OncoRx: An Integrative Approach to Pan-Cancer Biomarker Identification and Targeted Cancer Multi-Drug Regimen Prediction

Darsh Mandera[1] and Lara Shamieh[1]
[1]Jesuit High School, Portland, OR 97229 USA (corresponding author to e-mail: dsmandera@gmail.com)
[1]L. Shamieh is with Jesuit High School, Portland, OR 97229 USA (e-mail: lshamieh@jesuitportland.org)

***Abstract*— The current practice of treating cancer is a one-size-fits-all approach, in which patients with the same type and stage of cancer receive the same treatment. This approach is ineffective 75% of the time. With microRNA (miRNA) having been identified as a key biomarker of cancer, precision therapeutics based on miRNA should provide the highest specificity and sensitivity by virtue of their cancer-specific expression and stability. However, identifying particular miRNAs that play a key role in driving cancer remains a challenge, as the expression of some types of miRNAs is found to be significantly different between normal tissues and tumor tissues. The focus of this research is to create a pan-cancer solution using machine learning to identify key miRNAs as biomarkers of cancer, and predict drug combinations based on miRNA. Data from 23 cancer types from The Cancer Genome was used. Top miRNAs were identified as key biomarkers using ExtraTreesClassifier, and were validated through functional enrichment analysis and survival analysis. Three different models were implemented using Multi-label ML algorithms: K-NearestNeighbors, AdaBoostClassifier, and OneVsRestClassifier, with OneVsRestClassifier yielding the highest accuracy. The final model was tuned using cross validation and a novel Median Scoring Method, based on F1 Score, Jaccard Score, and Accuracy Score. The resulting solution overcomes the challenges of monotherapy and allows oncologists to prescribe anti-cancer drug combinations with high accuracy based on patients' miRNAs, yielding higher survivability.**

I. Introduction

Cancer is a highly heterogeneous disease with complex underlying biology, with the genomic and transcriptomic heterogeneity of tumor cells being very high [1]. Despite the inherent variability and complexity, the current approach of treating cancer patients is not patient-specific [2]. While chemotherapy has been the standard treatment for various cancer types, studies have shown that it contributed to only 4.3% survival rate in United States [3]. This is because cancer cell manifestation is unique across each patient, resulting into varied treatment response [4].

The goal of precision medicine is to treat patients based on their unique biological attributes. In cancer, this entails predicting which treatment will be the most effective on a patient based on various biomarkers. Research has shown that miRNAs, which are small, single-stranded non-coding RNAs, have been found to be heavily dysregulated in cancer cells [5]. With microRNA (miRNA) being identified as a key biomarker of cancer in 2008 [6], miRNAs can easily be extracted from a patient's blood or urine and miRNA-based molecular diagnostics in cancer should provide the highest specificity and sensitivity by virtue of their cancer-specific expression and stability [7]. However, not all miRNAs play the same role in cancer, making it critical to identify miRNAs that drive cancer.

Machine learning, a branch of artificial intelligence that allows computers to "learn" from experience, has been used in various aspects of cancer prediction and prognosis [8, 9]. The majority of computational precision oncology research focuses on using machine learning programs with data from cancer cell lines – sets of cancer cells that grow and divide in a laboratory. Others approaches use humanized mice models – human xenografts and immune cells grafted into lab mice [10]. Although these may be effective models, it is impossible for any model to fully capture the complexity of real patient data. In addition, most research focuses on predicting single drugs for patients, as opposed to drug regimens, which can offer improved therapeutic outcomes.

This research is focused on offering oncologists a machine learning-based tool that uses patient data from The Cancer Genome Atlas (TCGA), identifies individual miRNAs as key biomarkers of cancer, and predicts targeted drug regimens that will be effective for a cancer patient based on these identified miRNAs.

II. Methods

The overall methodology is a four-stage pipeline model: data preparation, identification of key miRNAs using machine learning, validation of biomarkers through functional enrichment analysis and survival analysis, and implementing of a multi-label machine learning model to predict drug combinations. This machine learning pipeline was built using Python and its libraries.

Data of 23 major cancer types from TCGA, which is a public cancer data repository curated by the National Institutes of Health and Broad Institute, was retrieved. The data included drugs administered, patients' drug response, patient lifestyle, patient follow-up data, patient demographics, and temporal data of the study, as well as the miRNA profile of each patient. From this data, the drugs each patient received and their responses to those drugs was extracted. The drug responses are categorized into four groups: complete response, partial response, clinical progressive disease, and stable disease, which were encoded as 1, 0.5, 0, and 0 respectively in this study. Each patient's miRNA alteration was also calculated from the data, by comparing each patient's miRNA count to that of a healthy human. Overall, the data contained 6,280 patients, 705 miRNAs, and 230 drugs.

In order to identify which particular miRNAs were playing a role in driving cancer, feature selection was employed through the ExtraTreesClassifier. This algorithm selects the miRNAs that have the strongest impact on which drug is selected for a patient. Feature selection helped reduce the features from 705 miRNAs to 84 key driver miRNAs. To validate that the miRNAs selected through feature selection were biologically significant, functional enrichment analysis (which demonstrates the role of miRNAs in cancer pathways) and survival analysis (which demonstrates the role of a miRNA alteration in patient survival) was undertaken. These analyses demonstrated that the 84 selected miRNAs played a critical role in the biology of cancer, as well as in a patient's survival over time.

To build a predictive model, the patient data was split into two parts: 70% for model training, and 30% for model testing. A predictive model using multi-label algorithms was built. Multi-label algorithms allow for predictions involving multiple outputs, which was critical for a program seeking to predict combinations of two drugs, as opposed to single drugs. Three such algorithms were implemented: K-NearestNeighbors, AdaBoostClassifier, and OneVsRestClassifier. Cross validation was implemented through the K-Fold and GridSearchCV algorithms, which allowed for iterative training and hyperparameter optimization, contributing to a higher final prediction accuracy.

A novel statistical method, called the Median Scoring Method, was devised and implemented in order to reduce the impact of outlier accuracy values during hyperparameter optimization. While using a single accuracy metric is prone to outlier measurements, combining three metrics and finding their median value reduces the impact of outliers. The Median Scoring Method calculates the median value of three different accuracy metrics – accuracy score (which measures the ratio of correct predictions to total predicts), Jaccard score (which measures the similarity between the set of correct predictions and all predicted and true values), and F1 score (which balances precision and recall), and thereby reduces the impact of outliers.

III. Results and Discussion

The machine learning model implemented with three learning algorithms (K-NearestNeighbors, AdaBoostClassifier, or OneVsRestClassifier), alongside cross validation through K-Fold and GridSearchCV, and the Median Scoring Method, yielded varying degrees of prediction accuracy based on the learning algorithm used. AdaBoostClassifier and OneVsRestClassifier are tree-based algorithms, meaning they base their predictions based on a series of tree-like flowcharts. Hence, the prediction accuracies for these two algorithms are measured at varying tree depths, which is how far along the flowchart the accuracy is obtained. On the other hand, K-NearestNeighbors takes number of neighbors as a parameter, and hence the accuracy is measured at varying number of neighbors. While the models based on K-NearestNeighbors and AdaBoostClassifier were implemented and tuned, OneVsRestClassifier achieved the highest prediction accuracy of 92% at a tree depth of 95 (Fig. 1). Hence, the final pipeline was implemented with the OneVsRestClassifier.

TABLE I. Prediction Accuracy by Algorithm

Maximum Accuracy by Algorithm		
K-NearestNeighbors	***AdaBoost Classifier***	***OneVsRest Classifier***
68.7% (neighbors = 9)	82.5% (depth = 40)	92.1% (depth = 95)

Figure 1. The maximum accuracy for each algorithm, and with which parameters it was achieved.

IV. Conclusions

OncoRx offers oncologists a precision cancer therapeutics tool that will improve clinic practice by improving patient outcome, as it offers real-time prediction with over 92% accuracy, compared to the 25% effectiveness of the current treatment approach. OncoRx takes a unique approach of training on miRNA data of real patients across 23 cancer types, as well as predicting drug regimens as opposed to single drugs, improving on current research methods which focus on using cancer cell line data and making single-drug predictions.

V. Acknowledgment

D.M. would like to thank Joyesh Mishra for his guidance and support.

VI. References

[1] S. C. P. Williams, "News feature: Capturing cancer's complexity." *Proceedings of the National Academy of Sciences of the United States of America*, Apr. 2015.

[2] P. Krzyszczyk, A. Acevedo, E. J. Davidoff, L. M. Timmins, I. Marrero-Berrios, M. Patel, C. White, C. Lowe, J. J. Sherba, C. Hartmanshenn, K. M. O'Neill, M. L. Balter, Z. R. Fritz, I. P. Androulakis, R. S. Schloss, and M. L. Yarmush, "The growing role of precision and personalized medicine for cancer treatment." *Singapore World Sci*ence, Sep. 2018.

[3] G. Morgan, R. Ward, and M. Barton, "The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies." *Clinical Oncology*, Dec. 2004.

[4] I. Burney and R. Lakhtakia, "Precision Medicine: Where have we reached and where are we headed?" *Sultan Qaboos University Medical Journal*, Oct. 2017.

[5] Y. Peng and C. M. Croce. "The role of MicroRNAs in human cancer." *Sig Transduction and Target Therapy*, Jan. 2016.

[6] C. H. Lawrie, S. Gal, H. M. Dunlop, B. Pushkaran, A. P Liggins, K. Pulford, A. Banham, F. Pazella, J. Boultwood, and J. Wainscoat. "Detection of elevated levels of tumour-associated micrornas in serum of patients with diffuse large B-cell lymphoma." *British Journal of Haematology*, May 2008.

[7] M. Ferracin, A. Veronese, and M. Negrini, "Micromarkers: miRNAs in cancer diagnosis and prognosis." *Expert Review of Molecular Diagnostics*, Jan. 2014.

[8] L. Bocchi, G. Coppini, J. Nori, and G. Valli, "Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks." *Medical Engineering and Physics*, May 2004.

[9] H. Yasrebi, "Comparative study of joint analysis of microarray gene expression data in survival prediction and risk assessment of breast cancer patients." *Briefings in Bioinformatics*, Sep. 2016.

[10] A. Kalamara, L. Tobalina, and J. Saez-Rodriguez, "How to find the right drug for each patient? Advances and challenges in pharmacogenomics." *Current Opinion in Systems Biology*, Aug. 2018.