitable clinical detection. Beyond technical advances, the dual-branch design reflects TCM principles of comprehensive multi-regional fundus assessment, laying a foundation for the standardization and objectification of TCM ophthalmological diagnosis and suggesting future directions for integrating expert knowledge with deep learning to improve diagnostic accuracy and patient care.

Fundings

Natural Science Foundation of Hunan Province (2025JJ90031), Key Research and Development Program of Hunan Province of China (23A0273), and Hunan Provincial Administration of Traditional Chinese Medicine (A2023048).

Author contributions

Xin He: conceptualization, methodology, software, data curation, and writing – original draft. Xiaohui Li: data

curation, validation, and formal analysis. Jun Peng: resources and validation. Lei Sun: software and visualization. Dan Shu: investigation. Li Xiao: methodology, supervision, and writing – review & editing. Qinghua Peng: supervision, funding acquisition, project administration, and writing – review & editing. Xiaoxia Xiao: conceptualization, methodology, and writing – review & editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

Qinghua Peng is an editorial board member for *Digital Chinese Medicine* and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

[1] JIANG PF, PENG J, ZHOU YS, et al. Ophthalmic syndrome differentiation system and digital Chinese medicine. Digital Chinese Medicine, 2018, 1(1): 9–13.

[2] GONG D, LI WT, LI XM, et al. Development and research status of intelligent ophthalmology in China. International Journal of Ophthalmology, 2024, 17(12): 2308–2315.

[3] XIE LK, CHEN ZY, HAO XF. Discussion on the current status and prospects of traditional Chinese medicine ophthalmology research based on the advantageous disease research model. Chinese Journal of Traditional Chinese Medicine Ophthalmology, 2023, 33(3): 201–205.

[4] WANG J, YANG L, HUO ZQ, et al. Multi-label classification of fundus images with EfficientNet. IEEE Access, 2020, 8: 212499–212508.

[5] ABDELMAKSOUD E, EL-SAPPAGH S, BARAKAT S, et al. Automatic diabetic retinopathy grading system based on detecting multiple retinal lesions. IEEE Access, 2021, 9: 15939–15960.

[6] SENGAR N, JOSHI RC, DUTTA MK, et al. EyeDeep-Net: a multi-class diagnosis of retinal diseases using deep neural network. Neural Computing and Applications, 2023, 35(14): 10551–10571.

[7] HE JJ, LI C, YE J, et al. Multi-label ocular disease classification with a dense correlation deep neural network. Biomedical Signal Processing and Control, 2021, 63: 102167.

[8] GOUR N, KHANNA P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. Biomedical Signal Processing and Control, 2021, 66: 102329.

[9] HE JJ, LI C, YE J, et al. Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification. Biomedical Signal Processing and Control, 2021, 67: 102491.

[10] OU XY, GAO L, QUAN XW, et al. BFENet: a two-stream interaction CNN method for multi-label ophthalmic diseases classification with bilateral fundus images. Computer Methods and Programs in Biomedicine, 2022, 219: 106739.

[11] LI ZW, XU MY, YANG XL, et al. Multi-label fundus image classification using attention mechanisms and feature fusion. Micromachines, 2022, 13(6): 947.

[12] SUN K, HE MJ, HE ZC, et al. EfficientNet embedded with spatial attention for recognition of multi-label fundus disease from color fundus photographs. Biomedical Signal Processing and Control, 2022, 77: 103768.

[13] BHATI A, GOUR N, KHANNA P, et al. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset. Computers in Biology and Medicine, 2023, 153: 106519.

[14] LI ZW, XU MY, YANG XL, et al. A multi-label detection deep learning model with attention-guided image enhancement for retinal images. Micromachines, 2023, 14(3): 705.

[15] AL-FAHDAWI S, AL-WAISY AS, ZEEBAREE DQ, et al. Fundus-DeepNet: multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. Information Fusion, 2024, 102: 102059.

[16] RODRÍGUEZ MA, ALMARZOUQI H, LIATSIS P. Multi-label retinal disease classification using transformers. IEEE Journal of Biomedical and Health Informatics, 2023, 27(6): 2739–2750.

[17] ZHAO JQ, ZHU JF, HE JN, et al. Multi-label classification of retinal diseases based on fundus images using Resnet and Transformer. Medical & Biological Engineering & Computing, 2024, 62(11): 3459–3469.

[18] SISWADI AAP, BRICQ S, MERIAUDEAU F. Multi-modality multi-label ocular abnormalities detection with transformer-based semantic dictionary learning. Medical & Biological Engineering & Computing, 2024, 62(11): 3433–3444.

[19] CHENG YL, MA MN, LI XY, et al. Multi-label classification of fundus images based on graph convolutional network. BMC Medical Informatics and Decision Making, 2021, 21(2): 82.

[20] SUN K, HE MJ, XU Y, et al. Multi-label classification of fundus images with graph convolutional network and LightGBM. Computers in Biology and Medicine, 2022, 149: 105909.

[21] LI N, LI T, HU CY, et al. A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection. Benchmarking, Measuring, and Optimizing. Cham: Springer, 2021: 177–193.

[22] PACHADE S, PORWAL P, THULKAR D, et al. Retinal fundus multi-disease image dataset (RFMiD): a dataset for multi-disease detection research. Data, 2021, 6(2): 14.

[23] SHI D. TransNeXt: robust foveal visual perception for vision transformers. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 17773–17783.

[24] DELGADO R, TIBAU XA. Why Cohen's kappa should be avoided as performance measure in classification. PLoS ONE, 2019, 14(9): e0222916.

[25] BISWAS S, KHAN MIA, HOSSAIN MT, et al. Which color channel is better for diagnosing retinal diseases automatically in color fundus photographs? Life, 2022, 12(7): 973.

[26] AZIZI S, CULP L, FREYBERG J, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering, 2023, 7(6): 756–779.

(Editor-in-Charge Siyi Wei)

基于双分支深度学习的眼底疾病多标签分类：一种受中医五轮理论启发的智能诊断框架

何昕[a], 李晓辉[a], 彭俊[b], 孙磊[a], 舒丹[a], 肖莉[c], 彭清华[c*], 肖晓霞[a*]

a. 湖南中医药大学信息科学与工程学院, 湖南 长沙 410208, 中国
b. 湖南中医药大学第一附属医院眼科与耳鼻咽喉科, 湖南 长沙 410007, 中国
c. 湖南中医药大学中医学院, 湖南 长沙 410208, 中国

【摘要】目的 针对现有自动化诊断方法在互补特征提取不足及跨模态特征融合能力欠缺等关键问题，开发一种用于眼底疾病精准多标签分类的双分支深度学习框架。**方法** 通过整合 ODIR 和 RFMiD 数据集中的互补样本，构建了包含 6 936 幅眼底图像、涵盖 12 类视网膜病变的 FMLC-12 数据集，并在 FMLC-12 和 ODIR 两个数据集上对所提框架进行验证。受中医眼科五轮学说整体多区域观察原则的启发，本研究开发了双分支多标签网络（DBMNet），该框架将互补视觉特征提取与病理关联建模相结合。网络架构采用 TransNeXt 作为骨干网络，设计了双分支结构：一条分支处理红绿蓝（RGB）图像以捕获血管形态、病灶结构等色彩依赖性特征，另一条分支处理灰度转换图像以增强细微纹理细节和对比度变化。特征交互模块（FIM）有效融合了两条分支的多尺度特征。通过全面的消融实验评估双分支架构和 FIM 的贡献。将 DBMNet 与四种先进方法进行了性能比较，包括 EfficientNet 集成方法、基于迁移学习的卷积神经网络（CNN）、BFENet 和 EyeDeep-Net，评估指标包括平均精度均值（mAP）、F1 分数和 Cohen's kappa 系数。**结果** 与单分支 TransNeXt 基线模型相比，引入双分支结构后，模型的 mAP 提升了 15.44 个百分点，由 34.41% 提高至 44.24%；在此基础上进一步引入 FIM 模块，使 mAP 进一步提升至 49.85%。在 FMLC-12 数据集上，DBMNet 的 mAP 达到 49.85%，Cohen’s kappa 系数为 62.14%，F1 值为 70.21%。与 BFENet（mAP：45.42%，kappa：46.64%，F1 值：71.34%）相比，DBMNet 在 mAP 和 kappa 指标上分别提高了 4.43 和 15.50 个百分点，但 BFENet 在 F1 值上略高。在 ODIR 数据集上，DBMNet 的 F1 值达到 85.50%，与当前先进方法的性能相当。**结论** DBMNet 通过双分支架构有效整合 RGB 和灰度视觉模态，显著提升了眼底疾病多标签分类的性能。该框架不仅解决了现有方法中缺乏有效特征融合的问题，还展现了卓越的跨疾病类别平衡性，特别是在对常见和罕见疾病的均衡检测方面，为智能化、标准化的眼底疾病分类提供了一条具有临床应用前景的技术路径。

【关键词】多标签分类；眼底图像；深度学习；双分支网络；中医眼科；五轮学说