Identifying Types of Cancer from Cancer Patients' miRNA

Brad Shook and **Natnael Mulat**
{brshook,namulat}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

Identifying and classifying cancer is an impactful way that machine learning can benefit the health care industry and save lives. In this research, we attempted to create a classification model that could accurately classify a patient's type of cancer based on their miRNA profile. We tuned and evaluated three types of classification models: a random forest classifier, adaboost classifier, and support vector classifier. These classifiers achieved test set F1 scores of 0.965, 0.830, and 0.912, respectively. Despite these high F1 scores, some cancer types were difficult to classify and there is further work to be done on this classification problem.

1 Introduction

Over the past decades, cancer research has undergone a continuous evolution. In order to find types of cancer before they cause symptoms, scientists have applied various methods, such as early stage screening. With the availability of large amounts of cancer data from cancer organizations such as the International Collaboration of Cancer Reporting (ICCR), The Cancer Genome Atlas (TCGA) repository, and other databases, new strategies for the early prediction of cancer treatment outcome have been developed.

For this research, the dataset used is from The Cancer Genome Atlas (TCGA 2021) repository, which has microRNA (miRNA) profiles for a number of cancer patients where the tissue samples provided represent six different cancer types:

1. Breast invasive carcinoma (BIC)
2. Kidney renal clear cell carcinoma (KRCCC)
3. Lung adenocarcinoma (LA)
4. Lung squamous cell carcinoma (LSCC)
5. Pancreatic adenocarcinoma (PA)
6. Uveal melanoma (UM)

The accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, machine learning methods have become a popular tool for medical researchers. These techniques can discover and identify patterns in complex datasets which allows for the accurate predictions of future outcomes of a cancer type.

This problem has been studied by Telonis et al. using the TCGA dataset (Telonis et al. 2017). In their work, they used the presence and absence of isomiR to classify 32 cancer types using support vector machines (SVM) and random forest classifiers. They found that an SVM-based classifier using binarized isomiR profiles as features can label datasets accurately with a sensitivity score of 90%. They also experimented with binarized miRNA as features but found that the sensitivity score was lower with 83%. According to their research, the false discovery rate (FDR) for a SVM-based classifier using binarized isomiR was 3% while using binarized miRNA yielded a 6% FDR.

Similar to Telonis' work, we developed supervised machine learning models to predict cancer types using miRNA profiles as features. However, we developed our models as random forest classifiers, adaptive boosting classifiers, and support vector classifiers to get a high classification metric. Each models' hyperparameters were tuned through grid searches to find the best set of hyperparameters that yield the best possible classifications of cancer types using cancer patients' miRNA.

The rest of this paper will show the various data preprocessing techniques that we employed, the process of tuning model hyperparameters, and the evaluation of several different classification models to determine which model classifies the cancer types most accurately.

2 Data Preparation

The data corresponding to each cancer type mentioned earlier were in different folders in the dataset from TCGA. After consolidating each data point into one dataset with the miRNA markers as the features and the cancer types as the labels, we created a visualization of the dataset to show the number of examples belonging to each type of cancer. Figure 1 shows the bar chart with the number of samples for each cancer type. BIC has the highest count with 1,096 samples, while UM has the lowest count with 80 samples. Overall, the bar chart highlights the imbalance of labels in the dataset.

In order to expedite the convergence of our machine learning models, we performed feature scaling using $L2$ normal-

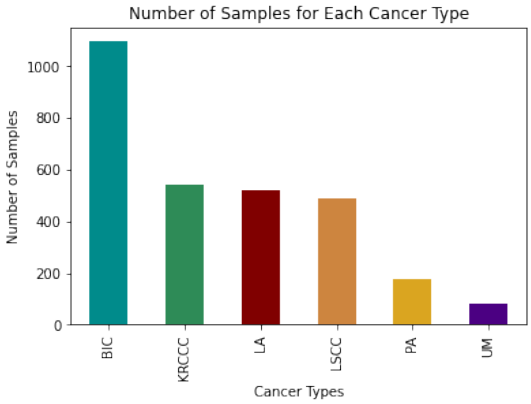

Figure 1: Bar graph showing the number of samples for each classification of cancer types.

ization on the miRNA features in the dataset. We randomly shuffled the dataset, then divided the dataset into training and test sets, where the training set was comprised of 80% of the dataset while the test set was only 20% of the dataset.

3 Experiments

For our experiments, we created three distinct models: a random forest classifier (RFC), an adaptive boosting classifier (ADAC), and a support vector classifier (SVC).

A SVC can be created using four kernel functions. These kernels are linear, RBF, polynomial, and sigmoid. These respective kernel functions are defined as follows:

$$\langle a, b\rangle \tag{1}$$

$$exp(-\gamma||a-b||^2) \tag{2}$$

$$(\gamma\langle a, b\rangle + r)^d \tag{3}$$

$$tanh(\gamma\langle a, b\rangle + r) \tag{4}$$

where a and b are training example vectors. Then, γ is specified by the parameter gamma, r by coef0, and d by degree.

The following experiments focused on tuning the hyperparameters of each of these models with grid searches and comparing the efficacy of each model.

Experiment 1

In the first experiment, the hyperparameters of each model were tuned using a grid search over different sets of values (see Section 8). Each grid search used weighted F1 scores to compare models. Additionally, the grid searches used 5-fold cross-validation to find an average measure of performance in each model that was fitted. The mean weighted (MW) F1 scores were measured on the cross-validation folds.

In the RFC, the number of estimators were tuned over a range of values. Then, in the ADAC, both the number of estimators and learning rate were tuned.

Next, the SVC was tuned over five different hyperparameters: kernel, C, degree, gamma, and coef0. The hyperparameters kernel, degree, gamma, and coef0 were introduced earlier and C is the regularization hyperparameter. Note that the strength of the regularization is inversely proportional to C.

These grid searches attempted to find the hyperparameters that yielded the highest weighted F1 scores for each model.

Experiment 2

For the second experiment, we compared the F1 scores of each tuned model and the individual F1 scores for each type of cancer, as well as how our models compared to Telonis' model in terms of sensitivity scores. The F1 and sensitivity scores were evaluated on the test set. By calculating the F1 scores, we were able to determine which model performed the best and whether some cancer types were more easily identifiable by the models. Then, we analyzed whether the models were overfit and by how much in terms of F1 score differences between the train and test sets.

4 Results

Results of Experiment 1

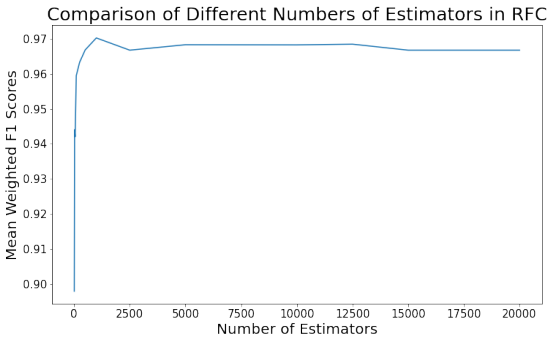

Figure 2: Line chart showing how different numbers of estimators affect the MW F1 scores. F1 scores are from a 5-fold cross-validation grid search.

For the RFC, our grid search found that 1,000 estimators was the value that yielded the best MW F1 scores. In Figure 2, we see that the scores peaked at 1,000 estimators and then plateaued as the number of estimators increased.

Then, the grid search for the ADAC determined that a learning rate of 0.1 and 5,000 estimators resulted in the highest MW F1 scores. A heatmap displaying the full results from this grid search is shown in Figure 3. In this heatmap, we see that higher numbers of estimators generally produced higher MW F1 scores. Moreover, we

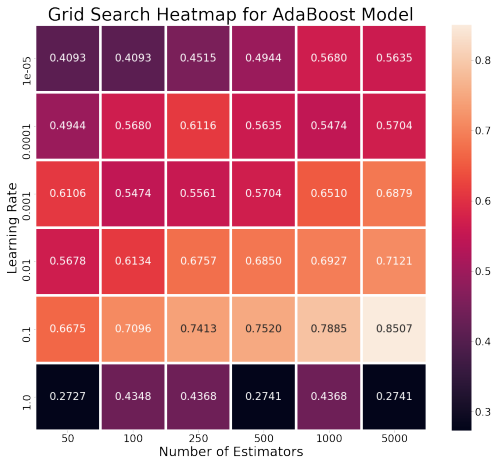

Figure 3: Heat map showing showing the MW F1 scores of each model created in the grid search. Each MW F1 score is from 5-fold cross-validation.

see that a learning rate of 0.1 yielded the highest MW F1 scores. Interestingly, a learning rate of 1.0 made the models perform much worse, even though this learning rate is one magnitude off of 0.1, the best learning rate.

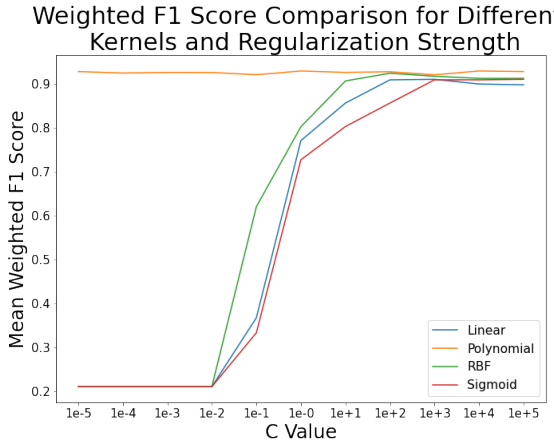

Figure 4: Comparison between different SVC kernel functions and C values, in terms of mean weighted F1 scores from 5-fold cross-validation.

The last grid search involving the SVC found that the following hyperparameter settings created the model with the highest MW F1 scores:

- Kernel Function: Polynomial
- C: 1
- Degree: 2
- Gamma: 10
- Coef0: 1

This means that the other kernel functions did not yield higher MW F1 scores. Additionally, we can infer that a linear SVC was not able to classify as well as a polynomial SVC which has a more curved decision boundary. Furthermore, the higher degree polynomial SVCs likely resulted in overfitting, so a degree of 2 was the best choice between underfitting and overfitting.

As shown in Figure 4, the C values had little to no effect on the polynomial models searched through in the grid search. All of the polynomial SVCs had high MW F1 scores while SVCs using the other kernel functions were greatly affected by the strength of regularization. Generally, these other SVCs performed better when the strength of regularization was minimal, in other words, when C was large. Contrarily, those SVCs performed poorly when the strength of regularization was large.

Results of Experiment 2

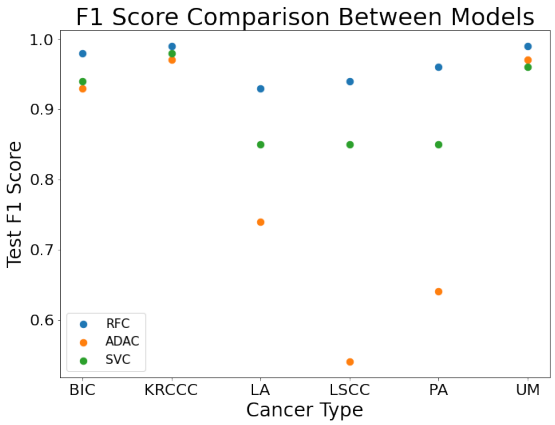

Figure 5: Scatter plot comparing the test F1 scores for each model when classifying each type of cancer.

After fitting each model with their tuned hyperparameters on the training set, the F1 scores for each model were calculated on the test set. We found that the F1 scores for the RFC, ADAC, and SVC were 0.965, 0.830, and 0.912, respectively. Also, the sensitivity scores of each model were 0.96, 0.84, and 0.91, respectively. All three models had sensitivity scores greater than Telonis' model which had a sentivity score of 0.83 for the binarized miRNA dataset. In Figure 5, we see that the RFC was able to classify each type of cancer better than the ADAC and SVC. Specifically, when classifying LA, LSCC, and PA, both the ADAC and SVC had F1 scores much lower than the RFC. For these cancer types, the RFC F1 scores were an average of 0.303 greater than the ADAC scores and an average of 0.093 greater than the SVC scores. These cancer types were the hardest to classify for each model. This trend likely means that these three cancer types are closely related in terms

of the miRNA expressed, while the other types are more distinguishable in miRNA. Despite the imbalance in the number of samples of each cancer type, there was not a distinguishable correlation between the number of samples and the F1 scores of the cancer types. For example, as mentioned previously, BIC has the most samples while UM has the least samples. Despite this, all three models were able to achieve high F1 scores when classifying these two cancer types. Thus, it is unlikely that lower numbers of samples caused some cancer types to be harder to classify by the models.

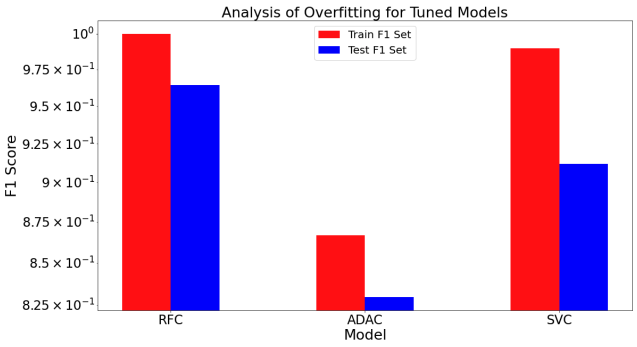

Figure 6: Comparison of F1 scores on test and train sets for each model.

As we saw in Figure 5, F1 scores on the test set for each model vary in efficacy. In Figure 6, we see that all three models were overfit by varying amounts in terms of F1 scores. The SVC was the most overfit model with the train F1 score being 0.078 greater than the test F1 score. Both the RFC and ADAC were less overfit, with train-test F1 score differences being 0.036 and 0.037, respectively. Despite being overfit, these models were tuned through grid search so the grid search determined that there were no other hyperparameter combinations that yielded better F1 scores on held-out data points. Thus, the models being overfit was worth the F1 score gains on the test set.

5 Broader Impacts

According to researchers, the accuracy of cancer predictions has significantly improved by 15%–20% with the application of machine learning techniques (Cruz and Wishart 2006). The use of machine learning in disease prediction is part of a growing trend towards personalized, predictive medicine. This movement towards predictive medicine is important, not only for patients in terms of lifestyle and quality-of-life decisions but also for physicians in choosing treatment. It also benefits health economists and policy planners in implementing large scale cancer prevention or cancer treatment policies (Weston and Hood 2004).

However, this reliance comes with some drawbacks. There are around 100 known types of cancer just based off of organs or tissues where the cancers form (Aroef, Rivan, and Rustam 2020). The models that we built using RFCs, ADACs, and SVCs only classify six of these cancers when given the imbalanced data from TCGA. One of the main problems is that the model is limited to only predicting those cancer types defined in the model initially. If a cancer patient has a type of cancer besides those defined, the model will misclassify that undefined cancer type. Cancer is diagnosed by relying heavily on the data. If a cancer patient relies on such models frequently, they pay the price of wrong diagnosis with their lives and well-being (Fotouhi, Asadi, and Kattan 2019). As such, including all the known cancer types is vital from a patient's health perspective.

Assuming a model was built using all the known cancer types, there is still a problem of applying such models in a real-world scenario since all of our data belonged to cancer patients and there is a risk of misclassifying unknown cancer types not yet discovered by medical researchers. Thus, one of the issues of deploying these models in health clinics or in the hands of consumers is the inability of the models to predict cancers in non-diagnosed patients. Further, the fact that the model will incorrectly diagnose a cancer type that has not yet been discovered makes the predictions problematic. If physicians follow the model's diagnosis and prescribe particular medical procedures for that incorrectly predicted cancer type, one can imagine the risks and medical malpractice that may follow.

In addition, this problem of misclassification of unknown cancer types could also stifle discovery of new types of cancers that are not yet known to humans. If the patient has a cancer type that is new and not yet diagnosed, the ideal model will approximate the miRNA values to the closest cancer type defined in the model. The new cancer type will not be identified but rather misclassified. In other words, misclassifying an unknown cancer type is a roadblock in the field of oncology if there is a heavy reliance on a model like those mentioned in this paper since physicians and researchers will not have the opportunity to discover the new cancer type.

6 Conclusions

We attempted to classify miRNA data belonging to cancer patients into six different cancer types using three classification models. These models were the RFC, ADAC, and SVC. We found that the RFC was the model that yielded the highest F1 scores on held-out data. Specifically, the F1 scores for the RFC, ADAC, and SVC were 0.965, 0.830, and 0.912, respectively. All three models were able to classify BIC, KRCCC, and UM with high F1 scores, however, there was a large decrease in F1 scores when classifying LA, LSCC, and PA. We theorize that this is due to the miRNA expressed in LA, LSCC, and PA being less distinguishable from each other and the other cancer types, leading to misclassifications. Moreover, we found that each model was overfit, but the models being overfit were acceptable since these models were tuned through a grid search. Overall, the RFC was an exceptional classifier of the cancer types, but there is further work to be done to more accurately classify each cancer type, especially LA, LSCC, and PA.

Further Work

Our paper primarily focused on the use of RFCs, ADACs, and SVCs to solve this cancer classification problem. However, another possible avenue for classifying this data could be through neural networks. Neural networks are prominent machine learning models that can solve complex classification problems. Additionally, it would be beneficial to add miRNA profiles belonging to non-cancer patients to the dataset. This would allow for the creation of models that can identify cancer types in cancer and non-cancer patients alike. Lastly, it may be worth increasing the number of cancer types included in the dataset. This would address the limits of the current models only being able to classify six cancer types.

7 Contributions

B.S. and N.M. both worked on data normalization and exploratory data analysis. B.S. and N.M. built random forest, adaptive boosting, and support vector classifiers. B.S. ran grid searches and generated experimental figures. For the paper, N.M. wrote the abstract, introduction, data preparation, and broader impacts sections. B.S. wrote the experiments, results, conclusion, and further work sections. Both B.S. and N.M. edited and reviewed the final paper.

References

Aroef, C.; Rivan, Y.; and Rustam, Z. 2020. Comparing random forest and support vector machines for breast cancer classification. https://core.ac.uk/download/pdf/295538238.pdf.

Cruz, J., and Wishart, D. 2006. Applications of Machine Learning in Cancer Prediction and Prognosis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/pdf/cin-02-59.pdf.

Fotouhi, S.; Asadi, S.; and Kattan, M. W. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. https://www.sciencedirect.com/science/article/pii/S1532046418302302.

TCGA. 2021. The cancer genome atlas. https://portal.gdc.cancer.gov/.

Telonis, A.; Magee, R.; Loher, P.; Chervoneva, I.; Londin, E.; and Rigoutsos, I. 2017. Knowledge about the presence or absence of mirna isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389567/pdf/gkx082.pdf.

Weston, A., and Hood, L. 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. https://pubmed.ncbi.nlm.nih.gov/15113093/.

8 Appendix

The following shows the values searched over in the grid searches mentioned in Section 3.

- RFC
  - Number of Estimators: 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 12500, 15000, 20000
- ADAC
  - Number of Estimators: 50, 100, 250, 500, 1000, 5000
  - Learning Rate: .00001, .0001, .001, .01, .1, 1
- SVC
  - Kernel: Linear, Polynomial, RBF, Sigmoid
  - C: .00001, .0001, .001, .01, .1, 1, 10, 100, 1000, 10000, 100000
  - Degree: 2, 3, 4, 5, 6
  - Gamma: .00001, .0001, .001, .01, .1, 1, 10
  - Coef0: .00001, .0001, .001, .01, .1, 1, 10