Application of Deep Learning Neural Network Architectures in Ocular Disease Classification Models

Jie-Xin Li[1], Syuan-Hao Li[1]*, Po Sen Tsou[1]

[1]Department of Computer Science and Information Engineering, National Taitung University, Taiwan;

*Corresponding Author: hibana2077@gmail.com

DOI: https://doi.org/10.30211/JIC.202503.007

Submitted: Apr. 07, 2025 Accepted: Jun. 09, 2025

ABSTRACT

Early diagnosis of ocular diseases is critical to prevent vision impairment, yet traditional methods rely heavily on specialists' expertise and face resource allocation challenges. This study systematically evaluates the performance of diverse deep learning architectures (including CNNs, Transformers, and lightweight models) for multi-class ocular disease classification. Utilizing public ocular image datasets with transfer learning and data augmentation to address class imbalance, our experiments demonstrate that self-attention-based models achieve superior accuracy for complex conditions (e.g., diabetic retinopathy), while lightweight architectures significantly improve computational efficiency. The findings provide empirical guidelines for clinical decision-support systems and propose optimization strategies for architecture design, contributing to practical AI-driven healthcare solutions.

Keywords: deep learning, data augmentation, transfer learning, ocular disease detection

1. Introduction

According to the latest statistics from the World Health Organization, approximately 2.2 billion people suffer from various forms of visual impairment. Nearly half of these cases could be prevented from progressing to irreversible vision loss if detected and treated early [1]. Prior studies have shown that many ocular diseases can lead to irreversible vision loss or blindness if not diagnosed in time, despite the availability of effective treatments [2]. However, the current ophthalmic diagnostic systems face a significant imbalance in resource distribution. The training of professional ophthalmologists typically requires over a decade of specialized education, and in developing countries, the number of qualified eye specialists is significantly lower than in developed nations. At present, the works of ocular disease screening are mainly performed on optical coherence tomography (OCT) images and fundus images. With the development of artificial intelligence in the field of medical image processing, some related methods have achieved pleasing results.[3]

In recent years, the rapid advancement of Artificial Intelligence (AI) technologies has opened new avenues for addressing this issue. Deep Learning has achieved remarkable success

in medical image analysis, particularly in tasks such as chest X-ray interpretation and MRI analysis, where AI systems have matched or even surpassed expert-level performance. However, in the specific domain of ocular disease diagnosis, existing technologies still face two major challenges. First is the severe class imbalance problem. In the Kaggle ocular disease dataset [4] used in this study, there are as many as 3,444 images of diabetic retinopathy but only 102 images of pterygium—making the former 33 times more prevalent. This extreme imbalance significantly impairs the model's ability to recognize minority classes. Second is the issue of fine-grained inter-class differences. Some diseases, such as retinitis pigmentosa and retinal detachment, exhibit highly similar visual characteristics in color fundus photography, making accurate differentiation difficult even for experienced ophthalmologists without the aid of Optical Coherence Tomography (OCT) or other auxiliary diagnostic tools.

Given this background, the present study aims to systematically compare the performance of state-of-the-art deep learning architectures in the task of ocular disease classification, and to explore innovative training strategies that enhance model applicability in real-world clinical environments.

2. Literature Review

2.1 Machine Learning Approaches

The research on automated diagnosis of ocular diseases can be broadly divided into two major phases. Prior to the emergence of deep learning technologies, most studies relied on traditional machine learning methods. These approaches typically required handcrafted feature extraction algorithms—such as Gabor filters or morphological operations—to capture critical characteristics in fundus images, including vascular structures and hemorrhagic spots. For instance, the method proposed in [5] utilized a combination of Gabor filters to extract retinal texture features, achieving a classification accuracy of 78% in diabetic retinopathy detection tasks. Although such methods could deliver reasonable performance on specific datasets, their generalization capability was often severely constrained by the limitations of manual feature design.

2.2 Deep Learning Approaches

With the rapid advancement of deep learning technologies—particularly the breakthrough developments of Convolutional Neural Networks (CNNs) in the field of computer vision—the research on automated ocular disease diagnosis has entered a new phase. [6] Image classification was one of the first areas in which deep learning made a principal contribution to medical image analysis.

In recent years, convolutional neural networks (CNNs) have been used to analyze medical images, and they have achieved impressive performance on medical image datasets.

In a landmark study in 2017, Esteva et al. [7] demonstrated that deep neural networks pre-trained on large-scale natural image datasets (such as ImageNet) could be effectively

transferred to medical image analysis tasks through appropriate fine-tuning.

In a landmark study in 2025, Benny Sukma Negara et al. [8] illustrate that CNN-based models can automatically learn hierarchical representations from raw fundus images, eliminating the need for manual feature extraction.

In recent years, increasingly sophisticated network architectures and training strategies have been introduced to this domain. For instance, the model proposed in [9] incorporates a channel attention mechanism, integrating a Multi-scale Feature Representation (MFR) module to capture features at varying scales, along with a Channel-Spatial Dual Attention (CSDA) module to emphasize salient regional features. Compared with traditional methods, this approach achieves higher classification accuracy and computational efficiency. Experimental results on datasets such as Retinal fundus image (RFI), Online Retinal fundus Image database for Glaucoma Analysis and research (ORIGA) database., and High-resolution fundus (HRF) Image Database have demonstrated its superior performance.

Another notable approach is presented in [10], which applies a Vision Transformer (ViT)-based framework to grade Diabetic Retinopathy (DR). The method has shown excellent performance across fundus image datasets with varying resolutions, outperforming CNN-based and other baseline models in terms of accuracy (91.4%), AUC (0.986), sensitivity (0.926), and specificity (0.977). Compared to CNNs [11], Vision Transformers (ViTs) are free from convolution-induced biases, enabling them to learn global features and capture complex relationships in the data more effectively. ViTs [12] also leverage a self-attention mechanism that enhances their ability to model long-range dependencies. As a result, ViTs [13] achieve competitive or even superior performance compared to state-of-the-art convolutional networks, while requiring significantly fewer computational resources during training.

2.3 Existing Challenges

Despite the remarkable progress achieved through deep learning in automated ocular disease diagnosis, several critical challenges remain unresolved. Among the most pressing is the issue of data scarcity, particularly concerning clinically rare disease categories. For example, in the dataset used in this study, the number of samples for pterygium accounts for only 0.6% of the total data. This extreme class imbalance hinders the model's ability to effectively learn representative features of such minority classes, ultimately compromising diagnostic accuracy.

Another major challenge lies in the limited interpretability of model decision-making processes. In the highly specialized field of medical diagnosis, clinicians must be able to understand the rationale and reasoning behind AI-generated predictions. However, most current research focuses predominantly on improving performance metrics such as accuracy, often at the expense of exploring and visualizing the internal decision mechanisms of models. This lack of transparency significantly limits the practical applicability of AI systems in real clinical settings.

Addressing these challenges requires interdisciplinary collaboration across both technical

and clinical domains. For the issue of data scarcity, Generative Adversarial Networks (GANs) offer a promising solution. A study by Frid-Adar et al. (2018) [14] demonstrated that synthetic medical images could improve CNN classification performance by up to 15%. As for model interpretability, the integration of explainable AI techniques such as attention mechanisms [15] and Gradient-weighted Class Activation Mapping (Grad-CAM) [16] is likely to become a key area of future research. Only by overcoming these critical barriers can automated ocular disease diagnosis truly bridge the gap from laboratory research to clinical deployment.

3. Research Design

3.1 Dataset and Preprocessing

This study utilizes a publicly available ocular disease dataset from the Kaggle platform as the foundation for experimentation. The dataset comprises approximately 15,328 high-quality fundus images, covering 10 common types of ocular diseases. All images are meticulously annotated, encompassing a range of conditions from prevalent diabetic retinopathy to rarer diseases such as pterygium. The dataset is partitioned into training, validation, and test sets using a 7:2:1 split ratio.

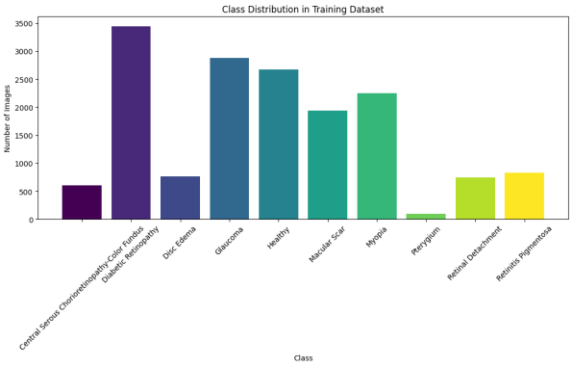

Figure1. Dataset Image Statistics

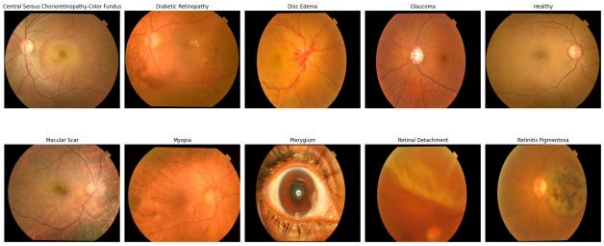

Figure2. Image Categories

To ensure data quality, we implemented a rigorous preprocessing pipeline, including the removal of low-quality images, uniform resizing of all images to a resolution of 224×224 pixels, and pixel normalization. Specifically, pixel values were linearly scaled from the original range [0, 255] to the interval [-1, 1] using mean = 0.5 and standard deviation = 0.5. The normalization formula is defined as:

$$I_{norm} = \frac{\mathrm{I} - 127.5}{127.5}$$

I: input image

To enhance the model's generalization ability, address class imbalance, and mitigate overfitting, we employed various data augmentation techniques. These included targeted geometric transformations (random rotations within ±15 degrees and vertical/horizontal

flipping) as well as subtle adjustments to brightness and contrast (controlled within ±10% and ±20%, respectively). These strategies effectively increased the data diversity of minority classes and helped alleviate overfitting.

3.2 Model Architecture

In selecting the model architecture, we focused on comparing five deep learning models, as shown in Table 1.

Table 1. Model architecture

Model name	Parameters	Input size
EfficientNet-B7	66.3M	224×224×3
MobileNetV3-Large	5.4M	224×224×3
DenseNet121	8.0M	224×224×3
GhostNet_100	5.2M	224×224×3
EVA-02_large	300M	224×224×3

EfficientNet-B7, recognized as one of the most efficient models to date, adopts a compound scaling method to balance network depth, width, and resolution. In this study, we froze the parameters of its first five MBConv(Mobile Inverted Bottleneck Convolution) feature extraction blocks and redesigned only the top classification head, incorporating a fully connected network with a 512-dimensional hidden layer and a dropout rate of 0.4.

DenseNet121 utilizes a densely connected structure to promote feature reuse, with each layer directly connected to all subsequent layers, effectively alleviating the vanishing gradient problem. We froze parts of its intermediate layers and optimized the classifier structure to better adapt to the specific task.

EVA-02, a novel Vision Transformer architecture, captures global feature relationships through a self-attention mechanism, demonstrating strong long-range dependency modeling capabilities.

GhostNet employs ghost modules to generate more feature maps with minimal computational cost, making it particularly suitable for deployment on mobile devices.

MobileNetV3 combines neural architecture search techniques with the h-swish activation function, significantly reducing the number of parameters while maintaining accuracy.

As lightweight models, GhostNet and MobileNetV3 are particularly suitable for deployment in resource-constrained environments, such as edge computing devices like Raspberry Pi. For MobileNetV3, we only unfroze the last two inverted residual blocks to balance transfer learning performance with computational cost.

3.3 Training Strategy

The batch size during training was uniformly set to 16 to ensure a fair comparison. We

adopted the cross-entropy loss function and applied class-specific weighting based on the inverse of the number of samples per class, enabling the model to pay more attention to minority classes.

The cross-entropy (CE) loss is defined in Equation (1), where yi represents the true label and pi denotes the predicted probability.

$$CE = -\sum(yi * \log(pi)) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{[Formular 1]}$$

The Adam optimizer was employed with parameters set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate(lr) was configured at 0.001, and adjusted dynamically using the ReduceLROnPlateau strategy, which reduced the learning rate by a factor of 10 upon stagnation of the validation loss.

To prevent overfitting, we implemented an EarlyStopping mechanism, terminating training if the validation metric did not improve for five consecutive epochs.

Epoch, in machine learning, refers to the one entire passing of training data through the algorithm. It's a hyperparameter that determines the process of training the machine learning model.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4070 GPU.

3.4 Evaluation Metrics

The evaluation metrics included not only traditional measures such as accuracy, precision, and recall, but also incorporated confusion matrix analysis to reveal the model's error patterns for specific classes.

The confusion matrix, as shown in Table 2, defines True Positive (TP) as the number of samples correctly predicted as positive, False Negative (FN) as the number of positive samples incorrectly predicted as negative, False Positive (FP) as the number of negative samples incorrectly predicted as positive, and True Negative (TN) as the number of samples correctly predicted as negative. Based on the confusion matrix, the evaluation metrics for classification tasks, including accuracy (2), precision (3), recall (4), and F1-score (5), were computed.

Table 2. Confusion Matrix

(Confusion Matrix)		
Actual\Predict	**Positive**	**Negative**
Positive	**True Positive (TP)**	**False Positive (FP)**
Negative	**False Negative (FN)**	**True Negative (TN)**

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \cdots\cdots [\text{Formular 2}]$$

$$Precision = \frac{TP}{(TP+FP)} \cdots\cdots [\text{Formular 3}]$$

$$Recall = \frac{TP}{(TP+FN)} \cdots\cdots [\text{Formular 4}]$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \cdots\cdots [\text{Formular 5}]$$

Experimental Results

Through systematic experimental comparisons, we obtained extensive result data. In terms of overall performance, EfficientNet-B7 demonstrated the best results, achieving a classification accuracy of 87.16%, with a precision of 86.32%, a recall of 85.74%, and a stable F1-score of 86.03%. These results validate the effectiveness of the compound scaling strategy for medical image classification tasks. Although MobileNetV3 exhibited slightly lower absolute performance (accuracy of 85.82%), it showed significant advantages in training efficiency, requiring only 350 seconds per epoch—42% less time compared to EfficientNet-B7—which makes it particularly valuable for real-time application scenarios. DenseNet121 achieved comparable performance to MobileNetV3 (accuracy of 85.96%), but demanded more computational resources. A summary of the experimental results is presented in Table 3.

Table 3. The experimental results

Model	Accuracy	Precision	Recall	F1	Epoch
EfficientNet-B7	**87.16**	**86.32**	**85.74**	86.03	650s
MobileNetV3	**85.82**	84.91	84.12	84.51	**350s**
DenseNet121	85.96	85.07	84.35	84.71	340s
GhostNet_100	82.41	81.23	80.67	80.95	440s
EVA-02_large	79.83	78.92	77.56	78.23	720s

In terms of specific pathological categories, the performance differences among different models were more pronounced. For diseases with sufficient sample sizes, such as diabetic retinopathy and retinal detachment, all models performed excellently, with EfficientNet-B7 achieving recall rates of 96.3% and 97.4%, respectively, for these two categories. However, for rare conditions like pterygium, even with data augmentation and class weighting, the recall rate of the best-performing model was only 93.6%. This highlights the ongoing challenge of small sample learning.

The visual analysis of the training process revealed the learning characteristics of different models. The loss curves indicated that MobileNetV3 converged the fastest, reaching a stable state within 20 epochs. In contrast, EVA-02_large, due to its large parameter size (over 300M), exhibited significant overfitting, with validation loss fluctuations exceeding 15%. This study demonstrates that, in medical image analysis tasks, a more complex model is not always better; an appropriately designed deep learning network may achieve better generalization performance.

Discussion and Conclusion

The systematic experiments in this study provide important empirical references for the field of automated diagnosis of ocular diseases. From a clinical application perspective, the choice of model for different scenarios requires a comprehensive consideration of multiple factors. In settings where computational resources are abundant and the highest diagnostic accuracy is prioritized, EfficientNet-B7 is undoubtedly the best choice. However, in scenarios requiring mobile deployment or real-time analysis, the lightweight characteristics of MobileNetV3 make it a more suitable candidate. It is worth noting that current research still has several limitations, the most prominent being insufficient data coverage, particularly for late-stage diseases such as proliferative diabetic retinopathy. This limitation somewhat restricts the clinical applicability of the model.

Looking ahead, we believe several important directions are worth further exploration. First, the integration of multimodal data is a promising avenue. Current research is primarily based on color fundus photography, while other modalities such as OCT imaging could provide complementary diagnostic information. Second, improving model interpretability is essential. By integrating techniques like Grad-CAM to generate heatmaps for lesion localization, clinical trust in AI systems could be significantly enhanced.

In conclusion, through systematic comparative experiments, this study comprehensively evaluated the application of deep learning in ocular disease classification. We not only verified the outstanding performance of EfficientNet-B7 but also provided MobileNetV3 as an efficient alternative for resource-constrained scenarios.

References

[1] World Health Organization. Blindness and visual impairment. Available online: https://www.who.int/zh/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] Dipu, N.M., Shohan, S.A. and Salam, K. Ocular disease detection using advanced neural network based classification algorithms. Asian Journal for Convergence in Technology, 2021, 7(2), 91–99. DOI: 10.33130/AJCT.2021v07i02.019.

[3] Li, N., Chen, H., Xu, Y., Wang, Y. and Li, Y. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. BenchCouncil International Symposium, 2021.

[4] Ruhul Amin Sharif. Eye disease image dataset. Available online: https://www.kaggle.com/datasets/ruhulaminsharif/eye-disease-image-dataset/data.

[5] Kuri, S.K. Automatic diabetic retinopathy detection using Gabor filter with local entropy thresholding. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), Kolkata, India: IEEE, 2015, 411–415. DOI: 10.1109/ReTIS.2015.7232914.

[6] Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. and Hajirasouliha, I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine, 2018, 27, 317–328.

[7] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542, 115–118. DOI: https://doi.org/10.1038/nature21056.

[8] Irsyad, M. and Negara, B.S. Classification of diabetic retinopathy using EfficientNet-B7 with hyperparameter optimization. Engineering and Technology Journal, 2025, 10(3), 3984–3989. DOI: 10.47191/etj/v10i03.06.

[9] Das, D., Roy, S., Mondal, R., Pal, A. and Dey, N. CDAM-Net: Channel shuffle dual attention based multi-scale CNN for efficient glaucoma detection using fundus images. Engineering Applications of Artificial Intelligence, 2024, 133, 108454.

[10] Wu, J., Sun, Z., Zhao, H., Zhang, Y. and Yang, Y. Vision transformer-based recognition of diabetic retinopathy grade. Medical Physics, 2021, 48(12), 7850–7863. DOI: 10.1002/mp.15312.

[11] Deininger, L., Seibold, H., Zettwitz, S. and Maier, A. A comparative study between vision transformers and CNNs in digital pathology. arXiv preprint, 2022. arXiv:2206.00389.

[12] Xiao, Z., Li, Y., Zhang, Y., Zhu, X. and Zheng, Y. MM-UNet: A mixed MLP architecture for improved ophthalmic image segmentation. OMIA@MICCAI, 2024.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv, 2020. arXiv:2010.11929.

[14] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 2018, 321, 321–331.

[15] Bai, N., Zhou, W., Hu, S., Lin, D., Tang, C. and Zhang, Y. Deep learning methods with the improved attention for explainable image recognition. IEEE Access, 2024, 12, 70559–70567. DOI: https://doi.org/10.1109/ACCESS.2024.3397323.

[16] Selvaraju, R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 2020, 128, 336–359. DOI: https://doi.org/10.1007/s11263-019-01228-7.