Multi-label classification of retinal disease images using transfer learning-based techniques

Georgios Botsoglou[1], Theano Drosoglou[2], and Christos Psychalas[1]

[1]AI M.Sc., School of Informatics, Aristotle University of Thessaloniki, Greece

[2]DWS M.Sc., School of Informatics, Aristotle University of Thessaloniki, Greece

gbotso@csd.auth.gr, tdrosog@csd.auth.gr, cpsyc@csd.auth.gr

Abstract

Early detection and accurate diagnosis of retinal disorders are critical in preventing visual impairment. This study explores the application of transfer learning techniques for multi-label classification of retinal disease images. Utilizing a publicly available dataset, we implemented and fine-tuned pre-trained image classification models to compare their performance. Our approach included handling imbalance through resampling methods and weighted loss functions while leveraging attention mechanisms to interpret model decisions where applicable. Experimental results demonstrated that vision transformers outperform the state-of-the-art ResNet architecture offering improved diagnostic capabilities.

1 Introduction

Early detection and diagnosis of retinal disorders is of great importance in the prevention of permanent or partial visual impairment. Over the last two decades, several color fundus image datasets have been collected for both binary classification and multi-label tasks. In addition, many deep learning algorithms have been developed for ophthalmological disorder classification, in the past few years, utilizing mainly Convolutional Neural Network (CNN) techniques (e.g., Cen et al., 2021).

Most of the research focuses on single-disease detection and/or frequent retinal fundus pathologies, such as age-related macular degeneration, diabetic retinopathy, and glaucoma (e.g., Ting et al., 2017). Real-world applications, however, require diagnostic tools capable to support simultaneous detection of multiple retinal diseases affecting a single patient and deal with the issue of underrepresentation of a wide range of quite rare sight-threatening pathologies.

The advances in transfer learning using the attention mechanism and the recent application of such approaches on computer vision (Dosovitskiy et al., 2021) have enabled researchers to experiment with transformer-based architectures on multi-label retinal disease tasks, leading to promising results when compared to state-of-the-art CNN models (e.g., Rodriguez et al, 2023; Wang et al., 2024).

In the present work, pre-trained vision transformer-based models are optimally configured and fine-tuned on fundus images for multi-label retina disease classification and their performance is investigated in comparison to state-of-the-art CNN-based predictions and findings from other relevant studies. For this purpose, a publicly available and widely used fundus image dataset is selected (ODIR-2019, 2019). Appropriate data cleaning and preprocessing are applied to this dataset to ensure data quality and a variety of disorders to predict. In order to overcome any class imbalance issue and the underrepresentation of rare retinal disorders, various approaches were investigated, such as resampling methods and weighted loss function. We report multi-label image classification results and present the inherent attention layer-feature of a vision transformer, selected based on its performance, as a first attempt to explain the model classification decisions.

2 Related work

Deep learning analysis of fundus images can be helpful in diagnosing and controlling ocular diseases at an early stage (Wang et al., 2020a). In recent years, there has been an advanced research effort on collecting colored fundus images for binary or multi-label classification tasks on retinal fundus pathologies. In the following, we emphasize on a few publicly available datasets of retina images and we discuss some computer vision multi-label classification studies on such data collections.

2.1 Fundus image datasets

In the frame of this study, an extensive literature review was performed in order to identify open access fundus image datasets for retinal disease classification. Over the last two decades, several color fundus image datasets have been published for both binary classification and multi-label tasks. Some of these image repositories are publicly available and have assisted several retina disease classification studies in the medical computer vision domain.

For example, STructured Analysis of the REtina (STARE; Hoover et al., 2000), and ORIGA (Zhang et al., 2010) are publicly available databases for retinal vessel segmentation and glaucoma binary classification, respectively. Other publicly available fundus image sources like Drishti-GS1 (Sivaswamy et al., 2015), AREDS (Davis et al., 2005), the Kaggle-EyePacs database for diabetic retinopathy, and RIGA (Almazroa et al., 2018) are curated for two- or three-class classification problems with mutually exclusive labels for each image.

The Automated Retinal Image Analysis (ARIA) dataset was collected in the United Kingdom between 2004 and 2006 from adult males and females (Farnell et al., 2008). The images consists of three annotated groups, i.e., healthy, age-related macular degeneration (ARMD), and diabetic patients.

The Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR) database (ODIR-2019, 2019) is one of the largest publicly available fundus datasets and one of the most commonly used for ocular disease detection. It comprises of about 5000 pairs of left and right fundus images from patients, labeled for eight categories of ophthalmological diseases.

Another popular retina image resource is the Retinal Fundus Multi-disease Image Dataset (RFMiD), which consists of 3200 fundus images captured by three different cameras (Panchal et al., 2023). It includes 46 retina pathological conditions that appear in routine clinical settings, annotated by adjudicated consensus from two senior retinal experts.

Recently, Rodriguez et al. (2023) introduced Multi-label Retinal Disease (MuReD) dataset, which is a collection of a few publicly available retina image sources, i.e., the ARIA, STARE, and RFMiD datasets. It consists of 2208 images of 20 different labels, with varying image quality and resolution. MuReD was constructed to address some common problems in the existing fundus image datasets, such as the narrow range of pathologies to classify, high class imbalance, low number of samples for the underrepresented labels, lack of assurance in image quality, etc.

2.2 Retina disease classification

Although early studies focused on single-disease detection and classification (e.g., Ting et al., 2017; Hemelings et al., 2021; Sun et al., 2021), there is an increased effort from researchers to address the multi-label classification problem of ocular diseases using deep learning techniques. The majority of such works proposed CNN architectures. In addition, a large amount of such studies are conducted on the publicly available ODIR dataset.

For example, He et al., (2020) used ResNet models (He et al., 2016) and a special attention-based and feature weighting and fusion network to perform classification on ODIR images, reporting AUC values of about 92-93% and F1 scores ranging from 89.6 to 91.3%. Li et al. (2020) proposed a so-called dense correlation network comprising a ResNet backbone model along with a trainable spatial correlation module to find similarities between the paired ODIR images, and a classifier, reporting AUC scores ranging between 91 and 93%, and F1 values of about 89-91%. Gour et al. (2021) experimented on ODIR dataset classification using four different pre-trained CNN architectures and two different optimizers, achieving AUC scores ranging from about 50.5% to 84.9% and F1 score values between 46.6 and 85.6%. In Wang et al. (2020b) an ensemble of two

EfficientNet models (Tan et al., 2019) were used for classification on ODIR images preprocessed using grey and colour histogram equalization, obtaining maximum AUC and F1 scores of 74% and 89%, respectively.

In terms of private fundus image datasets, there is a variety of published works leveraging CNN techniques. For example, Cen et al. (2021) developed a multi-disease deep learning platform for the automatic detection of 39 common ocular diseases and conditions based on CNN models and a customized two-step strategy. They used a combination of private and public color fundus image datasets, reporting a frequency-weighted average F1 score of 0.923 and an AUC score of 0.998. In another study, Ju et al. (2021) proposed a hybrid multiple knowledge distillation approach to train a ResNet50 model (He et al., 2016). More specifically, they managed to extract knowledge from two teachers, each trained with different sampling strategies. They employed two large private fundus image datasets, containing 100K and 1 million images, respectively, and classified 50 ocular disease categories, reporting a mean Average Precision score of 64.14% and 64.69% for the above datasets, respectively.

Although CNN-based models have been extensively used for multi-label classification of ocular abnormalities, only a few studies have proposed image transformer architectures for such tasks. Rodriguez et al. (2023) were the first to employ a vision transformer, namely, the C-Tran model (Lanchantin et al. 2020) which is specifically designed for multi-label tasks, on a newly constructed fundus image dataset. They reported better performance by 8.1% in terms of AUC score for disease classification compared to other state-of-the-art studies on similar tasks. Very recently, Wang et al. (2024) presented a novel vision transformer architecture called retinal ViT for multi-label classification on ODIR image dataset, outperforming state-of-the-art CNN-based algorithms.

3 Datasets and pre-processing

3.1 ODIR database

ODIR (also known as ODIR-2019 and ODIR-5K) database is a benchmark structured fundus image dataset utilized by researchers for multi-label retina pathology classification tasks (ODIR-2019, 2019). It is a collection of color retina images obtained from various hospitals located in China compiled by Shanggong Medical Technology Co., Ltd. It provides desensitized fundus images from about 5000 patients, including additional information such as the patient's age, sex, and doctors' diagnostic keywords. ODIR dataset is unique in comparison to other publicly accessible fundus image datasets, since it contains color images of both left and right eyes for each patient with diagnostic keywords for abnormalities for each image, separately. The dataset contains a common target label for each pair of eye images. The annotation task of the dataset was performed by trained human readers with quality control management based on the above information.

ODIR is a multi-class multi-label dataset comprising one label indicating normal fundus condition (N) and a total of seven ocular disease categories, i.e., diabetic retinopathy (D), glaucoma (G), cataract (C), age-related macular degeneration (A), hypertension (H), pathological myopia (M), and other diseases/abnormalities (O). It should be noted that the publicly available ODIR is highly imbalanced. Another challenging part of the ODIR dataset is that the images labeled as O (other diseases/abnormalities class) contain features related to 12 different ophthalmological diseases. Thus, it is not easy to learn appropriate features in such a case. Moreover, the fundus images were captured with various cameras such as Canon, Zeiss, and Kowa and are characterized by different dimensions and quality due to the different configuration of these cameras. Although the ODIR dataset is more applicable to real-life clinical situations, the above issues negatively affect the accuracy and loss of any models trained for classification.

The dataset is split into training, validation and testing sets of about 4000, 500 and 500 pairs of fundus images, respectively. In this study, we employed the data samples of three main ocular diseases, i.e., diabetic retinopathy (D), cataract (C) and pathological myopia (M), along with the normal-labeled (N) fundus images. The left and right eye images of each patient were treated separately along with their corresponding ground truths. The annotation schema and the number of samples related to each label are presented in Table 1. The class distribution of the images is illustrated in Figure 1.

Some examples of image pairs, i.e., both the left and right fundus for each patient, can be seen in

Label	Class	Training	Validation	Testing	Total
N	Normal	1337	325	402	2064
D	Diabetic Retinopathy	1079	285	354	1718
C	Cataract	195	49	58	302
M	Pathological Myopia	163	39	53	255

Table 2: The ocular disease categories employed in this classification study from ODIR dataset. The number of samples corresponding to the training, testing and validation sets as well as the total number per category are presented.

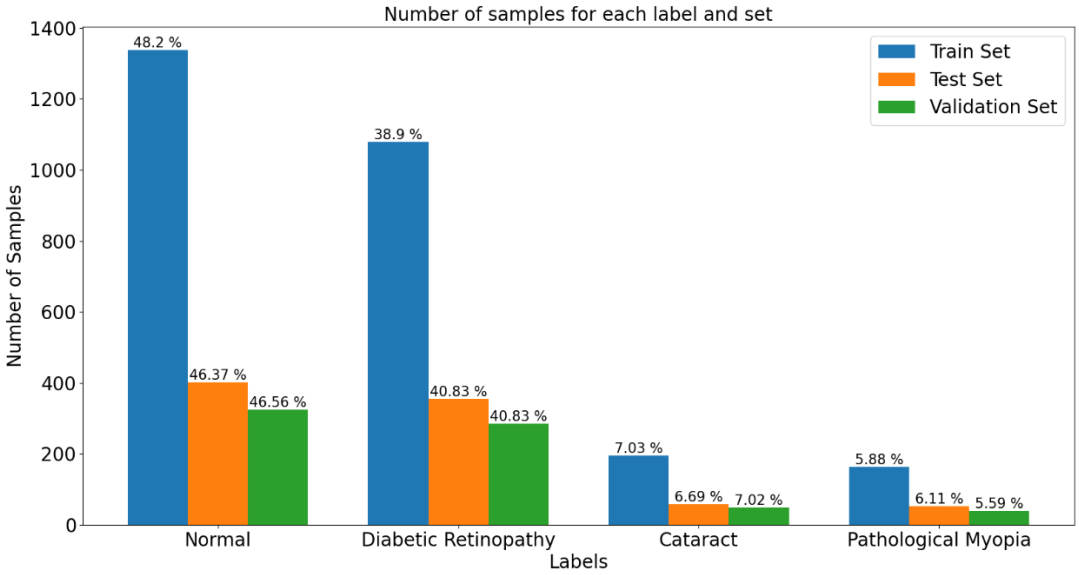

Figure 1: Distribution of ocular disease categories corresponding to the training, testing and validation sets.

Figure 2, along with the corresponding classes in the caption. In general, an image of a normal fundus is characterized by a low optic cup to optic disc ratio, i.e. less than 0.5 and normal vasculature structure with no atrophies and extra membranes.

Abnormalities such as micro-aneurysms and hemorrhages, observed via red spots, and exudates, which are observed through yellow spots, are related to diabetes retinopathy. Cataract is a retina disease commonly found among elderly patients and is characterized by blurred – or, in severe cases, by absent – basic anatomical structures such as blood vasculature, optic disc, and fovea. Finally, myopia causes vision loss due to progressive thinning of retinal pigment epithelial layer and its main observed feature is a hyperpigmentation around the optic disc, which is called peri-papillary atrophy. A brief description of the retina disease categories of the ODIR dataset and their main abnormalities as well as the related fundus image features can be found in Gour and Khanna (2021).

3.2 Pre-processing

As already mentioned, ODIR dataset contains pair images of both left and right eyes for each patient. In this work, the left and right eye images of each individual were treated as separate samples in the model architecture. Since the paired fundus images for each patient share the same combined list of ground truth labels, appropriate classes were assigned separately to each image based on the combined list and the diagnostic keywords appearing for a particular eye part in the dataset. Then, only the image samples containing at least one of the 4 chosen labels, which were described previously (section 3.1), were kept for the analysis.

Moreover, any samples containing keywords like “low image quality”, “optic disk photographically invisible”, “lens dust”, and “image offset” in the description were removed from the final dataset to reduce the false recognition rate.

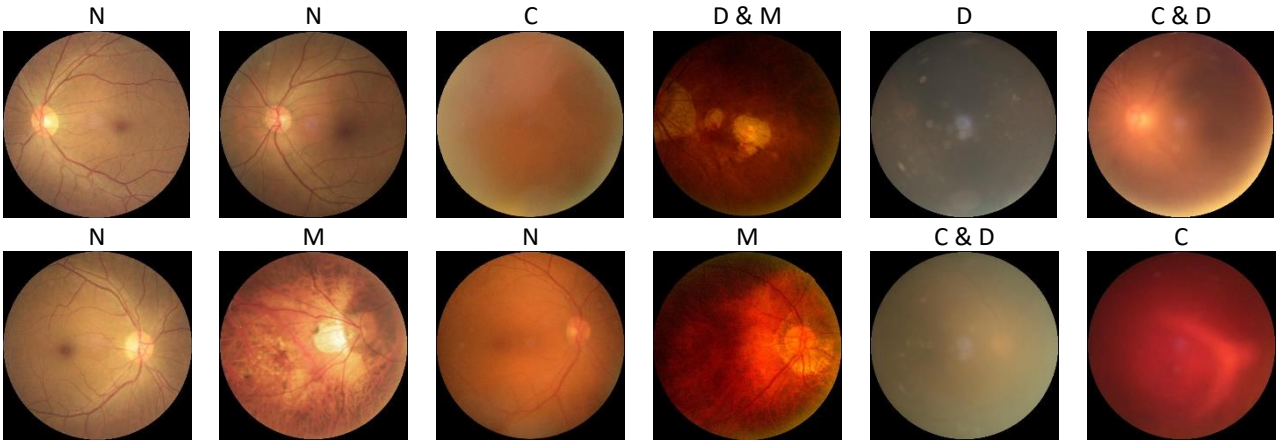

Figure 2: Fundus image samples in the ODIR dataset. Both the left (upper panels) and right (lower panels) eyes for each patient are presented. The corresponding labels can be seen above images. N, D, C and M denote normal retina, diabetic retinopathy, cataract and pathological myopia, respectively.

The original ODIR dataset consists of fundus images with different resolutions as they are captured with various cameras. For this project, the already preprocessed ODIR images were obtained. These images were cropped and resized at 512×512 pixels. In order to make these images appropriate for model training, further crop operations were utilized, making sure that only the background black pixels were cropped and the displayed retina remained untouched. To support various models, the size of the cropped images is initially set to 224×224 pixels, which is a commonly accepted size by most of the models. Then, this resolution is adjusted explicitly for each model according to its needs.

A stratification method was implemented to split data in train, validation and test sets while maintaining proportionate distributions. The class-wise distribution obtained after data pre-processing is illustrated in Figure 1. The corresponding number of images are reported in Table 1, both in terms of total and separately for the training, testing and validation subsets. It can be observed that two classes out of four, i.e., normal (N) and diabetic retinopathy (D), have a significantly higher number of samples compared to the other two ocular disease categories, i.e., cataract (C) and pathological myopia (M). In order to address this high class imbalance issue, an extensive literature review was performed and several oversampling and undersampling methods were investigated. However, in general, the commonly used naive resampling strategies do not work well in a multi-label setting. Instead, they usually lead to a new relative imbalance. Ju et al. (2024) used an experimental analysis to explain in detail why such sampling techniques fail on multi-label datasets and they proposed an alternative approach called "instance-wise class-balanced sampling" to address this issue. In their experiments, they used fundus image datasets and their approach is based on the simple assumption that the image samples belonging too many retina disease categories should be avoided as much as possible in the resampling procedure. Here, we followed the above methodology by computing the proposed sampling factors δ and applying them as weights (one for each class) to the binary cross-entropy (BCE) loss function.

4 Modeling and experimentation setup

4.1 Models

In this study, five pre-trained vision transformer models are used as backbone for ocular disease image classification. In particular, the vision models utilized for transfer learning are: ViT, BEiT, DeiT. LeViT and Swin.

The Vision Transformer (ViT) model was introduced by Dosovitskiy et al. (2021), who trained the model on large amounts of data from the ImageNet and JFT datasets and applied transfer learning on several popular small or mid-sized benchmark datasets for image classification, like ImageNet, CIFAR, and VTAB. It is the first transformer encoder that revealed very good performance in comparison to state-of-the-art convolutional architectures such as ResNet and EfficientNet.

The Bidirectional Encoder representation from Image Transformers (BEiT) model was proposed in Bao et al. (2022) and was inspired by the - very

successful on Natural Language Processing (NLP) tasks - BERT model (Devlin et al., 2018). Following BERT, BEiT is the first model that uses a self-supervised pre-training methodology, i.e., auto-encoding with masked image patches, outperforming previous supervised pre-training techniques on downstream tasks, such as image classification and semantic segmentation, using benchmark datasets like ImageNet and CIFAR-100.

Touvron et al. (2021) published the Data-efficient image Transformer (DeiT) model, a convolution-free transformer which is more efficiently trained for image classification, requiring far less data and computing resources compared to the original ViT model. In addition, DeiT introduces a teacher-student strategy, using a so-called distillation token to effectively learn from a teacher through attention mechanism.

Graham et al. (2021) proposed LeVIT, a hybrid neural network for fast inference image classification. LeViT is an improved version of the original ViT model in performance and efficiency with a few architectural modifications. In particular, LeViT utilizes activation maps with decreasing resolutions, a concept known from CNN architectures, and introduces an attention bias to integrate positional information.

The Swin (Shifted Windows) Transformer was introduced by Liu et al. (2021), who proposed a hierarchical transformer design with a shifted windowing approach in order to address the challenges in adapting transformer architecture from NLP to vision tasks.

In addition, a popular deep convolutional neural network, namely, the ResNet50 model (He et al., 2016), is adopted as a convenient baseline in order to evaluate the performance of vision transformer architectures. The ResNet model introduced residual connections, which allow to train networks with an unseen number of layers (up to 1000). In particular, the ResNet50 v1.5 model is utilized here, which was proposed by Nvidia as a modified version of the original ResNet50 v1 model (Nvidia NGC Catalog, 2023).

4.2 Experimental setup

In this subsection, we present in detail the experimental setup employed to evaluate the performance of the various models of this study. The focus is on ensuring a consistent comparison across all models by maintaining uniform conditions and procedures.

We trained each model using a standardized pipeline to ensure that the results are directly comparable. Key aspects of this setup include the training procedure, data augmentation and the metrics used for evaluation of performance.

Regarding the transformer architectures the classifier was removed from each model and was replaced by a custom 2 linear layer classifier, which was able to handle the high-level features extracted from the encoder.

For optimization, we employed the Adam optimizer and tried to minimize a weighted Multi-label Binary Cross-Entropy loss. The learning rate was constant throughout training with a value of 5×10^{-4} and batch size was set to 16.

To prevent overfitting, early stopping criteria were applied. Training was halted if the validation loss did not improve for 10 consecutive epochs, ensuring that the model did not overtrain and degrade in performance on unseen data.

Data augmentation played a crucial role in improving model generalization. More specifically we applied several augmentations including random horizontal flip, random vertical flip, random rotation, color jittering and Gaussian blurring to increase the diversity of the input data.

We evaluated the model performance using a set of standardized metrics. We used macro F1 score for the multi-label problem and F1 score per label to increase the granularity of the monitoring. We also used Label Ranking Average Precision proposed by Tsoumakas et al. (2010) to capture the ranking quality of our predictions across multiple labels for each sample.

All the experiments were conducted on a RTX6000 GPU. The environment built used Python 3.12.2 and Pytorch version 2.2.2.

5 Results and discussion

The model evaluation results are summarized in Table 2. We report the ranking average precision and F1 macro to investigate the overall performance on the multi-label classification. In addition, the F1 score for each label, separately, is included to show the performance of each model on the detection of the different ocular diseases. The prediction evaluation of the proposed models is performed on the ODIR testing subset, i.e., on fundus images which were completely unseen

Metric	Class	BEiT	DeiT	LeViT	Swin	ViT	ResNet50
F1 per class [%]	Normal	70.2	78.9	79.3	**80.8**	76.8	80.1
	Diabetic Retinopathy	**72.1**	63.5	56.5	61.1	63.4	53.7
	Cataract	**89.1**	85.1	82.9	86.3	83.3	79.1
	Pathological Myopia	96.1	**98.0**	94.3	95.9	95.9	**98.0**
F1 Macro [%]		**81.8**	81.4	78.2	81.1	79.9	77.7
Ranking Avg. Precision [%]		79.0	80.8	79.4	**81.0**	78.6	79.4

Table 2: Evaluation results for each model on the ODIR fundus imaging test set. The ranking average precision and F1 macro are reported for the overall multi-label classification. The F1 score for each label separately is also included. The best value for each metric is indicated in bold.

during model training and hyperparameter tuning. The best value for each metric is indicated in bold.

In general, all the vision transformers utilized in our experiments show comparable or better performance against the state-of-the-art convolutional model ResNet50. BEiT seems to perform better in predicting retinal diseases effectively on unseen fundus images in terms of F1 score, giving the best values for diabetic retinopathy (72.1%), cataract (89.1%) and the macro-averaged score (81.8%). In particular, BEiT seems to successfully detect the abnormalities (features) related to diabetic retinopathy, which is the most misclassified label from all other models. It may worth noticing that the BEiT F1 score for diabetic retinopathy is higher compared to the one computed for the normal fundus condition.

In terms of label ranking average precision, Swin performs better with a value of 81.0%, followed closely by DeiT model (80.8%). Swin outperforms the other models also in detecting the normal retina images, giving an F1 score of 80.8%. ResNet50 gives the highest score, along with DeiT, only in case of pathological myopia (98%), which seems to be the easiest abnormality to detect in the ODIR dataset according to our results.

Rodriguez et al. (2023) report a lower F1 macro for their proposed transformer-based C-Tran model on retina disease multi-label classification using the in-house MuReD dataset (57.3%), compared to those computed here, ranging from 77.7% to 81.8%. For normal, diabetic retinopathy, and myopia, they computed an F1 score of 82.4%, 85.9%, and 87.2%, respectively. Our best result for normal condition is comparable (80.8%), while we report lower best score for diabetic retinopathy (72.1%), and much higher for myopia (98.0%). It should be noted, though, that they utilized a lot more disease labels in their experiments, i.e., a total number of 20 different ocular abnormalities excluding cataract.

The proposed vision transformer-based approach by Wang et al. (2024) for multi-label classification on ODIR dataset resulted in an F1 score of 91.9 ± 0.02, which is much higher compared to our F1 macro scores. They also report an F1 of about 85.0% utilizing the ResNet50 model, which is higher against the 77.7% computed in the current work. However, it should be noted that the results cannot be directly compared, since Wang et al. (2024) employed all 8 classes available in the ODIR database.

To further illuminate the explanation behind the performance of our best vision transformer model, i.e. BEiT, we employ attention maps as a tool for visual explanation (Figure 3). These maps provide insights into the decision-making process of the transformer by highlighting which regions of the fundus images were deemed more important during classification. Through the depiction of the spatial distribution of attention weights across the image we can try to understand where the model focuses on different parts when identifying ocular diseases.

The attention maps were generated from the self-attention of the final layer of BEiT. We extracted the attention weights associated with each patch of the input image. These weights are then visualized as heatmaps. The produced heatmaps, low in interpretability, are transformed to 14×14 heatmaps and superimposed on the original image to provide a more intuitive result of the model's perception.

In Figure 3, selected images from Figure 2 are illustrated (upper panels) along with the corresponding BEiT attention maps (lower panels). The ground truth label is reported at the top of each original image whereas the BEiT-predicted label is

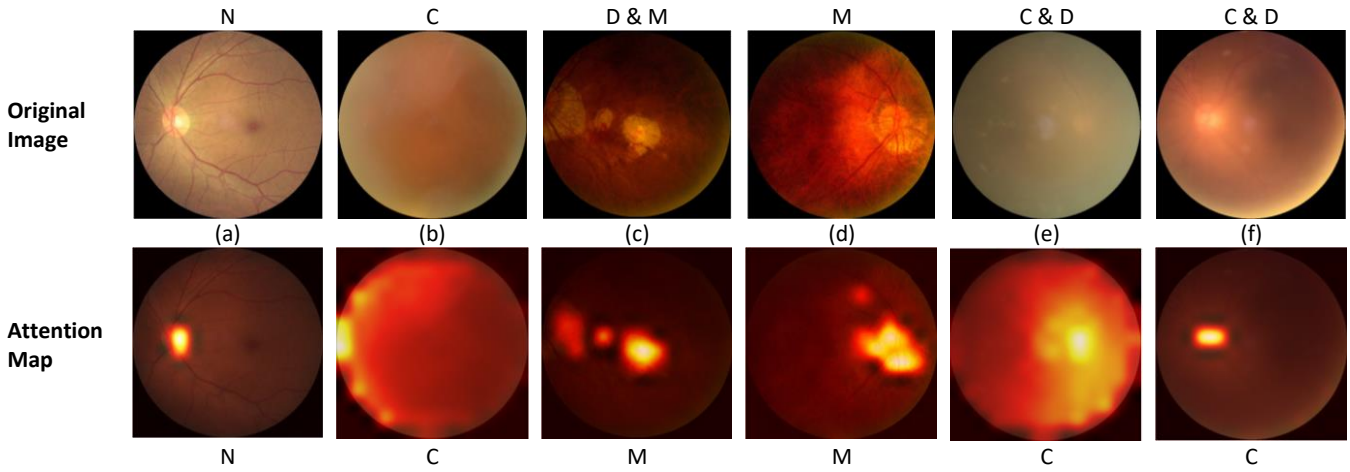

Figure 3: Selected images from Figure 2 (upper panels) and the corresponding BEiT attention maps (lower panels). The ground truth label is reported at the top of each image whereas the predicted label is at the bottom of the corresponding attention map.

noted at the bottom of the corresponding attention map. For a brief description of the features with the retinal disease categories of the dataset see subsection 3.1

In general, the BEiT model seems to successfully recognize the optic cup and disc in a normal fundus (Figure 3a). This is also the case for the rest of normal-labels fundus of Figure 2 and their attention maps (not shown here). In addition, we observe that the model is capable of detecting the general blurring in a fundus image of a patient eye diagnosed with cataract (see Figure 3b, 3e, 3f). However, it may be difficult for the model to detect abnormalities related to diabetic retinopathy, which are observed through red and yellow spots, within the general blurring of the basic anatomical structures in a cataract-diagnosed retina (Figure 3e, 3f). BEiT also seems capable to classify correctly features related to pathological myopia (Figure 3c, d), but when myopia is combined with diabetic retinopathy, it may miss the red and yellow spots features related to the latter pathology (Figure 3c). The interpretation of the selected fundus images labeled with diabetic retinopathy along with another ocular disease in Figure 3, may explain partly the tendency of all models to misclassify diabetic retinopathy and the relatively low performance in such cases (see Table 2).

6 Conclusions

Early detection and accurate diagnosis of ocular diseases is crucial for the prevention of permanent or partial vision impairment. Several studies, utilizing deep learning techniques for both binary and multi-label classification of retinal abnormalities, have been published, and various public color fundus image datasets have been collected.

Most of the research focuses on single-disease detection and/or frequent ocular pathologies. However, real-world applications require diagnostic tools capable to support simultaneous detection of multiple ophthalmological diseases affecting a single patient and deal with the issue of underrepresentation of a wide range of quite rare sight-threatening pathologies.

In the current work, we utilized transfer learning from pre-trained vision transformer models on fundus images for multi-label ocular disease classification. The models were optimally configured and their performance is investigated in comparison to the CNN-based ResNet50 predictions. In general, all the vision transformers employed in our experiments outperform the state-of-the-art convolutional model. BEiT performed better in classifying unseen fundus images effectively in terms of F1 score, giving the best scores for diabetic retinopathy (72.1%) and cataract (89.1%), as well as the best macro-averaged score (81.8%) compared to the other models. It worth noticing that BEiT performs very well in detecting the abnormalities (features) related to diabetic retinopathy, which is the most misclassified label from all other models.

In a future work, the multi-label classification utilizing all 8 classes of ODIR dataset and/or images from other databases could be investigated.

For this attempt, advanced image pre-processing and/or image reconstruction techniques could be employed in order to improve image quality. In addition, the trained models could be externally validated against other unseen datasets of retina images.

Code Availability

The code used in this study is available at https://github.com/Xritsos/RetViT

Acknowledgments

The authors would like to acknowledge the National Institute of Health Data Science at Peking University (NIHDS-PKU), the Institute of Artificial Intelligence at Peking University (IAI-PKU), the Shanggong Medical Technology Co. Ltd. (SG) and the Advanced Institute of Information Technology at Peking University (AIIT-PKU) for co-organizing the International Competition on Ocular Disease Intelligent Recognition (ODIR) and providing the respective fundus image dataset (https://odir2019.grand-challenge.org/). Results presented in this work have been produced using the Aristotle University of Thessaloniki (AUTh) High Performance Computing Infrastructure and Resources.

References

Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V., Retinal fundus images for glaucoma analysis: the riga dataset, in: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, Vol. 10579, International Society for Optics and Photonics, 2018, p. 105790B.

Bao, H., Dong, L., Piao, S., and Wei, F.: BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 (2022).

Cen, L.-P., Ji, J., Lin, J.-W., Ju, S.-T., Lin, H.-J., Li, T.-P., Wang, Y., Yang, J.-F., Liu, Y.-F., Tan, S., Tan, L., Li, D., Wang, Y., Zheng, D., Xiong, Y., Wu, H., Jiang, J., Wu, Z., Huang, D., Shi, T., Chen, B., Yang, J., Zhang, X., Luo, L., Huang, C., Zhang, G., Huang, Y., et al., Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature Communications, 12(1):4828, Aug 2021.

Davis, M.D., Gangnon, R.E., Lee, L.Y., Hubbard, L.D., Klein, B., Klein, R., Ferris, F.L., Bressler, S.B., Milton, R.C., The age-related eye disease study severity scale for age-related macular degeneration: Areds report no. 17, Arch. Ophthalmol. (Chic. Ill.: 1960) 123 (11) (2005) 1484–1498.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

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. 2021. arXiv:2010.11929.

Farnell, D.J.J., Hatfield, F.N., Knox, P., Reakes, M., Spencer, S., Parry, D., and Harding, S.P., Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. Journal of the Franklin institute, 345(7):748–765, 2008.

Gour, N., and Khanna, P., Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network, Biomed. Signal Process. Control 66 (2021) 102329, http://dx.doi.org/10.1016/j.bspc.2020.102329.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., and Douze, M. 2021. LeViT: A vision transformer in ConvNet's clothing for faster inference. arXiv preprint arXiv:2104.01136 (2021).

He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

He, J., Li, C., Ye, J., Wang, S., Qiao, Y., and Gu, L., Classification of ocular diseases employing attention-based unilateral and bilateral feature weighting and fusion. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pages 1258–1261, 2020.

Hemelings, R., Elen, B., Barbosa-Breda, J. et al. Deep learning on fundus images detects glaucoma beyond the optic disc. Sci Rep 11, 20313 (2021). https://doi.org/10.1038/s41598-021-99605-1

Hoover, A., Kouznetsova, V., Goldbaum, M., Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210.

Ju, L., Wang, X., Yu, Z., Wang, L., Zhao, X., and Ge, Z., Long-tailed multi-label retinal diseases recognition using hierarchical information and hybrid knowledge distillation, 2021.

Ju, L., Yu, Z., Wang, L., Zhao, X., Wang, X., Bonnington, P., Ge, Z., Hierarchical Knowledge

Guided Learning for Real-World Retinal Disease Recognition, in IEEE Transactions on Medical Imaging, vol. 43, no. 1, pp. 335-350, Jan. 2024, doi: 10.1109/TMI.2023.3302473.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y., General multi-label image classification with transformers. arXiv:2011.14027, 2020.

Li, C., Ye, J., He, J., Wang, S., Qiao, Y., and Gu, L., Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pages 1–4, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

Nvidia NGC Catalog: ResNet v1.5 for PyTorch, Latest Version: 21.03.9, Modified: April 4, 2023, Nvidia, https://catalog.ngc.nvidia.com/orgs/nvidia/resources/resnet_50_v1_5_for_pytorch (last access: 7 June 2024), 2023.

ODIR-2019: Peking University International Competition on Ocular Disease Intelligent Recognition, 2019, https://odir2019.grand-challenge.org/ (last access: March 16, 2024).

Panchal, S., Naik, A., Kokare, M., Pachade, S., Naigaonkar, R., Phadnis, P., et al. (2023). Retinal fundusmulti-disease image dataset (RFMID) 2.0: a dataset of frequently and rarely identified diseases. Data 8. doi: 10.3390/data8020029.

Rodriguez, M.A., AlMarzouqi, H., Liatsis, P. Multi-Label Retinal Disease Classification Using Transformers. IEEE J. Biomed. Health Inform. 2023, 27, 2739–2750.

Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S. et al., A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis, JSM Biomed. Imaging Data Pap. 2 (1) (2015) 1004.

Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y., Lesion-aware Transformers for diabetic retinopathy grading, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10938–10947.

Tan, M., and Le, Q., EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 6105–6114. PMLR, 09–15 Jun 2019.

Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318, 2211–2223 (2017).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. arXiv:2012.12877.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. In Data mining and knowledge discovery handbook (pp. 667-685). Springer US.

Wang, Z., Keane, P.A., Chiang, M., Cheung, C.Y., Wong, T.Y., Ting, D.S.W., Artificial intelligence and deep learning in ophthalmology, Artif. Intell. Med. (2020a) 1–34.

Wang, J., Yang, L., Huo, Z., He, W., and Luo, J., Multi-label classification of fundus images with efficientnet. IEEE Access, 8:212499–212508, 2020b.

Wang D, Lian J and Jiao W (2024) Multi-label classification of retinal disease via a novel vision transformer model. Front. Neurosci. 17:1290803. doi: 10.3389/fnins.2023.1290803.

Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y., Origa-light: An online retinal fundus image database for glaucoma analysis and research, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE, 2010, pp. 3065–3068.