Early Screening for Diabetic Retinopathy with Retinal Imaging Data

Michael Anderson

University of Kentucky

12/5/2019

Section 1

Introduction

Diabetic Retinopathy

- Diabetic retinopathy (DR) is essentially damage to the retina caused by diabetes
- Diabetes causes poor blood circulation and damage to blood vessels that can lead to permanent damage to the retina and other parts of the eye, resulting in vision loss and, if untreated, blindness
- Can display in many different ways, but often in the form of one or more of the following:
  - microaneurysms - swelling in small blood vessels due to weak vessel walls
  - exudates - mass of cells and fluid that have seeped out of a blood vessel
  - neovascularization - new, weak blood vessels created due to poor blood flow
  - edema or hemorrhaging - build up of fluid or internal bleeding

Examples

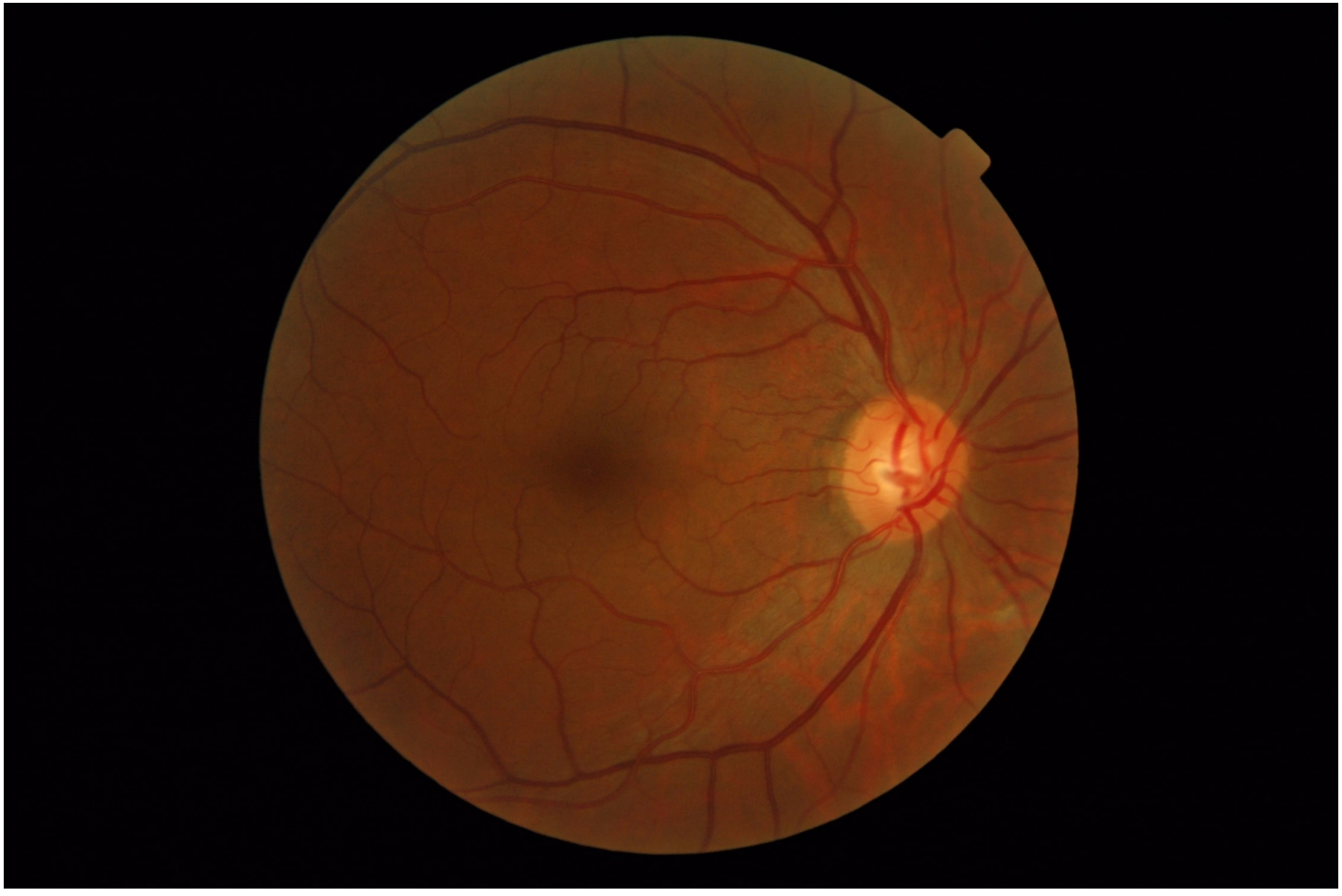


Figure 1: Normal Eye Fundus

Examples

Figure 2: DR with Microaneurysms and Exudates

Examples

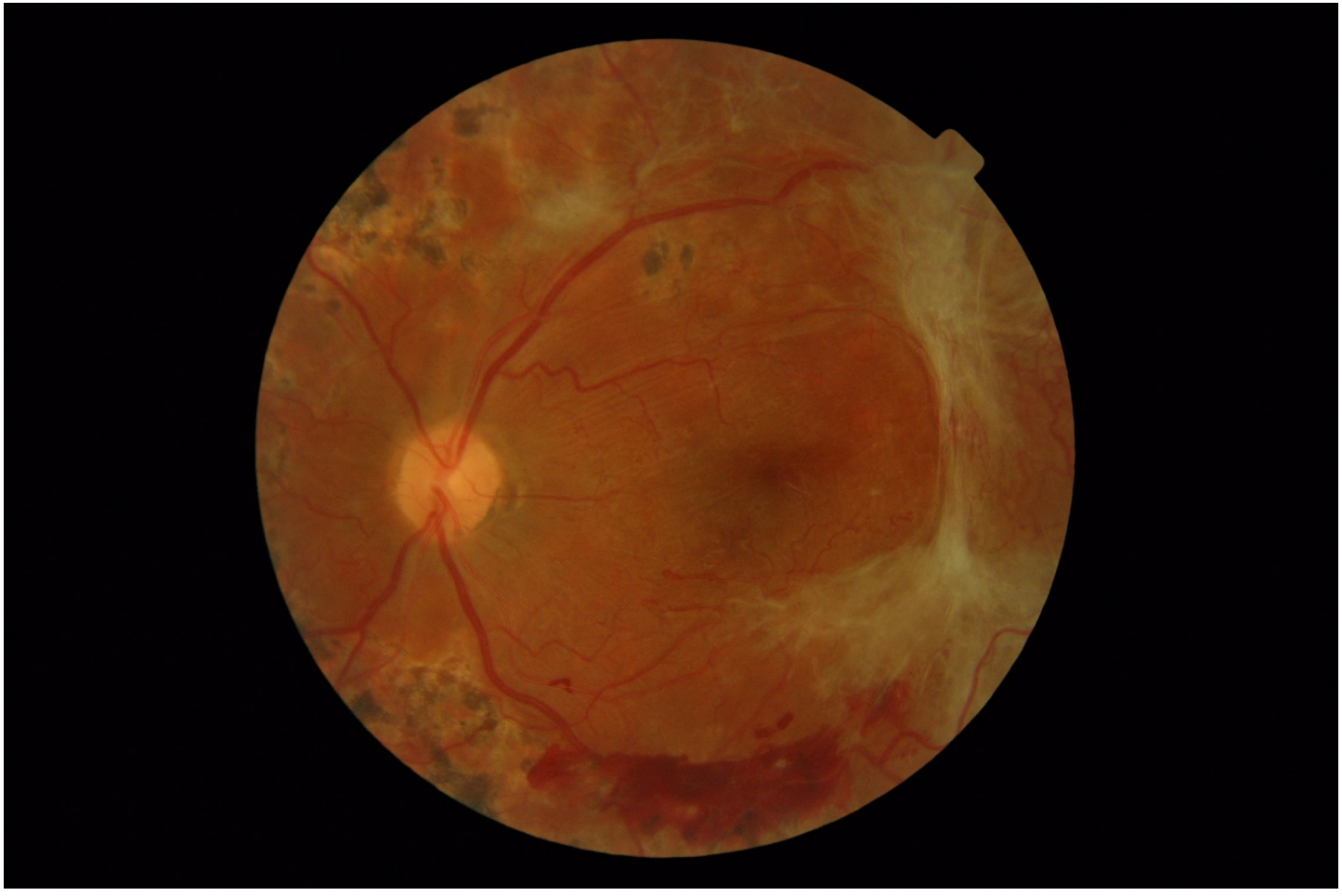


Figure 3: Severe DR with Hemorrhaging

Data Source

- Data came from Antal and Hajdu (2014), which used several machine learning algorithms and ensemble techniques to produce a screening system for DR from eye images
- The authors used the publicly available Messidor image dataset (Decencière et al. 2014) as an example of the applicability of their method
- Messidor set contains 1200 eye fundus images, with 4 levels of DR, by physician diagnosis:
  - R0 = No DR
  - R1 = Mild, Non-Proliferative DR
  - R2 = Severe, Non-Proliferative DR
  - R3 = Proliferative DR
- Antal dataset has data on 1151 patients, with various features extracted from the underlying Messidor images

Variables

- Outcome:
  - *DR*: Diabetic Retinopathy status (binary)
- Predictors:
  - *AMFM*: Prediction of DR status from AM/FM based feature extraction (binary)
  - *PRE*: Pre-screening indicator of severe retinal abnormality (binary)
  - *DMD*: Estimated distance between centers of macula and optic disk
  - *ODD*: Estimated diameter of the optic disk
  - *MA1* - *MA6*: Estimated counts of microaneurysms, at increasing confidence levels (higher numbers are more conservative)
  - *EX1* - *EX8*: Estimated counts of exudates, at increasing confidence levels (higher numbers are more conservative)
- Sidenote: these are all my own chosen names for the variables, since the variables didn't have any labels to begin with

Data Pre-Processing

- There were no missing values in the dataset, but there were four observations that were denoted as having "poor image quality"
- For simplicity, those four observations were dropped from the analysis, leaving 1147 observations total
- All of the *MA* and *EX* variables were log-transformed, due to their being count variables and having a fairly heavy right skew
- The overall percentage of patients with DR in the dataset was about 53%, so 0.5 was used as the split point for classification of all predictions
- Finally, the data was split into train/validation/test sets, using a 60/20/20 split, maintaining a roughly equal distribution of the outcome in all three sets

Section 2

Model Training

General Training Framework

- All models were trained on the 60% training dataset only
- The `caret` package (Kuhn 2008) was used to implement 5-fold cross-validation (CV) within the training set for the selection of any tuning parameters
- Prediction accuracy of final model was evaluated using the 20% test dataset
- The 20% validation dataset was set aside for the exploration of model ensembling techniques later in the project

Penalized Logistic Regression

- Penalized logistic regression (PLR) was performed on all predictors, using the `glmnet` package's implementation of elastic net regularization (Friedman et al. 2010)
- Elastic net regression utilizes a penalty parameter of λ, and a parameter of α to measure the amount of mixing between lasso and ridge regression
- CV produced an optimal α of 1, which represents pure lasso regression
- After fixing $\alpha = 1$, the optimal λ was calculated using `glmnet`'s built-in CV implementation, which produced a value of $\lambda \approx 0.0004$
- Utilizing this final tuning parameter combination, the final PLR model was fit to the full training dataset, in which none of the β coefficients were reduced to 0

Generalized Additive Model

- A generalized additive model (GAM) was trained on all predictors using the `mgcv` package (Wood 2011)
- The effect of each continuous and count predictor was estimated using an individually spline-smoothed fit, and the binary predictors were included linearly
- Originally, all *EX* and *MA* variables were to be smoothed together into their own multivariate groups, but the estimated degrees of freedom required to do so were very high (nearly the size of the training set), which indicated overfitting would have likely been an issue for that model
- Since GAMs do not have any explicit tuning parameters, once the functional form of the model described above was finalized, the GAM model was trained on the entire training set.

Extremely Randomized Trees

- Extremely randomized tree (ERT) models (Simm et al. 2014) are similar to random forest models, except:
  1. ERTs do not use the bagging technique of random forests, instead fitting every tree to the full training set
  2. The splits in a given tree are chosen from a small number of random splits among all selected predictors for that tree, leading to more variable tree performance
- CV was used to tune the number of trees, number of candidate variables selected per tree, and number of random splits selected per tree
- Notably, tree size was not tuned here, which could improve performance
- Once the tuning parameter values were selected, an ERT model was fit to the full training set

Section 3

Model Ensembling & Performance

Model Averaging

- The first model ensembling technique was just a simple average of the test predictions of the three models described previously
- Essentially, if two or more of the three models "voted" one way, then that was the vote that represented the entire group
- A slightly more sophisticated ensembling technique was also utilized, which was a weighted average of the three models' predictions.
- The weight used for each model's prediction was its validation set accuracy, so that more accurate models would get more of a say in the final vote, while less accurate models would be downweighted
- However, the three models didn't really perform dramatically differently, so the three weights ended up being pretty similar

Model Stacking

- The final model ensembling technique used was a simple model stacking algorithm
- Model stacking usually involves training a "meta-model" to decide the joint prediction of a group of individual models
- In this case, logistic regression was used to combine the predictions of the three training models
- A logistic regression model was trained, using the validation data, to predict the true validation outcome from the three sets of predicted probabilities produced by our three training models
- This logistic regression model would then be utilized to combine the three individual model predictions for each test set case by predicting the test set outcomes from the test set predictions of the training models

Model Performance

	ACCU	*SENS*	*SPEC*
PLR	0.773	0.779	0.766
GAM	**0.782**	0.779	**0.785**
ERT	0.716	0.730	0.701
sAVE	0.773	**0.795**	0.748
wAVE	0.773	**0.795**	0.748
GLM	0.777	0.779	0.776

Section 4

Conclusion

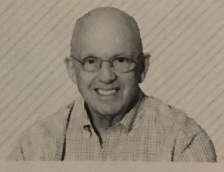
Pros and Cons

- Overall, I'm okay with the performance of these models
- Didn't do as well as the original paper, which had 90% prediction accuracy, but also used more sophisticated ensemble methods and combined 8 different individual modeling techniques
- Could certainly use more advanced techniques
- Also could use some more time on proper model/tuning parameter selection
- Regardless of how my techniques performed, I gained a lot of knowledge about the practical implementation of tuning and fitting various models, and knowing is half the battle!

Public Health Impact

Smart, But Not Smart Enough

AI image interpretation will not solve the diabetic retinopathy epidemic



By Steve Charles, CEO and Founder, Charles Retina Institute, Clinical Professor of Ophthalmology, University of Tennessee, USA

The use of artificial intelligence (AI) for interpreting digital retinal images is currently believed by many ophthalmologists and researchers to be a solution to the well documented worldwide diabetic retinopathy epidemic. However, there are both tactical and strategic issues that are seemingly overlooked by those promoting this "solution" – and I would like to explore some of them here.

Diabetic retinopathy is a function of poor serum glucose control as shown by the DCCT and many other worldwide, multi-center clinical trials. The worldwide obesity explosion both in developed and developing countries is a core problem driving the rapid growth of diabetic populations. Diet education starting in early childhood is a crucial part of addressing this epidemic. Access to preventive medicine, diabetes medications, and frequent or real time blood sugar monitoring is a huge financial burden but less costly than waiting until diabetic retinopathy and other complications develop to initiate care.

Socioeconomic and cultural issues are far more important than a perceived shortage of ophthalmologists to read digital fundus images. Assuming AI was effective – who would pay for technicians to acquire images, who would purchase the imaging devices, who would pay for transporting patients to the imaging device or the device to the patient? And even if all these issues were addressed, who would inject anti-VEGF agents or implant sustained delivery devices, and who would pay for the treatment? Anti-VEGF therapy has been shown to be more effective than laser photocoagulation; who would pay for the laser treatment and lasers if we moved back to this outmoded therapy? These socioeconomic and cultural issues far outweigh a perceived image interpretation burden. In reality, there is not a backlog of unread images.

"These socioeconomic and cultural issues far outweigh a perceived image interpretation burden."

All ophthalmologists frequently see diabetic retinopathy patients that fundus examination and digital fundus images would be interpreted as inactive or stable, but OCT demonstrates diabetic macular edema. Widespread availability of low-cost, portable and reliable OCT devices that can be operated by minimally-trained individuals is mandatory for screening to be effective. Trained graders perform as well or better than ophthalmologists in reviewing fundus images, as has been shown in clinical trials; this approach could be applied to OCT images as well.

Acknowledgements

I want to thank Kennie Anderson, COA, OSC (my wife) for inspiring me to pursue this topic and for providing her subject area expertise along the way. . .

and for, like, being my wife or whatever

References I

Antal, B., and Hajdu, A. (2014), "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Systems*, 60, 20–27. https://doi.org/10.1016/j.knosys.2013.12.023.

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., and Klein, J.-C. (2014), "Feedback on a publicly distributed image database: The MESSIDOR database," *Image Analysis & Stereology*, 33, 231. https://doi.org/10.5566/ias.1155.

References II

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33. https://doi.org/10.18637/jss.v033.i01.

Kuhn, M. (2008), "Building Predictive Models in *r* Using the **caret** Package," *Journal of Statistical Software*, 28. https://doi.org/10.18637/jss.v028.i05.

Simm, J., Magrans De Abril, I., and Sugiyama, M. (2014), "Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression," *IEICE Transactions on Information and Systems*, E97.D, 1677–1681. https://doi.org/10.1587/transinf.E97.D.1677.

References III

Wood, S. N. (2011), "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.