Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048307

ARTICLE

HCSP-Net: A Novel Model of Age-Related Macular Degeneration Classification Based on Color Fundus Photography

Cheng Wan[1], Jiani Zhao[1], Xiangqian Hong[2], Weihua Yang[2,*] and Shaochong Zhang[2,*]

[1]College of Electronic and Information Engineering/College of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China

[2]Shenzhen Eye Institute, Shenzhen Eye Hospital, Jinan University, Shenzhen, 518040, China

*Corresponding Authors: Weihua Yang. Email: benben0606@139.com; Shaochong Zhang. Email: zhangshaochong@gzzoc.com

Received: 04 December 2023 Accepted: 14 February 2024 Published: 25 April 2024

ABSTRACT

Age-related macular degeneration (AMD) ranks third among the most common causes of blindness. As the most conventional and direct method for identifying AMD, color fundus photography has become prominent owing to its consistency, ease of use, and good quality in extensive clinical practice. In this study, a convolutional neural network (CSPDarknet53) was combined with a transformer to construct a new hybrid model, HCSP-Net. This hybrid model was employed to tri-classify color fundus photography into the normal macula (NM), dry macular degeneration (DMD), and wet macular degeneration (WMD) based on clinical classification manifestations, thus identifying and resolving AMD as early as possible with color fundus photography. To further enhance the performance of this model, grouped convolution was introduced in this study without significantly increasing the number of parameters. HCSP-Net was validated using an independent test set. The average precision of HCSP-Net in the diagnosis of AMD was 99.2%, the recall rate was 98.2%, the F1-Score was 98.7%, the PPV (positive predictive value) was 99.2%, and the NPV (negative predictive value) was 99.6%. Moreover, a knowledge distillation approach was also adopted to develop a lightweight student network (SCSP-Net). The experimental results revealed a noteworthy enhancement in the accuracy of SCSP-Net, rising from 94% to 97%, while remarkably reducing the parameter count to a quarter of HCSP-Net. This attribute positions SCSP-Net as a highly suitable candidate for the deployment of resource-constrained devices, which may provide ophthalmologists with an efficient tool for diagnosing AMD.

KEYWORDS

Computer-aided diagnosis; deep learning; age-related macular degeneration; transformer

1 Introduction

According to a study published in *The Lancet Global Health*, the number of patients with severe visual impairment or blindness due to age-related macular degeneration (AMD) is anticipated to reach 288 million by 2040, bringing a significant burden to ophthalmologists [1]. As reported in some studies, AMD is influenced by age, genetics, and complex environmental factors (such as smoking and diet) [2,3]. However, it remains unclear about the specific pathogenesis of AMD. Aging of macular tissues

is believed to be the primary cause of AMD, which is more prevalent in individuals over the age of 50 years [4–6]. There are serious challenges to global eye health in the aging population [1,3,7]. Although AMD can exert serious effects on patient's health, it has not yet attracted a high degree of societal attention or investment in medical resources. Insufficient knowledge and medical resources pose dire threats to patients and hinder limited ophthalmologists from providing convenient and comprehensive diagnostic service for such a large patient population.

AMD can be categorized into Normal Macula (NM), Dry Macular Degeneration (DMD), and Wet Macular Degeneration (WMD) based on its clinical manifestations and imaging features, as shown in Fig. 1 [3,8,9]. Over 80% of patients with AMD present with DMD, defined by the early development of drusen and later development of geographic atrophy, resulting in a gradual loss of vision and vision distortion [6,10]. In contrast, WMD develops from DMD, in which drusen of varying sizes form rapidly in the maculae. As the affected area rapidly invades the surrounding tissues, this condition can cause severe vision loss or even blindness [11]. In this study, the classification based on clinical manifestations could aid ophthalmologists in making quick diagnoses and selecting effective therapies.

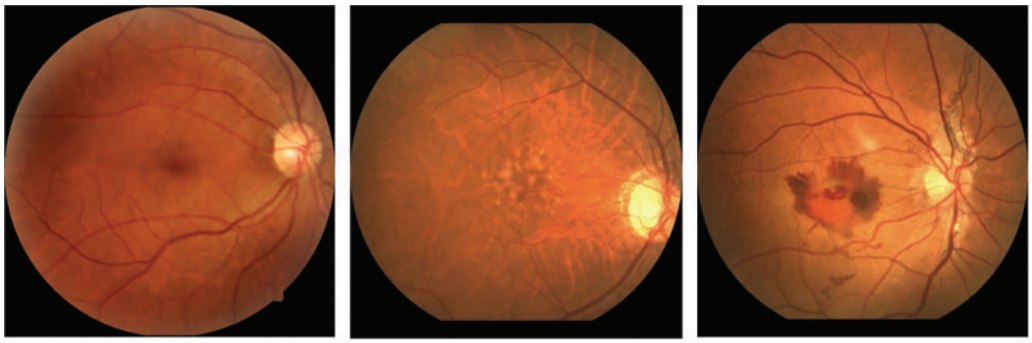

Figure 1: (A) Normal macula; (B) dry macular degeneration; (C) wet macular degeneration

It is important to note that conventional diagnostic techniques have been improved by deep-learning (DL) technology. DL techniques can be employed to avoid misdiagnosis in subjective assessments and make efficient and accurate diagnoses by objectively analyzing numerous data. This can reduce the burden on physicians and facilitate better therapeutic outcomes. A convolutional neural network (CNN) is the main model used in DL. As the depth of the CNN increases, the back-propagation algorithm can be utilized to solve the contribution distribution problem of each layer of the network, so that the model can be applied to prediction. Medical images feature uniform specifications, ease of use, and high quality in long-term clinical practice. The use of DL to process medical images exhibits extremely broad application possibilities and outperforms human experts in diagnosing certain diseases [12–16]. Priya et al. proposed a probabilistic graphical model and a series of image-preprocessing techniques to classify AMD [17]. To improve the visibility of lesions, they first extracted the green channel. Then, the discrete wavelet transforms in combination with the Kirsch operator were used to locate vessels and identify potential lesions. This optimized preprocessing pipeline effectively extracted pathological discriminative features, achieving a classification accuracy of 96%. In contrast to single-model approaches, Grassmann et al. proposed an integrated framework of convolutional neural networks based on random forests for the automated detection of macular degeneration [18]. Through the integration of multiple independently trained CNNs, the integrated

model demonstrated superior classification performance compared with individual human experts. Motozawa et al. constructed another dual-model approach for diagnosing AMD [19]. The first model was designed to differentiate between NM and AMD, and the second one was used to detect exudative changes in AMD, differentiating between DMD and WMD. However, the construction of two separate convolutional neural networks was required, and a diagnostic accuracy of 93.9% was only achieved for AMD. Vague et al. explored the application of multimodal image analysis methods in three cohorts: Young normal, old normal, and dry macular maculopathy [20]. The results showed a diagnostic accuracy of 96% for this task when combined with multimodal training, thus validating the excellence of multimodality for the diagnosis of macular lesions. However, multimodal models are more complex than normal ones and require more computational resources and time for training and optimization.

Nonetheless, there are still limitations in accurately diagnosing AMD through existing models. Firstly, manual feature engineering and image preprocessing are highly dependent, and more automation and intelligence are required. Secondly, sample imbalance still needs to be addressed. Thirdly, the architecture of existing models lacks lightweight and efficient deployment in real-world applications. Given that, a more automated and efficient diagnostic framework was introduced in this study to overcome these challenges. The proposed HCSP-Net, a DL model with a hybrid architecture, aims to tackle the issues resulting from uneven data distribution. Notably, HCSP-Net achieves an impressive diagnostic accuracy of 99%. Furthermore, knowledge distillation was employed to develop SCSP-Net—a lightweight network with enhanced feasibility on resource-limited platforms just based on 1.05 MB parameters.

2 Structure

2.1 Data Acquisition

This study has obtained ethical approval to collect retinal images from the Shenzhen Eye Hospital for research purposes, following the principles outlined in the *Declaration of Helsinki* [21]. To address privacy concerns associated with patient data usage, careful steps were taken before color fundus photographs were incorporated into the dataset. Utilizing the OpenCV toolkit, all patient-related information, including patient name, age, and date of diagnosis, was systematically removed from fundus photographs [22]. During the data preprocessing phase, which involved transforming fundus photographs into binary images, a 10×10 operator was applied to eliminate any remaining patient-related information. Subsequently, contour screening (extracting the contour with the most significant area) was performed to determine the retinal area and cropping the retinal area. Therefore, it is important to highlight that this study does not disclose patient-related statistical information.

Recognizing the critical role of image quality in influencing the data analysis of models, 745 data points were carefully selected, excluding those with inferior exposure. Besides, a meticulous approach using random number seeds was used for dataset division. This ensured a balanced proportion of each macular data in both Datasets A and B, which achieved equitable data distribution and mitigated the impact of subjective factors, thus ensuring fairness in the experimental process. Additional details about the data division are presented in Table 1. Owing to the limited number of dry maculae (only half as many as the other two maculae), the network's generalization and convergence may be severely hampered by the uneven spread of data [23]. Section 2.2 of this paper provides details of the approach we used.

Table 1: Number of fundus images in Datasets A and B from clinical classification of AMD

Clinical classification of AMD	No. of fundus images (% total)	
	Dataset A	Dataset B
NM	255 (39.5)	40 (40.0)
DMD	123 (19.1)	19 (19.0)
WMD	267(41.4)	41 (41.0)
Total	645(100.0)	100 (100)

2.2 Composition of Model Structure

Convolutional neural networks have been extensively applied to image processing and are constantly evolving, with many excellent networks emerging [24–26]. However, when an image is processed locally by CNNs via convolution and pooling, the influence of the context surrounding the image is ignored. This study combines a convolutional neural network with a transformer to access the specific and general contextual information of an image. The core feature extraction component of HCSP-Net involves an improved convolutional neural, CSPDarknet53. CSPDarknet53 reduces the number of model parameters while resolving redundant gradients [24]. The structure of HCSP-Net is shown in Fig. 2. The improved CSPDarknet53 is the backbone of YOLOv5 (You Only Look Once-v5) and comprises CSPDarknet53 and SPPF. YOLOv5 comprises four models with various depths and widths: S, M, L, and X. In this study, the backbone of the S model, which has the smallest width and depth, was used to construct a lightweight model.

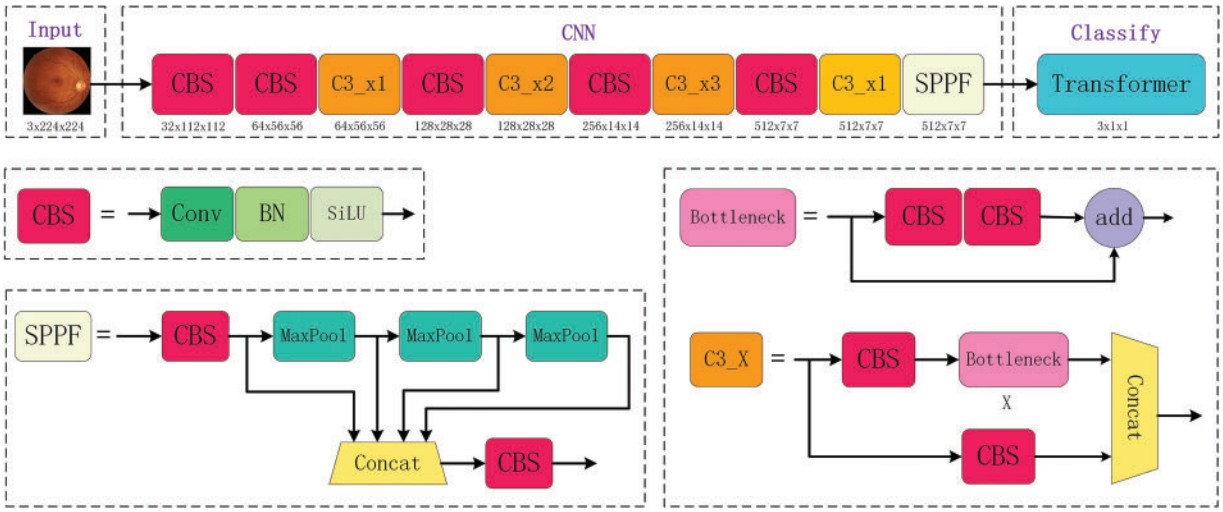

Figure 2: The structure of HCSP-Net

CSPDarknet53 reduces the number of model parameters while resolving redundant gradients. In addition, the data will be deformed when the model input part uniformly scales the image to a fixed resolution. In this study, the SPPF module was retained in the design of the model, making the HCSP-Net more resistant to object deformation [27]. Yet, the original SPPF module relies solely on the maximum pooling layer to emphasize the most salient or active features within a given

region, disregarding others. In this study, it was maintained that after maximum pooling, there was a need for further integration of features to ensure the effective capture of information from the entire feature map. Consequently, a more intricate integration strategy was adopted in the SPPF module. The convolutional operation was incorporated after each maximal pooling layer to enhance the network's perceptual capabilities. Additionally, in consideration of the model's parameter count, the conventional convolution operation was substituted with grouped convolution, aiming to reduce parameters without compromising performance. The enhanced SPPF module is illustrated in Fig. 3. The introduction of grouped convolution provided HCSP-Net with more potent feature extraction and processing capabilities. This established a robust foundation for the model to adeptly handle diverse and dynamic data in practical applications.

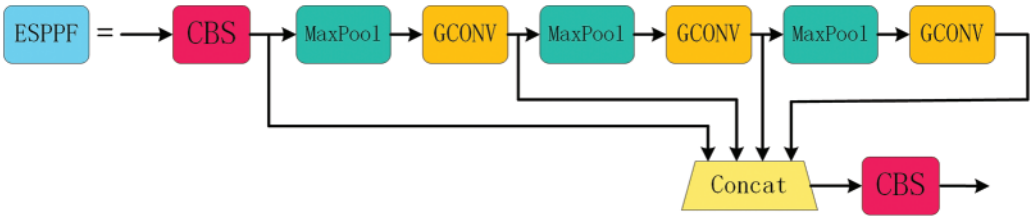

Figure 3: ESPPF: Enhanced SPPF module with GCONV (grouped convolution)

The transformer, a model based on a self-attention mechanism originally proposed by Google for machine translation tasks, has demonstrated superiority over conventional convolutional neural networks in processing long sequence data and modeling global dependencies [28]. It has become a pivotal technique in various domains, including image processing [28–31]. The classification module employs a Transformer block to capture long-term dependencies in the data while preserving sequence information [28]. The structure of the Transformer block is illustrated in Fig. 4. Besides, in light of the uneven data distribution and the shared characteristics between DMD and WMD, this study was conducted to enhance the model's ability to distinguish between the two by incorporating a transformer module. This addition seeks to mitigate the impact of data imbalance on the model's performance.

Of note, due to the small size of the feature map output from the convolutional neural network (7 $\times$ 7), the single-headed self-attention mechanism was directly used for calculations. The self-attention can be calculated as follows:

$$Attentation\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

After the calculation by the Transformer module, its output was transformed into a 3 $\times$ 1 vector using a linear layer, where each vector represents the probability of corresponding to NM, DMD, and WMD, respectively.

2.3 SCSP-Net Boosted by Knowledge Distillation

The Teacher-Student Training (TST) method is a widely recognized approach to knowledge distillation [32]. TST involves the construction of a teacher network with significant depth and width, alongside generating a comparatively lightweight student network. Then, the trained teacher network is used to supervise the training of the student network, aiming to enhance its overall performance. In this study, the well-trained HCSP-Net served as the teacher network, while a lightweight student

network, SCSP-Net, was constructed (refer to Table 2 for model structures). HCSP-Net and SCSP-Net exhibited notable distinctions in channel configuration and module settings. Notably, SCSP-Net featured half the number of channels across the entire network compared with HCSP-Net. In addition, differences in the number of modules between the two models were observed at specific layers. For example, in the third C3 layer, HCSP-Net incorporated three duplicate C3 modules, whereas SCSP-Net employed only two duplicate stacked modules. HCSP-Net's model parameters amounted to 4.19 MB, while SCSP-Net, through a reasonable reduction in width and depth, achieved a parameter count of 1.05 MB—approximately one-fourth of that of HCSP-Net. The amalgamation of lightweight design and knowledge distillation rendered the more pragmatic and viable deployment of SCSP-Net on embedded devices. In the knowledge distillation process, the generalization performance of the student model was improved by transferring the soft labels of the teacher model (HCSP-Net). This enabled SCSP-Net to effectively incorporate knowledge from HCSP-Net.

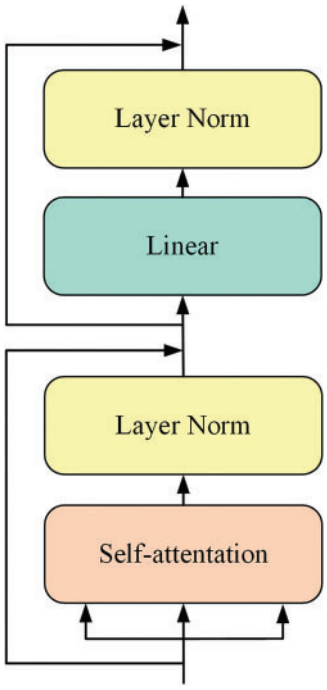

Figure 4: The structure of the Transformer block

The loss function comprises two key componesnts, namely, KL divergence and cross-entropy loss. These components can be employed to measure the difference in the output distribution between the teacher-student networks and the matching degree between the student network and the real label, respectively. The specific expressions for KL divergence and cross-entropy loss are shown as follows:

$$\text{Loss} = \propto \text{KL}(\gamma, \text{p}) + (1 - \alpha)\text{CE}(\text{y}, \text{p}) \tag{2}$$

In the above equation, the symbol $\propto$ represents the weighting coefficient in the loss function, the symbol γ represents the soft labeling of the output of the teacher's network, the symbol p represents the predicted value of the student's network, and the symbol y represents the true labeling of the input image.

Table 2: Architectures for HCSP-Net and SCSP-Net. In the table, 'n' represents the number of modules, 'c' represents the number of channels, 'w' represents the width of the feature map, and 'h' represents the height of the feature map

Layer name	Output size [n, c, w, and h]	
	HCSP-Net	SCSP-Net
CBS	1, 32, 112, 112	1, 16, 112, 112
CBS	1, 64, 56, 56	1, 32, 56, 56
C3	1, 64, 56, 56	1, 32, 56, 56
CBS	1, 128, 28, 28	1, 64, 28, 28
C3	2, 128, 28, 28	1, 64, 28, 28
CBS	1, 256, 14, 14	1, 128, 14, 14
C3	3, 256, 14, 14	2, 128, 14, 14
CBS	1, 512, 7, 7	1, 256, 7, 7
C3	1, 512, 7, 7	1, 256, 7, 7
ESPPF	1, 512, 7, 7	1, 256, 7, 7
Transformer	2, 3, 1, 1	2, 3, 1, 1

2.4 Implementation

The HCSP-Net model was built in PyTorch based on Python 3.7.11, and a GPU (NVIDIA GeForce RTX 1080) was used for the experiments [33]. Due to the limited data collected, Dataset A was expanded by skimming the data horizontally and flipping them vertically, with the final amount of data expanded to four times the original size. To prevent the model from overfitting, data enhancement techniques, including color space variation, random luminance contrast variation, translation scaling, and random orientation rotation, were applied to the fundus map with probabilities of 0.2, 0.2, 0.5, and 1.0, respectively. In the training process, Dataset A was divided according to the ratio of 8:2, which was used for model training and validation, respectively, and the weight file with the lowest epoch number and the highest accuracy was saved as the optimal model.

2.5 Statistical Method

The Scikit-learn toolkit was used to conduct the statistical study [34]. The precision, recall, F1-Score, PPV (positive predictive value), and NPV (negative predictive value) of the HCSP-Net for NM, DMD, and WMD were determined using binary indicators. The areas under the readings and receiver operating characteristic curves (ROC) were also computed. AUC values were classified as poor diagnostic values if they were between 0.5 and 0.70, average diagnostic values if they were between 0.75 and 0.85, and excellent diagnostic values if they were above 0.85.

The multi-classification indicator Kappa value was used to assess the degree of agreement between the true diagnostic findings and the CSPDarknet53 and HCSP-Net results. The Kappa value ranged from 0 to 1, and a higher Kappa value indicated greater agreement between the model's predicted and actual results. The Kappa values were calculated as follows:

$$\text{Kappa} = \frac{\mathrm{p}_o - p_e}{1 - p_e} \tag{3}$$

where p_0 represents the overall classification accuracy. p_e can be calculated as follows:

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \cdots + a_c \times b_c}{n \times n} \tag{4}$$

where a_i represents the number of actual samples in class i, b_i represents the number of samples predicted for class i, and n represents the total number of samples. The Jaccard similarity coefficient is also a simple and effective statistical indicator of similarity and diversity. It is defined as the ratio of the number of elements in the intersection of two sets to the number of elements in the concatenation, which can be expressed as follows:

$$\text{Jaccard (A, B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5}$$

where A and B represent the real and predicted label sets, respectively.

3 Results

3.1 Model Performance Evaluation

The confusion matrix provided a more intuitive view of the classification performance of the model. CSPDarknet53 can be utilized to identify 100% of the NM; however, owing to a lack of data and similarities between DMD and WMD in terms of lesion features, it is difficult for the model to distinguish between them, as shown in Fig. 5A. In Dataset B, 26.3% of DMD cases and 2.4% of WMD cases were incorrectly categorized. Notably, with the addition of the transformer module, the misdiagnosis of DMD was reduced by the HCSP-Net (SPPF), indicating the improved ability of this model to capture the difference between DMD and WMD. Following a careful integration of features extracted from the SPPF through the addition of grouped convolution, only one DMD lesion was misdiagnosed as a WMD lesion by HCSP-Net (ESPPF). The corresponding confusion matrices for the two improved networks based on CSPDarknet53 are shown in Figs. 5B and 5C, respectively. Hence, the problem of uneven data distribution was resolved, resulting in improved model efficiency.

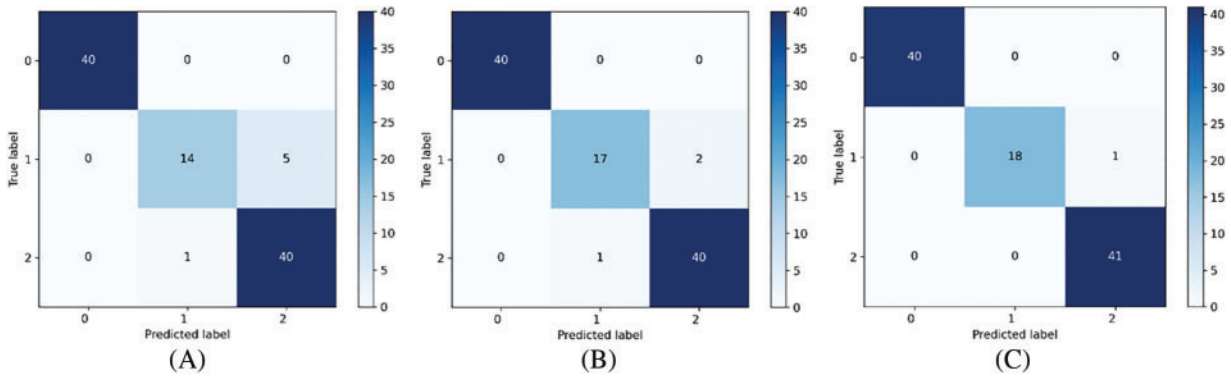

Figure 5: Confusion matrix. (A) CSPDarknet53; (B) HCSP-Net (SPPF); (C) HCSP-Net (ESPPF). 0 for normal macula, 1 for dry macular degeneration, 2 for wet macular degeneration

To objectively evaluate the performance of HCSP-Net, the receiver operating curve (ROC) was used in this study. Fig. 6 illustrates the AUC values for NM, DMD, and WMD, which were 1.0000, 0.9974, and 0.9880, respectively.

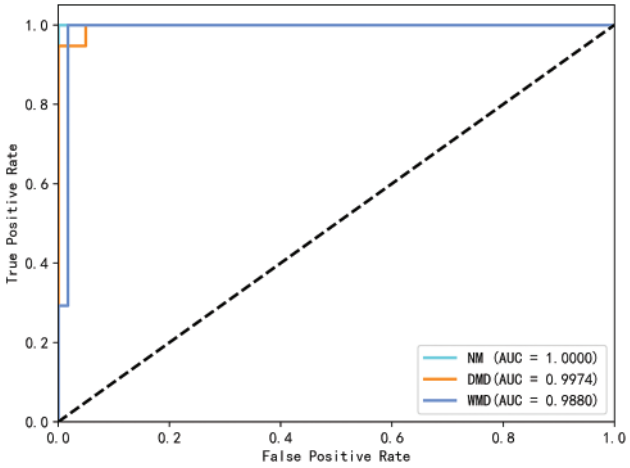

Figure 6: Roc curve

3.2 Enhanced SPPF Performance

In the comparison between SPPF and ESPPF, identical parameter settings, such as the learning rate, optimizer, and weight decay, were utilized in this study to ensure fair experimental results. As shown in Table 3, HCSP-Net (ESPPF) outperforms HCSP-Net (SPPF) model, with a 3.1% relative improvement in the kappa value (increasing from 95.3% to 98.4%) and a 4.7% relative improvement in the Jaccard value (increasing from 92.8% to 97.5%).

Table 3: Comparison of Kappa and Jaccard values between CSPDarknet53 and HCSP-Net

Model	Accuracy	Kappa	Jaccard
HCSP-Net (SPPF)	97.0%	95.3%	92.8%
HCSP-Net (ESPPF)	99.0%	98.4%	97.5%

In addition, the impact of SPPF and ESPPF on the convergence of the model was explored in this study. As depicted in Fig. 7, the initial 30 training epochs reveal a noteworthy observation: HSCP-Net (ESPPF) demonstrates a significantly higher initial accuracy compared with HCSP-Net (SPPF). Furthermore, throughout the entire training process, the accuracy curve of HSCP-Net (ESPPF) steadily ascends, ultimately converging. This indicates that HCSP-Net (ESPPF) is adept at capturing underlying patterns and features in the data during the early stages of model training, effectively mitigating the impact of data noise.

3.3 SCSP-Net Performance Insights

The compression of the depth and width of HCSP-Net resulted in an increase in the training time of the model and a significant decrease in the diagnostic accuracy to only 94%, as shown in Fig. 8A. This study introduced the HCSP-Net (ESPPF) with an accuracy of 99% as a teacher network to guide the SCSP-Net training to address this issue. Under the supervision of the teacher network, SCSP-Net successfully reduced the confusion between DMD and WMD of this model, and the accuracy was improved to 97% in Dataset (as shown in Fig. 8B). Drawing on the knowledge of the teacher network cannot only alleviate the performance degradation caused by HCSP-Net compression but also provide

more precise guidance for SCSP-Net to perform better when facing complex disease classification tasks.

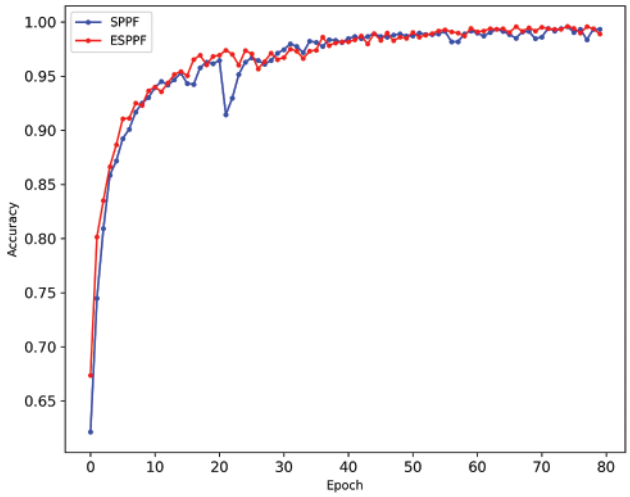

Figure 7: Comparison of the training set accuracy between HSCP-Net (SPPF) and HCSP-Net (ESPPF)

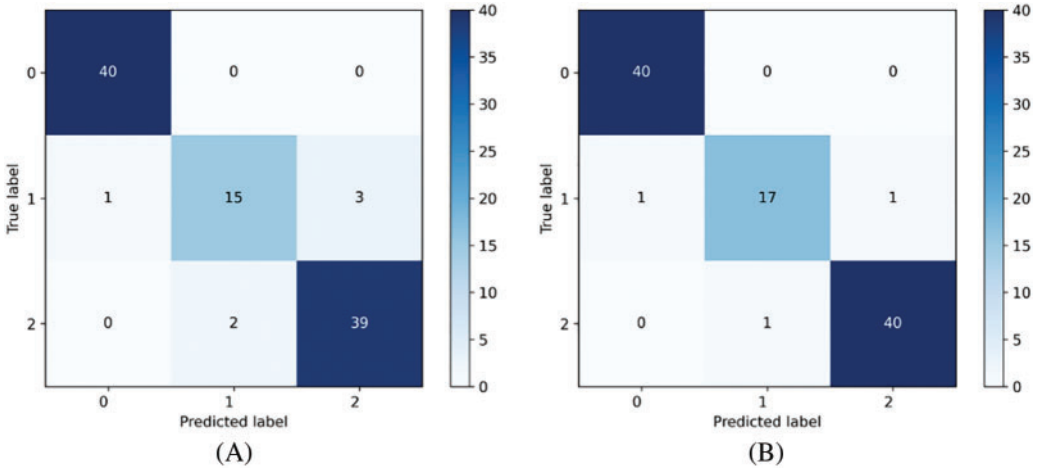

Figure 8: Confusion matrix. Subfigure (A) illustrates SCSP-Net without knowledge distillation, while subfigure (B) depicts SCSP-Net after knowledge distillation

According to the statistics in Table 4, the number of parameters and time complexity of different models were comprehensively analyzed in this study. Notably, in HCSP-Net (SPPF), replacing conventional fully connected layers with Transformer modules reduced the number of parameters. However, despite the decrease in the number of parameters, the time complexity did not show a significant declining trend. This indicated that introducing Transformer modules may face particular challenges in improving computational efficiency, requiring further in-depth research and optimization. In addition, after grouped convolutions were introduced, HCSP-Net (ESPPF) showed increased parameters and time complexity compared with the baseline model and HCSP-Net (SPPF). Although the increase was inevitable, this increase was acceptable considering the superior performance of HCSP-Net (ESPPF).

It was noteworthy that by compressing the width and depth of the model, the number of parameters and time complexity of SCSP-Net were reduced by 74.90% and 72.97%, respectively, demonstrating its advantages in lightweight design.

Table 4: Comprehensive analysis of model parameters and computational complexity

Model	Parameters (M)	FLOPs (G)
CSPDarknet53	4.1853	0.6430
HCSP-Net (SPPF)	4.1732	0.6429
HCSP-Net (ESPPF)	4.1867	0.6449
SCSP-Net	1.0509	0.1743

3.4 Ablation Experiment

In this study, the HCSP-Net model was constructed based on an innovative amalgamation of the enhanced CSPDarknet53 and the Transformer module. This model exhibited substantial distinctions compared with conventional counterparts, such as Resnet50, EfficientNetV2, InceptionV3, and ViT. The underlying rationale for this fusion design lay in the deliberate exploitation of the robust feature extraction capabilities of CSPDarknet53 and the utilization of the Transformer module to capture intricate long-term dependencies within the data, thereby facilitating a more profound exploration of abstract features within images. Regarding feature engineering, incorporating the Transformer module empowered HCSP-Net to discern and capture long-term dependencies inherent in image data more effectively, thus extracting expressive and discriminative features. Furthermore, the ESPPF module played a pivotal role during the initial stages of model training, significantly reducing noise impacts on the model. This aspect was of great significance when addressing challenges associated with data heterogeneity and inhomogeneity prevalent in medical images. To further demonstrate the superior performance of HCSP-Net, the performance of several models was evaluated using binary classification indicators. The experimental results are listed in Table 5. All models in this experiment were trained and validated on Dataset A and tested on Dataset B. The obtained test results were considered the ultimate performance indicators for each model. HCSP-Net (ESPPF) performed optimally in all other metrics compared with other models, achieving a precision of 99.2%, recall of 98.2%, F1-Score of 98.7%, PPV of 99.2%, and NPV of 99.6%. The ablation experiment also confirmed the superior performance of HCSP-Net, as the addition of the transformer and ESPPF module led to an improvement in model performance.

Moreover, to further validate the generalization capability of HCSP-Net (ESPPF), a validation experiment was performed on a publicly available dataset containing 11 types of retinal fundus images (https://www.kaggle.com/datasets/kssanjaynithish03/retinal-fundus-images/data). NM, DMD, and WMD images were extracted from this dataset to construct an external validation set. Then, HCSP-Net (ESPPF) was retrained and tested on this validation set. As shown in Fig. 9, HCSP-Net (ESPPF) achieves a classification accuracy of 100% on the test set with 256 images. This validation result demonstrated that HCSP-Net (ESPPF) had a strong generalization ability in distinguishing NM, DMD, and WMD fundus images and can adapt to unseen new datasets, laying a foundation for subsequent clinical applications.

Table 5: HCSP-Net performance evaluation: Precision, recall, F1-Score, PPV, and NPV

	Model	Precision	Recall	F1-Score	PPV	NPV
Train	ResNet50 [35]	98.448	98.736	98.589	98.448	99.430
	EfficientNetV2 [25]	98.437	99.166	97.543	98.437	99.469
	InceptionV3 [36]	98.145	98.023	98.083	98.146	99.278
	ViT [29]	98.940	99.359	99.143	98.340	99.635
	HCSP-Net (SPPF)	99.016	99.268	99.140	99.016	99.648
	HCSP-Net (ESPPF)	**99.260**	**99.387**	**99.323**	**99.260**	**99.730**
Test	ResNet50 [35]	95.714	93.924	94.709	95.713	98.210
	EfficientNetV2 [25]	94.845	94.865	94.845	94.845	98.066
	InceptionV3 [36]	94.845	94.865	94.845	94.845	98.066
	ViT [29]	94.778	92.169	93.239	94.778	97.828
	HCSP-Net (SPPF)	96.561	95.678	96.092	96.561	98.612
	HCSP-Net (ESPPF)	**99.206**	**98.246**	**98.697**	**99.206**	**99.593**

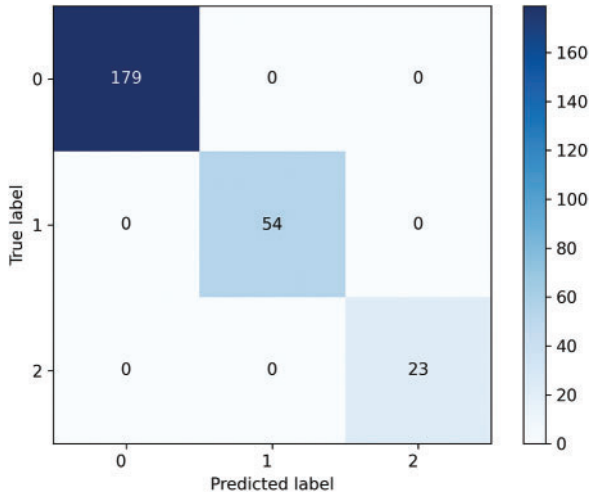

Figure 9: Confusion matrix. 0 for normal macular, 1 for dry macular degeneration, 2 for wet macular degeneration

4 Discussion

Age-related macular degeneration (AMD) is a major cause of irreversible damage to vision in individuals over the age of 50, affecting the health of millions of people worldwide [5,7,10]. Early detection and treatment of AMD contribute to slow disease progression. However, with the increasing number of patients with AMD, it is difficult for ophthalmologists to provide a comprehensive diagnostic service for this large patient population; therefore, the use of DL techniques to intelligently analyze AMD is of great clinical importance.

T-distributed stochastic neighbor embedding (T-SNE) is a nonlinear dimensionality reduction algorithm utilized for visualizing high-dimensional data [37]. As there is a lack of interpretability

in neural networks, T-SNE was employed in this study to reduce the dimensionality of the last hidden layer in both CSPDarknet53 and HCSP-Net (ESPPF) models. Characteristic probability distribution maps for various types of macular degeneration were generated to aid in interpreting the neural network models, as shown in Fig. 10. When applied to the macular degeneration triple classification task, the CSPDarknet53 model exhibited a noticeable overlap between the sample data of DMD and WMD. This suggested that CSPDarknet53 did not effectively learn the distinctive features necessary for differentiating between the two conditions. Conversely, the visualization results of HCSP-Net showed a low percentage of sample data overlap, indicating that this model had an improved ability to distinguish between DMD and WMD. This underscored the enhanced capacity of this model to learn the distinctions between the two conditions, successfully differentiating them. Additionally, it was apparent from the probability distribution plots that the probability plots of DMD and WMD overlapped, further emphasizing the relationship between DMD and WMD, namely that WMD developed in the context of DMD. Clinically, reliably differentiating DMD from WMD is critical for appropriate treatment decisions. Thus, the difficulty of CSPDarknet53 in separating these classes could lead to delays in treating WMD or unnecessary treatment for DMD due to incorrect classification. Meanwhile, the more apparent separation achieved by HCSP-Net demonstrated its potential to support more accurate clinical diagnosis and management between these AMD subtypes. These visualization results complement our analysis, offering a more comprehensive understanding of the model's behavior and its implications for clinical decision-making.

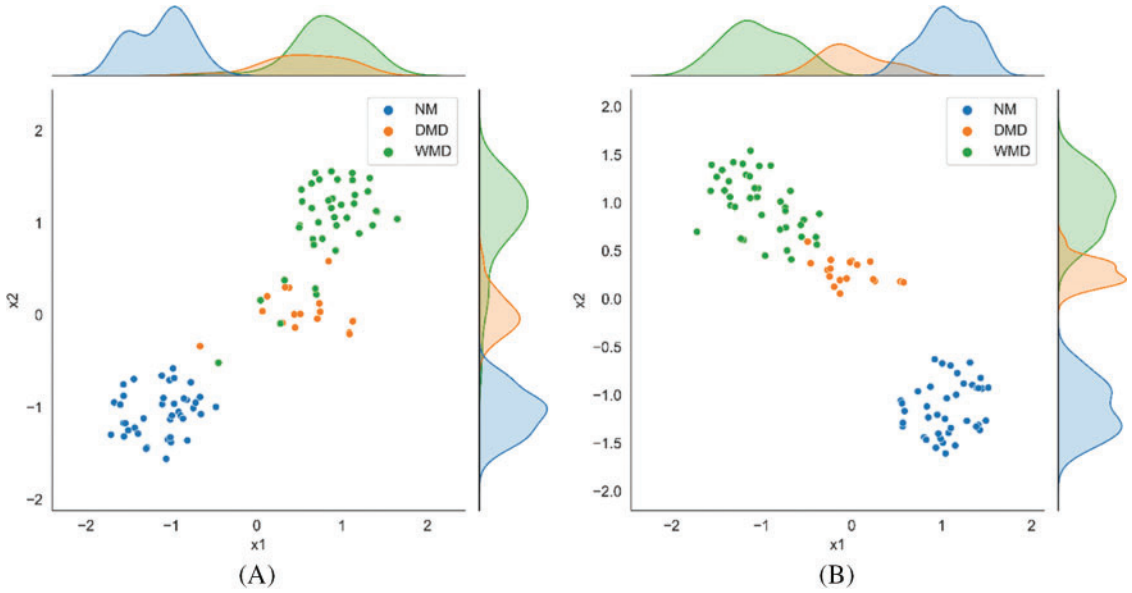

Figure 10: t-SNE dimensionality reduction visualization. (A) CSPDarknet53; (B) HCSP-Net (ESPPF)

Moreover, the HCSP-Net architecture was examined and evaluated experimentally. The Transformer model was tested using various model architectures, such as replacing only one C3 module at a time at a different location from the transformer module and replacing all C3 modules with the transformer module. However, none of these models performed as well as HCSP-Net, which may be explained from two aspects. First, the small amount of data used in this study prevented the transformer model from learning all the characteristics of each macula, which was necessary for it to obtain superiority over the convolutional neural network model [22]. Second, after pre-training on the

COCO dataset, the improved CSPDarknet53 was loaded with weights and was able to derive general features from the images. In this instance, changing the intermediate CSPDarknet53 architecture at random would nullify the impact of the pre-trained weights. The Transformer module successfully integrated its effect on the backbone feature extraction module's information without erasing the pre-trained weights, thereby enhancing the performance of HCSP-Net and allowing the model to learn new features. This allows the model to detect minute variations between various maculae, thereby improving the model's performance while preserving its lightweight characteristics.

HCSP-Net, constructed in this study, had a parameter count of 4.19 MB. Further reduction of model parameters to enhance its deployment on mobile devices represented the challenge addressed in this study. Building upon HCSP-Net (ESPPF), knowledge distillation was employed to guide the training of SCSP-Net in this study, resulting in an accuracy improvement from 94% to 97%, with parameters accounting for less than a quarter of HCSP-Net. In CSPDarknet53 and HCSP-Net, there was no misdiagnosis of DMD as NM, indicating that the model can capture the differences between the two conditions. However, even with the assistance of knowledge distillation, SCSP-Net exhibited misdiagnosis, incorrectly identifying DMD as NM. The experimental results highlighted the significant impact of model depth and width on learning data features. This indicated that larger models were more prone to extracting abstract features from the data. Consequently, striking a balance between model detection accuracy and operational speed while reducing model parameters is a key point for the future optimization of SCSP-Net.

An innovative approach was also introduced in this study by replacing the conventional fully connected classification method with a Transformer model based on the self-attention mechanism. Additionally, improvements were achieved by incorporating group convolution to expedite model convergence and enhance robustness, leading to the development of HCSP-Net (ESPPF) and achieving favorable diagnostic outcomes. All training images were sourced from individuals diagnosed with AMD, and professional ophthalmologists meticulously labeled the dataset. HCSP-Net (ESPPF) exhibited exceptional classification precision (99.2%), recall (98.2%), F1-Score (98.7%), PPV (99.2%), and NPV (99.6%). Compared with other studies, our research obviated the need for complicated data preprocessing steps, such as single-channel data extraction or vascular localization [17]. Furthermore, HCSP-Net avoided the complexity of ensemble learning or the construction of dual models to enhance accuracy [18,19]. While multimodal approaches were explored for accuracy improvement, acquiring diverse data types posed significant challenges in the medical domain [20]. Therefore, a straightforward and efficient diagnostic method that circumvented the complexities associated with data preprocessing and model assembly was presented in this study.

Nevertheless, there are still some limitations in this study. Firstly, although DL has been immensely successful in many domains, acquiring large amounts of data is frequently challenged due to ethical and privacy concerns. This study used data augmentation to increase image diversity during the data preprocessing step and data inversion to extend the dataset. However, the experimental results showed that it was still necessary to improve HCSP-Net (SPPF) by increasing the number of training rounds to deal with fluctuations in accuracy during the pre-training period. This ongoing need for improvements highlighted the complexity of working with limited data in the context of DL-based research. Secondly, a three-class classification study was conducted based on the leading criteria. Expanding the dataset with more diverse cases may help further verify the robustness of the proposed methods. Thirdly, after compressing the depth and width of the model, the diagnostic accuracy of SCSP-Net decreased significantly. Further balancing the accuracy and efficiency of the model was another limitation of this study. Finally, it is required to perform the integration of the model into clinical workflows and user studies to assess its acceptance. This will be an essential direction for further research.

This initial study focuses on model development and evaluation on a limited dataset. An important direction for future work is integrating the model into clinical workflows and evaluating its acceptance by healthcare professionals through user studies. Besides, as more AMD data with refined grading become available, the model can be retrained and modified to enhance accuracy and maintain long-term reliability. In meantime, SCSP-Net can be enhanced by leveraging the success achieved in HCSP-Net, thus providing improved support for medical professionals in diagnosing AMD.

5 Conclusion

In conclusion, AMD is a prevalent retinal disease that can cause blindness, making prompt diagnosis critical. In this study, the viability of using DL technology to aid in the classification of AMD was demonstrated. Besides, an HCSP-Net, a classification model combining a convolutional neural network and a transformer, was constructed to achieve the AMD diagnosis accuracy of 99%. This model can fulfill functions in the early diagnosis of AMD based on color fundus photography, offering valuable assistance to clinicians. In particular, it can offer strong support for the early diagnosis and screening of AMD in primary care, assist in detecting early AMD, and provide reasonable treatment recommendations, thereby enhancing the visual quality of patients.

Acknowledgement: We gratefully acknowledge the support of the Shenzhen Fund for Guangdong Provincial High-Level Clinical Key Specialties, the Sanming Project of Medicine in Shenzhen, and the Shenzhen Science and Technology Planning Project.

Funding Statement: Shenzhen Fund for Guangdong Provincial High-Level Clinical Key Specialties (SZGSP014), Sanming Project of Medicine in Shenzhen (SZSM202311012) and Shenzhen Science and Technology Planning Project (KCXFZ20211020163813019).

Author Contributions: CW and JZ: Analyzed, discussed the data, and drafted the manuscript. XH: Analyzed, discussed the data, and collected and labeled the data. SZ and WY: Designed the research, collected and labeled the data, and revised the manuscript.

Availability of Data and Materials: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics Approval: This study was approved by the Medical Ethics Committee of Shenzhen Eye Hospital (Approval Code: 2023KYPJ015, Approval Date: February 24, 2023).

Conflicts of Interest: The authors declare that this study was conducted without commercial or financial relationships that could be construed as potential conflicts of interest.

References

[1] W. L. Wong *et al.*, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis," *Lancet Glob. Health*, vol. 2, no. 2, pp. e106–e116, 2014. doi: 10.1016/S2214-109X(13)70145-1.

[2] X. Ding, M. Patel, and C. C. Chan, "Molecular pathology of age-related macular degeneration," *Prog. Retinal Eye Res.*, vol. 28, no. 1, pp. 1–18, 2009. doi: 10.1016/j.preteyeres.2008.10.001.

[3] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," *Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012. doi: 10.1016/S0140-6736(12)60282-7.

[4] C. Skevas, H. Weindler, M. Levering, J. Engelberts, M. van Grinsven and T. Katz, "Simultaneous screening and classification of diabetic retinopathy and age-related macular degeneration based on fundus photos—a prospective analysis of the RetCAD system," *Int. J. Ophthalmol.*, vol. 15, no. 12, pp. 1985–1993, 2022. doi: 10.18240/ijo.2022.12.14.

[5] R. van Leeuwen, C. C. Klaver, J. R. Vingerling, A. Hofman, and P. T. de Jong, "Epidemiology of age-related maculopathy: A review," *Eur. J. Epidemiol.*, vol. 18, no. 1, pp. 845–854, 2003. doi: 10.1023/A:1025643303914.

[6] Z. Zhang *et al.*, "Efficacy and safety of intravitreal HLX04-O, an anti-VEGF monoclonal antibody, for the treatment of wet age-related macular degeneration," *Int. J. Ophthalmol.*, vol. 15, no. 9, pp. 1549, 2022. doi: 10.18240/ijo.2022.09.20.

[7] R. Rubner, K. V. Li, and M. V. Canto-Soler, "Progress of clinical therapies for dry age-related macular degeneration," *Int. J. Ophthalmol.*, vol. 15, no. 1, pp. 157–166, 2022. doi: 10.18240/ijo.2022.01.23.

[8] P. L. Penfold, M. C. Madigan, M. C. Gillies, and J. M. Provis, "Immunological and aetiological aspects of macular degeneration," *Prog. Retinal Eye Res.*, vol. 20, no. 3, pp. 385–414, 2001. doi: 10.1016/S1350-9462(00)00025-2.

[9] J. Ambati and B. J. Fowler, "Mechanisms of age-related macular degeneration," *Neuron*, vol. 75, no. 1, pp. 26–39, 2012. doi: 10.1016/j.neuron.2012.06.018.

[10] E. Buschini *et al.*, "Recent developments in the management of dry age-related macular degeneration," *Clin. Ophthalmol.*, vol. 9, pp. 563–574, 2015.

[11] T. Lim and A. Laude, "Age-related macular degeneration: An asian perspective," *Ann.-Acad. Med. Singapore*, vol. 36, no. 10, pp. S15–S21, 2007.

[12] W. Zhou *et al.*, "Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images," *Nature Commun.*, vol. 12, no. 1, pp. 1259, 2021. doi: 10.1038/s41467-021-21466-z.

[13] P. Tschandl *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019. doi: 10.1016/S1470-2045(19)30333-X.

[14] A. Hekler *et al.*, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *Eur. J. Cancer*, vol. 118, no. 10151, pp. 91–96, 2019. doi: 10.1016/j.ejca.2019.06.012.

[15] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics," *Genome Med.*, vol. 11, no. 1, pp. 1–12, 2019. doi: 10.1186/s13073-019-0689-8.

[16] W. H. Yang, Y. Shao, and Y. W. Xu, "Guidelines on clinical research evaluation of artificial intelligence in ophthalmology (2023)," *Int. J. Ophthalmol.*, vol. 16, no. 9, pp. 1361, 2023.

[17] R. Priya and P. Aruna, "Automated diagnosis of age-related macular degeneration from color retinal fundus images," in *2011 3rd Int. Conf. Electronics Comput. Technol.*, Kanyakumari, India, 2006, vol. 2, pp. 227–230.

[18] F. Grassmann *et al.*, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmol.*, vol. 125, no. 9, pp. 1410–1420, 2018. doi: 10.1016/j.ophtha.2018.02.037.

[19] N. Motozawa *et al.*, "Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes," *Ophthalmol. Ther.*, vol. 8, no. 4, pp. 527–539, 2019. doi: 10.1007/s40123-019-00207-y.

[20] E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: A feasibility study," *J. Ophthalmol.*, vol. 2020, pp. 1–7, 2020. doi: 10.1155/2020/7493419.

[21] World Medical Association (WMA), "Declaration of Helsinki. Ethical principles for medical research involving human subjects," *Jahrbuch für Wissenschaft und Ethik*, vol. 14, no. 1, pp. 233–238, 2009. doi: 10.1515/9783110208856.233.

[22] G. Bradski, "The openCV library," *Dr. Dobb's J.: Softw. Tools Prof. Program.*, vol. 25, no. 11, pp. 120–123, 2000.

[23] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, no. 7, pp. 249–259, 2018. doi: 10.1016/j.neunet.2018.07.011.
[24] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR) Workshops*, Seattle, WA, USA, 2020, pp. 390–391.
[25] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, pp. 10096–10106, 2021.
[26] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Seattle, WA, USA, 2020, pp. 1580–1589.
[27] M. Qiu, L. Huang, and B. H. Tang, "ASFF-YOLOv5: Multielement detection method for road traffic in UAV images based on multiscale feature fusion," *Remote Sens.*, vol. 14, no. 14, pp. 3498, 2022. doi: 10.3390/rs14143498.
[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, USA, 2017, pp. 6000–6010.
[29] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," *Int. Conf. Learn. Rep.*, vol. 2010, no. 11929, pp. 1–18, 2021.
[30] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 10012–10022.
[31] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6575–6586, 2022. doi: 10.1109/TPAMI.2022.3206108.
[32] H. E. Davidson, "Filling the knowledge gaps," *Consult. Pharm.*, vol. 30, no. 5, pp. 249, 2015.
[33] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, Canada, 2019, pp. 8026–8037.
[34] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Boston, USA, 2016, pp. 770–778.
[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Las Vegas, USA, 2016, pp. 2818–2826.
[37] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.