A Data-Driven Technique for Identifying Disease-Causing Genes in Patients with Retinal Dystrophy

1 Introduction and Related Work

Retinal dystrophies are genetic conditions associated with reduced or deteriorating vision that may lead to blindness. Current diagnosis techniques require a specialist to manually identify and test specific genes for probable disease-causing variants. While these techniques are frequently able to uncover causative gene mutations [1, 2], gene sequencing of known retinal dystrophy genes can be prohibitively expensive and require specialists to interpret the results [3]. As a result, many patients lack a conclusive genetic diagnosis, which are critical to providing proper treatment as many therapies are gene specific [4].

To address the high cost of gene sequencing and the small number of specialists able to interpret the results, we propose a data-driven statistical model of retinal dystrophies to provide physicians with actionable diagnostic data. The model is built from a database of patients who visited a specialized retinal dystrophy clinic at a major university research hospital. The data driven model was observed to significantly outperform a baseline.

2 Overview

Our model takes as input features collected from patients during clinical visits, such as demographic data (e.g. age, sex), clinical test data (e.g. electroretinogram, visual acuity, visual field), and inheritance pattern information based on a patient's family history (e.g. autosomal dominant, x-linked). The output of the model is several genes, together with the probability that each contains a disease-causing mutation.

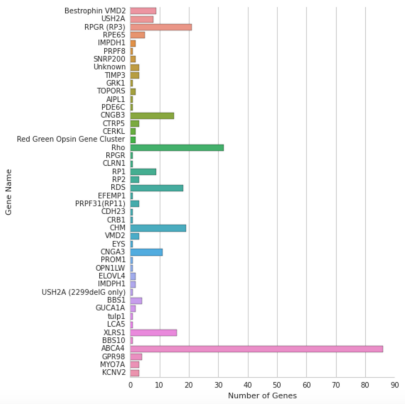

Figure 1: The distribution of causative mutated genes in the original data set.

The dataset used to build the model is composed of 509 retinal dystrophy patient records from a specialized retinal dystrophy clinic at a major research university hospital. Each patient record is labeled with a confirmed disease-causing mutated gene (ground truth), with a total of 42 unique genes. Preprocessing removes duplicate patient records, reduces the data set to 313 samples, and normalizing electroretinogram test results based on patient age and institution-specific equipment calibration.

Genes with fewer than five patients in the data set are removed to maintain the relative prevalence of each gene in the 80/20 training/testing data splits. Following the removal of these genes, there are 249 unique patient records with 12 distinct mutated genes.

One key challenge in building this model was the balance of genes (labels for our data). The prevalence of some labels relative to others, as illustrated in Figure 1, resulted in an imbalanced data set. For support vector machine (SVM) models, this was overcome by giving each class an error cost inversely proportional to its frequency.

The likelihood of each gene being mutated useful for physicians to assess ambiguous genetic sequencing results. Thus, algorithms able to output probabilities associated with each class are selected. We tested multiple algorithms, including SVM with linear and radial basis function (RBF) kernels using the one vs. one multi class approach, random forest, adaboost, bagging with a *k*-nearest neighbors classifier, and bagging with a decision tree classifier. All hyper parameters in these models are selected with 4-fold cross validation and grid search. Several imputation techniques are applied to compare

the classification performance, including imputation by median and mean in continuous and binarized values. Continuous values are binarized by bucketing values into logical groupings based on physiological similarity.

3 Results

In our experimental setup, 20 unique 80/20 training/testing splits of the data were created. Models were compared based on their logistic loss in order to determine which models associated the highest probabilities with the true mutated gene. Our results showed that an SVM with RBF kernel with imputation by mean and binarized model inputs performed best with regards to logistic loss, with a mean trial value of 1.507 and standard deviation of 0.079 across all testing folds. A baseline classifier is used for comparison, which outputs the frequency of each label within the training set. Compared to the baseline classifier's logistic loss of 23.325, the p-value of this result is < 0.0001. Figure 2 shows that when conducting classification, this model is able to discriminate reasonably well between several classes, but it predicts the class comprising a disproportionate amount of the data in many of the misclassifications. This, in addition to a lower mean logistic loss of 1.115 on training data for the linear SVM suggests that there exists a linear model which can better distinguish between these less frequent classes. This implies that through collecting additional training data, the more powerful RBF model should be capable of learning better predictions as well.

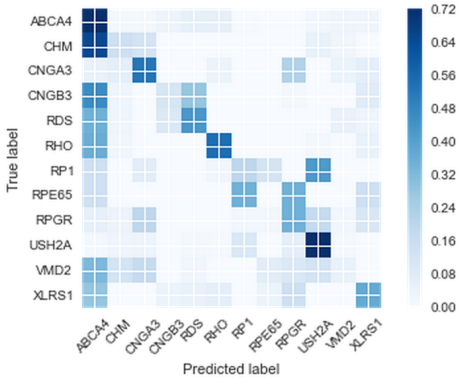

Figure 2. The True vs. Predicted label is shown for the RBF kernel SVM using the preprocessing technique of binarizing data values and imputation by mean, where darkness of color represents the percentage of times that particular class on the x axis was considered most probable when an instance of the class on the y axis was observed. Results along the diagonal are instances of correct classification.

4 Conclusion and Future Work

The significantly lower logistic loss values imply our machine learning model is able to provide probabilistic predictions that are substantially more accurate than a baseline. We found that using an SVM with RBF kernel yields the greatest performance according to this metric. Having shown through proof of concept that the ability to predict disease-causative mutated genes in retinal dystrophy patients is learnable, future work will include increasing the size of the current data set. Additionally, features from specialized retinal tests such as fundus autofluorescence will be added to the model, as domain experts correlate these features with retinal dystrophy diagnosis.

4 References

[1] Zernant, Jana, et al. "Analysis of the ABCA4 genomic locus in Stargardt disease." Human molecular genetics 23.25 (2014): 6797-6806.

[2] Flanagan, Sarah E., et al. "Next-generation sequencing reveals deep intronic cryptic ABCC8 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation." The American Journal of Human Genetics 92.1 (2013): 131-136.

[3] Trzupek, Karmen. "The current status of molecular diagnosis of inherited retinal dystrophies." Current opinion in ophthalmology 26.5 (2015): 346-351.

[4] Koenekoop, Robert K., et al. "Genetic testing for retinal dystrophies and dysfunctions: benefits, dilemmas and solutions." Clinical & experimental ophthalmology 35.5 (2007): 473-485.