Early Screening for Diabetic Retinopathy with Retinal Imaging Data

Michael Anderson

Contents

Acknowledgement

I would like to thank my wife, Kennie Anderson, COA, OSC, for sparking my initial interest in this research topic, as well as for providing her subject area expertise along the way.

Introduction

The goal of this project will be to predict whether a diabetic patient is suffering from diabetic retinopathy, using fundus imaging of the patient's retina. First, there will be some background information on diabetic retinopathy and the specific dataset being used here. Then the data will be described, including a discussion of any abnormalities or required transformations. After that, the three different modeling techniques used in this project will be described - one parametric, one semiparametric, and one nonparametric - including a discussion of any tuning parameter selection methods used. Next, three different model ensembling techniques will be used to combine the predictions of the three models described, and the predictive ability of all six methods will be evaluated. Lastly, there will a discussion of the results and possible public health implications of this project.

Background

Diabetic retinopathy (DR) is a condition characterized by damage to the retina that is caused by symptoms of diabetes mellitus. The prevalence of DR among patients with diabetes is approximately 40%, and nearly 75 people go blind every day from DR (Antal and Hajdu 2014). Normally, it takes years of untreated progression of DR before irreparable vision loss sets in. And yet, DR is still the leading cause of blindness among adults under 65 years old, despite the fact that highly successful treatments are available for patients in which the disease is diagnosed before permanent retinal damage occurs. For these reasons, early detection of DR remains one of the biggest obstacles to a large-scale reduction in the public health impacts of this disease.

Early detection of DR can usually be achieved by an ophthalmologist using techniques such as 2-dimensional fundus imaging or 3-dimensional optical coherence tomography (OCT). While OCT and other, more recent, techniques can provide a much more detailed view of the eye and its pathologies, not all eye care professionals have access to the technology required to carry out these diagnostic procedures. Fundus imaging, however, is a relatively simple process, and virtually all ophthalmologists have access to the cameras required to carry it out. The fundus is the interior portion of the back of the eye, and includes the retina, optic disc, and macula. The image in figure 1 (Decencière et al. 2014) shows a picture of a normal eye fundus, where the bright, circular area to the right is the optic disc, the dark spot near the center is the macula, and the surrounding regions of the photo contain the retina, and the blood vessels that run through it.

The method that ophthalmologists usually use to diagnose DR involves the analysis of fundus images for retinal abnormalities that can often be caused by diabetes. In most patients, diabetes leads to poor blood

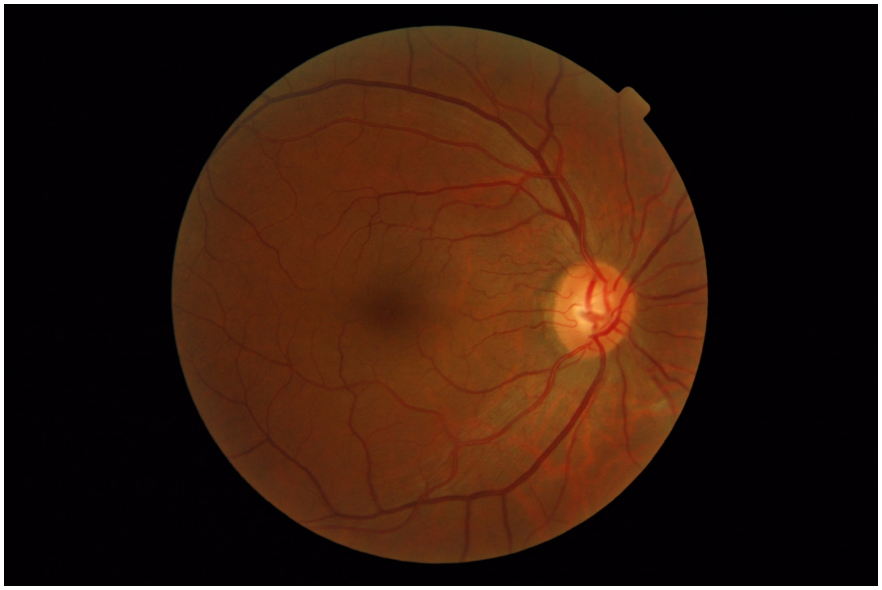

Figure 1: Normal Eye Fundus

circulation and weakening of smaller blood vessels throughout the body. This effect can often be seen in the blood vessels of the retina, where the circulatory symptoms of diabetes can lead to various signs of DR, such as: microaneurysms (swelling in blood vessels), exudates (mass of cells that have seeped out of a blood vessel), neovascularization (creation of new, weak blood vessels due to poor blood flow), edema (buildup of fluid), and hemorrhaging (internal bleeding). The above symptoms are roughly in order of increasing severity, and by the time neovascularization occurs, permanent damage to the retina is very likely. This stage of the disease is referred to as proliferative diabetic retinopathy (PDR), and often patients won't even begin to notice symptoms of the disease until long after PDR has set in. Figure 2 (Decencière et al. 2014) shows the fundus of a patient with severe PDR, which includes visible hemorrhaging along the lower part of the image.

Data

The dataset that's being used for this project is the Debrecen dataset (Antal and Hajdu 2014), which aims to develop a method for accurate diagnosis of DR using fundus images. The authors of this paper used images from the Messidor dataset (Decencière et al. 2014), which consists of 1200 fundus images from diabetic patients, for the use of DR screening. The authors of the Debrecen paper (named for the associated institution of the authors) used computer processing techniques to extract relevant features from the raw images, in the form of numerical data, for 1151 of the 1200 original images. Some of the original images were removed due to duplicate images and patients in the original dataset.

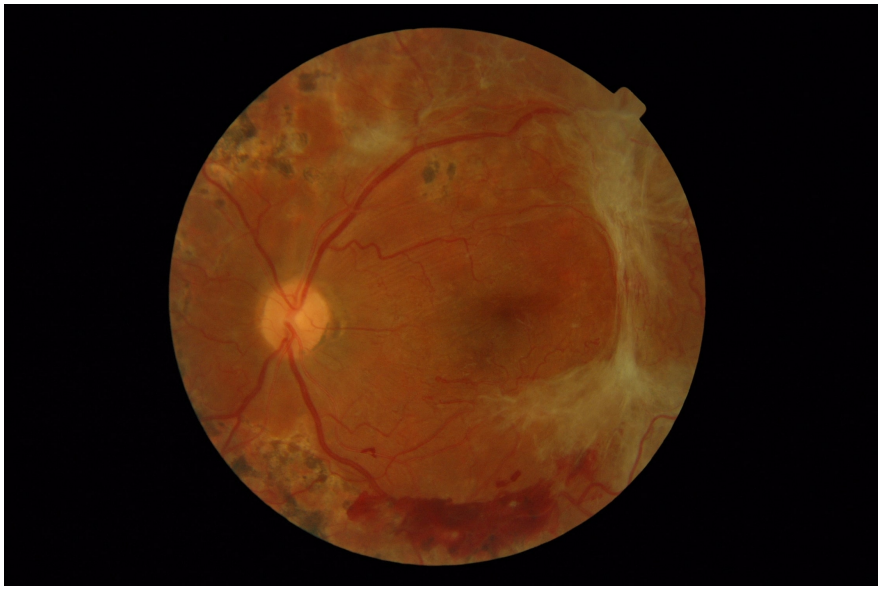

Figure 2: Eye Fundus with Severe PDR

The original Messidor data also contained a DR diagnosis outcome variable, as decided by a team of ophthalmologists, with 4 categories of increasing severity, from no DR to severe PDR. The Debrecen dataset recategorizes this variable into a binary Healthy/DR outcome. The predictors in the Debrecen dataset consist of: a prediction of DR status based on an AM/FM based feature extraction method (binary), a pre-screening indicator of severe retinal abnormality prior to fundus imaging (binary), the estimated distance between the center of the macula and the center of the optic disc in the image, the estimated diameter of the optic disc in the image, six variables estimating the count of microaneurysms in the image at different confidence levels, and eight variables estimating the count of exudates in the image at different confidence levels.

Modeling

Before any analyses were carried out, four observations were dropped due to being flagged by an indicator of poor image quality. Since these four observations were the only ones with this distinction, it was decided that the quality of the measurements extracted from these images would likely be poor, and could reduce prediction accuracy. Other than this, there weren't any missing values, extreme outliers, or questionable observations that were worth exploring any further.

As for the predictors, the two binary variables were left unchanged, as were the two continous variables, which appeared approximately normal. The microaneurysm and exudate variables were log transformed due to their being counts. This helped make the heavily skewed predictors much more symmetrical. The overall percentage of patients with DR in the dataset was about 53%, so it was decided that a cutoff of

0.5 would be used for classification based upon class probabilities. Finally, the data was split into training, validation, and test sets of size 60%, 20%, and 20%, respectively. The splits were made with respect to the outcome, so that each dataset would have an accurate amount of patients with and without DR, based on the overall proportion. The `createDataPartition` function from the `caret` package (Kuhn 2008) was used to accomplish this.

Penalized Logistic Regression

The first model trained was the parametric model, which was a penalized logistic regression model (PLR) using all 18 predictors. Since correlation was expected between the count variables, this method would allow for more stable estimation of the regression parameters. The `glmnet` package's implementation of elastic net regression was chosen for this model (Friedman et al. 2010). Elastic net is essentially a mixture between lasso and ridge regression, using the penalty parameter λ, with an additional tuning parameter α that represents the mixing proportion between the two methods. The `caret` package (Kuhn 2008) was used to train the model, using 5-fold cross-validation within the 60% training set to optimize the selection of α .

Cross-validation produced an α value of 1, which represents 100% lasso regression. After the α level was fixed at 1, the `glmnet` package's built in cross-validation procedure was used to optimize the choice of λ, which produced a value of approximately 0.0004. Finally, this combination of tuning parameters was used to fit the final model. Figure 3 shows the coefficients for differing values of λ, with the selected λ denoted by the vertical gray line. For this value of λ, none of the coefficient estimates were reduced completely to zero.

Generalized Additive Model

The second model trained was a semiparametric generalized additive model (GAM), based on the implementation of Simon Wood's `mgcv` package (2011). This specific implementation was chosen due to its flexibility with regard to the specification of the relationship between smoothed terms. In the `gam` package (Hastie 2019), for example, only one variable at a time can be entered into the `s()` function, which represents a cubic smoothing spline term. The `s()` function in the `mgcv` package, on the other hand, allows multiple variables to be entered into it together, which allows for the creation of a single term in a GAM model which represents the multivariate smooth of multiple predictors.

The reason this functionality was desired is because the microaneurysm and exudate variables were to be combined into two cubic spline smoothed terms, each representing the composite effect of either all 6 microaneurysm variables or all 8 exudate variables. However, given the size of the training set, the estimated degrees of freedom required to fit such high-dimensional smoothed terms was prohibitive. Technically

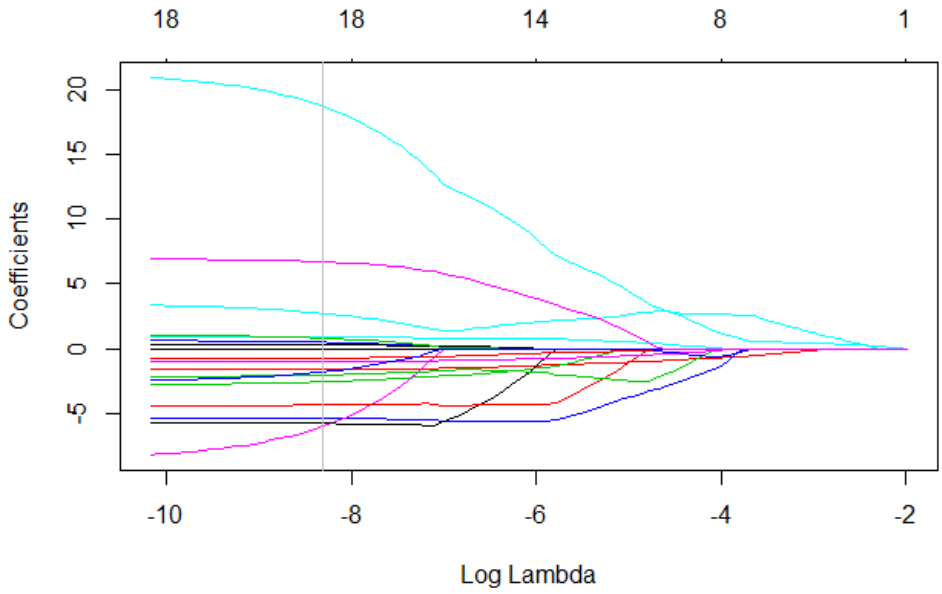

Figure 3: Lasso Plot

the computation could have been carried out, but the extremely low residual degrees of freedom left over suggested that this model would have been dramatically overfitting the training data, and this initial model was disposed of before any predictions were obtained from it.

Instead, each of the numerical predictors were entered into the GAM with their own independent cubic spline smoothed term (which means the `gam` package could have been used after all), while the binary variables were included linearly. Since GAMs do not have any tuning parameters of their own, and it was decided that the default cubic spline estimation technique was sufficient for the `s()` terms, no cross-validation was performed for this technique, and the specified model above was fit to the entire training set.

Extremely Randomized Trees

The third and final modeling technique used was an extremely randomized trees (ERT) model from the `extraTrees` package (Simm et al. 2014). ERT models are similar to random forest models (Liaw and Wiener 2002), except for two main distinctions: (1) ERT models do not use the bagging technique employed in random forests. Bagging in a random forest involves fitting each constituent tree to a random bootstrap sample of the original training data, instead of the whole training dataset. In ERT, all trees are trained on the entire training set. (2) Random forest trees select the best possible split at each decision node based on the predictors available to it, which is normally a small number of the entire predictor pool. ERT trees

also only allow each tree access to a small random sample of the predictor pool, but it only lets the tree split at one of a small random sample of all possible splits for the available predictors. So ERT trees are first given random splits, then they calculate the best split of that small sample of splits, instead of for all possible splits, like in random forests. The main practical difference in ERT and random forest models is that ERT tree tend to be less computationally intensive to train, and more variable. In terms of model accuracy, random forest models and ERT models tend to perform pretty similarly, though ERT models may perform better when there are very noisy variables included in the predictor pool.

Again the `caret` package (Kuhn 2008) was used to carry out 5-fold cross validation for selecting the various tuning parameters of this model. Four tuning parameters were selected, including the number of trees, the minimum node size within trees, the number of candidate predictors selected per tree, and the number of random splits selected per candidate predictor. The final parameter values used in the model training were 500 trees, with a minimum node size of 20, 5 candidate predictors per tree, and 2 random splits per predictor. This model was then fit to the full training set.

Ensembling

The first and most basic model ensembling technique used was just a simple average of the predictions of the three models. The predicted probabilities were used from each model, instead of just the class prediction. So, the result of this ensembling technique is essentially the average predicted probability across all three models that a given patient in the test set has DR. A slightly more sophisticated variation of this idea was also used, which was a weighted average of the three model's predicted probabilities. The weights chosen were based on the accuracy of each of the models in the 20% validation set. This way, the models that were more accurate at predicting new observations would have their predictions upweighted in relation to those that were less accurate.

The final model ensembling technique used was a model stacking algorithm. Model stacking involves training a "meta-model" to produce predictions based on the predictions of other models. In this case, a logistic regression model was used to predict the validation set DR outcome, using the predictions of the validation set from the three models as predictors. This way, the meta-model can be used to combine the predictions of future test data, based on parameters estimated from the validation data. This is ultimately fairly similar to taking a weighted average based on validation data accuracy, but isn't equivalent in general.

Prediction

The table below summarizes the prediction accuracy, sensitivity, and specificity of the three models and three ensembling techniques used in this project, where PLR, GAM, and ERT are used as before, and sAVE, wAVE, and GLM represent the simple model average, weighted model average, and logistic regression model stacking method, respectively. In general, there wasn't much difference between any of the methods presented. In terms of overall accuracy, the GAM model was the best by a little, the ERT model was the worst by a lot, and the other four models were very similar to one another. For specificity, the GAM model is also definitely the best, and the ERT is also the worst. However, in terms of sensitivity, the three ensemble models are all the best. Sensitivity is probably the most important of the three, given we're discussing a diagnostic tool, where detection of true DR patients is the ultimate goal, and a false positive isn't all that bad compared to a false negative. It's interesting that the weighted average and model stacking methods produced identical prediction in this test set. So, based on these results, either one of them would likely be the best choice here, since they have the highest sensitivity, and higher specificity than the simple average method.

	ACCU	*SENS*	*SPEC*
PLR	0.773	0.779	0.766
GAM	**0.782**	0.779	**0.785**
ERT	0.745	0.746	0.748
sAVE	0.773	**0.787**	0.757
wAVE	0.777	**0.787**	0.766
GLM	0.777	**0.787**	0.766

Discussion

Overall, I'm pretty satisfied with the results from this project. The original authors were able to get a model with 90% accuracy. And, while I couldn't break 80%, they definitely used more complex modeling, parameter tuning, and model ensembling techniques, whereas I stuck with relatively simple and computationally easy techniques. Given this project was just a learning tool, I'm definitely happy with how it went, seeing as I learned a lot about new modeling techniques, and issues surrounding the implementation of those modeling techniques. If I were to spend more time on this project, I would likely include more modeling techniques, as well as spend more time on the proper training of all the modeling techniques chose. I would

also look at more sophisticated ensembling techniques, since the ones I used were relatively straightforward, and simple to implement.

As for the public health impact of this project, I don't see my own research having much of an effect on anything. But, this area of research is a fairly popular one. There's a lot of discussion around ways to catch patients with early signs of DR before the damage becomes permanent. Machine learning techniques are being used for similar predictive tasks in many areas of medicine today. However, in the area of fundus image analysis, there is some pushback, since the difficulty isn't necessarily in interpreting the retinal images, since most ophthalmologists can do that themselves pretty quickly and accurately. The main issue is in getting diabetic patients to maintain regular eye exams to monitor the status of their retinal health, especially in patients in poorer and less developed areas. Having a model that can give you good prediction accuracy doesn't actually put doctors in communities where they need to be, and it certainly doesn't help patients afford the care they need to prevent and treat DR. Ultimately, better healthcare funding and infrastructure is needed to properly address this problem. And until that time, DR will continue to be a serious burden for long-term diabetic patients around the world.

References

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.

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.

Hastie, T. (2019), *gam: Generalized additive models*.

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.

Liaw, A., and Wiener, M. (2002), "Classification and regression by randomForest," *R News*, 2, 18–22.

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.

Wood, S. N. (2011), "Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric 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.