A Clinically Interpretable Deep Learning Framework for the Detection and Grading of Diabetic Retinopathy

Frank Bogan[1], Hunter Mathias Gill[2], Doaa Hassan Salem[2], Michael Happe[1], Sarath Chandra Janga[1,3,4], Amir Reza Hajrasouliha[1].

[1]Glick Eye Institute, Indiana University School of Medicine; [2]Department of BioHealth Informatics, Luddy School of Informatics, Computing and Engineering, Indiana University Indianapolis; [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine; [4]Centre for Computational Biology and Bioinformatics, Indiana University School of Medicine.

Objective: One of the leading causes of blindness in working age adults is diabetic retinopathy (DR) which can result in vision loss if uncontrolled. DR can be detected and graded by fundus retinal imaging, however the amount of images requiring grading creates a burden on ophthalmologists. Thus, a growing demand exists for an optimized DR image reading process. To accomplish this, we propose the use of a deep learning artificial intelligence model to detect and grade DR using lesion feature extraction with clinical interpretation.

Methods: Retinal fundus images were collected from two sources for training (n=608): E-Ophthalmology, and a private Indiana University Eugene & Marilyn Glick Eye Institute dataset. Each dataset was divided into 70% (training), 20% (validation), and 10% (testing) subgroups. An external dataset, the UKBB, was also used (n=944) for evaluation. The AI model assigned images to 5 categories based on lesion features. The model operated through 2 stages: a multi-scale DeepLabV3+ to segment retinal lesions from input fundus images, followed by segmentation predictions for lesion. A classifier incorporates data to predict whether DR is present and the grade for images with DR.

Results: AI performance was analyzed using several metrics to compare against human grading. Developmental results for segmentation were: 0.88 (precision), 0.70 (recall), 0.78 (f1 score), 0.99 (accuracy), 0.94 (AUC), and 0.76 (IOU). Classification of DR results were: 0.89 (precision), 0.6 (recall), 0.78 (f1 score), 0.62 (accuracy), and 0.8 (AUC). External dataset results were: 0.47 (precision), 0.58 (recall), 0.52 (f1 score), 0.58 (accuracy), and 0.65 (AUC).

Conclusions: The proposed deep learning AI framework demonstrates good DR lesion detection and grading performance. Further improvement may allow AI to replace human grading, saving ophthalmologists time and standardizing the grading process. The current approach may be improved by including more lesion features and larger sample sizes.