OPHTHALMIC BIOMARKER DETECTION USING INCEPTION NET

Aaseesh Rallapalli [1] *Lokesh Badisa* [2] *Nithish S* [1] *Utkarsh Doshi* [1] *Soumya Jana*[1]

[1]Department of Electrical Engineering, Indian Institute of Technology Hyderabad
[2]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad

ABSTRACT

Optical Coherence Tomography (OCT) imaging has emerged as an indispensable tool in the screening and management of retinal diseases, including age-related macular degeneration (AMD) and diabetic retinopathy (DR). OCT facilitates cross-sectional visualization of structural changes within the posterior segment of the eye, encompassing layers such as the retina and choroid. OCT images have proven instrumental in the detection of various biomarkers associated with these diseases. As part of this challenge, our objective was to identify the presence or absence of six distinct biomarkers within a given OCT B-Scan (Multi-Label Classification). In pursuit of this goal, we explored different models based on convolution neural networks and transformers. A comparative analysis of the performance metrics across these models was conducted. Our findings elucidate that the optimal performance was achieved by leveraging the InceptionNet V3 architecture [1] as the backbone. This work was done as part of the IEEE SPS VIP CUP 2023.

Index Terms— OCT, InceptionNet, ResNet,

1. INTRODUCTION

Ophthalmic clinical trials, designed to assess the effectiveness of treatments, are meticulously conducted with predefined objectives and a structured set of procedures established prior to trial initiation. This meticulous approach carried out during [2] ensures controlled data collection, tracking gradual changes in the condition of afflicted eyes. The dataset comprises both 1D clinical measurements and volumetric 3D optical coherence tomography (OCT) images. These 3D OCT images are analyzed by medical professionals to discern structural biomarkers for each patient. In conjunction with clinical measurements, these findings inform personalized treatment decisions for individual patients.

The OCT Dataset provided by the organizers, [2], stands out due to its comprehensive inclusion of multiple biomarkers as output labels, facilitating further medical diagnostic procedures. This distinguishes it from other OCT datasets, such as the Kermany Dataset [3]. The latter aims as classifying the image as one of the following biomarkers - Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen, and Normal.

In recent times, the remarkable advancements in deep learning have exerted a substantial influence on the domain of medical science. While the standard practice primarily revolves around assessing the generalizability of these models, the focus of this competition is to focus both on generalization and personalization capabilities. Generalization aims to develop algorithms that can show good performance across diverse set of patients and scenarios, providing standard solutions that can be applied widely whereas personalization, in contrast, tries to tailor algorithms to individual patients based on their unique characteristics, diagnosis and treatment planning.

The intricacies of personalization, as opposed to generalization, pose a distinctive set of challenges in the context of healthcare algorithms. Within optical coherence tomography (OCT) scans, variations between patients during different visits can be minimal, while the presentation of the same disease can vary significantly among individuals.

The evaluation metric employed in this challenge is the macro-average F1 Score, encompassing all biomarkers, and applied to both generalization and personalization tasks. A more detailed explanation of this metric can be found in Section 7.

In Section 2, a comprehensive discussion on the dataset is provided. 3 delves into a review of pertinent literature, showcasing key works that have previously explored this domain. In Section 4, we expand on different approaches we used. Section 5 offers an in-depth examination of the InceptionNet v3 architecture, dissecting its components and functionalities. Our training methodologies, including the techniques and parameters adopted, are outlined in Section 6. Subsequently, in Section 7, we delineate the outcomes of our experiments, placing particular emphasis on the most optimal results achieved. The report culminates in Section 8, where we point out our observations and potential avenues for future research.

2. DATASET

The OLIVES OCT dataset [2] provided by the organizers is derived from two distinct clinical trials, namely PRIME

We thank the Department of AI, IITH for computational support.

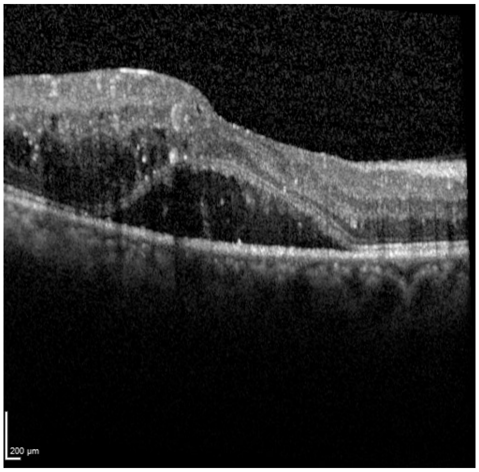

Fig. 1. a diseased oct image.

and TREX-DME. Common Biomarkers and Clinical Labels formed the basis for training and testing.

2.1. Structural Biomarkers

In this competition, our primary goal is the detection of 'biomarkers,' objective indicators of a patient's medical condition distinct from subjective symptoms. These biomarkers may serve as endpoints for diagnoisis or may be used for furthur medical diagnosis. The evaluation was based on six biomarkers which are as follows : Intraretinal Hyperreflective Foci (IRHRF), Partially Attached Vitreous Face (PAVF), Fully Attached Vitreous Face (FAVF), Intraretinal Fluid (IRF) and Diffuse Retinal Thickening or Diabetic Macular Edema (DRT/ME) Vitreous Debris (VD). In addition, 4 clinical labels namely, CST (Central subfield thickness),Best-corrected visual acuity (BCVA) , Eye ID and Patient ID were provided were provided

2.2. Train and Test

Each patient visit yields a collection of 49 OCT B-scans. In total, we have access to 78,819 B-scans; however, only 9,408 of these images are accompanied by biomarker outputs, which we employed for training InceptionNet V3. The test dataset, sourced from a distinct trial, comprises 3,871 images with corresponding biomarker output. All images in both the training and testing datasets were accompanied by Clinical Labels

3. RELATED WORK

We chose to use InceptionNet V3 based on the work of [3]. They effectively used it to classify eye images into four types. Another growing trend in this field is the use of self-supervised learning, wherein other modes of data are used as 'pseudo-labels' for training.

In the study by [4], the authors employed contrastive learning. They minimized the loss function for samples with the same pseudo-label. These labels were derived from clinical measures such as BCVA, CST, and the eye ID.

Another study by [5] took a different approach to pseudo-labeling. They framed it as an anomaly detection problem, using a severity score as the pseudo-label. They trained an autoencoder on the Kermany dataset [3] using the GradCON method [6]. They defined the severity score as $-L_{\text{recon}} + \alpha L_{\text{grad}}$ Where L_{recon} represents the autoencoder's reconstruction loss, and L_{grad} denotes a gradient-based loss function. With these scores, they segmented the OLIVES dataset [2] into distinct bins. These bins were used as pseudo-labels for training. Lastly, they trained a linear layer atop the pre-trained model while preserving the initial weights.

4. OUR APPROACH

We experimented pre-training various model backbones using the Kermany dataset[3] followed by training a linear layer while keeping the backbone's weights fixed. However, these models didn't produce the expected improvements. One potential reason might be the inherent noise and irregular padding present in the Kermany dataset[3]. This is in contrast to the OLIVES dataset[2], which doesn't exhibit such noise.

As part of our extensive experimentation, we explored a diverse array of model architectures, encompassing well-established ones such as ResNet, EfficientNet, ResNext, as well as venturing into the realm of transformer-based models, including ViT, DeiT, DiNo, among others. In our pursuit of improved performance, we even devised custom attention-augmented architectures built upon the foundation of existing CNN-based models, which yielded modest enhancements over the baseline. Furthermore, we diligently embraced the approaches recommended by the competition organizers. This included the adoption of the multi-modal guided loss approach and the integration of semi-supervised Contrastive Learning into our framework. Remarkably, our experimentation revealed that the architecture yielding the most impressive results was InceptionNet V3 [1]. This model distinguished itself through its unique architecture, capable of capturing features at multiple scales by employing kernels of varying sizes and executing parallel convolutions.

In our preliminary experimentation, we modified the first convolutional layer to accommodate a single-channel input. Subsequently, to leverage the temporal coherence between sequential B-Scans, we assembled three consectuive B scans, creating a tri-channel image for model input. When we integrated this approach with InceptionNet V3 and ResNet18 [7], and initialized the architectures with ImageNet Pretrained Weights, there was a discernible enhancement in the F1 Score, with an increment of around 2%. Pursuing this line of inquiry further, we duplicated the same image thrice, resulting in an-

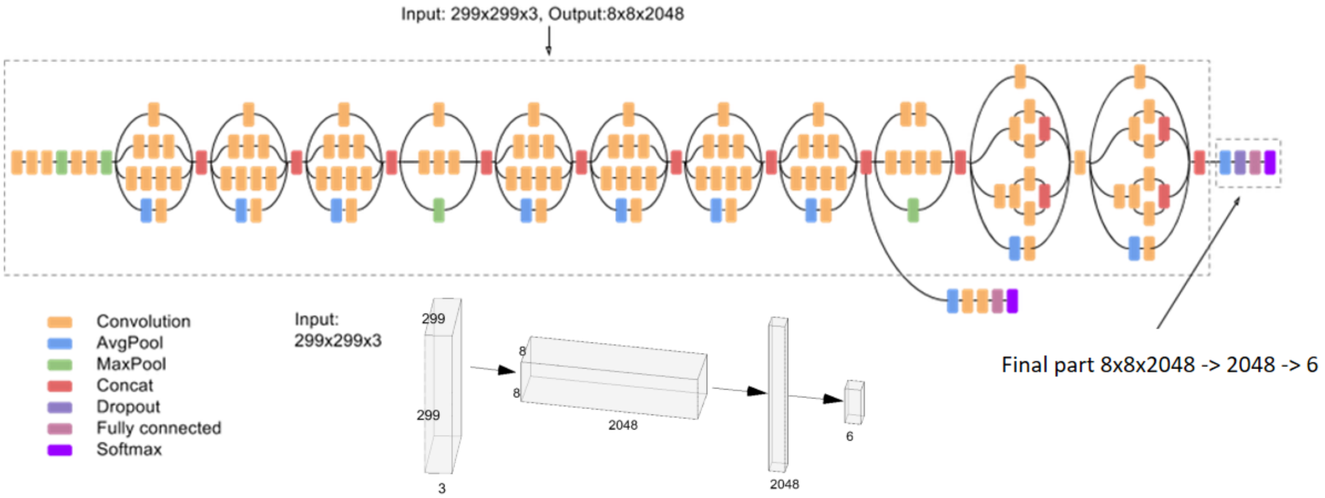

Fig. 2. Network architecture. [1]

other tri-channel input image. The outcome exhibited an even superior performance, yielding an increase of 3.5% ResNet18 Baseline. This suggests that the tri-channel input principally facilitates the optimization of pretrained weights, rather than capitalizing on temporal coherence.

5. ARCHITECTURE

InceptionNet V3 [1], represents an advanced and optimized iteration of the Inception Net model. This architecture comprises multiple inception modules, with the fundamental module consisting of four parallel layers. Among these layers, three are convolutional layers employing kernel sizes of 1, 3, and 5, while the fourth involves a 3x3 max pooling operation. The noteworthy enhancements incorporated in InceptionNet V3 are as follows:

a) Factorization of Convolution: In this approach, larger convolution kernels, such as a 5x5, are replaced with two consecutive 3x3 convolutional layers.

b) Spatially Separable Convolution: Instead of using standard nxn convolutions, spatially separable convolutions are employed, comprising nx1 and 1xn convolution operations. These help in decreasing the parameter count.

c) Efficient Grid Size Reduction: A novel strategy is adopted for grid size reduction. Rather than employing pooling layers directly, the model utilizes two parallel blocks involving convolutional and pooling layers. Subsequently, the output features from these blocks are concatenated.

In its entirety, the final model encompasses 42 layers, a bit more than it's predecessor. It has around 25 million parameters. The output of the backbone architecture is a 2048-dimensional vector, which is then connected to a Fully Connected Neural Network with 6 nodes corresponding to the number of biomarkers.

6. TRAINING

All models were trained on a single Nvidia Tesla V100-SXM3 GPU, equipped with 32 GB of RAM. The training process spanned 75 epochs, utilizing the Stochastic Gradient Descent (SGD) optimization algorithm with a learning rate of 1e-3, a momentum value of 0.9, and a weight decay parameter set to 1e-4. During training, a batch size of 64 was employed. The entire codebase was developed in PyTorch, using the starter code provided by the organizers.To quantify the model's performance, the Binary Cross Entropy Loss metric was adopted, applied after the application of the sigmoid function to the model outputs. Notably, the Training Loss for the best performing InceptionNet model achieved during the training phase was recorded at 0.06.

7. EVALUATION AND RESULT

To assess the performance of our biomarker detection models, the evaluation criterion employed is the macro-averaged F1-score. This metric is computed as follows:

$$F1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In terms of True Positives (TP), False Positives (FP), and False Negatives (FN), it can be expressed as:

$$F1 \text{ score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

The macro-averaged F1 score, often referred to as the macro F1 score, is subsequently determined through the calculation of the arithmetic mean, or unweighted mean, of all the individual class-specific F1 scores.

Table 1. Comparison across various models

Model	Phase 1 score	Phase 2 score
InceptionNet v3 (single channel)	0.649	-
InceptionNet v3 (Tri-channel)	**0.672**	**0.7682**
ResNet18 (single channel)	0.63	-
ResNet18 (Tri-channel)	0.652	0.7616
InceptionNet v4 (Tri-Channel)	0.659	0.7617

In phase 1, the F1 score was calculated across the entire dataset and averaged across the 6 biomarkers whereas in phase 2, the f1 score was averaged with respect to each set of slices associated with an individual patient. This approach tests the personalisation of the model.

Results from the test set revealed a good performance from the InceptionNet V3 model. During the phase 1 evaluation, the model achieved a Macro Average F1 Score of 0.672. Additionally, across all patients during the phase 2 evaluation, the model registered an Average F1 Macro Score of 0.7682. Other architectures like ResNet [7] and InceptionNet V4 [8] also demonstrated significant efficacy. The comprehensive results, encompassing the top-performing models (inclusive of single-channel inputs), are detailed in Table 1

8. REMARKS AND FUTURE WORK

In our exploratory analysis, we observed that adopting a tri-channel input, as opposed to a single-channel configuration, yielded superior results, primarily because this facilitated the effective utilization of pretrained weights. Notably, the InceptionNet v3 architecture demonstrated a substantially enhanced performance relative to the ResNet18 models. A significant portion of our time was directed towards the integration of the Supervised Contrastive Learning Method [4] as all of the 78,000 images could be utilized. However, for reasons yet to be ascertained, its empirical performance did not parallel our anticipations, falling short of the supervised learning algorithm. Our forays into Transformer-Based Techniques, albeit rigorous, did not culminate in noteworthy results. Even the incorporation of pretrained models failed to rectify this. A plausible hypothesis for this outcome could be the relatively diminutive size of our dataset, which might not be adequate for the complexities inherent to transformer architectures. Looking forward, a promising avenue for elevating the efficacy of Self-Supervised Learning could be the adoption of the Dino methodology [9]. This approach, grounded in recent advances, may provide the breakthrough needed for the next phase of our research.

9. REFERENCES

[1] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[2] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Corona, Ghassan Alregib, and Charles Wykoff, "Olives dataset: Ophthalmic labels for investigating visual eye semantics," 09 2022.

[3] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[4] Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib, "Clinical contrastive learning for biomarker detection," in *Thirty-sixth Conference on Neural Information Processing Systems*, 2022.

[5] Kiran Kokilepersaud, Mohit Prabhushankar, Ghassan AlRegib, Stephanie Trejo Corona, and Charles Wykoff, "Gradient-based severity labeling for biomarker classification in oct," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3416–3420.

[6] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib, "Backpropagated gradient representations for anomaly detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *AAAI Conference on Artificial Intelligence*, vol. 31, 02 2016.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," 2021.