ECE8803 Final: Multimodal Biomarker Detection Based on OLIVES Dataset

1st Zihan Diao
Graduate student
College of Engineering
Georgia Institute of Technology
zdiao7@gatech.com

1st Junyi Zhang
Graduate student
College of Engineering
Georgia Institute of Technology
jzhang3585@gatech.edu

1st Mingyu Zhang
Graduate student
College of Engineering
Georgia Institute of Technology
mzhang745@gatech.edu

***Abstract*—The detection of biomarkers is of great significance for the diagnosis and management of retinal diseases. This study uses the OLIVES dataset, which integrates OCT images and clinical information, to develop an Attention U-Net-based multimodal deep learning model for four key biomarkers: Diabetic Macular Edema (DME), Intraretinal Hyperreflective Foci (IRHRF), Intraretinal Fluid (IRF), and Fully Attached Vitreous Face (FAVF). Furthermore, we analyze in detail the challenges of multimodal data fusion and local biomarker feature extraction and validate the effectiveness of the model in biomarker detection tasks through experiments. The implementation code is available at https://github.com/Allakikun/FMLFinalproject.**

***Index Terms*—OLIVES dataset, Multimodal Learning, Biomarker Detection, Attention U-Net, Retinal Disease**

I. Introduction

Retinal diseases, such as diabetic macular edema (DME) and Intraretinal Fluid (IRF), are among the leading causes of vision loss world [1]. Optical coherence tomography (OCT), a non-invasive imaging technology, has become a critical tool for identifying biomarkers associated with retinal diseases [2]. These biomarkers include Intraretinal Hyperreflective Foci (B1), Intraretinal Fluid (B3), Diabetic Macular Edema (B4), and Fully Attached Vitreous Face (B5), which are essential for clinical diagnosis.

The OLIVES dataset is a specialized resource designed for retinal disease research [3]. It integrates high-resolution OCT images with detailed clinical labels, characterized by multimodality, diverse annotations, and large-scale data. This study focuses on four biomarkers (B1, B3, B4, B5) for in-depth analysis.

II. Research objectives

This study aims to develop a multimodal deep learning model using the OLIVES dataset to accurately detect the aforementioned four key biomarkers. By integrating OCT images and clinical information, the study explores multimodal data fusion methods to improve biomarker detection accuracy.

III. Model Selection

A. Baseline Model: ResNet

1) Model Design:

In this study, ResNet18 was chosen as the backbone network for image feature extraction due to its lightweight and efficient design. To reduce training time, the model was pretrained on the ImageNet dataset. Since the images in the *biomarker_detection* dataset are single-channel medical images, the following modifications were made to adapt ResNet18 to the dataset:

- **First Convolutional Layer Adjustment:** The input channels of the first convolutional layer were changed from 3 to 1, while the output channels were kept at 64:

$$\text{conv1: input channels} = 1, \text{ output channels} = 64$$

- **Replacement of Fully Connected Layer:** The fully connected classification head of ResNet was replaced with an identity layer to output 512-dimensional features for subsequent processing.

For clinical features, a fully connected network was designed with the following architecture:

$$\text{Clinical Features} \rightarrow 16 \text{ neurons} \rightarrow \\ \text{ReLU activation} \rightarrow 16 \text{ neurons}$$

The final output is a 16-dimensional vector to capture high-level representations of clinical data.

The extracted image features and clinical features were concatenated using `torch.cat`, resulting in a combined representation with a dimension of:

$$512 + 16 = 528$$

This fused representation was passed through a two-layer fully connected network for prediction:

$$\text{Fused Features} \rightarrow 128 \text{ neurons} \rightarrow \\ \text{ReLU activation} \rightarrow 4 \text{ neurons}$$

The model ultimately outputs predictions for the four biomarkers.

2) Loss Function and Optimizer:

- **BCEWithLogitsLoss:** Used as the loss function, suitable for multi-label classification tasks where each target (biomarker) is an independent binary classification problem.
- **Adam:** Chosen as the optimizer for its fast convergence and suitability for sparse gradients and complex nonlinear networks.

3) Experimental Results:

- The **Macro F1 Score** was 0.6021, which was lower than the **Micro F1 Score** (0.6491), indicating that the model's performance varied across different biomarkers.
- The F1 score for B5 was the 0.5732, likely due to insufficient data (only 2% of the data in B5 were labeled as 1).
- The results showed that the model performed significantly better for B3 but exhibited notable declines for B1, B4, and B5.

As referenced in [4], B1, B4, and B5 are local features, while B3 is a global feature. Therefore, incorporating attention mechanisms to optimize feature fusion may improve the model's classification capabilities for local features.

B. Multi-Modal Learning Model (MML)

To address the limitations of the ResNet-based model, an improved Multi-Modal Learning Model (MML) was developed using EfficientNet-B0. The improved model was optimized in terms of image feature extraction, global and local feature modeling, multimodal fusion, and loss function design.

ResNet18 was replaced with EfficientNet-B0. Using a compound scaling strategy, EfficientNet extracts higher-quality image features at the same computational cost. Adjustments were made to accommodate single-channel medical images by modifying the conv stem to have 1 input channel and 32 output channels. Furthermore, the classification head was replaced with `nn.Identity`, allowing the model to directly output 1280-dimensional features for subsequent processing.

1) Local and Global Feature Modeling:

The process of local and global feature modeling aims to capture both detailed local features and global contextual information from images. The following designs were implemented:

- **Local Feature Enhancement:** A multi-scale convolution module (3×3 and 5×5 convolutions) and a Squeeze-and-Excitation module were added to further extract edge details and local features from the images [5].
- **Global Feature Modeling:** A fully connected network was used to reduce the dimensionality and model the overall image features, mapping the 1280-dimensional features to 64 dimensions.

By separating the modeling of local and global features, the model can simultaneously focus on detailed patterns and global dependencies in the images.

2) Cross-Modal Attention and Fusion:

To effectively integrate image and clinical features, the following modules were designed:

- **Dynamic Feature Fusion:** Learnable weight parameters (`fusion_weight`) were used to weight local and global features, allowing the model to adaptively adjust the contributions of these two types of features.
- **Cross-Modal Attention:** A Multi-Head Attention mechanism was applied to enable interaction between image and clinical features, capturing correlations between the two modalities [6].
- **Bi-Directional LSTM (Bi-LSTM):** The fused feature sequences were further modeled using Bi-LSTM to extract temporal and contextual features [7].

3) Adaptive Multi-Task Loss:

To address the issue of class imbalance, an adaptive multi-task loss function (`MultiTaskLoss`) was introduced to dynamically assign loss weights to each biomarker:

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} \left(\exp(-\sigma_i) \mathcal{L}_i + \sigma_i \right)$$

where σ_i is a learnable parameter for the i-th task, representing the importance of that task.

4) Experimental Results:

The improved MML model demonstrated significant improvements in overall prediction performance, as follows:

- The Micro-Average F1 Score increased to 0.6658.
- The Macro-Average F1 Score increased to 0.6037.

However, the detection performance for B1, B4, and B5, despite improvements, remained suboptimal. These biomarkers are characterized by local features, and the model still struggles to capture local details effectively. This suggests that further optimization of local feature modeling is required.

C. Attention U-Net

Although MML achieved certain breakthroughs in multimodal feature fusion and regional feature modeling, it still exhibited deficiencies in detecting complex boundary targets such as IRF. In the field of medical image segmentation, U-Net is a classic convolutional neural network architecture widely used for biomedical image segmentation tasks [8]. However, traditional U-Net may face limitations in handling complex boundaries and small target detection.

In this study, we ultimately chose Attention U-Net over traditional U-Net, primarily based on the characteristic requirements of biomarkers [9]. IRHRF is sparse and easily obscured by noise, and the attention mechanism in Attention U-Net dynamically weights and enhances the focus on small objects, improving detection accuracy. The attention mechanism can be mathematically expressed as:

$$\mathbf{Z} = \text{softmax}(\mathbf{Q}\mathbf{K}^{\top})\mathbf{V}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively. The resulting matrix $\mathbf{Z}$ adjusts feature importance dynamically.

FAVF, as a regional structure, relies on prioritizing modeling of the retinal surface area, and the attention module effectively filters out irrelevant regional features, enhancing the ability to recognize surface areas [10]. For morphologically irregular and boundary-ambiguous IRF, Attention U-Net outperforms traditional U-Net in multi-scale feature capturing and boundary detection. Multi-scale features are extracted using convolutional filters of varying sizes, such as:

$$f_{ij}^{(k)} = \sum_{m=-1}^{1} \sum_{n=-1}^{1} w_{mn}^{(k)} x_{i+m,j+n}$$

where $f_{ij}^{(k)}$ represents the feature value at position (i, j) in the k-th feature map, and $w_{mn}^{(k)}$ are the convolutional weights applied to the input x.

DME detection heavily depends on features in the macular central region, and the attention module enhances detection performance through regional prioritization while better integrating clinical information [11]. This prioritization can be modeled as:

$$\mathbf{A}_r = \sigma(\mathbf{W}_r \mathbf{F})$$

where $\mathbf{F}$ represents the feature embeddings, $\mathbf{W}_r$ are learnable weights, and σ is the activation function. Thus, based on the characteristic requirements of the above biomarkers, Attention U-Net demonstrates higher accuracy and robustness in detecting retinal disease-related biomarkers compared to traditional U-Net.

In this model, OCT images are processed through Attention U-Net for feature extraction. The model's encoder uses five convolutional layers to extract multi-scale features, and the decoder dynamically adjusts attention weights on lesion regions through attention modules, using layer-by-layer upsampling and skip connections to generate high-dimensional image feature embeddings. For clinical data, a fully connected neural network is used for nonlinear embedding, mapping it into the same high-dimensional feature space as the image features. After obtaining the feature representations of the image and clinical data, we perform weighted fusion of the two modalities:

$$\mathbf{F}_{\text{fusion}} = \alpha \mathbf{F}_{\text{image}} + (1 - \alpha) \mathbf{F}_{\text{clinical}}$$

where $\mathbf{F}_{\text{image}}$ and $\mathbf{F}_{\text{clinical}}$ represent the feature embeddings of image and clinical data, respectively, and α is a learnable parameter to balance the contributions of the two modalities. The fusion strategy adopts a weighted average approach, enabling the model to dynamically adjust attention to different modality information and fully utilize the complementarity of the two data types. Finally, binary cross-entropy loss is used to record the model's loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

where y_i is the ground truth, $\hat{y}_i$ is the predicted probability, and N is the total number of samples.

In experiments, we compared the detection performance of standard U-Net and Attention U-Net on four biomarkers. The experimental results show that for sparse targets such as HRF and IRF, Attention U-Net has significant advantages, whereas in standard U-Net, important features may be obscured by larger background regions, reducing detection performance. In the detection of complex biomarkers (e.g., DME), Attention U-Net better integrates image and clinical features, improving prediction accuracy.

IV. Experiments and Results

A. Dataset and Experimental Setup

We used the *biomarker_detection* dataset from the OLIVES dataset. Out of the initial 78,822 records, samples with missing features were removed, resulting in 17,444 records retained for model training and 3,871 records used for testing. To ensure a fair comparison, all three models were trained under identical experimental conditions, with each model trained for 30 epochs.

B. Evaluation Metrics

To comprehensively evaluate the model's performance, we primarily used the F1 Score [12], Accuracy [13] and ROC-AUC [14] as evaluation metrics. The F1 Score, which combines Precision and Recall, is particularly suitable for multi-label classification tasks with imbalanced classes, as it objectively reflects the model's performance on each biomarker detection task. The F1 Score is calculated as:

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

where Precision is defined as $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, and Recall is defined as $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$. Here, TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively.

Accuracy, on the other hand, represents the proportion of correctly predicted samples out of the total samples, reflecting the overall performance of the model. Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TN represents the number of true negatives. These metrics provide a comprehensive evaluation of the model's capability to detect biomarkers effectively.

C. Experimental Results and Analysis

In the experiments, we first evaluated the performance of each model on the test set. The results showed that the Attention U-Net model achieved the best performance in detecting all four biomarkers.

TABLE I
Performance Comparison

Metric	ResNet18	MML	U-Net	Attention U-Net
Micro	0.6491	0.6735	0.6812	0.7143
Macro	0.6021	0.612	0.6172	0.6514
B1 (IRHRF)	0.6019	0.652	0.6196	0.6579
B3 (FAVF)	0.7702	0.7873	0.8064	0.8105
B4 (IRF)	0.4632	0.4954	0.5254	0.5657
B5 (DME)	0.5732	0.5133	0.5173	0.5714

1) Analysis of F1 Score:

In the detection of IRHRF, the F1 Score of the Attention U-Net model significantly outperformed other models, indicating that the attention mechanism effectively adjusts its focus on critical regions. This highlights its advantages in detecting small targets and emphasizing key regions.

For FAVF and DME, which are regional features, the model needs to capture global regional characteristics while suppressing irrelevant background information. Additionally, DME is closely related to clinical data, as both BCVA and CST influence its feature extraction. The F1 Score of Attention U-Net was noticeably higher than other models, demonstrating that multi-scale feature extraction and attention mechanisms effectively enhance the model's recognition ability.

However, it is worth noting that all models performed moderately in detecting IRF and DME, indicating that handling ambiguous boundaries remains a challenge for these models. This suggests that the models may struggle to accurately capture the extent of fluid accumulation. Moreover, DME often correlates strongly with IRF, and the models may fail to effectively distinguish these two features. In terms of multimodal data fusion, we employed a simple weighted fusion strategy for combining biomarkers and clinical data, which may have negatively impacted the model's predictive capability. Additionally, the dataset for DME is relatively small, with only 2% of valid data, significantly less than other biomarkers. This data imbalance likely caused the model to favor the prediction of other majority classes during training.

2) Analysis of Accuracy Results:

According to the experimental results, Attention U-Net achieved higher accuracy across the selected four biomarkers, particularly showing superior performance on the complex target IRF. This indicates that Attention U-Net is more effective in capturing regional features and handling ambiguous targets. Additionally, the detection of DME was relatively easier, with the accuracy of all models approaching 97%. This further demonstrates that the regional features of DME are very distinct, leaving limited room for performance improvement through the Attention module.

TABLE II
ACCURACY COMPARISON FOR BIOMARKERS

Biomarker	ResNet18	MMI	Attention U-Net
B1	0.7569	75.41	76.49
B3	0.6673	75.79	78.38
B4	0.7641	72.46	82.51
B5	0.9579	96.23	96.67

We also observed that IRF and DME exhibited low F1 Scores but high Accuracy. This discrepancy can be attributed to the fact that positive samples for IRF and DME (e.g., regions with fluid accumulation or macular edema) typically account for a small proportion, while negative samples (e.g., regions without fluid or macular edema) dominate the dataset. Due to this class imbalance, models tend to predict the majority of samples as negative, resulting in high Accuracy. However, the F1 Score, which emphasizes the positive class, significantly decreases.

3) AUC Analysis:

With the AUC results, we can see the advantages of Attention U-Net in detecting biomarkers. For a small target such as IRHRF, the high AUC values indicate that the model is able to focus well on key regions, which is consistent with the results of the F1 score analysis.

For two biomarkers, IRF and DME, the increase in AUC suggests that the attention mechanism and multimodal fusion play an important role.The increase in AUC for IRF suggests that the model has improved its handling of boundaries, which solves the problems identified in the previous F1-score analysis.The better performance of AUC for DME suggests that the model successfully combines OCT images with clinical data.

TABLE III
PERFORMANCE COMPARISON ON BIOMARKER DETECTION (AUC)

Biomarker	ResNet18	MML	Attention U-Net
B1 (IRHRF)	0.80	0.82	0.83
B3 (FAVF)	0.71	0.80	0.81
B4 (IRF)	0.77	0.80	0.84
B5 (DME)	0.92	0.92	0.95

V. CONCLUSIONS

In this study, we trained a multimodal deep learning model for biomarkers in retinal disease diagnosis. By combining OCT images and clinical information, we evaluated and compared models such as ResNet18, MML and Attention U-Net. The experimental results show that Attention U-Net performs optimally in the detection task of four biomarkers.

Attention U-Net performs better in region feature recognition and fuzzy boundary processing through dynamic weighting mechanism and multi-scale feature capture, which improves the detection accuracy of targets such as FAVF and IRF. Meanwhile, Attention U-Net effectively integrates the complementary nature of image and clinical information to further improve the prediction ability of DME.

However, this study also found limitations of the methods used. In particular, there is still some room for improvement in the F1 Score of the model in the detection tasks of IRF and DME. Future work will focus on optimization strategies and explore the use of more advanced network architectures, such as Transformer, to further improve the performance of the model.

REFERENCES

[1] World Health Organization, "Blindness and visual impairment," 2023. Accessed: 2024-12-01.

[2] American Academy of Ophthalmology, "What is optical coherence tomography?," 2023. Accessed: 2024-12-01.

[3] M. Prabhushankar, K. Kokilepersaud, Y.-y. Logan, S. Trejo Corona, G. AlRegib, and C. Wykoff, "Olives dataset: Ophthalmic labels for investigating visual eye semantics," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 9201–9216, Curran Associates, Inc., 2022.

[4] M. Tan and Q. Le, "An efficient compound scaling strategy for model performance optimization," *arXiv preprint arXiv:2310.14005*, 2023.

[5] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[9] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention u-net: Learning where to look for the pancreas,” 2018.

[10] P. K. Karn and W. H. Abdulla, “Advancing ocular imaging: A hybrid attention mechanism-based u-net model for precise segmentation of sub-retinal layers in oct images,” *Bioengineering*, vol. 11, 2024.

[11] B. Pang, L. Chen, Q. Tao, E. Wang, and Y. Yu, “Ga-unet: A lightweight ghost and attention u-net for medical image segmentation,” *Journal of Imaging Informatics in Medicine*, vol. 37, 03 2024.

[12] D. M. W. Powers, “Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation,” *ArXiv*, vol. abs/2010.16061, 2011.

[13] N. Japkowicz and M. Shah, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[14] A. P. Bradley, “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.