A Data Driven Technique for Diagnosing Retinal Dystrophies

Benjamin Katz, Vittorio Bichucher, Richmond Starbuck, Wei Xu, Jacob Durrah, Dana Schlegel, Thiran Jayasundera, Andrew DeOrio

University of Michigan College of Engineering and Kellogg Eye Center, Ann Arbor, MI

University of Michigan
Kellogg Eye Center

Background

Retinal dystrophies: are genetic conditions associated with reduced or deteriorating vision that may lead to blindness.

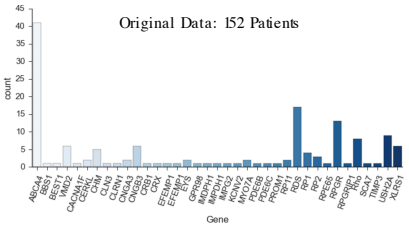
Current diagnosis technique: specialists test specific genes using gene-sequencing techniques for probable disease-causing variants.

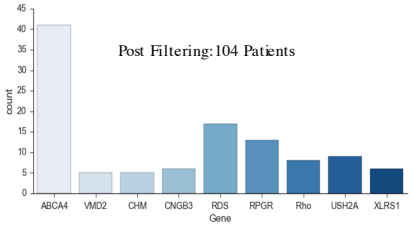
Problem: can be prohibitively expensive and require specialists to interpret results. Therefore, many patients lack a conclusive molecular diagnosis critical to providing proper treatment, as many therapies are gene specific.

Solution: A supervised model that predicts the likelihood of a particular disease-causing gene being mutated. This may help inform providers about appropriate genetic testing panels to order as well as assist ophthalmologists in analyzing inconclusive genetic testing results.

Data set

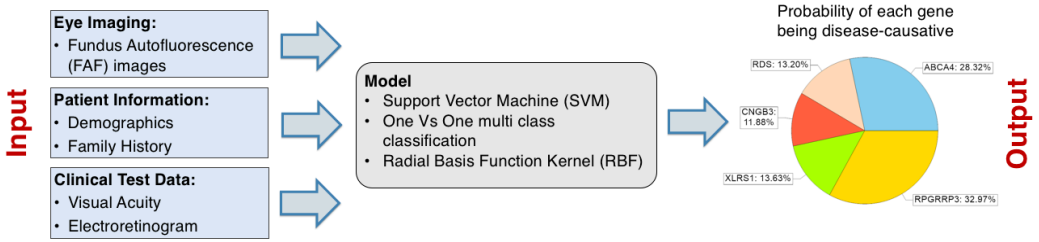
- Labeled data-set: mutated gene confirmed by specialist for each patient through genetic testing.
- Only patients with a single mutation previously reported to be disease causing were included.
- Genes with fewer than five occurrences in the data set were filtered out.





Prediction Model

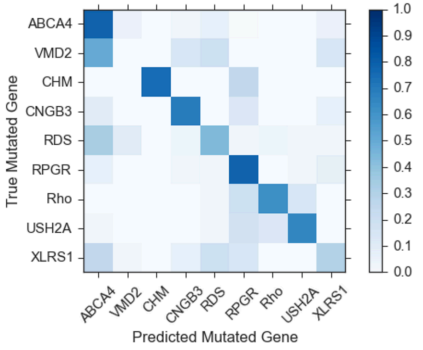
- Algorithms evaluated included: linear regression, adaboost, random forest, bagging with *k*-nearest neighbor, bagging with decision tree, and support vector machine with linear and radial basis function kernel (RBF SVM).
- Hyper parameters tuned using cross validation.
- RBF SVM was found to produce results with the lowest Brier score.
- Probabilities of each gene in RBF SVM calculated using "Probability Estimates for Multi-class Classification by Pairwise Coupling" by Wu, et al. 2004.



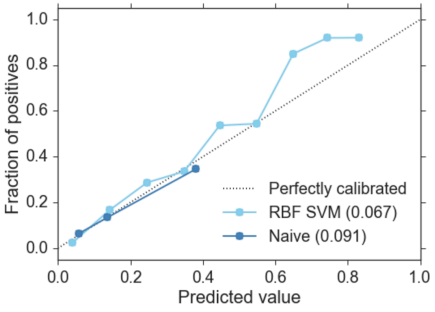
Experimental Results

Procedure

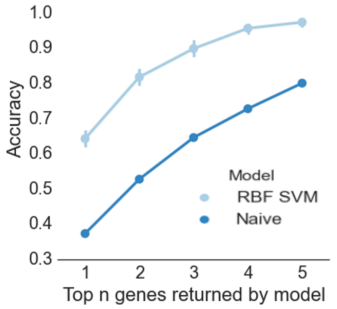
Results were obtained by randomly selecting 80% of the data to train the model and validating the results by testing on the remaining 20%. Both sets were stratified. This procedure was repeated 20 times, and results averaged over all trials.



Confusion matrix of top prediction for RBF SVM. Cells on the diagonal are correct predictions. Darker line on the diagonal indicates better discrimination between mutated genes.



Calibration plot and Brier score (in legend) of RBF SVM and naïve model. Lower score implies predictions can be more directly interpreted as a confidence level.



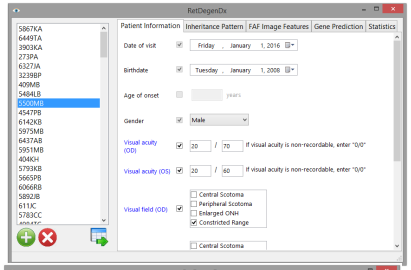
Probability that one of the top *n* genes returned by the model is the correct gene for RBF SVM and naïve model. Higher values imply more accurate relative ordering of returned probabilities.

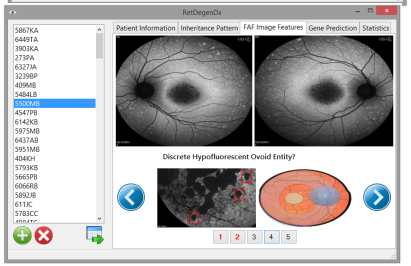
Results

- Results compared to naïve model that predicts the class priors from the training data.
- RBF SVM predicted the disease-causing mutations with lower Brier score than naïve model (P value < 0.0001).
- RBF SVM had a higher accuracy than the naïve model when considering the top n predictions returned by the model (P value < 0.0001).
- The majority of genes were predicted with greater than 50% accuracy by the RBF SVM's top prediction, despite the data set being highly imbalanced.

Prototype

- A prototype was developed to be utilized at the Kellogg Eye Center.
- Allows physicians to quickly enter in clinical data and receive predictions.
- Determines probable inheritance pattern through series of simple questions.
- Will facilitate data transfer between partner institutions.





Prototype deployed to Kellogg Eye Center for inputting clinical data and labeling retinal images. Output screen is shown in prediction model flow diagram.

Conclusion

- 90% of the time the true disease-causing gene is predicted as one of the top 3 model outputs.
- Model gives effective predictions for a majority of genes, despite small, unbalanced data set.

Future Work

- Collect more data to improve predictions and to include more genes in the model.
- Refine features extracted from FAF images.
- Improve features extracted from patient family history.

Acknowledgements

We would like to thank our sponsors at Kellogg Eye Center and the U-M Multidisciplinary Design Program for their generous support.