Multi-Modality Semi-Supervised Learning for Ophthalmic Biomarkers Detection

Yanming Chen[a], Chenxi Niu[a], Chen Ye[a], Shengji Jin[a], Yue Li[a], Chi Xu[a], Keyi Liu[a], Haowei Gao[a], Jingxi Hu[a], Yuanhao Zou[a], Huizhong Zheng[a], and Xiangjian He[a]

[a]University of Nottingham Ningbo, China

ABSTRACT

Ophthalmic Biomarkers, as an objective and quantifiable approach to identifying the ophthalmological disease process, are proven to be useful not only in assisting healthcare professionals in disease diagnosis but also in the identification of phenomena and risk factors in the early stages, which greatly contribute to disease prevention and better treatment of patients. In this study, a deep learning method is introduced to achieve simultaneous automatic recognition of six prevalent ophthalmic biomarkers in the OLIVES dataset. To enhance identification accuracy, semi-supervised learning techniques are adopted in this research and different data modalities are jointly optimized using a guided loss function. The experimental results reveal that the method reaches an F1 score of 0.70 on a test set with 3,872 images.

Keywords: semi-supervised learning, ophthalmic biomarkers, multi-modality, disease diagnosis

1. INTRODUCTION

The human eye, a complex organ, offers significant diagnostic opportunities, particularly through optical coherence tomography (OCT) scans. These scans offer comprehensive insights into ocular health by objectively detecting specific biomarkers, which serve as objective indicators of medical conditions independent of a patient's subjective perception of their health.[1] The integration of Machine Learning, particularly deep learning, into ocular diagnostics, has enabled automated interpretation of biomarkers in OCT images.[2] However, challenges like data generalization and personalization persist. OCT datasets often lack comprehensive biomarker labels and have limited images per biomarker, hindering model generalization. This is further compounded by the minimal variance seen in OCT scans from the same patient across different visits, contrasted by pronounced differences when comparing patients with identical diseases.

In the intricate domain of ophthalmology, numerous biomarkers play pivotal roles in diagnostics. Our study specifically zeroes in on six of these: Intraretinal Hyperreflective Foci (IRHRF), Partially Attached Vitreous Face (PAVF), Fully Attached Vitreous Face (FAVF), Intraretinal Fluid (IRF), Diffuse Retinal Thickening or Diabetic Macular Edema (DRT/ME), and Vitreous Debris (VD). These are rigorously assessed against the OLIVES dataset,[3] which is furnished with 1,268 near-IR fundus images and 49 OCT scans per image. To provide insights into these biomarkers: IRHRF appears as bright spots in the retina, similar to microaneurysms or exudates.[4] PAVF and FAVF, on the other hand, signal the attachment state of the vitreous to the internal limiting membrane (ILM). While IRF becomes evident with hyporeflective areas within the retina, DRT/ME signifies abnormal retinal thickness.[5,6] The identification of VD hinges on the detection of hyperreflective elements in the vitreous or the shadowing without intraretinal bleeding.[7]

The research described here incorporates a comprehensive approach to biomarker detection in ophthalmology by utilizing advanced deep learning techniques and leveraging the capabilities of the VGG-16 model[8] and the OLIVES dataset. The key contributions of this study can be summarized as follows.

1. **Utilization of the OLIVES Dataset.** We use the OLIVES dataset to train a deep learning model for detecting six biomarkers.
2. **Incorporation of Multi-Modality Learning.** We integrate patient-personalized clinical labels with OCT scans and optimize both models using a guided loss function.
3. **Application of Semi-Supervised Learning Techniques.** We apply these techniques for model optimization, achieving an F1 score of 0.70, which surpasses the baseline by approximately 7%.

International Workshop on Advanced Imaging Technology (IWAIT) 2024, edited by Masayuki Nakajima, Phooi Yee Lau, Jae-Gon Kim, Hiroyuki Kubo, Chuan-Yu Chang, Qian Kemao, Proc. of SPIE Vol. 13164, 1316433
 · 0277-786X · doi: 10.1117/12.3019655

2. RELATED WORK

In the past, the establishment and annotation of ophthalmic biomarkers rely on manual completion, based on the scholarly knowledge of experts as well as their professional experience. However, this might lead to several downsides: 1) Manual labeling and annotation are time-consuming and labor-intensive, greatly diminishing the efficiency of diagnosis and treatment of ophthalmic diseases; 2) Variability and inaccuracy in labeling due to bias in recognition by diverse specialists;[9] 3) Specialists limit in discovering new biomarkers due to current dogma.[10]

Nowadays, the advent of deep learning has caused drastic changes in the field of ophthalmology. With its proven accuracy in classification comparable to expert results and its outperformance of other machine learning in image analysis, it has become the hottest technology in ophthalmic disease recognition.[9] Convolutional neural networks (CNN) represent the predominant deep learning models extensively applied across a spectrum of image processing domains, including medical imaging. It employs an image processing filter to learn the image features of pathological indicators from the training dataset and implement automatic extraction.[11] According to Ting et al.,[12] the CNN model consistently demonstrates outstanding outcomes in the realm of medical image analysis. The notable experiment result of Liu et al.[13] indicates the super sensitivity of CNNs in tasks in grading, density, and precise localization of pediatric cataracts when contrasted with the performance of pediatric ophthalmologists. Moreover, CNNs have also been applied to detect intraretinal fluid at an impressive cross-validated dice coefficient of 0.911.[11]

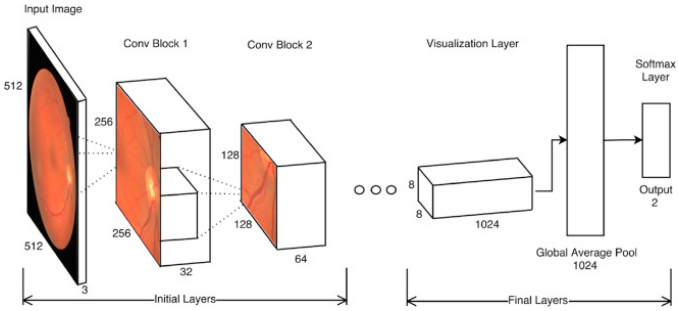

Figure 1. Typical flow chart of CNN[11]

The remarkable achievements of deep learning in medical image analysis have paved the way for its applications in OCT (Optical Coherence Tomography) image processing. Kim et al.[14] develop ML models with strong predictive power and interpretability for glaucoma diagnosis based on RNFL thickness and VF, with sensitivities, accuracies, specificities, and AUCs of 0.983, 0.98, 0.975, and 0.979, respectively. Using single wide-field OCT, a hybrid deep learning approach could effectively distinguish between healthy glaucoma suspects and early glaucoma by generating probability maps. This method achieves an impressive accuracy of up to 93.1%, surpassing the performance of existing OCT parameters.[15] Schlegel et al.[16] design a fully automated deep-learning-based diagnostic method for IRC detection and quantification of three macular lesions with an average accuracy (AUC) of 0.94 (range, 0.91-0.97), an average precision of 0.91, and an average recall of 0.84. In an experiment using deep learning to detect age-related macular degeneration (AMD) from OCT images, 100% accuracy is achieved on a new dataset after training on 1.2 million images of AMD patients using the TensorFlow deep learning approach.[17]

3. DATASET AND METHODOLOGY

This section provides an overview of the OLIVES dataset utilized in our study and outlines our methodological approach, including image pretreatment, dataset division, model evaluation metrics, and the specifics of our training process.

3.1 Dataset

The dataset used in this study is the Ophthalmic Labels for Investigating Visual Eye Semantics (OLIVES) dataset, introduced in.[3] OLIVES contains 96 eye's data, which consist of 1268 near-IR fundus images each with

at least 49 OCT scans, and 16 biomarkers, along with 4 clinical labels and a disease diagnosis of DR or DME. Table 1 provides the summary statistics and overview of the OLIVES dataset. The advantage of the OLIVES dataset is that it researches the relationships and interactions between all relevant data over a treatment period, the dataset contains time-series data, which spans an average of 66 weeks of treatment and 7 injections per eye, and 1D clinical labels can be combined with 3D OCT scans to provide a more personalized diagnosis.

Detail	**OCT**	**Fundus**	**Clinical**	**Biomarker**
Per Visit	49	1	4	16
Per Eye	$N_p \times 49$	N_p	$N_p \times 4$	1568
Total	78189	1268	5072	150528
General Overview				
96 Eyes, Visits 4-16 weeks, Avg. 16 visits/eye, Avg. 7 injections/patient				
Clinical Labels				
BCVA, CST, Patient ID, Eye ID				
Biomarkers				
IRHRF, FAVF, IRF, DRT/ME PAVF, VD, Preretinal Tissue, EZ Disruption				
IR Hemorrhages, SRF, VMT, Atrophy, SHRM, RPE Disruption, Serous PED				

Table 1. OLIVES Dataset Summary and Overview

3.2 Image Pretreatment

A combination of standard and advanced preprocessing steps is applied in the pretreatment of our OCT images. Before applying the transformation, the images are normalized based on the mean and standard deviation computed over the entire dataset. The OCT images are resized to a dimension of 224×224 pixels and basic transformation including horizontal flipping and color adjustments are applied. Some of the images are transformed to grayscale to ensure the model's robustness to color variations. Elastic transformation is a type of data augmentation method that introduces local distortions into images. In our implementation, elastic transform is employed to simulate the natural deformations in different human eye structures. Given an image I, the elastic transformation produces deformed image I':

$$I'(x, y) = I(x + \Delta x, y + \Delta y), \tag{1}$$

where Δx and Δy are displacement fields generated by Gaussian random fields. Random fields f_x and f_y are generated from Gaussian distribution N:

$$f_x, f_y \sim N(0, \sigma^2). \tag{2}$$

Random fields f_x and f_y then convolved with a Gaussian kernel G to produce smooth displacement fields:

$$\Delta x = G * f_x, \tag{3}$$

$$\Delta y = G * f_y. \tag{4}$$

Based on our experiment result, applying the elastic transformation to OCT images can improve the model accuracy by two points.

3.3 Test set and Analysis Method

The OLIVES dataset is divided into a training set and a test set. The training set comprises 9402 OCT images and the test set contains 3872 images. Both of the sets are supported with biomarkers labels and 1D clinical labels across 40 patients. To assess the models' performance on binary classification of biomarkers, the F1 score is employed as our primary evaluation metric. Given:

$$\text{Precision (P)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

and

$$\text{Recall (R)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

The F1 score is defined as:

$$F1 = \frac{2 \times P \times R}{P + R}$$

where Precision is $\frac{TP}{TP+FP}$ and Recall is $\frac{TP}{TP+FN}$, with TP, FP, and FN representing true positives, false positives, and false negatives, respectively.

3.4 Training Process

In our research, we develop a training framework that adeptly combines various methodological advancements for the enhancement of models. A guided loss mechanism[3] is used for concurrent training of models that process biomarkers and clinical labels. This guided loss in Eq.5, the mean square error between MLP logits and VGG-16 logits, functions by minimizing the logits between the Multi-Layer Perceptron (MLP) model handling biomarkers/clinical labels and the VGG-16 model analyzing OCT features. This minimization of disparity between the logits $\phi_{\text{MLP}}(x_i)$ and $\phi_{\text{VGG-16}}(x_i)$ at each training epoch until the stopping criteria for training is met. In addition to the guided loss in Eq.6, the other two terms $\mathcal{L}_{\text{VGG-16}}$ and $\mathcal{L}_{\text{MLP}}$ are binary cross-entropy losses computed between the ground truth labels and logits from each model respectively. By using these three terms to jointly optimize two models, we enable the transfer of knowledge from the MLP to the VGG-16 model.

$$\mathcal{L}_{\text{Guided}} = 1[\hat{y}^{\text{MLP}} = y] \frac{1}{2} \left\| \phi_{\text{VGG-16}}(x_i) - \phi_{\text{MLP}}(x_i) \right\|_2^2 \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{\text{VGG-16}} + \mathcal{L}_{\text{MLP}} + \mathcal{L}_{\text{Guided}} \tag{6}$$

We also implement a dynamic learning rate schedule in conjunction with the data augmentation pipeline. This combination of strategies, featuring a warm-up period followed by cosine annealing intertwined with random resizing, cropping, horizontal flipping, color jittering, and occasional grayscale conversion, enriches the training dataset with varied transformations. We employ a five-fold cross-validation technique to ensure robustness and prevent over-fitting. This method divides the dataset into five distinct subsets, each in turn serving as the test set, while the others form the training set. This iterative process, coupled with the averaging of outcomes from each fold, provides a more comprehensive assessment of the model's predictive capabilities. Building upon this protocol, we introduce a semi-supervised learning component. Starting with the best-performing model, we engage it in training on the OLIVES dataset to generate pseudo-labels for the RECOVERY dataset. These pseudo-labels, combined with the true labels from OLIVES, create an expanded training set. In successive training phases, the model refines its ability to discern and predict with heightened accuracy, continually leveraging and enhancing these pseudo-labels for ongoing improvement.

4. EXPERIMENT AND RESULTS

This section details our experiments and findings, beginning with the selection of a baseline model from various state-of-the-art image processing models, and then evaluating our proposed method using the VGG16 model. The results include performance metrics and analyses for different biomarkers, evaluate the effectiveness of our approach.

4.1 Baseline Model Selection

To establish the baseline, we train multiple variants of the state-of-art image processing models including ResNets,[18] VGGNets,[8] Inceptions,[19] and EfficientNets.[20] We find that parameter size is not a reliable indicator of the models' performance. Table 2 shows the baseline models' accuracy on our test dataset and VGG16 proves to be the most effective model with the highest F1 score across six biomarkers and a relatively lower parameter size than VGG19. Given its superior performance, we have chosen VGG16 as our baseline model for further improvement and development in our research.

Model	F1 score	Parameters
Resnet101	0.62	44.5 million
Resnet50	0.61	25.6 million
VGG16	0.63	138 million
VGG19	0.63	143 million
InceptionV3	0.58	23.8 million
Efficient_B3	0.59	12 million
Efficient_B2	0.59	9 million
Efficient_B1	0.56	7.8 million

Table 2. Comparison of baseline model accuracy

4.2 Evaluation of Proposed Method

This section evaluates our proposed method using the VGG16 model on the test set based on the evaluation method detailed in section 3.3, comparing its performance against the baseline models.

Table 3 presents a comparative analysis of the accuracy for both the baseline and proposed methods across six biomarkers. The proposed method shows improved precision and F1 scores for all biomarkers. For instance, the Partially Attached Vitreous Face (PAVF) exhibits an increase of 0.16 in the F1 score, compared to 0.54 in the baseline and the Fully Attached Vitreous Face (FAVF) shows the highest F1 score of 0.88. Other biomarkers, such as Intraretinal Hyperreflective Foci (IRHRF), and Vitreous Debris (VD), also show increased F1 scores above 0.70. However, biomarkers like Diffuse Retinal Thickening or Diabetic Macular Edema (DRT/DME) and Intraretinal Fluid (IRF) still present challenges in classification accuracy, and the accuracy of IRF and DRT/DME have more than 20 percent gap compared to others.

	Baseline Method			Proposed Method			
Biomarker	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Support
IRHRF (B1)	0.64	0.80	0.71	0.85	0.66	0.74	1204
PAVF (B2)	0.72	0.43	0.54	0.68	0.73	0.70	1219
FAVF (B3)	0.73	0.98	0.84	0.81	0.97	0.88	2010
IRF (B4)	0.37	0.77	0.50	0.49	0.70	0.58	695
DRT/DME (B5)	0.45	0.51	0.48	0.65	0.44	0.53	143
VD (B6)	0.64	0.83	0.72	0.69	0.83	0.75	426
MACRO AVG	0.59	0.72	0.63	0.70	0.76	0.70	5697

Table 3. Comparison of Baseline and Proposed Method Accuracy on Biomarkers

5. CONCLUSION

In conclusion, our research focuses on exploring the OLIVES dataset to enable the auto-detection of ophthalmic biomarkers using deep learning techniques to assist the diagnosis process. Compared to the existing OCT scans dataset which is limited to a single data modality, the OLIVES dataset delves into the relationship between all relevant data such as 1D data modality clinical labels (e.g., Patient ID and Eye ID) and 3D data modality clinical (e.g., OCT scans). Rooted in these previous findings, our proposed method has applied a data augmentation pipeline and improved the training process with learning rate scheduling and five-fold cross-validation. A guided loss[3] has been implemented to jointly optimize different data modalities and semi-supervised learning has been embedded in our training process to further refine the model's performance on the test set. Overall, the achieved F1 score is approximately 7 percent higher than the baseline result, ranked fourth place in the IEEE Video and Image Processing Cup 2023.

Our future research will focus on further improving the model's accuracy on the OLIVES dataset by investigating and introducing attention mechanisms to certain biomarkers such as IRF which exhibit a lower discriminative capacity. Meanwhile, we will also explore the time-series data in OLIVES, since this dataset

provides data for each patient's different visits over a period of time, this offers a dynamic perspective on the progression of ophthalmic conditions.

REFERENCES

[1] Strimbu, K. and Tavel, J. A., "What are biomarkers?," *Current Opinion in HIV and AIDS* **5**(6), 463 (2010).

[2] Saha, S., Nassisi, M., Wang, M., Lindenberg, S., Kanagasingam, Y., Sadda, S., and Hu, Z. J., "Automated detection and classification of early amd biomarkers using deep learning," *Scientific reports* **9**(1), 10990 (2019).

[3] Prabhushankar, M., Kokilepersaud, K., Logan, Y.-y., Trejo Corona, S., AlRegib, G., and Wykoff, C., "Olives dataset: Ophthalmic labels for investigating visual eye semantics," *Advances in Neural Information Processing Systems* **35**, 9201–9216 (2022).

[4] Mohankumar, A., O'Keefe, G. D., Kim, L. A., Dedania, V. S., and Tsui, E., "Hyperreflective foci in optical coherence tomography," (June 2023).

[5] Bhende, M., Shetty, S., Parthasarathy, M. K., and Ramya, S., "Optical coherence tomography: A guide to interpretation of common macular diseases," *Indian journal of ophthalmology* **66**(1), 20 (2018).

[6] Bhagat, N., Zarbin, M. A., and Mukkamala, L., "Diabetic macular edema," *American Academy of ophthalmology* (2014).

[7] Strzalkowski, P., Schuster, A., Strzalkowska, A., Steinberg, J., and Dithmar, S., "Semi-automated quantification of vitreal hyperreflective foci in sd-oct and their relevance in patients with peripheral retinal breaks," *BMC ophthalmology* **23**(1), 324 (2023).

[8] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[9] Wu, J.-H. and Liu, T. Y. A., "Application of deep learning to retinal-image-based oculomics for evaluation of systemic health: A review," *Journal of Clinical Medicine* **12**(1), 152 (2022).

[10] Waldstein, S. M., Seeböck, P., Donner, R., Sadeghipour, A., Bogunović, H., Osborne, A., and Schmidt-Erfurth, U., "Unbiased identification of novel subclinical imaging biomarkers using unsupervised deep learning," *Scientific reports* **10**(1), 12954 (2020).

[11] Grewal, P. S., Oloumi, F., Rubin, U., and Tennant, M. T., "Deep learning in ophthalmology: a review," *Canadian Journal of Ophthalmology* **53**(4), 309–313 (2018).

[12] Ting, D. S., Peng, L., Varadarajan, A. V., Keane, P. A., Burlina, P. M., Chiang, M. F., Schmetterer, L., Pasquale, L. R., Bressler, N. M., Webster, D. R., et al., "Deep learning in ophthalmology: the technical and clinical considerations," *Progress in retinal and eye research* **72**, 100759 (2019).

[13] Liu, X., Jiang, J., Zhang, K., Long, E., Cui, J., Zhu, M., An, Y., Zhang, J., Liu, Z., Lin, Z., et al., "Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network," *PloS one* **12**(3), e0168606 (2017).

[14] Kim, S. J., Cho, K. J., and Oh, S., "Development of machine learning models for diagnosis of glaucoma," *PloS one* **12**(5), e0177726 (2017).

[15] Muhammad, H., Fuchs, T. J., De Cuir, N., De Moraes, C. G., Blumberg, D. M., Liebmann, J. M., Ritch, R., and Hood, D. C., "Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects," *Journal of glaucoma* **26**(12), 1086 (2017).

[16] Schlegl, T., Waldstein, S. M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B. S., Langs, G., and Schmidt-Erfurth, U., "Fully automated detection and quantification of macular fluid in oct using deep learning," *Ophthalmology* **125**(4), 549–558 (2018).

[17] Treder, M., Lauermann, J. L., and Eter, N., "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," *Graefe's Archive for Clinical and Experimental Ophthalmology* **256**, 259–265 (2018).

[18] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," (2015).

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," (2014).

[20] Tan, M. and Le, Q. V., "Efficientnet: Rethinking model scaling for convolutional neural networks," (2020).