Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 1, January 2026, pp. 200~209
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i1.pp200-209

A novel approach for detecting diabetic retinopathy using two-stream CNNs model

Pham Thi Viet Huong[1], Le Duc Thinh[2], Tran Thi Oanh[1], Tran Xuan Bach[2], Hoang Quang Huy[2], Tran Anh Vu[2]
[1]International School, Vietnam National University, Hanoi, Vietnam
[2]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

Article Info

Article history:

Received Oct 30, 2025
Revised Dec 4, 2025
Accepted Dec 13, 2025

Keywords:

Diabetic retinopathy
Fourier transform
Fundus photography
Loose pairing training
Two-stream CNN

ABSTRACT

Major causes of visual impairment, particularly diabetic retinopathy (DR) and aged-related macular degeneration (AMD), has posed significant challenges for clinical diagnosis and treatment. Early detection and prompt intervention can help prevent severe consequences for patients. The study presents a novel approach for detecting eye diseases using a two-stream convolutional neural network (CNN) model. The first stream processes pre-processed fundus images, while the second stream analyzes high-pass filtered fundus images in the spatial frequency domain. To assess the model's performance, we use the APTOS 2019 dataset, which was originally compiled for the Asia Pacific Tele-Ophthalmology Society 2019 Blindness Detection competition and is publicly available on Kaggle. Our method shows promise as an early screening tool for DR detection with an accuracy of 0.986.

Corresponding Author:

Tran Anh Vu
School of Electrical and Electronic Engineering, Hanoi University of Science and Technology
Hanoi, Vietnam
Email: vu.trananh@hust.edu.vn

1. INTRODUCTION

Retinal pathologies increasingly represent a global health concern, demanding significant attention and resources. Among these conditions, retinopathy is a leading cause of severe visual impairment and blindness worldwide [1]. By 2040, age-related macular degeneration (AMD) is predicted to affect nearly 300 million people. Meanwhile, one of the main causes of vision loss in individuals is diabetic retinopathy (DR), which has been identified as a worldwide epidemic. This underscores the critical need for preventive measures, early diagnosis, and innovative treatments to address these debilitating eye conditions [2], [3].

Medicine and biology increasingly depend on automated systems for analysis and diagnosis [4]. For instance, computer-aided diagnosis (CAD) of retinal images assists healthcare providers in identifying diseases, initiating early treatments, and reducing inconsistencies in image interpretation [5]. Automated analysis also offers numerous benefits over manual inspection, including cost efficiency, objectivity, reliability, and reduced dependence on skilled specialists for image evaluation [6], [7]. Before the growing of deep learning (DL) techniques, CAD systems were used for image restoration and enhancement, among other phases of the retinal diagnostics process. Some CAD techniques aim to replicate the procedures that physicians do to identify retinal disorders, which include segmenting images, extracting features, and using machine learning to classify the results [8].

DL techniques, especially convolutional neural networks (CNNs) and more recently transformer models, have shown remarkable success in the automated diagnosis of retinal disorders. Numerous efforts

have focused on identifying prevalent retinal conditions, such as AMD, DR, and glaucoma. For instance, a method utilizing two pretrained CNNs (VGG16 and a custom CNN) was developed to diagnose DR by analyzing the likelihood of lesion patches [9]. Similarly, a system which combined three CNN models—Inception-v3, ResNet152, and Inception-ResNet-v2 was built to detect DR in fundus images [10]. Additionally, a DL-based diagnostic tool has been created to screen patients for a range of common retinal diseases [11]. The approaches still have some weaknesses, such as loss of image information during processing or high computational demands and complexity.

The APTOS dataset has been widely used in various methods for classifying eye diseases. The research in [12] emphasized training on a focused subset of challenging cases while minimizing the influence of a large number of easy negatives that could overwhelm the detector during training. A neuron intrinsic learning framework was introduced in [13] to identify distractors in the CNN feature space. It employed a novel distractor-aware loss function to create a clear distinction between the original image and its distractor within the feature space. The approach in [14] leveraged contrastive learning for feature embedding to address class imbalance, replacing the traditional cross-entropy loss. The research in [15] is conducted through an iterative training process to get an identity matrix for the cross-correlation.

Despite significant advancements, a critical research gap persists in developing a model that can effectively integrate multi-spectral feature information while maintaining computational efficiency. Specifically, the primary scientific problem we address is the lack of a robust model architecture that can: (i) preserve the full context of the original fundus image; (ii) isolate and emphasize critical high-frequency pathological features (such as microaneurysms and hemorrhages) using a dedicated channel, and (iii) achieve superior performance compared to computationally heavy ensemble methods. An efficient strategy for fusing these distinct feature sets is essential to resolve this issue.

To overcome this research gap, a novel methodology is proposed for detecting DR using a two-stream CNN model. The key innovation and main contribution of this research is the integration of the proposed architectural design and the pre-processing/fusion methodology. The first stream processes the standard pre-processed fundus image (with data augmentation) to capture global, low-frequency features (e.g., optic disc, general vasculature). The second stream processes the high-pass filtered fundus image in the spatial frequency domain (derived via fourier transform). This step is crucial as it isolates and emphasizes the critical, high-frequency details related to DR pathology, forcing the network to focus on small, clinically relevant lesions. Feature fusion strategy after global average pooling (GAP): the outputs from the two streams are concatenated after the GAP layer. This strategic fusion point ensures that each stream has already extracted robust, refined, high-level semantic features before merging, leading to a richer final representation for classification. This combined dual-stream methodology successfully preserves image integrity while actively guiding the machine's focus toward specific pathological features, achieving an impressive accuracy of 0.986 on the APTOS 2019 dataset.

2. METHOD

2.1. Dataset

In this study, we employ the APTOS 2019 BD dataset [16] to train and evaluate the proposed model's performance. This dataset comprises 3,662 samples, including 1,805 normal images and 1,857 images depicting DR, collected from a large population in India. The collection process was organized by the Aravind Eye Hospital in India, featuring fundus photos taken under different conditions and over diverse periods. A team of trained medical experts carefully reviewed and labeled the samples based on the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDRSS). The APTOS 2019 BD dataset is categorized into five groups based on this criterion: proliferative DR, mild DR, moderate DR, severe DR, and no DR.

2.2. Workflow

The workflow of the model is presented in Figure 1. The detection system utilizes two input channels. As shown in the workflow diagram, data augmentation is applied first to address data limitations and mitigate overfitting issues. After that, the images are divided into two channels.

While the second channel offers additional information to improve classification accuracy, the first channel processes the original images. By transforming the augmented images into the spatial frequency domain in the second channel, we can isolate and remove the low-frequency data. Since these low frequencies encode general information (like the shape of the optic nerve head, macula, and main blood vessels), their removal serves to sharpen the focus on specific indicators of DR. Finally, the two channels are fed into a classification block, which is a two-stream CNN model, to detect the disease.

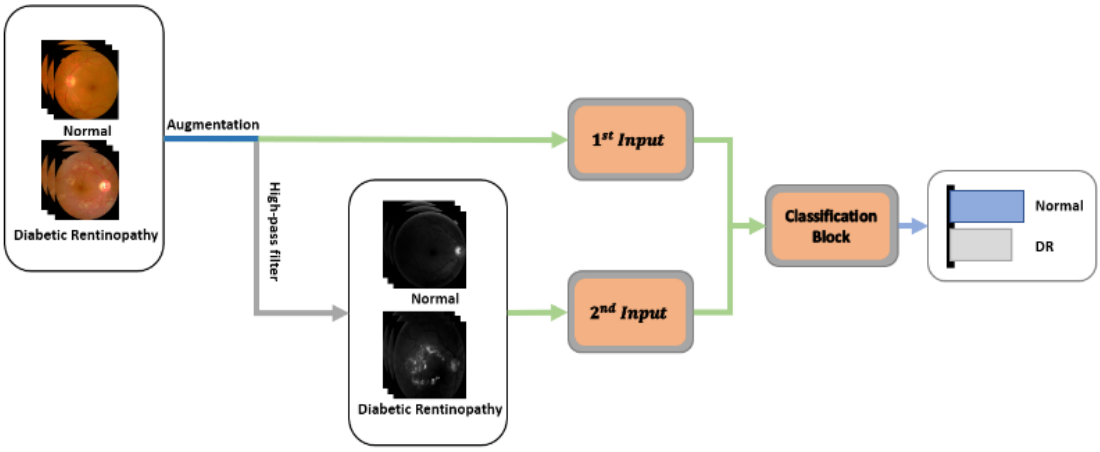


Figure 1. The workflow of the proposed method

2.3. Data augmentation

To guarantee the number of training examples is large enough for the two-stream model, we employed data augmentation during the training process [17], applying at least one augmentation to each training image before inputting it into the two-stream CNN model. These augmentations, implemented using the Albumentations library [18], included optical distortion, grid distortion, piecewise affine transformations, horizontal and vertical flips, random rotations, shifts, scaling, red, green, and blue (RGB) value shifts, and random adjustments to brightness and contrast. As a result, the dataset expanded to 29,296 samples, comprising 14,440 normal images and 14,856 disease images. This augmentation process significantly enhanced data diversity, reducing overfitting and improving the model's generalization capabilities for detecting and classifying DR.

2.4. High-pass filter

In the training and validation phases, modified versions of the original images were utilized. These photos were scaled and cropped. The APTOS2019 dataset shows erroneous associations between disease stages and a number of image meta-features, including brightness, zoom level, cropping type, and resolution. To prevent CNNs from overfitting to these meta-features and to minimize correlations between the image content and such attributes, extensive augmentations were applied. Subsequently, high-pass filtering using fourier transform was implemented on the first stream's inputs to generate inputs for the second stream:

2.4.1. Fourier transform

The fourier transform is capable of decomposing images by analyzing variations in intensity (gray levels) across spatial distances, rather than examining signals that change over time [19]. This process transforms the time domain of the images into the spatial domain. In the context of sound, frequency represents the (inverse) regularity of sine wave repetitions. Similarly, in images, spatial frequency describes the (inverse) regularity of fluctuations in intensity values. The spatial frequency is determined using the following formula:

$$F(u,v) = \iint_{-\infty}^{\infty} I(x,y) \cdot e^{-j2\pi(ux+vy)} dxdy \quad (1)$$

where $I(x,y)$: gray level value at a pixel in the spatial domain, $F(u,v)$: value in the spatial frequency domain, and (u,v): spatial frequency coordinates.

2.4.2. Spectrum

High spatial frequencies correspond to image features that show rapid intensity variations over short distances [19]. In contrast, features with gradual intensity changes over longer distances are associated with low spatial frequencies. Image popularity can be determined by applying the fourier transform and calculating the natural logarithm of the result.

$$Spectrum(u,v) = \log(1 + |F(u,v)|^2) \quad (2)$$

2.4.3. Analyzing spatial frequency domain

The spatial frequency is determined by the distance from the center of the discrete fourier transform (DFT) [20]. A spatial frequency of zero corresponds to the image's average brightness. A point located two pixels from the center represents a sinusoidal component of the image that completes two cycles of the sine wave. Consequently, the distance in the image, denoted as x_i, and its corresponding spatial frequency, f, can be expressed as:

$$x_i = \frac{\Delta x}{f} \tag{3}$$

where Δx: the image dimension (in pixel).

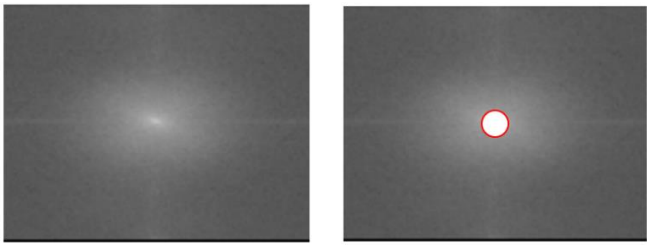
2.4.4. Highpass filtering

As analyzed, in fundus photography, the low frequencies correspond to regions near the center of the spectrum, while high frequencies are associated with the outer areas of the spectrum. To apply high-pass filtering, a threshold D_0 is defined, representing the distance from the center of the spectrum to a point in the spectral space. For each point in the spectrum, if its distance from the center is less than D_0, it will be removed. To calculate distances in a 2D space, the Euclidean distance is used. The formula for Euclidean distance between two points $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ is:

$$d(P, Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{4}$$

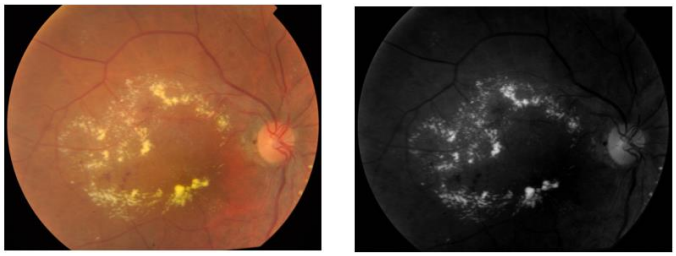
where -p_i, q_i :components of points P and Q respectively, and n: number of dimensions of space.

An illustration of using high-pass filtering for fundus images is presented in Figures 2 and 3. Figure 2(a) shows the frequency spectrum of the image before filtering, an Figure 2(b) shows the frequency spectrum after high-pass filtering. Figure 3(a) is the fundus image before processing. Figure 3(b) is the image after high-pass filtering, which clarifies important details. After high-pass filtering, all of the important information about DR is highlighted.



(a) (b)

Figure 2. Images of (a) original and (b) after high-pass filtering in spatial frequency domain



(a) (b)

Figure 3. Illustration of (a) before and (b) after high-pass filtering

2.5. Two-stream network architecture

In this study, a two-stream CNN is designed to process two distinct feature classes input. The architecture of this model is illustrated in Figure 4, with ResNet serving as its backbone. This design concept shares similarities with the two-stream network approach previously explored in the domain of human activity recognition in videos [21].

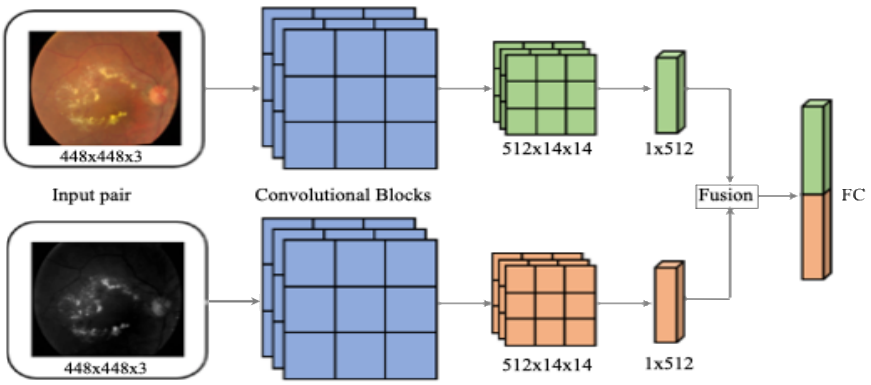


Figure 4. Two-stream CNNs model

The first ResNet stream focuses on extracting features from fundus images, while the second stream provides supplementary information to support the first channel. Both branches of the two-stream CNN are based on ResNet-34. While other CNN architectures could have been employed, ResNet-34 was selected for several reasons: it has fewer parameters, requires less data for training, and has proven effective in various fundus image identification tasks [22], [23].

A single-modal CNN creates successive 2D feature maps layer by layer in order to extract information from an image. These feature maps keep some spatial information from the original input image even if they get smaller and smaller. Thus, after the GAP layer, we choose to undertake spatially invariant fusion. By reducing each feature map to a single scalar value, the GAP layer efficiently eliminates spatial information.

In the first stream, the final convolutional block of ResNet-34 produces 512 feature maps, represented as $(F_f = \{F_{f,1}, \ldots, F_{f,512}\})$. Each feature map has dimensions of m×m, where the size of m depends on the input image size. For an input size of 448×448, m equals 14. To extract a feature value at a certain position (x, y) in a feature map $F_{f,i}$, we use the notation $F_{f,i}(x, y)$. Similarly, the second stream branch outputs feature maps represented as $(F_o = \{F_{o,1}, \ldots, F_{o,512}\})$. Both F_f and F_o are passed through the GAP layer simultaneously, resulting in two 512-dimensional feature vectors denoted as $\overline{F_f} = (\overline{F_{f,1}}, \ldots, \overline{F_{f,512}})$ and $\overline{F_o} = (\overline{F_{o,1}}, \ldots, \overline{F_{o,512}})$.

The fusion module layer combines the high-level output features from two separate streams, effectively utilizing the information from each to improve overall performance. This fusion layer architecture provides two key advantages. First, it captures a much richer representation of the original data without requiring the training of multiple classifiers, as seen in late-fusion methods. Second, in contrast to early-fusion approaches where fusion occurs at the initial layer, the CNN layers in each individual stream collect more useful and refined information from the raw features. As illustrated in Figure 4, the first stream processes fundus images, with each element of the output tensor corresponding to these images, while the second stream handles high-pass filtered images, with each tensor element representing this filtered data.

The fused feature tensor, represented as $f = concat(F_f, F_o)$, combines the strengths of both streams. After that, this tensor is sent into later CNN layers for additional processing. Lastly, the classification is carried out by a SoftMax layer and a fully connected (FC) layer. This method improves prediction accuracy by allowing adaptive tuning and balancing the significance of different feature sets.

In training and testing phase, the cross-entropy loss function shown below is used:

$$L = -\sum_{i=0}^{C-1} y_i \cdot \log(p_i) \quad (5)$$

Where,

- $p = [p_0, \ldots, p_{C-1}]$ represents a probability distribution, p_i denotes the probability that a sample belongs to class i.
- $y = [y_0, \ldots, y_{C-1}]$ denotes the one-hot representation of class labels, and C is the number of classes.

3. RESULTS AND DISCUSSION

In this part, we conduct experiments to validate the proposed model and demonstrate the effectiveness. Section 3.1 presents the experimental results using the APTOS dataset, while 3.2 explores additional experiments to evaluate the model's compatibility with other datasets.

3.1. Experiment results

For this part, the APTOS dataset was split into two parts: training and testing, with a 7:3 ratio, respectively. Initially, the performance of conventional CNN models was not particularly impressive, achieving a maximum accuracy of 0.88 and an F1 score of 0.87. To improve these results, we opted for a two-stream model, anticipating better performance compared to traditional single-stream CNN models. We tested several common architectures, including Inception, MobileNet, VGG, and ResNet. Results are shown in Table 1. High-pass filtering in one channel plays a critical role in training the two-stream model, as it directly influences the model's performance, generalization, and training efficiency.

Table 2 summarizes the model's robust performance. The classifier exhibits excellent precision, achieving 0.99 for the "Normal" category and 0.982 for "DR". Furthermore, its recall rates (0.985 for "Normal" and 0.988 for "DR") underscore the model's high sensitivity in capturing positive cases. The F1 scores, a comprehensive measure that harmonizes precision and recall, are uniformly strong at 0.986 for both classes. This consistency points to an optimal balance between minimizing false positives and false negatives, confirming the model's high effectiveness and reliability in accurately distinguishing between normal and DR case.

Table 1. Performance of different models

Methods		Accuracy	F1-score
1-stream conventional CNN model		0.88	0.87
2-stream CNN model	Inception_v3	0.984	0.983
	MobileNet	0.96	0.956
	VGG	0.959	0.96
	ResNet	**0.986**	**0.986**

Table 2. Classification performance

Diseases	Precision	Recall	F1-score
Normal	0.99	0.985	0.986
DR	0.982	0.988	0.986

As shown in Figure 5, the area under the curve (AUC) and receiver operating characteristic (ROC) curves highlights the model's stability. A high AUC score, typically paired with a well-formed ROC curve, demonstrates that the model maintains strong performance across various thresholds. This stability is critical for ensuring consistent classification performance, as it reflects the model's ability to reliably distinguish between positive and negative instances with minimal variance in predictions. Therefore, the strong AUC and ROC curves provide confidence in the model's robustness and reliability.

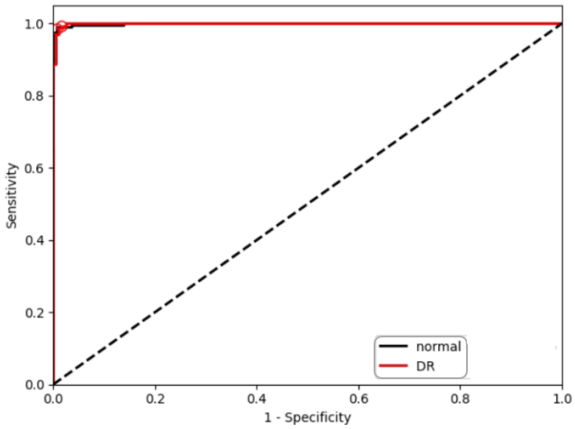


Figure 5. AUC and ROC curve

Table 3 compares the suggested approach to various state-of-the-art methods and shows that it performs better in terms of accuracy and F1-score. The suggested method shows remarkable accuracy in categorizing cases in the dataset, with an accuracy of 0.986. Furthermore, the model's capacity to strike an appropriate balance between accuracy and recall is demonstrated by its F1-score of 0.986.

Table 3. Comparison with the state-of-the-art approaches with APTOS dataset

Methods	Accuracy	F1-score
FG-SSL [15]	0.858	0.720
ProCo [24]	0.837	0.674
CL + resample [14]	0.816	0.608
CL [14]	0.825	0.652
DANIL [13]	0.825	0.660
Focal loss [12]	0.815	0.629
Proposed method	**0.986**	**0.986**

In contrast, alternative methods such as FG-SSL, ProCo, CL + resample, CL, and DANIL yield lower performance across these metrics. For instance, while FG-SSL achieves a reasonable accuracy of 0.858, its F1-score is significantly lower at 0.720. Similarly, the proposed method surpasses ProCo, CL + resample, CL, and DANIL in both accuracy and F1-score, further demonstrating its superiority in delivering accurate predictions.

3.2. Compatibility evaluation

In this section, we evaluate our model using a different dataset—the DR dataset [25]. The DR dataset, obtained from Kaggle and provided by EyePACS, a free platform for retinopathy screening, contains 35,126 fundus images. Of these, 25,805 are normal (disease-free), and 9,321 show signs of DR. To address the issue of class imbalance, 9,321 normal fundus images and 9,321 disease-affected fundus images are selected from the DR dataset. The dataset is then divided into training and testing sets, with a 70% training and 30% testing split.

Table 4 demonstrates that the proposed method outperformed all others, achieving the highest accuracy of 76%, with a 0.76 F1-score, indicating a well-balanced performance in both precision and recall. AlexNet, an early CNN, achieved an accuracy of 73%, while ResNet-50, a deeper CNN architecture, slightly exceeded it with 75%. A modified version of ResNet-50 maintained a competitive accuracy of 74%. These results suggest that the proposed approach provides a significant improvement over well-established architectures such as AlexNet, ResNet-50, and their revised versions.

Table 4. Comparison with the state-of-the-art approaches with DR dataset

Methods	Accuracy	F1-score
AlexNet [25]	0.73	_
ResNet-50 [25]	0.75	_
Revised ResNet-50 [25]	0.74	_
Proposed method	**0.76**	**0.76**

4. CONCLUSION

This paper successfully introduced a novel two-stream CNN model for diabetic retinopathy (DR) detection. The proposed architecture, which utilizes ResNet-34 as its backbone, uniquely integrates information from two channels: the original fundus image and a high-pass filtered image in the frequency domain. The high-pass channel enhances feature distinctiveness by emphasizing subtle disease-related details. Experimental results demonstrate the effectiveness of the proposed model, achieving an accuracy of 0.986 and an F1-score of 0.986 on the APTOS 2019 dataset, highlighting its reliability as an early screening tool.

Despite its strong performance, the model still faces several challenges, including its reliance on optimal filter parameter selection and its current focus on binary classification (normal versus DR) rather than full clinical severity grading. Nevertheless, this research shows significant clinical potential for improving early and automated screening, particularly in regions with limited access to ophthalmology specialists. Future work will focus on extending the model to multi-class DR grading in accordance with the ICDRSS severity levels, exploring adaptive feature fusion strategies such as attention mechanisms to dynamically balance spatial and frequency information, and developing automated filter optimization techniques to learn optimal high-pass filter characteristics during training. These improvements are expected to enhance the model's generalization capability across diverse imaging conditions and further increase its clinical applicability.

FUNDING INFORMATION

This research is funded by International School, Vietnam National University, Hanoi (VNU-IS) under project number CS.2024-06.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Pham Thi Viet Huong	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Le Duc Thinh		✓				✓		✓	✓	✓	✓	✓		
Tran Thi Oanh			✓	✓			✓		✓	✓	✓			
Tran Xuan Bach				✓		✓		✓			✓			
Hoang Quang Huy					✓		✓	✓						
Tran Anh Vu	✓	✓	✓		✓	✓		✓	✓			✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in https://www.kaggle.com/datasets/mariaherrerot/aptos2019.

REFERENCES

[1] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Progress in Retinal and Eye Research*, vol. 67, pp. 1–29, 2018, doi: 10.1016/j.preteyeres.2018.07.004.

[2] T. D. L. Keenan, C. A. Cukras, and E. Y. Chew, "Age-related macular degeneration: epidemiology and clinical aspects," *Advances in Experimental Medicine and Biology*, vol. 1256, pp. 1–31, 2021, doi: 10.1007/978-3-030-66014-7_1.

[3] Y. Zheng, M. He, and N. Congdon, "The worldwide epidemic of diabetic retinopathy," *Indian Journal of Ophthalmology*, vol. 60, no. 5, p. 428, 2012, doi: 10.4103/0301-4738.100542.

[4] J. Han, M. Kamber, and Jjtmks. Pei, "The Morgan Kaufmann series in data management systems," vol. 5, Morhgan Kaufmann Publishers, 2011, pp. 83–124.

[5] K. Mittal and V. M. A. Rajam, "Computerized retinal image analysis - a survey," *Multimedia Tools and Applications*, vol. 79, no. 31–32, pp. 22389–22421, May 2020, doi: 10.1007/s11042-020-09041-y.

[6] M. D. Abramoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010, doi: 10.1109/RBME.2010.2084567.

[7] M. M. Fraz *et al.*, "Blood vessel segmentation methodologies in retinal images - a survey," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 1, pp. 407–433, Oct. 2012, doi: 10.1016/j.cmpb.2012.03.009.

[8] M. A. Omar, M. A. Tahir, and F. Khelifi, "Multi-label learning model for improving retinal image classification in diabetic retinopathy," in *2017 4th International Conference on Control, Decision and Information Technologies, CoDIT 2017*, Apr. 2017, vol. 2017-January, pp. 202–207, doi: 10.1109/CoDIT.2017.8102591.

[9] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. Teatini Salles, "Diabetic retinopathy detection using red lesion localization and convolutional neural networks," *Computers in Biology and Medicine*, vol. 116, p. 103537, Jan. 2020, doi: 10.1016/j.compbiomed.2019.103537.

[10] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 2045–2048, doi: 10.1109/EMBC.2019.8857160.

[11] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[13] L. Gong, K. Ma, and Y. Zheng, "Distractor-aware neuron intrinsic learning for generic 2D medical image classifications," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 591–601, doi: 10.1007/978-3-030-59713-9_57.

[14] Y. Marrakchi, O. Makansi, and T. Brox, "Fighting class imbalance with contrastive learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 466–476, doi: 10.1007/978-3-030-87199-4_44.
[15] W. Park and J. Ryu, "Fine-grained self-supervised learning with jigsaw puzzles for medical image classification," *Computers in Biology and Medicine*, vol. 174, p. 108460, May 2024, doi: 10.1016/j.compbiomed.2024.108460.
[16] Karthik, Maggie, and S. Dane, "APTOS 2019 Blindness Detection." Kaggle, 2019, [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, May 2012, doi: 10.1145/3065386.
[18] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information (Switzerland)*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.
[19] X. Zhang, Y. Kuang, and J. Yao, "Detection of microaneurysms in color fundus images based on local fourier transform," *Biomedical Signal Processing and Control*, vol. 76, p. 103648, Jul. 2022, doi: 10.1016/j.bspc.2022.103648.
[20] A. M. Abdul-Rahman, T. Molteno, and A. C. B. Molteno, "Fourier analysis of digital retinal images in estimation of cataract severity," *Clinical and Experimental Ophthalmology*, vol. 36, no. 7, pp. 637–645, Sep. 2008, doi: 10.1111/j.1442-9071.2008.01819.x.
[21] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2016, vol. 2016-December, pp. 1933–1941, doi: 10.1109/CVPR.2016.213.
[22] Q. Wei, X. Li, H. Wang, D. Ding, W. Yu, and Y. Chen, "Laser scar detection in fundus images using convolutional neural networks," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision*, 2019, pp. 191–206, doi: 10.1007/978-3-030-20870-7_12.
[23] X. Lai, X. Li, R. Qian, D. Ding, J. Wu, and J. Xu, "Four models for automatic recognition of left and right eye in fundus images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11295 LNCS, Springer International Publishing, 2019, pp. 507–517.
[24] Z. Yang *et al.*, "ProCo: prototype-aware contrastive learning for long-tailed medical image classification," in *International conference on medical image computing and computer-assisted intervention*, 2022, pp. 173–182, doi: 10.1007/978-3-031-16452-1_17.
[25] C. L. Lin and K. C. Wu, "Development of revised ResNet-50 for diabetic retinopathy detection," *BMC Bioinformatics*, vol. 24, no. 1, Apr. 2023, doi: 10.1186/s12859-023-05293-1.

BIOGRAPHIES OF AUTHORS

Dr. Pham Thi Viet Huong obtained her B.S. in electrical engineering from Hanoi University of Science and Technology in 2007. She earned her M.Sc. and Ph.D., both in electrical engineering, from the University of Massachusetts Lowell, United States, in 2010 and 2012, respectively. From 2012 to 2015, she was a researcher at the Manning School of Business in Lowell, Massachusetts. From 2017 to 2020, she was a faculty member at Vietnam National University (VNU), University of Engineering and Technology (VNU-UET), Vietnam. Since 2020, she has been working at the International School – VNU. Her research interests include data mining and analytics, machine learning methodologies, with applications in biomedical engineering. She can be contacted at email: huongpv@vnu.edu.vn.

Le Duc Thinh is a lecturer at International School, Vietnam National University, Ha Noi, Vietnam. He teaches mathematics in English for business majors. He obtained his Bachelor degree in 2001 and Master degree in 2004, both in mathematics at Ha Noi National University of Education, Vietnam. He obtained his Ph.D. degree in 2012 at the Pennsylvania State University, USA. His research interests include machine learning, mathematical finance, economic statistics. He can be contacted at email: thinhld@vnu.edu.vn.

Tran Thi Oanh got the bachelor and master degrees in computer science at the University of Engineering and Technology, Vietnam National University, Hanoi in 2006 and 2009, respectively. She was awarded a Japanese Government Scholarship to pursue Ph.D. in Computer Science at Japan Advanced Institute of Science and Technology (JAIST) from 2011 to 2014. Currently, she is a lecturer at the International School of Vietnam National University, Hanoi (VNU-IS). Her main research interests are artificial intelligence and machine learning. Her contributions to the field include 50 publications in esteemed journals and conferences. She can be contacted at email: oanhtt@gmail.com or tranthioanh@vnu.edu.vn.

Mr. Tran Xuan Bach obtained his B.Sc. in the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam, in 2021. His main research interests include medical data analysis and classification, as well as research and development of applications in biomedical engineering. He can be contacted at email: sitizaiton@umk.edu.my.

Hoang Quang Huy is a lecturer of School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam. He earned his M.S. degree in electronics and telecommunications from Hanoi University of Science and Technology (Vietnam) in 2006, and his B.S. degree in electronics and telecommunications from the same university in 2002. His main research interests include medical data analysis and classification, research and development of applications for rehabilitation, smart health, and hospital information systems. He can be contacted at email: huy.hoangquang@hust.edu.vn.

Dr. Tran Anh Vu is a senior lecturer at School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam. He earned his Ph.D. degree in electrical engineering from the University of Massachusetts at Lowell (USA) in 2014, MS degree in biomedical engineering from Tufts University (USA) in 2010, MS degree in electronics and telecommunications from Hanoi University of Science and Technology (Vietnam) in 2002, and B.S. degree in electronics and telecommunications from Hanoi University of Science and Technology (Vietnam) in 2000. His main research interests include the medical data analysis and classification, research and development applications for rehabilitation, smart health. He can be contacted at email: vu.trananh@hust.edu.vn.