Exploration in Phenotype Clustering in High Dimensional Genomic Data

Ross Lawrence

1 Abstract Machine learning is poised as an analysis method uniquely positioned to discover hidden interactions in genomic data through its ability to highlight distinctions in datasets too large for manual analysis. However, the high feature to sample ratio inherent in training machine learning models on SNP data is an obstacle to these analyses. Previous studies have shown some success in the use of feature reduction methods, such as Principal Component Analysis (PCA), Isomap Embedding (ISO), and Random Feature Extraction (RFE). The success of these dimensional reduction techniques display the potential for there to be a lower-dimensional relationship between individuals with similar variants which cause a given phenotype to occur. In this study, I implement several feature reduction methods with SNP data from individuals with distinct phenotypes in an attempt to explore the relationships of phenotypes which influence different systems in the human body. Support vector machines (SVM), k-Nearest Neighbor (KNN), and Random Forest (RF) classification methods are utilized to determine the extent to which the feature-reduced phenotypes are clustered together, as each corresponds to the relative location of data points. The classifiers were evaluated using the following metrics: accuracy, precision, recall and area under the curve (AUC). Combinations of feature reduction methods and classifiers failed to achieve accuracy scores significantly above random chance, aside from combination of RFE and SVM, which trained an SVM on the 168 repeatedly most influential SNPs in the dataset. This combination resulted in an AUC for the binary classifier for astigmatism, ADHD, and lactose intolerance of 0.73, 0.80, and 0.76, respectively. When this model was used to predict data for samples with phenotypes of myopia or irritable bowel syndrome, the binary classifier for myopia had an AUC of 0.89. This significantly above random chance performance of the astigmatism SVM model to classify myopia is far from conclusive, but potentially displays some relationship between the two phenotypes which involve the eye.

2 Introduction Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation, being the variation of a single base pair in a genome. Study of these SNPs has allowed scientists to investigate correlations between genetic makeup and observed phenotypes, such as diseases, response to particular drugs, and physical attributes. In general, biological modeling studies for disease risk assessment using genetic information have been supplemented by several technologies and study designs. For example the genome-wide association studies (GWAS) use the prevalence of variants in different SNPs to determine their involvement in observed phenotypes [1]. Machine learning algorithms have also been applied to SNP data in order to classify particular phenotypes in samples. However, due to the curse of dimensionality inherent in using SNP data (i.e. the larger the dimension of a dataset is relative to the number of samples, the poorer many machine learning models perform)[2], other information is often required to preprocess the SNP data before training a model (literature, GWAS, etc.). Some studies have shown success in improving raw SNP data classification through treating an individual's SNP data as a vector and utilizing various machine learning algorithms in order to reduce the number of features/dimensions before classification. SVM, along with decision trees, have also been used to identify breast cancer cases using SNPs selected via information gain [3]. The k-Nearest Neighbors (KNN) method has also been employed to quantify SNP relevance and to perform classification task on breast cancer data [4]. This success of these modeling methods, which utilize distance calculations between data points with the same categorization, could hint that there exists some form of clustering that occurs between phenotypes, potentially between those that are "similar" in presentation. For example, are data points represented by the vector of the SNPs of individuals with astigmatism or myopia, which both effect the eyes, "closer" to each other than those with astigmatism and lactose intolerance?

Would these “closer” phenotypes be harder to distinguish from more “different” phenotypes? If this is true, then classifier algorithms which operate through relative distance, such as SVM and KNN would have improved success in classifying seemingly “dissimilar” phenotypes.

3 Methods

3.1 Data Collection and Preprocessing The SNP data used in this experiment was obtained from the openSNP database[5], a public repository for genome sequencing files. Only 23andMe files which were generated using the human genome assembly GRCh37 (build 37) were selected in order to keep the SNPs which were reported as consistent as possible. Each text file contains meta-information lines and a header line with the following field names: SNP ID, chromosome number, position, and genotype. The SNP data from individuals with only one of the following phenotypes were selected: astigmatism (N=91), attention deficit hyperactive disorder (ADHD) (N=88),irritable bowel syndrome (IBS) (N=54), lactose intolerance (N=82), and myopia (N=76). These phenotypes were selected due to the systems in the human body that they effect. Astigmatism and myopia are phenotypes which effect the eyes [6][7], IBS and lactose intolerance effect the digestive system [8][9], and ADHD is a neurological phenotype[10]. If there were to be any truth to the clustering of phenotypes which affect similar systems of the body, then the proper classification of the phenotypes for different systems would be easier to perform. These phenotypes were also selected due to the quantity of individuals on openSNP who stated that they had moderate to severe symptoms.

Once the data had been collected from openSNP, it was preprocessed using a similar quality control procedure from Hijaloo et.al. The frequency of different SNPs in the datasets were tabulated, along with their alleles. Any SNP that appeared in less than 95% of the dataset was omitted, along with the SNPs associated with the X and Y chromosomes (done to further standardize the genomic data between individuals). This resulted in 45,383 remaining SNPs being taken from each 23andMe file for analysis. Finally, the alleles of the remaining SNPs were encoded based on their allele structure. Using the tabulated alleles the major and minor alleles for each SNP were selected. Each sample file was then encoded, with homozygous (2 major alleles), heterozygous (1 major, 1 minor allele), and variant homozygous (2 minor alleles) SNPs being represented by 1, 2, and 3, respectively.

3.2 Feature Selection Several methods were used in an attempt to mitigate the effect of the curse of high dimensionality by reducing the number of dimensions from the original 45,383 SNPs. Even though the number of SNPs had been reduced from upwards of 650,000 down to 45,383 for each sample, this number of features was still significantly too large relative to the number of samples (261 samples total for astigmatism, ADHD, and lactose intolerance). Several different methods for feature/dimension reduction were selected for testing, with the criteria being their capability in handling high dimensionality.

Random Feature Extraction (RFE) Random feature extraction is a method by which a classifier model is recursively fit to a training set and used to calculate the importance each feature has to the proper classification of the training set. Each iteration, the n features with the lowest performance are omitted and the model is trained again. This process is repeated until the features which maximize the accuracy of the model is obtained. In this study, random forest was chosen as the classification model used.

Random forests (RF) is a supervised learning algorithm that consists of an ensemble of decision trees (hence it being a *forest*). Random forests are bagged decision tree models that split up a subset of features on each split in the tree. Each split in a decision tree is made with the intent to maximize the "purity", or the percentage of each classification group that results from the split. This method for feature reduction was chosen due to its ability to handle high dimensionality due to the fact that each decision tree is only given a subset of data to train/test on. It also had the capability of ranking the importance of a given feature in terms of classification, as that feature would more commonly result in significant increases in purity across the decision trees[11]. Using scikit-learn’s RFECV function [12] in combination with the RandomForestClassifier [12] method, a 10-fold feature reduction on the SNP

data of the astigmatism, ADHD, and lactose intolerance datasets were run using default parameters for those not specified in Table 1.

Principal Component Analysis Principle Component Analysis is commonly used for dimensionality reduction of higher dimensional data. This reduction is performed by first calculating the principal components, which are the series of unit vectors where the i^{th} vector is the direction of a line that fits the data while being orthogonal to the first $i-1$ vectors. By taking the values of a vector through which the averaged squared distance from the points to the line is minimized, each data point can be projected onto only the first n principal components (with principal components being ordered by how much of the data's variation they preserve). This allows for the creation of a lower-dimensional dataset while preserving as much of the variation in the data as possible [13].
scikit-learn's PCA function [12] was used with default parameters, aside from those specified in Table 1, in order to find the optimal parameters for the reduction of the astigmatism, ADHD, and lactose intolerance datasets. A limitation of less than 100 principal components being used for analysis was enacted due to the amount of variance being preserved by all components above 100 being less than 0.01%.

Isomap Isomap is another widely used method for dimensional reduction. It is one of the isometric mapping methods which extends metric multidimensional scaling (MDS) by incorporating the geodesic distances from weighted graph. While the theory behind this method of dimensional reduction is complicated, a simplified explanation is that the traditional scaling of metric MDS performs low-dimensional embedding based on the Euclidean distance between data points, trying to preserve the relative distances between each point[14]. Isomap, on the other hand, utilizes the geodesic distance induced by a neighborhood graph embedded. In other words if each data point is a node, instead of looking at the Euclidean distance between nodes, Isomap looks at shortest path between two nodes while traveling from node to node.[15] As with the other feature/dimensional reduction methods, scikit-learn's implementation of Isomap [12] was utilized using default parameters, aside from those specified in Table 1, in order to find the optimal parameters for each modeling method.

3.3 Classifier Models

In order to determine whether there was a marked difference between phenotypes, classifiers were used. Using the metric of prediction accuracy, SVM, KNN, and RF classifiers could be utilized as stand-ins for clustering. Meaning, if SVM and KNN (which measure relative location between data points) were able to successfully classify data based on phenotype, there would be grounds for claiming that there exists some lower-dimensional clustering.

Support-Vector Machines (SVM) Support-Vector Machines are a supervised learning model that can be used for classification. They perform classification by, with data points categorized into two groups, attempting to find a hyperplane that separates the classes with the highest possible margin [16]. SVM was selected as a potential classifier due to its ability to perform optimally on higher dimensional data, especially when using the "kernel trick". The kernel is a function that takes the original non-linear data and transforms it into linear data within a higher-dimensional space. A common kernel for the very high-dimensional data relative to number of samples is the radial basis function (rbf) kernel, which was selected for this analysis.

scikit-learn's SVC function [12] was used for the SVM classification process with default parameters selected, except for those specified in Table 1. In order to train the SVM classification method to a datasets with multiple categories, the OneVsRest classification method was selected. This method fit *one* SVM classifier for each phenotype while treating the other two phenotypes (rest) as the same category. The resulting collection of SVM classifiers were then used to vote on the classification of a given sample (i.e. each classifier determined whether it was their phenotype or not, with ties being viewed as incorrect classification).

3.4 K-Nearest Neighbors (KNN) K-Nearest Neighbors is a classification method which categorizes data points based on the category of the closest k training data points. Given a new data point, it calculates the k closest training data points, using either euclidean or minkowski distances (depending on the specified p value), and then assigns the new data point a category based on the which is in the majority [17]. The success of this classifier correlates with the extent to which different categories (i.e. phenotypes) are clustered together or set apart. scikit-learn's KNeighborsClassifier function [12] was used for the KNN classification process with default parameters selected, except for those speicifed in Table 2.

3.5 Feature and Model Combinations In order to maximize the distinctness of each phenotype, each of the feature selection methods was paired with a compatible classifier and tested with the optimal parameters. These combinations of methods were created through the use of scikit-learn's Pipeline functionality [12], which allowed for the output of the feature reducers to be fed directly into the classifier models and trained using 10-fold cross validation. These parameters were found through the use of scikit-learn's GridSearchCV function [12], which iterated through ranges of parameters (Table 1) for the pipelines until it found the parameters which maximized the multinomial classification accuracy for the astigmatism, ADHD, and lactose intolerance datasets. The optimal parameters for each combination can be found in Table 3. The receiver operating characteristic (ROC) curve was then graphed for each of the combinations using the optimal parameters and the area under curve (AUC) was calculated for each of the binary classifiers. In order to determine the multinomial classification capabilities of each combination, the macro-average and micro-average ROC curves were also plotted.

Method	Parameters
PCA	Number of components = 10 to 100 by steps of 5
ISO	Number of neighbors = 2 to 10 by steps of 1 Number of components = 10 to 100 by steps of 5
RFE	Number of trees = 100 to 1000 by steps of 100 step size = 50
SVM	C = 1 to 100 by steps of 10 gamma = 20 evenly spaced numbers from 0.0001 to 0.01
KNN	Number of neighbors 2 to 10 by steps of 1 p = 10 evenly spaced numbers from 1 to 2
RF	Number of trees = 100 to 900 by steps of 50 Maximum number of features given to a tree = 0.1 to 0.99 by steps of 0.1

Table 1: **GridSearchCV Parameter Selection** The ranges of parameters and the step size iterated over by GridSearch in order to find the parameters which resulted in the highest multinomial accuracy score.

3.6 Similar Phenotype Testing The combinations that were able to achieve classification accuracy significantly above random-chance with the astigmatism, ADHD, and lactose intolerance phenotype data had their classification performance tested using the "similar" phenotypes. The trained models were then tested on the myopia, ADHD, and IBS phenotypes, with the number of ADHD samples being reduced to 70 in order to minimize the effect of differences in sample size on performance measurements. ADHD was not removed from the testing dataset due to its ability to serve as a reference point for classification with the other two phenotypes. If its classifier's performance severely suffered, then it could be hypothesized that there was overlap in the positioning of ADHD data relative to myopia and IBS data.

4 Results After optimizing the model parameters (Table 3) in order to obtain the best multinomial accuracy scores for each combination of feature reduction and classifier, only one combination was capable of achieving an accuracy significantly above random-chance of approximately 33% (Table 2).

With an accuracy, precision, and recall score of 60.8%, 83%, and 63.8% respectively, the combination of RFE and SVM performed the best out of all model pairs.

	PCA	ISO	RFE
SVM	Accuracy = 0.202	0.289	0.608
	Precision = 0.4	0.485	0.830
	Recall = 0.202	0.290	0.638
KNN	0.326	0.152	0.370
	0.483	0.368	0.773
	0.326	0.152	0.369
RF	0.369	0.231	0.413
	0.708	0.692	0.792
	0.369	0.196	0.413

Table 2: **Accuracy, Precision, and Recall Scores** Multinomial classification accuracy, precision, and recall scores for each of the feature reduction and classifier combinations.

4.1 RFE SNP Results The RFE method, when trained on the astigmatism, ADHD, and lactose intolerance dataset using 10-fold cross validation, continually determined 168 SNPs to be the most influential (Table 4). While the number of most important SNPs varied with number of trees, these 168 appeared in every running of RFE on the data. Due to their prevalence, regardless of RFE parameters, they were isolated and used for training with the classification models. This series of SNPs resulted in the two combinations of feature reduction and classifier that achieved an AUC greater than 0.7. Of note was the fact that none of these SNPs appear to be related to either of the three phenotypes processed by the RFE.

4.2 SVM Performance When paired with PCA or Isomap, the SVM classifier method fails to have any of the binomial classifiers perform significantly above chance (Figure 1). The multinomial classification accuracy for the models using PCA or Isomap also do not significantly outperform random chance, with scores of 0.202 and 0.289, respectively (random chance being approximately 0.3 due to three phenotypes with similar sample size). However, when using the 168 SNPs determined by the RFE method, each of the binomial classifiers has an AUC of above 0.7, with the classifier for ADHD being greater than 0.8. The multinomial classification accuracy is also 0.608 when using this combination.

4.3 KNN Performance Similar to the SVM classifier, when paired with the either the PCA or Isomap feature reduction methods, the KNN classifier fails to achieve a classification performance significantly above random chance (Figure 2). The accuracy for the models using PCA and Isomap are 0.326 and 0.152, respectively. As with the SVM classifier, the 168 common SNPs determined from the RFE method did result in one of the binary classifiers performing significantly above random chance with an AUC above 0.7. However, this was the binary classifier for astigmatism and it, along with the other two classifiers, varied significantly based on which of the 10-fold portion of the dataset was being analyzed.

4.4 Random Forest Performance None of the combinations of feature reduction and Random Forest classifier resulted in a classification accuracy significantly above random chance (0.369 for PCA, 0.231 for ISO, and 0.413 for RFE). This pattern continued with regards to the binary classifiers, with none of them achieving an AUC score above 0.7 (Figure 3). The RFE model which used the RFE SNPs did achieve a higher accuracy than the other two methods, along with higher AUC values.

4.5 Generalized Models When tested with the "similar" phenotypes, only the SVM model using the SNPs from the RFE method performed significantly above random chance for all three phenotypes, as can be seen with the bottom plot of Figure 1. As the only model to perform this well, the generalizability of the model with "similar" phenotypes was tested. The model trained on astigmatism, ADHD, and

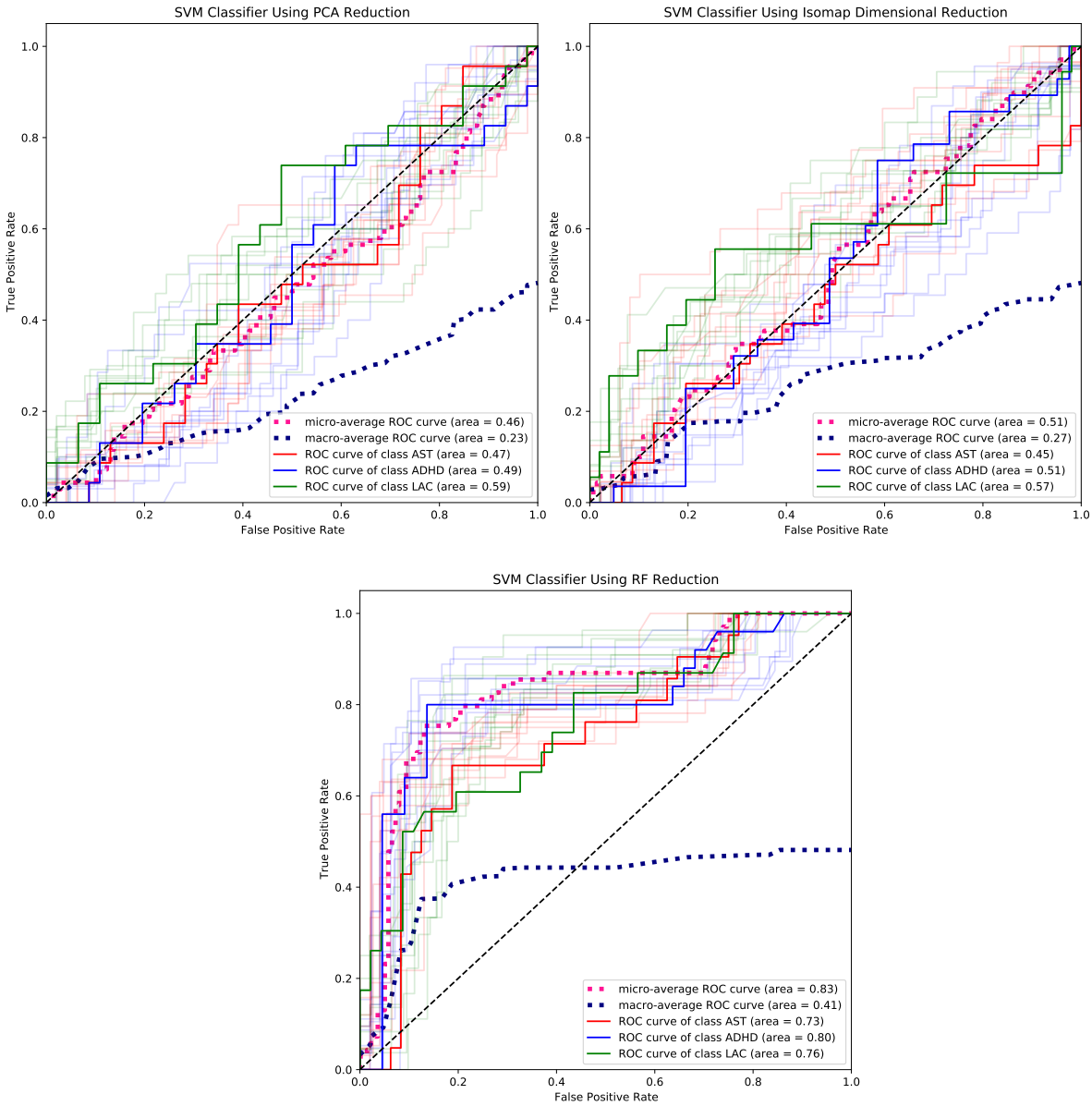

Figure 1: **SVM Classifier ROC Curves:** ROC curves of the optimal parameters for each combination of feature/dimension reduction technique and the SVM classifier using 10-fold cross-validation. The ROC curves denote the binary classifiers for each phenotype (yes or no to whether a sample is their particular phenotype), while the macro-average denotes the performance of all the binary classifiers voting on the multinomial classification of a sample. Of note is the performance of the classifiers trained using the RFE SNPs (Bottom), which is significantly above random chance for all three classifiers. *Note: AST = astigmatism, LAC = lactose intolerance.*

lactose intolerance phenotype data was tested on the myopia and IBS data (only using 70 samples of ADHD in order to minimize imbalance in classification). The multinomial model had an accuracy, precision, and recall rate of 0.45, 0.61, and 0.46, respectively, with the classification of myopia using the astigmatism classifier having an AUC greater than 0.8.

5 Discussion The results from this study go against the hypothesis that there exists some type of phenotype clustering based on system of the body that it effects. With the poor performance of the KNN

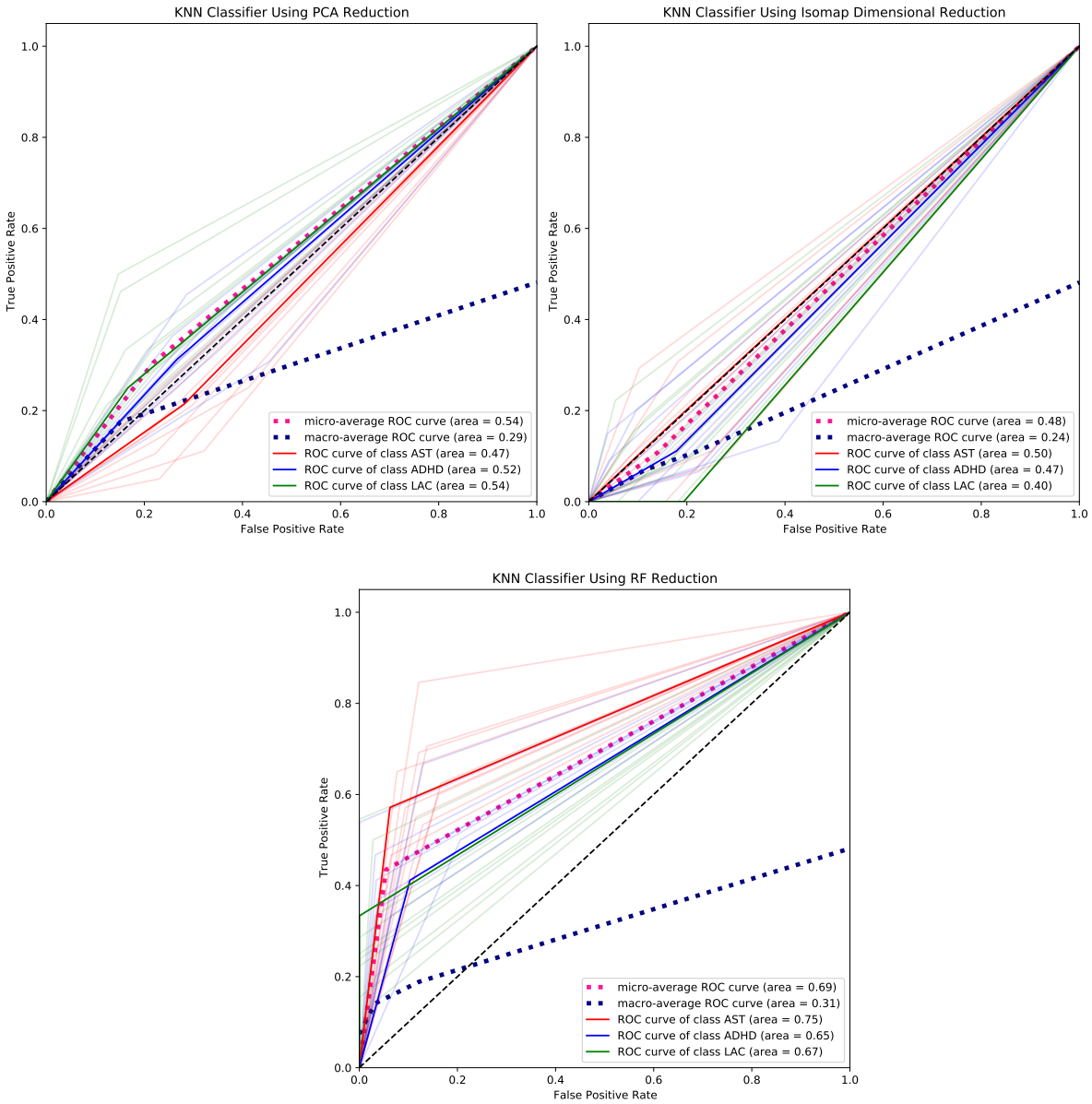

Figure 2: **K-Nearest Neighbors Classifier ROC Curves:** ROC curves of the optimal parameters for each combination of feature/dimension reduction technique and the k-Nearest Neighbors classifier using 10-fold cross-validation. The ROC curves denote the binary classifiers for each phenotype (yes or no to whether a sample is their particular phenotype), while the macro-average denotes the performance of all the binary classifiers voting on the multinomial classification of a sample. While significantly above random chance, the high variability in performance between folds for the astigmatism classifier for the KNN trained on RFE SNPs (Bottom) raises doubts about its accuracy. *Note: AST = astigmatism, LAC = lactose intolerance.*

models, as well as the SVM models when using common feature reduction methods, the clustering phenomena of phenotypes which are related to the same biological system seems unlikely. There are concerns about the performance of the SVM model using the 168 RFE SNPs, as none of the SNPs are known to be related to any of the phenotypes. The success witnessed in the application of the classifiers on the myopia data and IBS data may be due to the SNPs being selected correlating with some other unknown commonality between the individuals in either group. For example, both groups

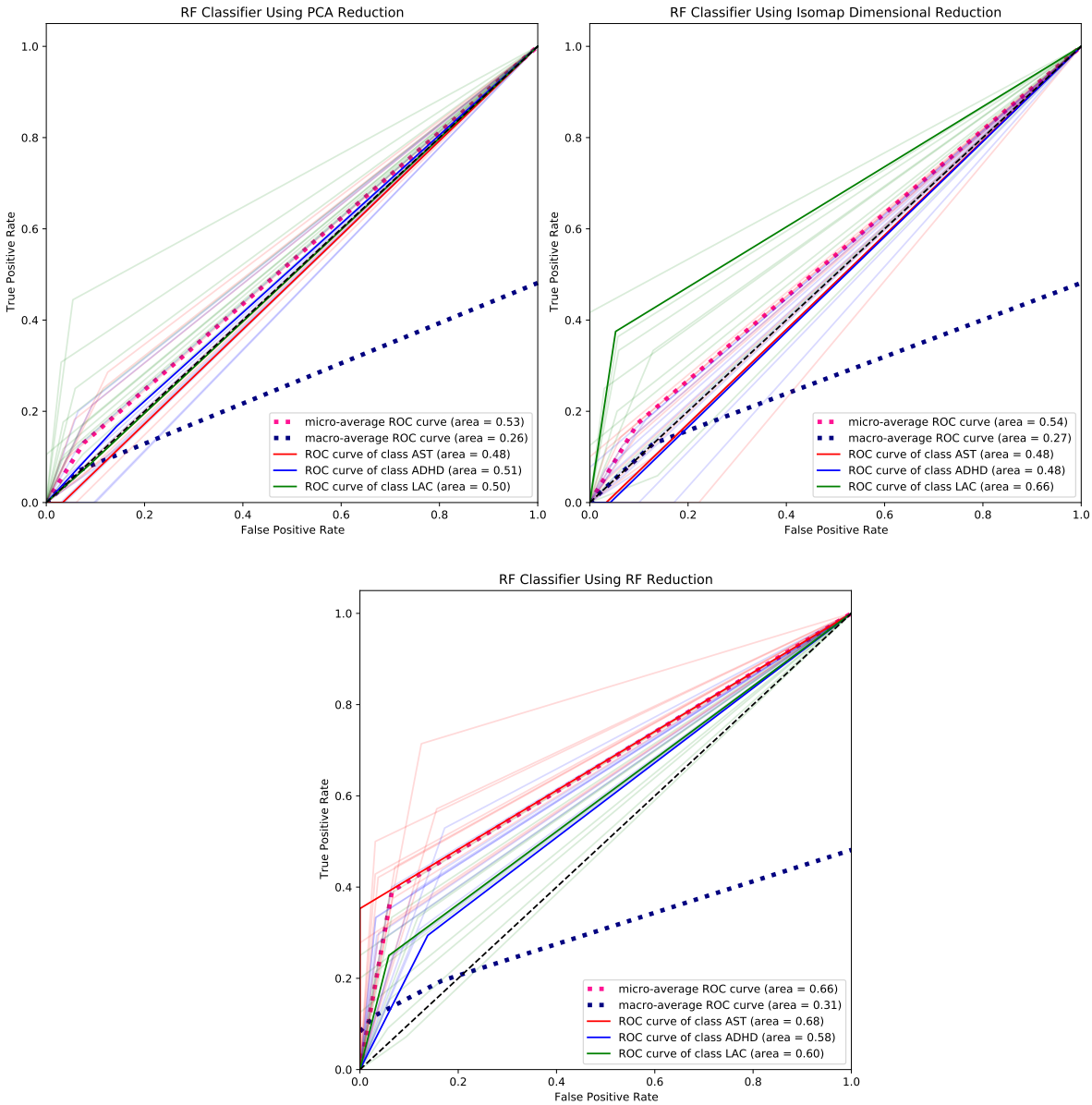

Figure 3: **Random Forest Classifier ROC Curves:** ROC curves of the optimal parameters for each combination of feature/dimension reduction technique and the Random Forest classifier using 10-fold cross-validation. The ROC curves denote the binary classifiers for each phenotype (yes or no to whether a sample is their particular phenotype), while the macro-average denotes the performance of all the binary classifiers voting on the multinomial classification of a sample. *Note: AST = astigmatism, LAC = lactose intolerance.*

may have a higher proportion of individuals of a certain ethnicity or physical attribute. Significantly more data is required before any confident claims can be made. Due to the poor performance of the classifiers, I was unable to use the feature reduced data to further explore the relationships between "similar" phenotypes as vectors.

There are several potential sources of error in this study, which draw several conclusions into doubt. First is the drastic reduction in SNPs during preprocessing due to the use of a 95% prevalence criteria. With this criteria, there was an approximate 90% decrease in the potential number of SNPs to analyzes,

	PCA	ISO	RFE
SVM	number of components = 10 C = 91 gamma = 0.0001	number of components = 10 number of neighbors = 2 p = 2 C = 81 gamma = 0.0001	number of SNPs = 168 C = 91 gamma = 0.0001
KNN	number of components = 15 number of neighbors = 3 p = 1.6	(Iso) number of components = 30 (Iso) number of neighbors = 2 (Iso) p = 1.7 number of neighbors = 2 p = 1.7	number of SNPs = 168 number of neighbors = 6 p = 1.5
RF	number of components = 70 max features = 0.47 number of trees = 400	number of components = 70 number of neighbors = 6 max features = 1 number of trees = 200	number of SNPs = 168 number of trees = 370 max features = 0.2

Table 3: **Optimal Parameter Selection** Optimal parameters for each combination of feature reduction method and classifier, as determined by 10-fold cross validation using scikit-learn's GridSearchCV function to iterate over the parameters described in Table 1.

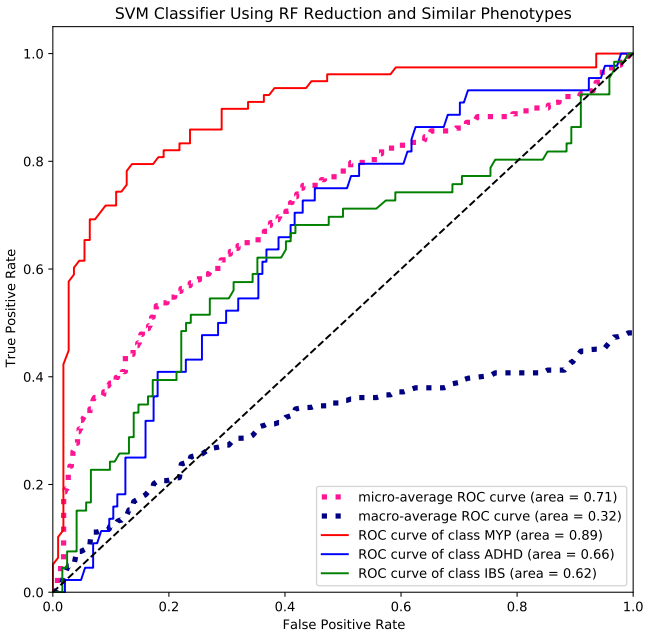

Figure 4: **Generalized Model:** ROC curve of the SVM model trained on the RFE SNPs for astigmatism, ADHD, and lactose intolerance attempting to classify myopia and IBS datasets. The ROC curves denote the binary classifiers for each phenotype, while the macro-average denotes the performance of all the binary classifiers voting on the multinomial classification of a sample. *Note: MYP = myopia*

from upwards of 650,000 in a few samples down to 45,383. It is likely that the information lost during this process has an impact on the feature reduction methods utilized in this study. Second, the number of samples used for each phenotype is very small relative to the number of features, making overcoming

the curse of dimensionality more difficult. The collection of samples might not be diverse enough to overcome unrelated correlations between individuals, such as ethnicity and height. Third, the use of only one feature reduction method in combination with a classifier limited how accurately the data could have its number of features reduced. There may exist a combination of the methods used in this study that would be more effective, but due to time and computational limitations, I was unable to test.

There are several paths for future investigation regarding the clustering of phenotypes. Additional testing of the SVM and KNN astigmatism classifier on other phenotypes involving the eyes should be performed in order to see if this generalizibility of modeling extends to further than just myopia. If such a trend could be observed with both the SVM and KNN models using the SNPs collected from RFE, it is possible that there exists an undiscovered relationship between the 168 SNPs and eye-related phenotypes. Further investigation in combining multiple methods of feature reduction aside from those covered in this study should also be done, including the expansion of which models are used. An expansion of the parameters cover by the grid search can also be performed, as there may be optimal parameters just outside of the ranges that were tested.

References

[1] M.C. Lopes, P.G. Hysi, V.J. Verhoeven, S. Macgregor, A.W. Hewitt, G.W. Montgomery, P. Cumberland, J.R. Vingerling, T.L. Young, C.M. van Duijn, B. Oostra, A.G. Uitterlinden, J.S. Rahi, D.A. Mackey, C.C. Klaver, T. Andrew, and C.J. Hammond. Identification of a candidate gene for astigmatism. Investigative ophthalmology visual science, 54:1260–1267, 2013. doi: 10.1167/iovs.12-10463.

[2] M Xu, KG Tantisira, A Wu, AA Litonjua, JH Chu, BE Himes, A Damask, and ST Weiss. Genome wide association study to predict severe asthma exacerbations in children using random forests classifiers. 2011. doi: 10.1186/1471-2350-12-90.

[3] J Listgarten, S Damaraju, B Poulin, L Cook, J Dufour, and A Driga. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical Cancer Research, pages 2725–2737, 2004.

[4] M Hajiloo, B Damavandi, M HooshSadat, F Sangi, JR Mackey, and CE Cass. Breast cancer prediction using genome wide single nucleotide polymorphism data. BMC bioinformatics, 14, 2013.

[5] B Greshake, PE Bayer, H Rausch, and J Reda. opensnp - a crowdsourced web resource for personal genomics. PLoS ONE, 9, 2014. doi: 10.1371/journal.pone.0089204.

[6] Scott A Read, Michael J Collins, and Leo G Carney. A review of astigmatism and its possible genesis. Clinical and Experimental Optometry, 90(1):5–19, 2007. doi: https://doi.org/10.1111/j.1444-0938.2007.00112.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1444-0938.2007.00112.x.

[7] D Fredrick. Myopia. BMJ, pages 1195–1199, 2002. doi: 10.1136/bmj.324.7347.1195.

[8] H Vahedi, R Ansari, M Mir-Nasseri, and E Jafari. Irritable bowel syndrome: a review article. Middle East Journal of Digestive Diseases, 2:66–77, 2008.

[9] Y Deng, B Misselwitz, N Dai, and M Fox. Lactose intolerance in adults: Biological mechanism and dietary management. Nutrients, 7:8020–8035, 2015. doi: 10.3390/nu7095380.

[10] A Singh, CJ Yeh, N Verma, and AK Das. Overview of attention deficit hyperactivity disorder in young children. Health Psychol Res., 2:2115, 2015. doi: 10.4081/hpr.2015.2115.

[11] A. Liaw and M. Wiener. Classification and regression by randomforest. R News, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[13] I.T. Jolliffe. Principle Component Analysis. New York: Springer-Verlag, 1986. doi: 10.1007/978-1-4757-1904-8_8.

[14] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2323, 2000.

[15] J.B. Tenenbaum and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

[16] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In ADVANCES IN LARGE MARGIN CLASSIFIERS, pages 61–74. MIT Press, 1999.

[17] N.S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3): 175–185, 1992. doi: 10.1080/00031305.1992.10475879. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879.

rs3751757	rs2580765	rs6430130	rs9835915	rs7023856	rs2233230	rs12201073
rs11615025	rs10498951	rs2694767	rs4148211	rs2566539	rs1851780	rs11775334
rs4646312	rs10800925	rs2328510	rs8109485	rs6792309	rs11120288	rs10185490
rs1421085	rs925042	rs2142090	rs3131888	rs3751812	rs9862002	rs8031322
rs9456841	rs561894	rs725745	rs2261712	rs2222186	rs9283527	rs2811425
rs15282	rs3916504	rs12927295	rs2417038	rs7151749	rs4601174	rs202963
rs338389	rs12154449	rs4774381	rs3786343	rs4738723	rs7520590	rs6736980
rs7713008	rs883616	rs3825725	rs2841277	rs820841	rs7542375	rs6704322
rs7590084	rs17186260	rs2700367	rs721936	rs7984972	rs6490188	rs749873
rs10768132	rs7542665	rs9556844	rs6465932	rs2062882	rs1447550	rs11174449
rs10245185	rs707482	rs6461202	rs1553114	rs7713338	rs439327	rs1893650
rs2966604	rs1460534	rs6935980	rs783258	rs7308671	rs10821808	rs4333997
rs4549077	rs7630698	rs726679	rs2928715	rs9787573	rs11105221	rs733694
rs6043738	rs17450685	rs3924917	rs522821	rs1800792	rs2627043	rs3924119
rs11587500	rs7358426	rs11072463	rs7717132	rs3732379	rs2322659	rs17212220
rs7998513	rs1880260	rs7580730	rs2622781	rs10211191	rs6798015	rs764855
rs12414155	rs7559985	rs4814441	rs9297594	rs4682644	rs11204328	rs4956987
rs734840	rs9889374	rs592048	rs11711838	rs1549765	rs407221	rs16925478
rs17239489	rs9461924	rs11634023	rs10946675	rs9373551	rs9512448	rs12883308
rs9321987	rs899842	rs12575710	rs8033470	rs1374111	rs4886321	rs7858174
rs7792246	rs737633	rs11143937	rs4497870	rs6864345	rs3764310	rs12030478
rs748944	rs2486948	rs6773379	rs9667991	rs13016342	rs10500337	rs9840992
rs12144885	rs7627615	rs619945	rs6426996	rs10828479	rs2295846	rs8083571
rs7794902	rs6936903	rs9296541	rs6862252	rs12491408	rs294856	rs10503809
rs34602						

Table 4: **Random Feature Extraction SNPs**

6 Additional Material