Cancer Classification by Gene Expression Monitoring

An Exploration in the Bias-Variance Tradeoff of Machine Learning

AS.553.450 Computational Molecular Medicine

Sam Dawley

Spring 2022

Abstract

Cancer classification is widely regarded as the first and most important step in the development of cancer diagnosis and treatment. For example, tumors which exhibit strikingly similar phenotypes may share an insiginificant number biological markers and lead to drastically different clinical treatments. This paper presents an approach to feature recognition and cancer classification while making use of standard statistical procedures and machine learning algorithms. Results of preliminary statistical analysis are convincingly effective at predicting significant features within the data set and provide a route to feature discovery in the future. Statistical learning algorithms prove less effective at class prediction while providing an excellent example of the bias-variance tradeoff within machine learning and the care researchers must take to consider the makeup of a data set as it applies to training a classifier and mirroring reality.

I. Introduction

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (i.e., cancer discovery) or for assigning tumors to known classes (i.e., class prediction). In this article an approach to cancer classification based on gene monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. Statistical analysis methods proved to have varying capability of differentiating between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without any previous knowledge of the classes. These results demonstrate the plausibility of cancer identification based solely on differential gene expression and suggests a general method of cancer classification without any prior biological knowledge.

This data set comes from a proof-of-concept study published in 1999 from Golub, et al. The research showed how new cases of cancer could be classified by gene expression monitoring and thereby provided a general approach for identifying new cancer classes and assigning tumors to known classes.

i. Background

In the past, cancer classification has been primarily based on morphological appearance of the tumor, however, this has serious limitations. Tumors

which exhibit similar histopathological or physical appearance can show immensely different responses to similar clinical treatment, suggesting that the tumors are markedly different at the molecular level. Thus, it is not unreasonable to believe that numerous classes of cancer exist but have yet to be classified.

In this article, acute leukemia will be used as a test case for developing a method of cancer classification. In particular, gene expression data from monitoring DNA microarrays is utilized in two distinct ways. Firstly, the molecular data is used to identify significantly differentially expressed genes that contribute largely to determining the class of cancer, whether it be ALL or AML. Secondly, a machine learning model is trained on the data with the intention of cancer classification without any prerequisite knowledge of biology or genomics.

As it stands, the distinction between ALL and AML is well established, although, no single test is currently sufficient to cement a diagnosis. Rather, current practice involves an expert's interpretation of the tumor's morphology, among other characteristics. While accurate, human error is invariably introduced to some of these cases leading to imperfect classification. Herein is presented a more systemic approach to ALL and AML classification.

ii. Methods

Standard statistical measures and tests were computed and performed prior to processing the data and building any predictors, all of which were taken from `sci-kit learn`. Firstly, the significance of each gene was determined in two ways. One method was to measure the mutual information between the expression data and the class label, defined in terms of the Kullback-Leibler divergence:[1]

$$I(X,Y) = D(p_{X,Y}||p_X p_Y) \tag{1}$$

Above, X and Y are random variables representing gene expression and class labels, respectively, and $p_{X,Y}$ is their joint distribution while p_X and p_Y are their respective probability distributions. Importantly, note that the mutual information between two random variables is zero if they are independent and that I grows monotonically as their dependence increases.

The most informative genes appeared often as significant determinants of class prediction. A histogram of mutual information for each gene is illustrated in Figure 3.

Further preprocessing was done by performing a Wilcoxon rank-sum test under the null hypothesis that the gene expression data is identically distributed regardless of the class label. The rank of the kth feature in the data set is determined by

$$r_k(X) = 1 + \sum_{j \neq k} \delta(x_j < x_k) \tag{2}$$

where δ is the Kronecker delta function and the corresponding statistic is measured as

$$W(X) = \sum_{k=1}^{n} y_k r_k(X) \tag{3}$$

for a rank r_k and class label y_k (note that the class labels are assumed to be binary, either 0 or 1).

Initial statistical analysis was concluded with principle component analysis (PCA), the process by which the data set is compressed into a sequence of vectors, each of which is the direction of a best fit line for the data. Importantly, the ith vector in the set of n is orthogonal to the other $n-1$ so that the new basis is orthonormal and each direction is linearly uncorrelated. Moreover, PCA identifies those features which contribute the greatest variance to the set.

Although the set of vectors returned by PCA is minimally related to the features present in the initial data set, the new basis can still be used to inform us about particular features within the data. An illustration of this use can be found in Figure 1.

The effect of the principle components on data analysis is best understood through an analogy. If we considered, say, the two most statistically significant features in the sample based on the Wilcoxon rank-sum test, we could train a classifier on solely

[1] The Kullback-Leibler divergence is defined by

$$D(p||q) = \sum_{\text{all } x} p_x \ln\left(\frac{p_x}{q_x}\right)$$

for probability distributions p and q.

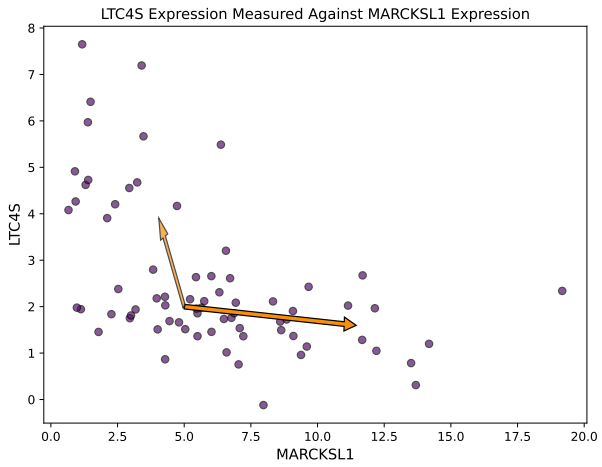

Figure 1: Expression data of the MARCKSL1 gene plotted against that of the LTC4S gene. The first two principle components of the data are shown.

the expression data of these genes for each sample before validating and testing the model. This method would be based on the assumption that only those features which are most statistically significant contribute to cancer classification while ignoring the majority of the other samples in the data set. Figure 2 illustrates the effectiveness of some machine learning algorithms on this approach of using only two features, fumarylacetoacetate (gene 2019) and oncoprotein 18 (gene 1927), to classify samples. The roughness of the classification illustrated in this figure is remedied when we remember that classification was performed using only those two genes, whereas in reality the models we train are free to use many more statistically significant features.

Upon completion of data preprocessing, statistical analysis was performed using different machine learning techniques, including, but not limited to, random forest and nearest neighbor classification, as well as quadratic discriminant analysis. All methods proved capable of predicting the the class label with fair accuracy, some better than others. Ultimately, these models were combined to yield a majority voting classifier.

II. Preliminary Analysis

Preliminary analysis of the data revealed the most informative genes of the set which corresponded to the most statistically significant genes from the Wilcoxon test. A few of the most statistically significant genes and their respective p-values are presented in Table 1. The majority of the genes in this sample overlap with the set of the most informative genes, most notably the genes ZYX, FAH, and CST3.

CST3 is interesting in an additional respect because it appears to introduce the most variability to the data set. Through principle component analysis it is seen that Cystatin C, the protein encoded by CST3, directs the most variation within the data set and therefore appears to be a large, if not the largest, contributor to the class label. Along with this facet of the gene within the data set is its connection to bone disease and some cancers. In particular, Cystatin C has been strongly associated with multiple myeloma and proven to be a great diagnostic metric for myeloma. In this way, CST3 may be a candidate for discovering new treatments for different types of cancers or a potential therapeutic target.

Gene	p-value/10^6	Gene	p-value/10^6
ZYX	1.755	CD33	4.504
CST3	2.412	LYN	5.252
LTC4S	3.859	FAH	5.252
ELA2	3.859	CHRNA7	5.252
CSTA	4.504	LEPR	5.252

Table 1: Most differentially expressed genes in the sample.

The other two genes, ZYX and FAH, have also been shown to contribute to carcinogenesis to varying degrees. The ZYX gene encodes Zyxin, a protein which has been shown to play a role cancer development, as well as apoptosis and wound healing. Moreover, some studies have revealed that ZYX can act as either an oncogene or a tumor suppressor, depending on the type of cancer. Potentially more significant is the role of ZYX in the *Hippo pathway*. Hippo signaling is an evolutionary pathway that controls cell proliferation and stem cell

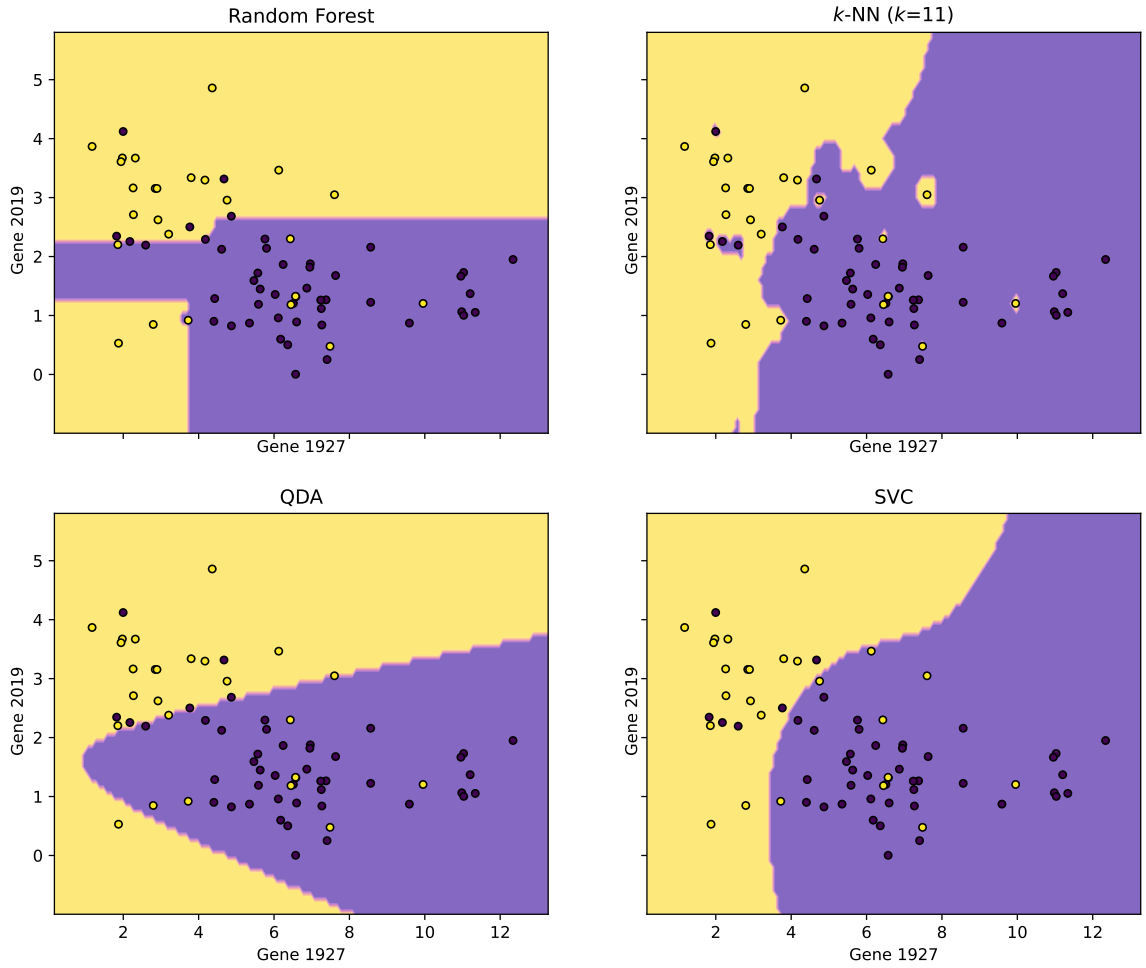

Figure 2: Mutual information of the genes with the highest mutual information across the entire data set.

self renewal. The role of ZYX, in conjunction with another pair of genes, may lead to deactivation of this pathway and cell proliferation. The FAH gene, which encodes fumarylacetoacetate, has been associated with liver cancer and general liver dysfunction.

III. Results

i. Random Forest Classification

A *random forest classifier* is a combination of decision tree predictors which uses averaging and ensemble voting to improve the predictive accuracy over a single decision tree alone. In fact, forests have been shown to be competitive with other classification algorithms, previously thought to have outclassed this nonparametric model.

While proven effective at different types of classification and regression, the random forest is confounding insofar as trying to understand its mechanism of prediction. In fields such as medicine, the interaction of features within a data set is critical to developing biological models which describe observed outcomes. Hence, random forests are suboptimal with regard to interpretability. Nevertheless, here we concern ourselves with

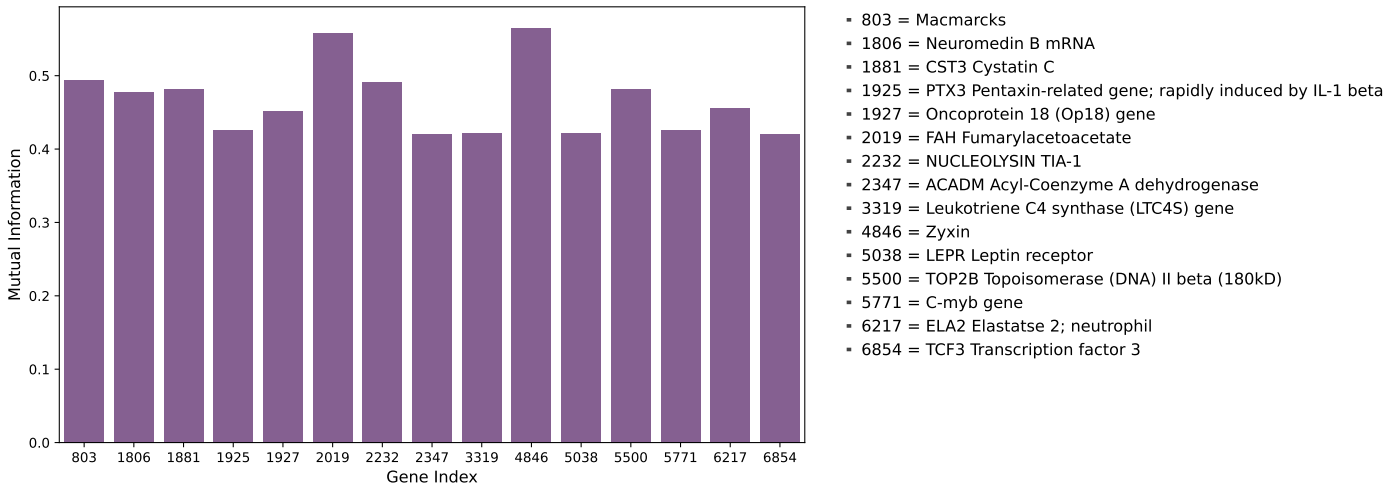

Figure 3: Mutual information of the genes with the highest mutual information across the entire data set.

classification and leave interpretability for another time.

Finding the optimal parameters for the random forest was done by a grid search using the built-in function within `sklearn.model_selection`. Grid searching allows us to fine tune hyperparameters of the model to optimize predictive ability. This model was trained and validated using leave-one-out cross-validation.

The importance of each feature within the random forest was calculated and the results are reported in Figure 6. Unsurprisingly, the same genes which were most statistically significant in the Wilcoxon rank-sum test are most important within classification. Evaluation of the model was done by plotting an ROC curve and also building a confusion matrix based on class predictions from the random forest. The area under the ROC curve for this classifier is 0.318.

The results of the random forest classifier on this data are disheartening. Though, it provides a nice illustration of the trade-off between bias and variance within machine learning. The cross-validation score of this model on the training set is excellent, above 0.90. Though, clearly, the classifier does not do as great a job at predicting the class labels of the testing set.

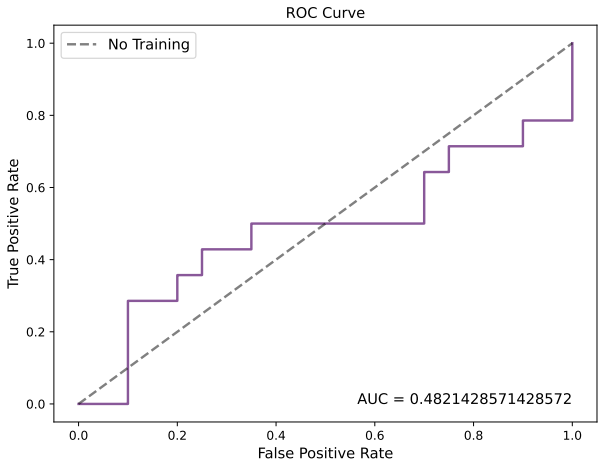

Figure 4: ROC curve for random forest classifier based on performance when modeling test set.

ii. Nearest Neighbor Classification

Nearest neighbor classification (or *k-nearest neighbor classification*) is a machine learning method which aims at labeling previously unseen sample points by inheriting the class label of the sample point(s) nearest it. The decision rule for this method combines labels of the k nearest sample points, either by majority, plurality, or distance-based voting.

Part of the novelty in nearest neighbor classification is that the it is a *lazy evaluation model* (in contrast to *eager* evaluation models). Unlike parametric estimators which allocate time and memory

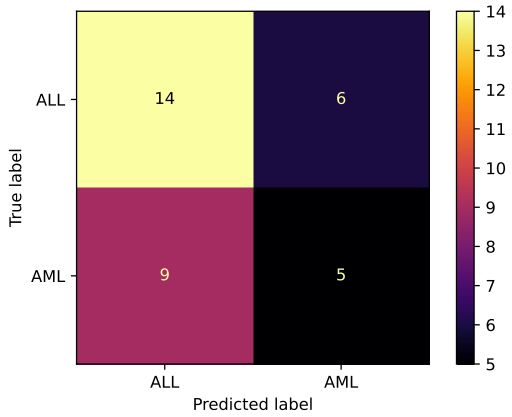

Figure 5: Confusion matrix based on predictions made by random forest classifier on testing set.

to build an appropriate model before classifying the data, lazy evaluators derive their decision-making ability from the training elements and therefore are sometimes referred to as *instance-based learners*.

Similar to how the random forest classifier was evaluated, the nearest neighbor model was assessed initially using a confusion matrix and an ROC curve. The area under the ROC curve showed a slight improvement over the random forest, granted, it remained abysmally low at 0.371. Unlike the random forest classifier, however, hyperparameters for the classifier were fine-tuned using an iterative method to determine the best value of k. Ultimately, $k = 11$ was chosen for the model and a plot of performance as k varies can be found in Figure 7.

The ROC curve for this model is illustrated in Figure 9. Both this model and the random forest classifier suffered from a greater false positive than true positive rate, indicating some potential bias within the models. To attempt to remedy this, the dimensionality of the training set was reduced greatly, bringing the number of features used to classify down from 7129 to just 38. In particular, the 38 most statistically significant based on the Wilcoxon rank-sum test. Performing the same analysis with the new training set afforded an ROC curve with more promising results and is illustrated in Figure 10.

a. *Observed Error Rate of Nearest Neighbor Classifier*

In the limit that the number of samples being trained and tested on grows increasingly large, the generalized error of the nearest neighbor classifier, $e(F_{NN})$, is shown to be

$$e^* \leq e(F_{NN}) \leq 2e^* \tag{4}$$

where e^* is the error rate of the Bayes classifier. In practice, these bounds imply that the nearest neighbor classifier is often hard to improve upon.

The observed error rate of the nearest neighbor classifier in this example, however, is exceedingly large and warrants some inspection. The Bayes estimator is a theoretically optimal classifier, assuming that the joint distributions of the features and class labels is known, and is typically unknown in practice. In particular, it minimizes the loss based on the expected value of the posterior distribution and the true class label. The reason for the observed error rate in this case likely stems from the fact that a large amount of features in the data set are redundant and likely lead to overfitting. This reason corroborates the effect of reducing the training set by considering only the most differentially expressed features, as shown in the ROC curve of Figure 10.

iii. Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) is a generative machine learning model based on linear discriminant analysis (LDA), with slightly less relaxed assumptions. Both LDA and QDA fall under the category of *Gaussian discriminant analysis*, a learning algorithm which attempts to fit a Gaussian distribution to each class of the data set. LDA results in a linear decision function because it assumes class-independent variances, whereas QDA results in a quadratic decision function because it does not.

Although this learning method is typically used only when the number of observations in a sample is large enough, the classifiers we've dealt with so far have suffered from suboptimal predictive power so why not try something new?

The QDA classifier was initially tested using both the training set and testing set containing

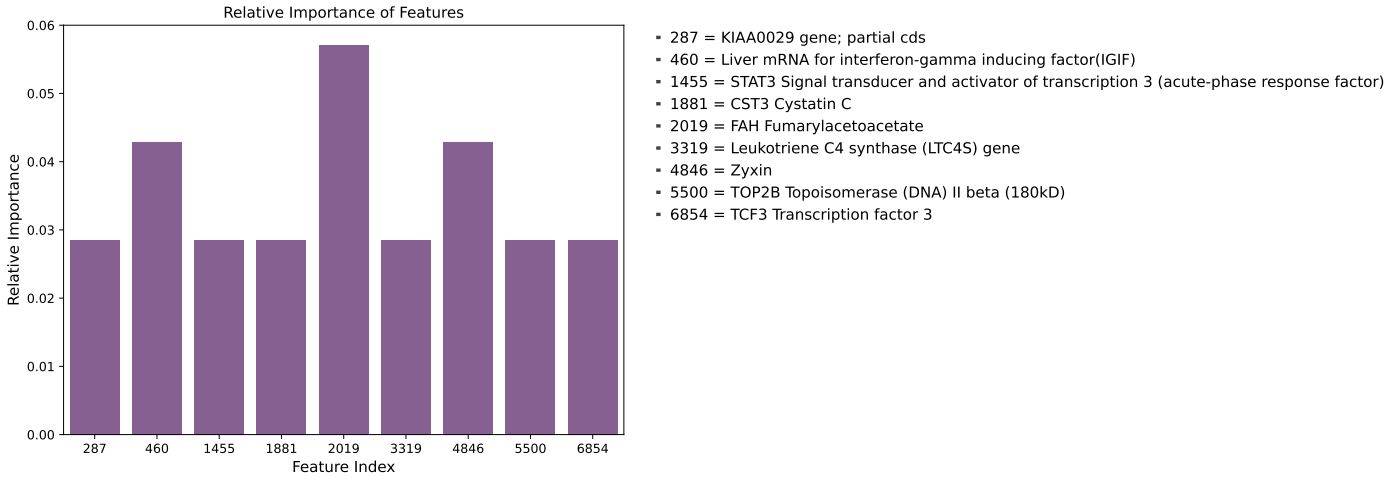

Figure 6: Relative importances of most important genes within random forest classifier.

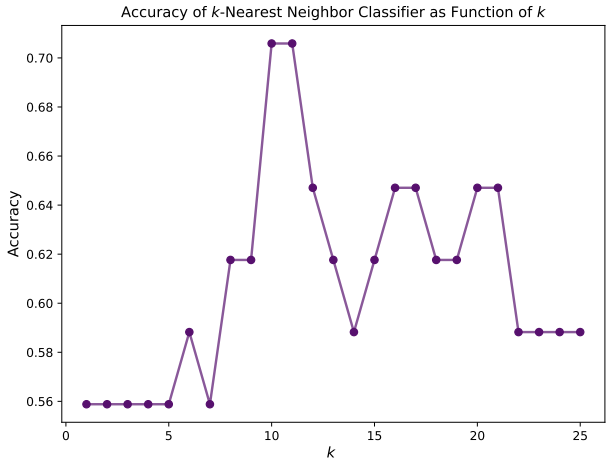

Figure 7: Plot of accuracy versus k for k-nearest neighbor classifier. This was the method used to determine the optimal value of k for the model.

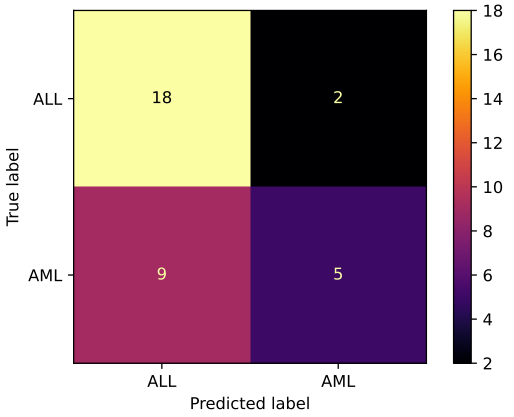

Figure 8: Confusion matrix based on predictions made by k nearest neighbor classifier on testing set.

all features of the set, yielding unfavorable results. While increasingly promising, the results for the model on the reduced training and testing sets still suffers from large false positive and false negative rates. An ROC curve of the latter, smaller training set can be found in Figure 12.

The misclassifications of ALL cases as AML cases is in contrast to the predictions made by the nearest neighbor model which did a finer job of predicting true class labels.

iv. Support Vector Machines

A *support vector machine* (SVM) is a supervised learning algorithm which finds the hyperplane in space with the largest margin between two classes (in the case of binary classification). SVMs are used widely in classification problems and can function as both linear and nonlinear classifiers, depending on the data set. Here, a *linear support vector classifier* (LSVC) is used to classify the data as belonging to either the ALL or AML classes.

The advantages and disadvantages to SVMs are well-recognized. Importantly, the advantages of SVMs include begin effective in high-dimensional spaces (such as our case, including over 7000 features) and being memory efficient, since only a sub-

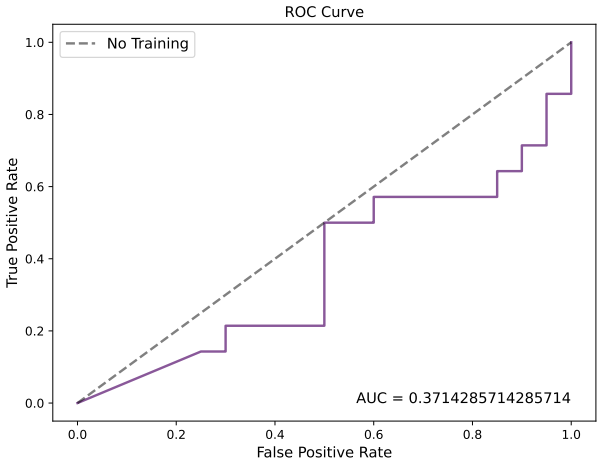

Figure 9: ROC curve for k nearest neighbor classifier based on performance when modeling test set.

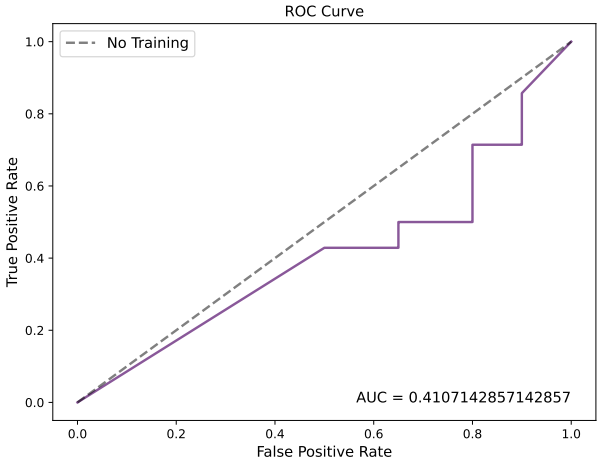

Figure 10: ROC curve for k nearest neighbor classifier based on performance when modeling test set using the reduced training set.

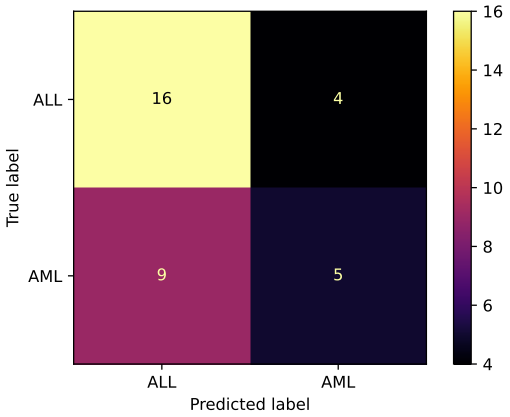

Figure 11: Confusion matrix based on predictions made by QDA classifier on testing set.

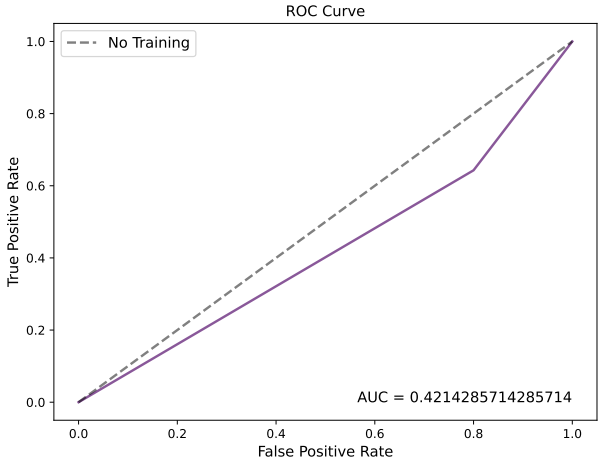

Figure 12: ROC curve for the QDA classifier based on performance when modeling test set using the reduced training set.

set of the training set is used to create the decision function (also known as the hyperplane in this case). However, disadvantages of SVMs occur when the number of features of the data set vastly outnumbers the number of samples (again, such as our case), leading to overfitting, as well as the fact that SVMs do not directly provide probability estimates of the decided class. Although probability estimates of the decision can be computed using algorithms such as Platt scaling, these calculations are expensive relative to the cost of building the classifier to begin with.

Figures 13 and 14 are confusion matrices evaluating the accuracy of an LSVC on the reduced feature set for both the training and testing class labels.

v. Ensemble Voting

The final technique employed to attempt to build an effective classifier involved compiling all of the models built so far into a *voting classifier*. The principle behind a voting classifier is to combine conceptually different classifiers and use a majority vote to predict class labels. Such classifiers work for

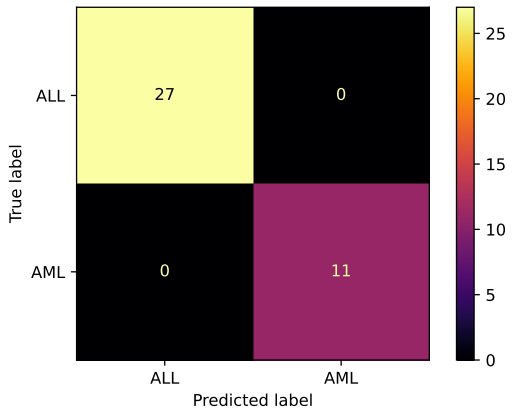

Figure 13: Confusion matrix based on predictions made by LSVC classifier on training set using reduced training set.

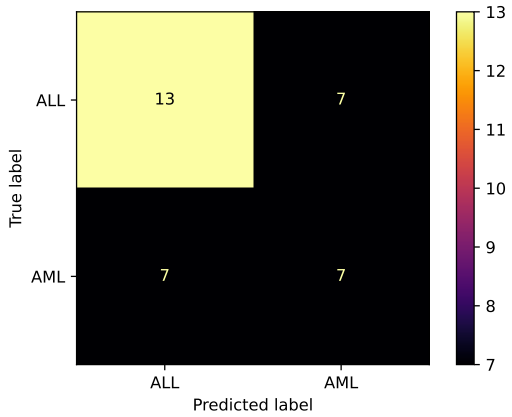

Figure 14: Confusion matrix based on predictions made by LSVC classifier on testing set using reduced training set.

equally well-performing models and for balancing models to minimize the weaknesses of individual contributors. The approach presented here is to try and address the latter.

A voting classifier was created using three of the four models presented herein, including LSVC, QDA, and the nearest neighbor classifier. The choice of these three classifiers was based on the performance of each model. Some relevant accuracy metrics for each of the four classifiers are illustrated in Table 2. These particular models were chosen in an attempt to avoid the weaknesses of each classifier and bolster the strengths because each showed poor performance in different respects.

Classifier	F1-score	AUC	Accuracy
Random Forest	0.519	0.482	0.529
Nearest Neighbor	0.676	0.371	0.500
QDA	0.414	0.421	0.489
LSVC	0.500	0.500	0.575

Table 2: Results for selected scoring metrics of classifiers used. The "Accuracy" column is a measure of the balanced accuracy.

Interestingly, the performance of the ensemble classifier based on the confusion matrix in Figure 15 is identical to that of the LSVC classifier alone.

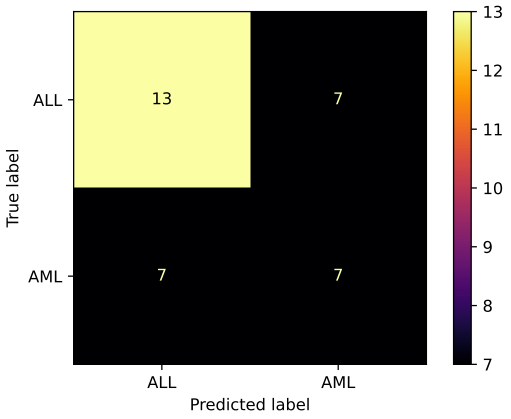

Figure 15: Confusion matrix based on predictions made by voting classifier on testing set using reduced training set.

IV. Discussion

Statistical analysis and machine learning algorithms have proven capable to some degree of binary class prediction. The distinction between ALL and AML was shown to vary with only a small subset of the features within the data, and by reducing the number of features, class prediction only grew more accurate. The most striking feature of the outcomes of these classifiers, however, is the large bias imparted into each model in spite of efforts to reduce it.

i. The Bias-Variance Tradeoff

No matter the particular machine learning algorithm, the goal of classification and regression is to maximize the accuracy of a model while minimizing the loss. The accuracy of a model can therefore be measured by the total error, in particular the mean-square error (MSE), determined as

$$\text{MSE} = B^2 + V \tag{5}$$

for a bias B and variance V. Hence, maximizing the accuracy of a model equates to minimizing the bias and variance.

Within each classifier studied earlier, performance on the training set was excellent, with each algorithm scoring well during cross-validation and earning high balanced accuracy scores. However, this performance dropped significantly on the testing set, suggesting a strong bias present within the training data.

Analyzing the samples used in the training set reveals the source of this bias as imbalanced data. The number of patients with ALL and AML in the training set was 27 and 11, respectively, and in the testing set the number of occurrences of each was 20 and 14, respectively. In both cases, the amount of ALL cases outnumbers the amount of AML cases, though, the ratio of cases is disproportionate across sets. Training on imbalanced data has generated two competing schools of thought regarding machine learning. Some argue that training on imbalanced data biases a classifier towards the majority class, exactly what is observed herein. Others argue that rebalancing sample selection affords an inaccurate view of reality and that imperfect training data is essential for real-world applications of machine learning to medicine.

The latter of these arguments is particularly applicable to cancer classification and prediction where proponents for the school of thought believe that bias inherent within the data is a characteristic component of the overall population. Especially within cancer prediction, false positives may be favorable over false negatives because the former is a more conservative prediction and less likely to lead to complications in the future.

ii. Biological Relevance

The biological relevance of the statistically significant genes is discussed briefly earlier, and commented on here as well. The most statistically significant genes in the sample are those which have been connected to carcinogenesis in previous studies and connected with important biological pathways related to cell proliferation.

Most importantly, it should be noted that learning and prediction from the model was done with no prior biological background. The significance of genes in certain pathways bore no impact on the statistical significance of the feature and did not impart any bias toward the impact that feature would have on classification. Thus, the outcomes presented here illustrate a method of prediction based purely on data. Hence, it is not unreasonable to believe that building a classification model *with* prior biological knowledge may bolster predictive power and create a stronger classification scheme.

Moreover, a classification model of the kind may function as a method of discovery for cancer-related genes in the future. Many features determined to be significant and informative based on only some of the classification models reviewed here have not been linked with leukemia, let alone different cancer types to begin with. In conjunction with the observation that the most highly correlated features of the sample, such as ZYX and CST3, have been shown to either directly or indirectly affect cancer development and tumorigenesis, future directions for this research may invariably include testing the effect that suppression of other highly correlated genes (which have not previously been focused on) have on cancer development.

V. Conclusion

The primary result of the statistical analysis and machine learning research done within this article is represented by numerous takeaways. Firstly, and most importantly, is the importance of inherent bias within a data set and how it plays out within classification.

The bias-variance tradeoff has been already laboriously discussed and its consequences to predictive power of different types of machine learning algorithms has been analyzed. Recognizing the

inherent bias within a training set is important when evaluating the predictive power of an estimator, however, altering this bias may not always be the optimal route to fast track real world applications. The advent of machine learning applied to medicine is promising, and while training classifiers on randomized data typically leads to unbiased algorithms, the bias inherent within a population is often a key characteristic of that data set. On the other hand, training a classifier on biased data oftentimes leads to a biased predictor, decreasing the variance of the model and weakening predictive power. This tradeoff is the primary reason machine learning applications to medicine are not as omnipresent as computational biologists might desire. The expertise of a specialized physician will always outclass a computer.

The other major takeaway of this project is the power of machine learning algorithms to discover new, biologically relevant features which significantly contribute to carcinogenesis, among other interests. Current research is motivated by previous studies on the effects of certain features, which is motivated by previous studies on the effects of certain features, and so on. Here, a method of detecting significant contributors to leukemia is only a preliminary to further statistical analysis, suggesting that applying a greater focus on gene expression data within patients with leukemia may lead to the discovery of features that were previously uncorrelated with leukemia. This hypothesis has already been shown to hold water for genes such as ZYX, which plays a role in breast cancer, lung cancer, and certain melanomas.

Within the context of the results presented here, machine learning algorithms appear to be ineffective at cancer classification and one may argue that the applications of statistical learning to medicine should be restricted to statistical analysis. This conclusion, however, is far removed from the particulars of machine learning and the significance that data plays in statistical analysis. While the eager student may remain ignorant to the consequences of the bias-variance tradeoff and standby standard practices of analysis in medicine (citing ineffective predictive power from certain machine learning algorithms), the results of this study offer a critical insight into how statistical learning must be implemented into medicine and why it is necessary to continue to develop methods which have potential to change the way cancer research is performed.

References

[1] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.), 286(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

[2] Timothy J. Triche (1990) Neuroblastoma and Other Childhood Neural Tumors: A Review, Pediatric Pathology, 10:1-2, 175-193, DOI: 10.3109/15513819009067106

[3] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Phys. Rev. E, 69, 066138. doi:10.1103/PhysRevE.69.066138

[4] Demirtaş, S., Akan, O., Can, M., Elmali, E., & Akan, H. (2006). Cystatin C can be affected by nonrenal factors: a preliminary study on leukemia. Clinical biochemistry, 39(2), 115-118. https://doi.org/10.1016/j.clinbiochem.2005.10.009

[5] Yijing Jiang, Jie Zhang, Chenlu Zhang, Lemin Hong, Yuwen Jiang, Ling Lu, Hongming Huang & Dan Guo (2020) The role of cystatin C as a proteasome inhibitor in multiple myeloma, Hematology, 25:1, 457-463, DOI: 10.1080/16078454.2020.1850973

[6] Partynska, A., Gomulkiewicz, A., Dziegiel, P., & Podhorska-Okolow, M. (2020). The Role of Zyxin in Carcinogenesis. Anticancer Research, 40(11), 5981–5988. doi:10.21873/anticanres.14618

[7] Jorquera, R., & Tanguay, R. M. (2001). Fumarylacetoacetate, the metabolite accumulating in hereditary tyrosinemia, activates the ERK pathway and induces mitotic

abnormalities and genomic instability. Human molecular genetics, 10(17), 1741–1752. https://doi.org/10.1093/hmg/10.17.1741

[8] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[9] Seidl T. (2009) Nearest Neighbor Classification. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.