An Explainable Vision Transformer Model for Staging Diabetic Retinopathy from Low-Quality Fundus Images

Jeannie She
jeanshe@mit.edu

Katie Spivakovsky
kspiv@mit.edu

I. Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness, affecting over 100 million people worldwide, with higher prevalence among socioeconomically disadvantaged groups. In the U.S., up to 9.6 million individuals suffer from DR [1], and those from lower-income communities face a 48% higher risk of developing advanced stages before diagnosis, primarily due to limited access to screening [2]. Comorbidities—other medical conditions that occur alongside diabetes, such as hypertension and chronic kidney disease—can also accelerate disease progression. We propose a novel pipeline combining retinal imaging, socioeconomic factors, and comorbidity profiles to improve DR staging accuracy. A significant barrier for underserved populations is the poor quality of retinal fundus images, leading to delays in diagnosis and treatment. Our solution leverages deep learning to accurately stage DR even from blurry, low-quality, or artifact-containing images, minimizing the need for repeat clinical visits. By maintaining diagnostic accuracy on suboptimal images and integrating critical health data, our system may improve early detection, particularly in underserved populations where advanced DR is often first identified. This approach has the potential to reduce healthcare costs, increase early detection rates, and address disparities in access to care, promoting healthcare equity.

II. Introduction

Diabetic retinopathy (DR) represents one of the leading causes of preventable blindness worldwide, affecting approximately 103 million people globally [3]. In the United States alone, an estimated 9.6 million individuals suffer from this sight-threatening complication of diabetes, with prevalence rates disproportionately higher among socioeconomically disadvantaged populations [1]. Studies indicate that patients from lower-income communities face up to a 48% higher risk of developing advanced DR stages before diagnosis, largely due to limited access to regular screening [2]. Furthermore, the presence of comorbidities significantly impacts disease progression, with hypertension increasing the risk of severe DR by 2-3 times and chronic kidney disease accelerating progression by up to 4 times [4], [5]. We propose a novel pipeline—from preprocessing data to training a vision model to interpreting results—that incorporates not only retinal imaging data but also socioeconomic determinants of health and comorbidity profiles to more accurately stage DR.

As of 2025, the FDA has approved IDx-DR, the first AI-driven tool for diagnosing diabetic retinopathy (DR), underscoring the growing need for such technologies. Although this model is in use globally [6], it is restricted to images captured with specialized digital fundus cameras and cannot process poor-quality images, such as grayscale, blurry, or scanned ones. Efforts are underway to enable smartphone-based retinal imaging, offering a more affordable alternative to expensive fundus cameras. However, a major challenge for socioeconomically disadvantaged populations is the poor quality of fundus images, which can delay DR diagnosis. In fact, 14% of imaging cases require patients to return for another session, further prolonging the time until treatment is initiated [7].

We aim to use deep learning to address this issue. We will train our model to maintain high accuracy on adversarial images that are blurry, low quality, or contain artifacts. This novel approach will enable DR staging even in situations that typically would require further clinical visits, accelerating the rate of treatment for many patients. Our computer vision pipeline will represent a significant advancement in DR screening by maintaining diagnostic accuracy even with lower-quality fundus images while incorporating critical comorbidity and socioeconomic data. Implementation of this system could reduce the estimated $93,000 lifetime cost per case of blindness due to DR [8] while potentially increasing early detection rates in underserved communities, thus also bringing awareness to the critical need for healthcare equity irrespective of socioeconomic factors.

III. Related Work

Many deep learning models predicting DR have been produced, but to our knowledge, none use data containing demographic features or comorbidities. Dai et al. (2021) impressively identify not only the lesions on the retinal image itself, but also the medical classification of the lesions [9]. However, they only utilize images as input, lacking the broader implications of demographic and medical history of the patient. Since the dataset we are using contains both images and tabular features, we are heavily inspired by Li et al. (2024), who discuss the advantages of different methods of multimodal data concatenation [10]. In our project, we will explore methods of fusing modes of data in order to achieve the highest performing model. Finally, we plan to embed explainability via a saliency map. We look to Szczepankiewicz et al. (2023), who visualize which image regions a CNN pays attention to [11]. This explainability will add a layer of credibility for clinical use and help us identify if the model has learned meaningful features differentiating stages of DR.

IV. Methods

Existing DR deep learning models utilize convolutional neural networks, vision transformers, and autoencoders in their architecture. However, these existing methods solely rely on images; for our project, we intend to train our model on additional tabular features (numerical and categorical). Our model will use the mBRSET dataset [12] which provides dozens of these additional features, ranging from demographic including age and sex, to medical including insulin usage and obesity. We will train a neural network to predict DR staging solely on these tabular features, then identify the 5 most predictive features using Shapley scores. We hope that this would better inform potential risk factors for DR patients and might prove more clinically relevant context for the prediction.

We next aim to train a residual neural network solely on mBRSET images to establish a baseline predictive accuracy for vision models via metrics such as as accuracy on held-out test data as well as a confusion matrix. We will use ResNet-18 pretrained on ImageNet due to its smaller architecture size apt for our small data set. Using Google Colab's GPU and potentially Satori cluster resources, we will train ResNet-18 for 10 epochs and test different hyperparameters.

After establishing these tabular-based and image-based baselines and identifying the 5 most predictive tabular features, we will develop a multimodal model integrating both data types. We have yet to finalize the exact architecture, but we are inspired by vision transformers; we may attempt to concatenate image embeddings with tabular feature embeddings, potentially identified through a separate autoencoder model, before passing this embedding through the rest of a traditional ViT architecture. We will then train the model and evaluate its performance on held-out test data.

On top of this main workflow, we hope to train an additional logistic regression model predicting the confidence of our multimodal model's predictions, which we assume will correlate with the quality of the image; we name this secondary aspect the *rejector model.* Inspired by deferral systems present in other medical software applications, we believe it is critical to produce an AI model that does not stand alone and instead can be incorporated into a doctor's routine. In the medical workplace, our rejector model will enable doctors to decide whether they can trust the decision model's diagnosis or whether the doctor must perform a manual diagnosis.

V. Results

We have downloaded mBRSET to a private GitHub repository in accordance with data privacy policies. We have not begun training models or rigorously analyzing mBRSET, but we have set up TA meetings and team meetings in the coming weeks to hold our team accountable and ensure gradual progress.

VI. Completion Plan

We have set aside over 30 hours throughout the coming weekends to work on our project, in addition to TA office hours during which we hope to receive feedback. By the end of the April 26/27 weekend, we hope to have both our tabular-based and image-based baseline models trained. Jeannie will work on data pre-processing and ensure that we have a reliable method of accessing the data while adhering to the data usage agreement. Katie will run the pretrained model on the mBRSET. Jeannie will train the pretrained model on a training set of mBRSET and evaluate the test set separately. Katie will run a neural network or other simple architecture to identify the most predictive demographic features for the DR outcome. Then, we will both explore methods of concatenating the demographic data to the images. We will work together to train the model on concatenated data and visualize the results. Jeannie will separately work on the rejector model and think about on what metrics to grade model confidence. We are meeting with a TA on April 25 and hope to have more concrete ideas for our multimodal model following this meeting so that we can begin implementation over the May 3/4 weekend. We will wrap up modeling and write up results over the May 10/11 weekend in advance of the May 13 deadline.

References

[1] "VEHSS Modeled Estimates: Prevalence of Diabetic Retinopathy (DR)." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, www.cdc.gov/vision-health-data/prevalence-estimates/dr-prevalence.html.

[2] Zheng, Y., et al. (2012). Socioeconomic status and diabetic retinopathy: The Singapore Epidemiology of Eye Disease Study. Ophthalmology, 119(9), 1877-1883.

[3] Wong TY, Tan TE. The Diabetic Retinopathy "Pandemic" and Evolving Global Strategies: The 2023 Friedenwald Lecture. Invest Ophthalmol Vis Sci. 2023 Dec 1;64(15):47. doi: 10.1167/iovs.64.15.47. PMID: 38153754; PMCID: PMC10756246.

[4] UK Prospective Diabetes Study Group. (1998). Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ, 317(7160), 703-713.

[5] Wong, C. W., et al. (2016). Chronic kidney disease and the risk of sight-threatening diabetic retinopathy: a systematic review and meta-analysis. Diabetes Care, 39(12), 2296-2303.

[6] DigitalDiagnostics (2020). Digital Diagnostics, formerly IDx, Expands Global Impact of Healthcare Autonomous AI with Acquisition of 3Derm Systems, Inc. DigitalDiagnostics. https://www.digitaldiagnostics.com/digital-diagnostics-formerly-idx-expands-global-impact-of-healthcare-autonomous-ai-with-acquisition-of-3derm-systems-inc/

[7] Tufail, A., et al. (2016). Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology, 123(5), 1136-1143.

[8] Javitt, J. C., Aiello, L. P., Chiang, Y., Ferris, F. L., Canner, J. K., and Greenfield, S. (2008). Preventive eye care in people with diabetes is cost-saving to the federal government: implications for health-care reform. Diabetes Care, 31(6), 1269-1271. https://doi.org/10.2337/dc07-2215.

[9] Dai, L., Wu, L., Li, H. et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 12, 3242 (2021). https://doi.org/10.1038/s41467-021-23458-5.

[10] Li, Y., et al. "A Review of Deep Learning-Based Information Fusion Techniques for Multimodal Medical Image Classification." Computers in Biology and Medicine, Pergamon, 22 May 2024, www.sciencedirect.com/science/article/pii/S0010482524007200.

[11] Szczepankiewicz, K., Popowicz, A., Charkiewicz, K. et al. Ground truth based comparison of saliency maps algorithms. Sci Rep 13, 16887 (2023). https://doi.org/10.1038/s41598-023-42946-w

[12] Wu, C., Restrepo, D., Nakayama, L.F. et al. A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. Sci Data 12, 323 (2025). https://doi.org/10.1038/s41597-025-04627-3.