Predicting Cancer Types Using Classifiers

{edshores,branthony}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Edward Shores and Bryce Anthony

Abstract

Predictive models have grown increasingly useful within the computer science world as people realize their use cases in a wider variety of industries. Our study focused on the medical industry, in that we aimed to predict cancer types based on miRNA marker data sourced from The Cancer Genome Atlas (TCGA). After performing some preprocessing on the data, it consisted of patient IDs, cancer types, and miRNA marker IDs with corresponding levels. To accomplish our goal, we used two classifiers: k-nearest neighbors (kNN) and random forest. The classifiers were trained on the dataset and evaluated using accuracy, precision, recall, and F1 score. Our results showed that both models predicted cancer types with higher accuracy than the baseline, but random forest outperformed kNN by a significant margin. Our study suggests that these classifiers, and likely more machine learning models, are useful for cancer type prediction and can therefore be utilized in the medical world to aid with cancer diagnosis.

1 Introduction

Cancer is one of the most concerning diseases in the modern world, with tens of millions of people worldwide suffering from various types. While doctors' abilities to correctly diagnose cancers have increased over the years due to various advancements in technology, there are still many misdiagnoses due to various factors. As the medical field has seen an increase in the quality and efficacy of instruments, so has the computer science world seen breakthroughs in machine learning technology.

With this innovation, medical practitioners have turned to using machine learning techniques in order to better their abilities at diagnosis and outcome prediction. Phillipe Loher, the director of Machine Learning at the Computational Medicine Center at Thomas Jefferson University, helped to pioneer research into the usage of predictive models in medicine with his 2017 study of the presence of miRNA isoforms and its relation to various cancer types (Loher 2017). In this study, Loher and his colleagues were able to build a classifier which took into account only the presence of the miRNA markers. The classifier successfully predicted the cancer type at a 90% rate, and had the same efficacy even when the number of markers was reduced drastically.

In our exploration, we hoped to build upon Loher's work by analyzing data sourced from TCGA, which contains thousands of patient profiles. Each profile consists of a patient-specific ID, the type of cancer the patient has, and the level of 1,882 unique miRNA markers in samples from the patient. We tested two classifiers, kNN and random forest, on the data against a dummy model to find how well the models could predict the cancer type of a given patient. Our results show that both classifiers proved more effective than the baseline, with the random forest approach outperforming kNN's accuracy score by nearly 9%.

2 Data Preparation

Before conducting our experiments we had to re-organize the data into a usable format. Originally, the Data was a single folder with 6 subfolders for each type of cancer 2. Each cancer folder contained a subfolder for an individual that had information on notably increased micro RNA id's present in that individual along with the micro RNA's read count, reads per million miRNA mapped, and whether it had been cross-mapped. Because we were focused on predicting the the type of cancer an individual has based on the presence and amount of micro RNA's observed in an individual, we only used the micro RNA ID and reads per million miRNA mapped columns in our models. The number of individulas diagnosed with each typoe of cancer can be seen in figure 1.

The specific micro RNA's observed varied between patients that had been diagnosed with different types of cancer, so we had to create a dataset that indicated the reads per million miRNA mapped of each micro RNA id in the entire dataset for each patient. Since each patient profile only indicated the notably increased micro RNA id's for that individual, the reads per million miRNA mapped of the micro RNA id's not specified for an individual were assigned a value of zero to indicate their lack of importance. After this was done for each patient our final dataset was a table with $3,186$ rows (for each individual) and $1,884$ columns (for the reads per million miRNA mapped of each micro RNA id). A portion of this re-organized dataset can be seen in figure 2.

3 Experiments

The classifiers that we tested were K Nearest Neighbors and Random Forest. At the beginning of our exploration, we created a dummy model using the DummyClassifier method from scikit-learn. With this model's accuracy score as our

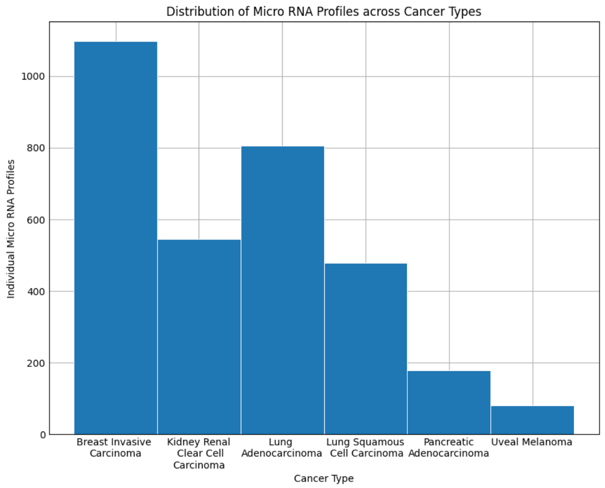

Figure 1: Distribution of Cancer diagnoses for micro RNA profiles

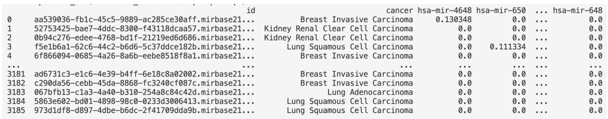
```
                                                  id                           cancer hsa-mir-4648 hsa-mir-650  ... hsa-mir-648
0     aa539036-fb1c-45c5-9889-ac285ce30aff.mirbase21...        Breast Invasive Carcinoma     0.130348         0.0  ...         0.0
1     52753425-bae7-4ddc-8300-f43118dcaa57.mirbase21...  Kidney Renal Clear Cell Carcinoma          0.0         0.0  ...         0.0
2     0b94c276-edee-4768-bd1f-21219ed6d686.mirbase21...  Kidney Renal Clear Cell Carcinoma          0.0         0.0  ...         0.0
3     f5e1b6a1-62c6-44c2-b6d6-5c37ddce182b.mirbase21...     Lung Squamous Cell Carcinoma          0.0    0.111334  ...         0.0
4     6f866094-0685-4a26-8a6b-eebe8518f8a1.mirbase21...        Breast Invasive Carcinoma          0.0         0.0  ...         0.0
...                                                 ...                              ...          ...         ...  ...         ...
3181  ad6731c3-e1c6-4e39-b4ff-6e18c8a02002.mirbase21...        Breast Invasive Carcinoma          0.0         0.0  ...         0.0
3182  c290da56-cebb-45da-8868-fc3240cf087c.mirbase21...        Breast Invasive Carcinoma          0.0         0.0  ...         0.0
3183  067bfb13-c1a3-4a40-b310-254a8c84c42d.mirbase21...              Lung Adenocarcinoma          0.0         0.0  ...         0.0
3184  5863e602-bd01-4898-98c0-0233d3006413.mirbase21...     Lung Squamous Cell Carcinoma          0.0         0.0  ...         0.0
3185  973d1df8-d897-4dbe-b6dc-2f41709dda9b.mirbase21...     Lung Squamous Cell Carcinoma          0.0         0.0  ...         0.0
```

Figure 2: Format of the dataset used in our experiments

starting point, we then created a baseline kNN model so that we could have yet another base of comparison. After creating the baseline kNN model, we then initialized and trained both the kNN and random forest models.

To optimize the performance of the kNN model on the data, we performed multiple iterations of hyperparameter tuning. We began the tuning by running the classifier with a range of k values from 1 to 31, and we used GridSearchCV to find the best performing value for the number of neighbors. After performing many experiments, we determined that 4 neighbors optimized the accuracy score of the model. We also used the GridSearchCV method to find the best value of the second hyperparameter, weights, which ended up being 'distance.'

The Random Forest Classifier performed better than the kNN model from the beginning, but it also required hyperparameter tuning. To find the optimal number of trees in the forest, we first tested the model with 50, 100, and 200 trees, and found that the best scoring classifier had 200 trees in the forest. In order to further tune the model, we decided to then run more experiments with 150, 200, and 250 trees. From this, we found that 150 trees performed better than the model with 200 trees. Therefore, we ran our final test with 125, 150, and 175 trees, and found that 150 was the optimal number of trees in the model. During these experiments, we also concluded that using 'sqrt' as the maximum features in the model outperformed the other options. Manipulating the other parameters did not yield better results, so we decided to leave them as their defaults.

To fit and test the data, we conducted a Train-Test-Split on

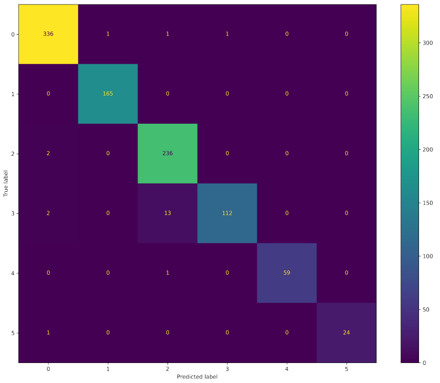

Figure 3: Confusion Matrix for Random Forest Classifier

the data set using a ratio of .3, so 70% of the data went into the training set, and 30% was split into the test set. After fitting both classifiers on the training set, we tested them on the test set in order to find their predictive abilities on the data.

4 Results

Our initial dummy model had an accuracy of 36.8%, indicating the need for more sophisticated models. Our baseline k nearest neighbors (kNN) model achieved an accuracy of 87.8%, which we used as a benchmark for comparison with more complex models.

We employed grid search cross-validation to optimize our models and found that our most accurate kNN model achieved an accuracy of 88.36%, while our most accurate random forest classifier (RF) achieved an accuracy of 97.69%. The confusion matrices for the kNN and RF models can be seen in their respective figures. Our kNN model achieved varying classification accuracies for each cancer type: Breast Invasive Carcinoma (87.02%), Kidney Renal Clear Cell Carcinoma (95.76%), Lung Adenocarcinoma (92.86%), Lung Squamous Cell Carcinoma (74.02%), Pancreatic Adenocarcinoma (85.00%), and Uveal Melanoma (96.00%). Our RF model achieved high classification accuracies for each cancer type: Breast Invasive Carcinoma (99.12%), Kidney Renal Clear Cell Carcinoma (100.00%), Lung Adenocarcinoma (99.16%), Lung Squamous Cell Carcinoma (88.19%), Pancreatic Adenocarcinoma (98.33%), and Uveal Melanoma (96.00%).

The confusion matrix for the KNN model, figure4 , shows that it performed best at correctly identifying cases of Kidney Renal Clear Cell Carcinoma and Uveal Melanoma, with accuracy rates of 95.76% and 96%, respectively. However, it struggled the most with identifying cases of Lung Squamous Cell Carcinoma, with an accuracy rate of only 74.02%.

On the other hand, the confusion matrix for the Random Forest classifier model, figure 3, shows that it performed extremely well overall, with accuracy rates above 98% for all

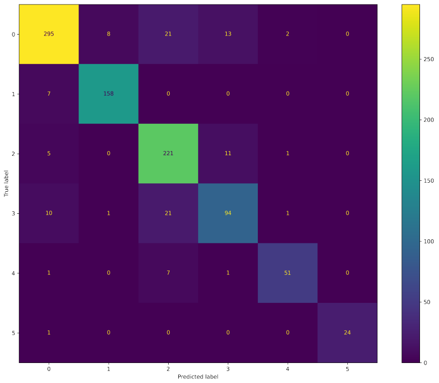

Figure 4: Confusion Matrix for kNN Classifier

categories except Lung Squamous Cell Carcinoma, which had an accuracy rate of 88.19%. This suggests that the Random Forest classifier may be a better choice for this particular classification problem.

It's worth noting that both models performed significantly better than our dummy model, which had an accuracy rate of only 36.89%. Additionally, the baseline KNN model had an accuracy rate of 87.8%, which means that our optimized KNN model was only able to improve on this by a relatively small margin. In contrast, our optimized Random Forest classifier model had an accuracy rate of 97.69%, which represents a substantial improvement over both the dummy model and the baseline KNN model.

Overall, these results suggest that the Random Forest classifier is a highly effective tool for classifying different types of cancer based on the available data, outperforming the KNN classifier by a significant margin. However, further research may be needed to determine whether these results generalize to other datasets and classification problems.

Model	Accuracy Score
Dummy Classifier	35%
KNN Baseline model	87.8%
Tuned KNN Classifier	88%
Random Forest Tree Classifier	97%

Figure 5: Accuracy Score for Each Model

5 Broader Impacts

Our results have important implications for the field of cancer diagnosis and treatment. The high accuracy achieved by our machine learning models suggests that these tools could be used as effective aids in the diagnosis of different types of cancer. In particular, the high accuracy of the random forest classifier model indicates that it may be particularly useful for identifying the specific type of cancer a patient has, especially for more rare cancers like Uveal Melanoma that medical professionals might not come across as often.

This could have several important benefits, such as reducing the time and cost of diagnosis, improving the accuracy of diagnoses, and ultimately improving patient outcomes. For example, early and accurate diagnosis is critical for successful treatment of cancer, and our models could potentially help clinicians identify the most effective treatment options for patients more quickly.

However, it is important to note that these tools should not replace the expertise of medical professionals, and any diagnosis made using these models should be confirmed by a physician. Additionally, there are potential risks associated with the use of machine learning models in medical contexts, such as privacy concerns and potential biases in the data used to train the models. These issues will need to be carefully considered and addressed as the use of machine learning in healthcare continues to evolve.

Despite these concerns, the potential benefits of machine learning in healthcare, particularly in cancer diagnosis, cannot be ignored. As the technology continues to improve and data collection becomes more comprehensive, it is likely that machine learning algorithms will play an increasingly important role in the diagnosis and treatment of cancer. It is important to continue monitoring and addressing potential concerns to ensure that the benefits of this technology can be realized while minimizing any potential negative impacts.

6 Conclusions

In conclusion, our results demonstrate the utility of machine learning models for cancer type prediction and the superiority of the RF model over the kNN model for this task. The high accuracies achieved by our models suggest that they may be valuable tools for clinicians in making accurate and efficient cancer diagnoses.

The goal of our exploration was to show the power of using machine learning models to predict cancer type given a patient's micro RNA profile. While neither the kNN or Random Forest models yielded perfect results, the models both proved that classifiers can predict the correct type of cancer at a high rate in comparison to a baseline. While our study slightly differed from Loher's 2017 investigation into this relationship, we found similar results, further implying the strength of predictive models for cancer diagnosis.

Our results suggest that this application of predictive models can potentially be expanded elsewhere in the medical field. Many other diseases are characterized by different factors such as symptoms, cellular inconsistencies, and more. Therefore, if researchers can hone in on the most important factors for other diseases, they could follow a similar structure to our investigation and hopefully yield similar results. While this may not work for each and every disease, there are definitely other diseases that can be diagnosed with the help of machine learning.

7 Contributions

Bryce and Edward did pair programming and worked on the data preparation together. Edward saved and added the figures. Bryce wrote the data preparation, experiments, and broader impacts sections. Edward wrote the abstract, introduction, results, and conclusion sections.

References

Loher, P. 2017. Knowledge about the presence or absence of mirna isoforms (isomirs) can successfully discriminate amongst 32 tcga cancer types. *National Library of Medicine* 45(6):2973–2985.