Advanced Data Augmentation and Additive Fusion for Predicting BCVA and CST

Alexander Snapp
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia
asnapp6@gatech.edu

Cac Phan
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia
ctphan@gatech.edu

John Donahoe
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia
jdonahoe8@gatech.edu

Hyochang Kim
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia
hkim931@gatech.edu

***Abstract*—We propose a deep learning framework for predicting Best Corrected Visual Acuity (BCVA) scores and Central Subfield Thickness (CST) measurements using multimodal medical data, including OCT scans. Our approach leverages advanced preprocessing techniques and state-of-the-art architectures, including ResNet and Vision Transformer (ViT), for feature extraction and multimodal fusion. The models demonstrate strong predictive performance, offering potential for improving clinical workflows in ophthalmology. The code used for this study is available at https://github.com/akisnapp/BiomarkerAnalysis .**

I. Introduction

Ophthalmology has greatly benefited from advancements in imaging technologies, enabling non-invasive assessment of retinal structures and aiding in the diagnosis of eye disorders. Key imaging modalities include Optical Coherence Tomography (OCT), which provides high-resolution cross-sectional views of the retina, and Fundus imaging, which captures detailed views of the retina, macula, and optic nerve.

BCVA measures a patient's best possible visual acuity with corrective measures and is traditionally assessed using a Snellen chart. CST, derived from OCT scans, quantifies retinal thickness and is a critical biomarker for conditions such as macular edema. Automating the prediction of these metrics can enhance clinical diagnostics by reducing time and resource constraints.

This study explores two deep learning architectures for predicting BCVA and CST: a ResNet-based model, which excels in feature extraction and multimodal fusion, and a Vision Transformer (ViT)-based model, leveraging self-attention mechanisms for processing imaging and clinical data. Using an 80/10/10 train-validation-test split and optimized with the Adam optimizer, the models integrate OCT scans to provide accurate predictions.

II. Related Work

A. IEEE SPS VIP Cup 2023

Recent advancements in deep learning have significantly enhanced ophthalmic biomarker detection through various strategies for processing Optical Coherence Tomography (OCT) images. In the 2023 SPS VIP Cup, based on the OLIVES dataset in [1], teams predicted whether a patient had Diabetic Retinopathy (DR) or Diabetic Macular Edema (DME). Team IITH [2] demonstrated the effectiveness of traditional Convolutional Neural Network (CNN) architectures by employing InceptionNet V3, achieving an F1 score 0.7682. Their work emphasized the importance of image preprocessing and highlighted the relationship between dataset size and model complexity. Similarly, Team Synapse [3] achieved state-of-the-art performance using an ensemble approach, underscoring the potential benefits of combining multiple models.

B. Residual Network Architecture

The ResNet architecture, introduced by He et al. [4], has become fundamental in medical image analysis due to its ability to mitigate the vanishing gradient problem through skip connections. Building upon this foundation, recent studies have demonstrated the effectiveness of ResNet-based models in OCT image classification. For instance, Pratap et al. [5] achieved 97.2% accuracy using ResNet-18 for automated diagnosis of retinal diseases from OCT images. Similarly, Li et al. [6] employed a modified ResNet50 architecture with transfer learning, achieving an overall classification accuracy of 99.48% in diagnosing retinal diseases.

C. Data Preprocessing and Augmentation

Data preprocessing and augmentation strategies specific to OCT images are crucial for robust model training. A comprehensive survey by Yang et al. [7] highlights various augmentation techniques including geometric transformations and color space augmentations to enhance dataset diversity and improve model generalization. Specifically for OCT images, Wang et al. [8] demonstrated that advanced augmentation techniques including elastic deformation and noise injection could effectively enhance model performance. These findings align with our preprocessing approach, which implements multiple augmentation strategies including elastic deformation, random sharpness adjustment, and various noise types (Gaussian, salt-and-pepper) to improve model robustness.

III. Methodology

A. Dataset

The OLIVES dataset that this model is trained and tested on has 4 different types of data: Scalar Clinical Labels (such as Patient ID and Visual Acuity Scores), Vectorized Biomarkers (including 16 different medically quantifiable characteristics that can show progression of diseases), 3D OCT scans, and 2D Fundus images. The dataset contains 78,189 individual OCT scans, 1,268 Fundus images, 5,072 Clinical Labels, and 150,528 Vectorized Biomarkers.

Each patient in the dataset is denoted by a unique Patient ID. The size of data associated with each patient is variable to the number of times a patient comes to the clinic. Per visit on a given patient's eye, a clinician takes 49 OCT scans, 1 Fundus image, and 16 vectorized biomarker analyses, which are all combined with the patient's Visual Acuity score and Central Subfield Thickness value (thickness of macula).

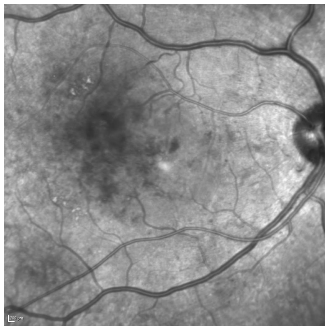

Fig. 1. Fundus Image

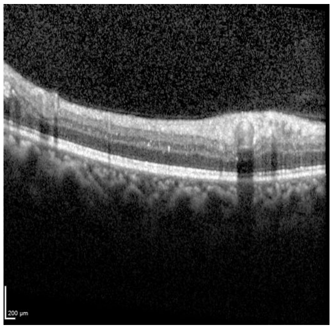

Fig. 2. OCT Scan

With the extensive images available for each patient, it becomes feasible to develop a Convolutional Neural Network to analyze the data and predict current diagnoses. The model will predict with high accuracy and can provide valuable insight into potential signs of future deterioration in the patient's eye and catch irregularities in the eye early, allowing for earlier predictions in potential regression in patient's vision.

Therefore, we used the OCT scans in the dataset in order to predict a patient's BCVA scores and Central Subfield Thickness values.

B. Data Preprocessing Procedure

The first step for setting up our model for training was the development of a data preprocessing pipeline which set each OCT scan in the dataset through a series of transformations in the following order:

1) Elastic Deformation
2) Random Sharpness
3) Random Erasing
4) Add Salt/Pepper Noise
5) Add Gaussian Noise
6) Random Brightness and Contrast

The first data augmentation technique used was Elastic Deformation of the images. In a similar fashion as [9], a slight elastic deformation of the OCT scans was used in the preprocessing pipeline to warp the scans. For our deformation parameters, we used an alpha value of $\alpha = 5.0$ to set the magnitude of the warping to be relatively small and a sigma value of $\sigma = 10.0$ to set the standard deviation of the Gaussian filter in order to control the smoothness of the deformation. These values were chosen after testing various values on the OCT scans as they provided minimal changes to the original image while still accentuating minor warps already present in the scan, allowing the CNN to pick up on these warps easier due to their increase in prominence following the elastic deformation procedure. Figure 4 shows an example of an elastically deformed image with an increased α value to demonstrate the effects of elastic deformation.

Fig. 3. Original Image

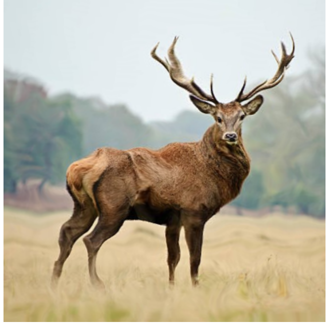

Fig. 4. Elastically Deformed Image using $\alpha = 200$ and $\sigma = 10$

Following the application of elastic deformation, random sharpness adjustments were used, utilizing the Pillow ImageEnhance Sharpness functions. The factors were selected randomly between 0.5 and 2, with 0.5 reducing overall image sharpness (blurring) and 2 significantly heightening details in a given image. This ensures that slightly blurry OCT/Fundus images will still have the potential for quality predictions.

The next three stages in the pipeline (Random Erasing, Salt/Pepper Noise, and Gaussian Noise) were implemented in order to reduce overfitting in the model. Random erasing works by splitting the image into a grid and randomly selecting a small portion of the squares from the grid to erase, thus forcing the model to pick up on multiple different factors in the scan as some areas in the scan may not be present due to being erased. This also reduces overfitting by randomly removing parts of each scan, thus introducing variance along with stopping the data from attempting to key in on certain small patterns. Adding Salt/Pepper Noise and Gaussian Noise help reduce overfitting by preventing the model from ”memorizing” small patterns in the data and by helping introduce variance.

The final stage of the preprocessing pipeline was done by randomly modifying the brightness and contrast of the images. The changes made were very minimal, as the level of brightness and contrast in each given image in the OLIVES dataset was highly similar.

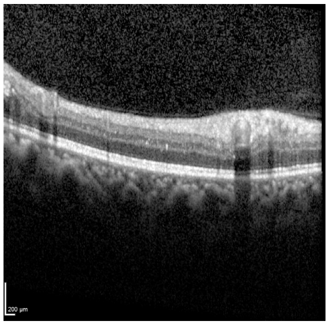

Fig. 5. Original Image

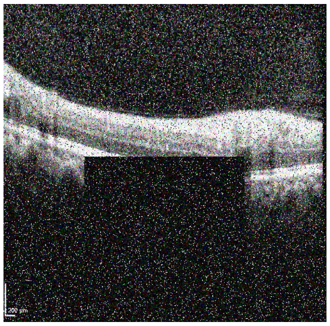

Fig. 6. Image after transformations have been applied (Elastic Deformation, Random Sharpness, Random Erasing, Salt & Pepper Noise, Gaussian Noise, Random Brightness & Contrast)

All images underwent standard preprocessing, including resizing to 224×224 pixels and normalization using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). This comprehensive augmentation strategy not only prevented overfitting but also improved the model's ability to generalize across slight differences in image qualities while maintaining its capacity to identify critical diagnostic features.

Techniques such as horizontal flipping and random rotations were not required transformations in this instance as all OCT scans were taken the same way in the dataset.

C. Model Architectures

To train and evaluate our models, we performed an 80/10/10 split of the dataset for training, validation, and testing, respectively. This ensured a sufficient training set while retaining unbiased evaluation metrics.

1) ResNet-Based Multimodal Fusion Model: The ResNet-based fusion model utilizes the **ResNet-50** architecture for extracting image features and incorporates additional biomarker features through a fusion mechanism. The architecture comprises:

ResNet-50 Backbone: The model uses a pre-trained ResNet-50 architecture on ImageNet to provide robust image feature extraction. The first convolutional layer is adjusted to handle RGB images. The final fully connected layer is replaced with an identity mapping, enabling feature extraction for downstream tasks.

Multimodal Fusion Layer: Biomarker inputs are normalized and concatenated with image features extracted from ResNet. A fully connected layer maps the combined features to the output space, supporting both regression and classification tasks.

2) Multimodal Transformer Model: The transformer-based multimodal model integrates vision and biomarker data with the following components:

Vision Transformer Backbone: The model uses the pre-trained ViT-Base architecture for extracting high-dimensional image features. The '[CLS]' token or the last hidden state is used as the image representation.

Transformer Encoder for Biomarkers: The biomarker data is processed as sequences using a transformer encoder with two layers and a hidden dimension matching the biomarker count. This component models the relationships between biomarkers through self-attention mechanisms.

Fusion Layer: The image and biomarker features are concatenated into a single representation. A two-layer feed-forward neural network generates predictions for both regression and classification tasks.

Next, we changed the fusion method—how different data modalities like images and biomarkers are combined. Traditional approaches often use concatenation, merging feature vectors into a single representation. However, this method may not effectively capture complex interdependence between modalities. Alternative fusion strategies, such as additive fusion (element-wise addition of feature vectors) which allow models to learn optimal combination strategies, can better capture intricate relationships between modalities, leading to improved performance.

D. Training Configuration

Both models were trained using the **Mean Squared Error (MSE)** loss and the **Adam optimizer**. We set the number of epochs to 25 to balance between minimizing loss and avoiding overfitting.

The MSE loss function is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the true label, and $\hat{y}_i$ is the predicted value. This function penalizes larger errors more heavily, ensuring minimal prediction deviation.

The Adam optimizer dynamically adjusts learning rates for network parameters, improving efficiency for large datasets.

E. Test Validation and Metrics

To evaluate the models, we used several metrics, including **MSE**, **F1 score**, **ROC-AUC**, and **Average Precision (AP)**.

1) F1 Score for BCVA Prediction: The **F1 score** evaluates the balance between precision and recall, specifically for Best Corrected Visual Acuity (BCVA) classification, which is defined as:

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

True Positive (TP): The predicted BCVA matches the true label within an acceptable threshold (e.g., ±1 line of visual acuity).

False Positive (FP): The predicted BCVA is within the threshold, but the true label is significantly different.

False Negative (FN): True BCVA lies within the threshold, but the prediction deviates substantially.

2) ROC-AUC and Average Precision: The **ROC-AUC** evaluates the area under the Receiver Operating Characteristic curve, assessing the model's ability to distinguish between BCVA classes. **Average Precision (AP)** measures the precision-recall trade-off, emphasizing true positive predictions.

IV. Experimental Results

A. Preprocessing Impact

One example of testing we ran below is a comparison of the Elastic Deformation preprocessing alpha and probability values. Alpha is the magnitude of the warping, and probability is whether or not the image is distorted at all in the preprocessing function. This is with using the main ResNet architecture. Below is a comparison of F1 scores with modifications:

TABLE I
Elastic Deformation Parameters in ResNet vs. F1 Score

	$\alpha = 1, p = .5$	$\alpha = 5, p = .5$	$\alpha = 10, p = .5$	$\alpha = 5, p = 1$
F1	0.9494	0.9499	0.9438	0.9381

The forced elastic deformation of every trained image (p = 1) shows worse overall results than the models with half of the training images deformed (p = .5). Even though the alpha value of 5 results in a very minimally augmented image, the result is still great enough to result in a worse trained model.

B. Model Performance Comparison

The results in Table II shown the improvement in accuracy of multimodal transformer model architectures for predicting BCVA and CST:

TABLE II
Different Models Architecture Results

Model	**MSE**	**MAE**	**R2**	**Accuracy**
ResNet50	0.029	0.105	0.972	0.927
Multimodal Transformer	0.01	0.062	0.99	0.964
Multimodal Transformer (w/ Add)	**0.009**	**0.063**	**0.991**	**0.972**

Model	**Precision**	**Recall**	**F1**
ResNet50	0.912	0.967	0.938
Multimodal Transformer	0.959	0.985	0.972
Multimodal Transformer (w/ Add)	**0.967**	**0.987**	**0.974**

The Multimodal Transformer (when adding features) outperformed other architectures, achieving the highest Macro Average F1 Score (0.974), ROC-AUC (0.998), and Precision (0.967), while recording the lowest MSE (0.009). This underscores the benefit of leveraging both image data and biomarkers through sophisticated fusion mechanisms.

The ResNet50 architecture, while a robust baseline model, showed slightly lower performance across all metrics, with an MSE of 0.029. Its simpler architecture and absence of advanced fusion layers likely contributed to this difference.

The standard Multimodal Transformer demonstrated significant improvement over ResNet50, confirming the advantage of transformer-based approaches in capturing complex interactions between OCT images and biomarker data.

These findings validate the effectiveness of multimodal learning frameworks in ophthalmic diagnostics. The superior performance of the Modified Multimodal Transformer suggests that integrating additional architectural enhancements can further refine predictions, making it a promising tool for clinical decision-making.

V. Discussion

These augmentations simulate the variability encountered in clinical settings, enabling models to better handle diverse real-world data. Additionally, we explored architectural enhancements by integrating additive fusion strategies within Vision Transformer (ViT) models. Traditional approaches often rely on simple concatenation for multimodal data fusion, which may not effectively capture complex interdependencies between modalities. Our additive fusion method allows the model to learn optimal combinations of features from different data sources, leading to improved performance in tasks such as medical image classification.

VI. Conclusion

A. Findings

Ultimately, our preprocessing and data augmentation techniques on top of the already robust ResNet architecture enables the model to accurately predict BCVA scores and CST given the OCT scans, Fundus images, and Vectorized Biomarkers. For our specific use case, an F1 score of 0.974 shows the viability of the model. Some preprocessing techniques that we pursued ended up producing worse results, such as horizontal flipping and random rotations. Given more time, there is likely more preprocessing parameter modifications that could have been done to produce even better results, however this would significantly increase overall time to train the model.

B. Future Direction

While our current implementation demonstrates promising results, several avenues remain unexplored. Advanced augmentation techniques such as mixup (generating linear combinations of images) and CutMix (replacing image patches) could improve model robustness. Similarly, contrast enhancement methods like CLAHE could help distinguish subtle pathological features.

Alternative CNN architectures also offer potential improvements. EfficientNet's balanced scaling approach and DenseNet's feature reuse capabilities could enhance performance while maintaining computational efficiency. Additionally, architectures like Inception (multi-scale feature capture) and MobileNet (lightweight design) might provide complementary benefits for medical image analysis.

Additionally, further testing of different Optimizers and Loss functions could show further increase in quality of prediction and an increase in F1 scores. Built-in optimizers in PyTorch such as Stochastic Gradient Descent or variations of the Adam architecture may result in better performing models.

Future work would focus on evaluating these techniques, architectures, and optimizers, with improvements assessed based on both accuracy and clinical practicality.

References

[1] M. Prabhushankar, K. Kokilepersaud, Y. Logan, S. Corona, G. AlRegib, and C. Wykoff ”OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics” Advances in Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks, Nov 29 - Dec 1, 2022.

[2] Aaseesh Rallapalli, Lokesh Badisa, Nithish S, Utkarsh Doshi, Soumya Jana, ”Ophthalmic Biomarker Detection Using Inception Net” IEEE SPS VIP CUP 2023: TEAM IITH.

[3] H.A.Z. Sameen Shahgir, Khondker Salman Sayeed, Tanjeem Azwad Zaman, Md. Asif Haider, ”Ophthalmic Biomarker Detection Using Ensembled Vision Transformers” IEEE SPS VIP CUP 2023: TEAM SYNAPSE.

[4] He et al., ”Deep Residual Learning for Image Recognition”, IEEE CVPR, 2016.

[5] Pratap et al., ”Automated Diagnosis of Retinal Diseases from OCT Images Using ResNet-18”, IEEE, 2021.

[6] Li et al., ”Deep Residual Network for Automatic Diagnosis of Retinal Diseases Using Optical Coherence Tomography”, Springer, 2021.

[7] Yang et al., ”Image Data Augmentation for Deep Learning: A Survey”, arXiv, 2022.

[8] Wang et al., ”Enhancing OCT Image Quality with GAN-Based Data Augmentation for Retinal Disease Diagnosis”, IEEE, 2021.

[9] Bar-David et al., ”Elastic Deformation of Optical Coherence Tomography Images of Diabetic Macular Edema for Deep-Learning Models Training: How Far to Go?”, IEEE, 2023.