参赛队员姓名：郑泓 Ethan Zheng

中学: Dulwich College Beijing

省份:北京

国家/地区:中国

指导教师姓名：Jason Wilson

指导教师单位：Science Head of Department

论文题目： In silico method for determining cancer diagnosis from patient blood

Author: Ethan Zheng

Abstract:

Cancer has been a prevalent medical concern among many scientists, and within cancer, the specific causes and treatment methods still have a comparatively low recovery rate. MicroRNAs (miRNAs) are endogenous non-coding functional RNAs that regulate gene expression by inhibiting/promoting certain signaling pathways.[5] They could be a potential indicator of cancer and can be detected from miRNA screening of patients' blood samples. This indicator could allow scientists to determine potential cancer victims at a very early stage and begin targeted therapy, or early treatment, which could be what makes the difference between a full recovery and no recovery. In this project, we aim to improve the understanding of gene expression in relation to cancer, using machine learning to identify miRNAs with a high relatedness to cancer and find pathways connected to this relatively novel field.

Key Words:

miRNA

Machine Learning

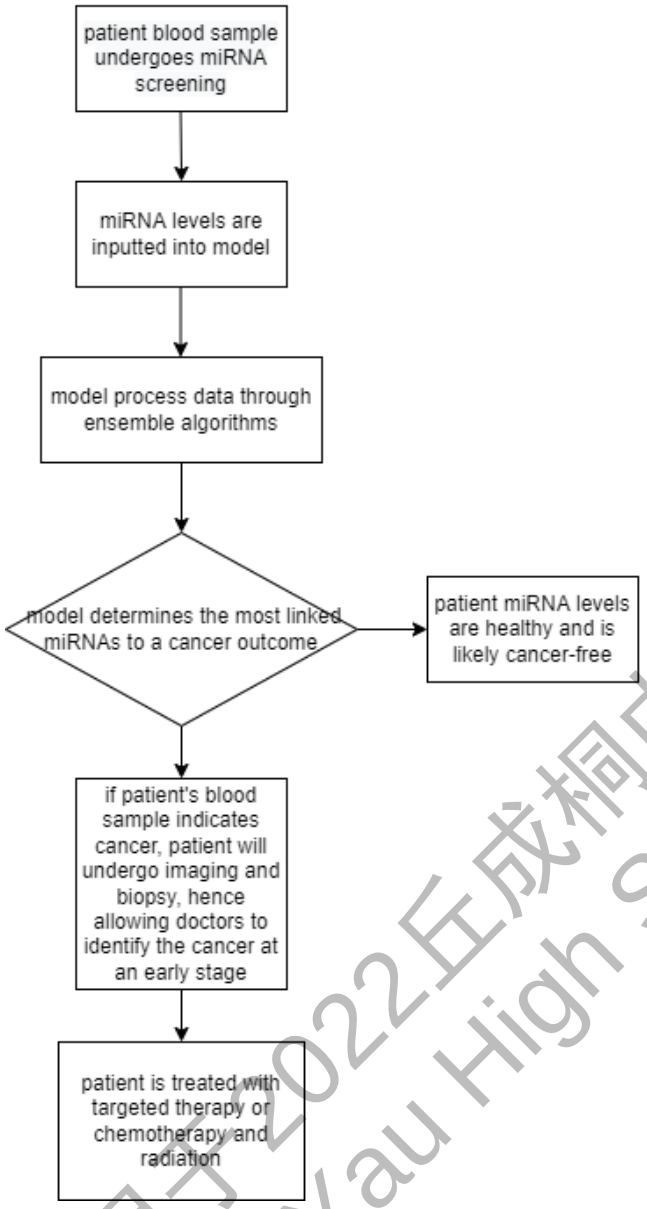
Liquid Biopsy

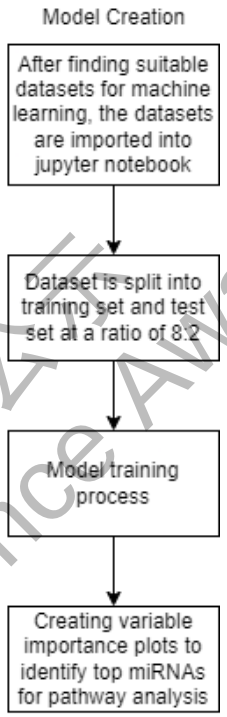
Lung Cancer

Esophageal Cancer

Gastric Cancer

Table of Contents





1. Introduction

1.1 Background

Lung cancer(25%),[9] Esophageal cancer(5%),[12] and Gastric cancer(8.3%)[3] are responsible for around 38.3% of cancer-related deaths, a large percentage of overall cancer-related deaths.

Lung cancer mainly refers to tumors growing in the parenchyma or within the bronchi.[9] It has proven to be a prevalent disease, being the second most common disease to occur both in men and women (above which is prostate cancer in men and breast cancer in women). It is the main holder of cancer-caused death (25% of all cancer-caused deaths), more than breast cancer and prostate cancer combined.[12]

The singular leading cause of lung cancer is smoking, which results in approximately 90% of lung cancer cases.[12] However, it is not the only cause, other notable factors include air pollution, radiation, secondary smoking, and other lung-affecting diseases, such as pulmonary tuberculosis. These lung diseases increase the chance of the patient developing lung cancer.

The reason lung cancer is so deadly is due to the ineffectiveness of treatments, mainly chemotherapy because lung cancer is particularly resistant to it. More than two-thirds of patients are only diagnosed at late stages of lung cancer, meaning that lung cancer has already spread too far past the point of treatment, leading to the survival rates of lung cancer being extremely poor.[9]

Esophageal cancer is a malignancy with a staggeringly low survival rate, even when the patient undergoes treatment. Currently, esophageal cancer is the sixth most common cancer in the world. The amount of people getting esophageal cancer each year has actually been increasing., and esophageal squamous cell carcinoma is generally caused by smoking, alcohol, or extremely unhealthy diets.[6]

Gastric cancer is the fifth most prominent cancer in the world, and third in cancer-caused deaths.[10] The reason the cancer-caused deaths are so high is due to the fact that there are not a lot of successful treatments for gastric cancer.

Due to this issue, most patients are instead offered life-prolonging palliative treatment. As of the moment, surgery is the only way to fully cure gastric cancer and is dangerous and expensive.

Another hindering factor is that cancer is also very difficult to diagnose.[12] The most obvious way to do it is a physical examination, where the doctor will check the outer body for abnormalities such as skin deformation or lumps, but it is hard to detect abnormalities in the early stages of cancer. Then there are screening methods such as CT and MRI scanning, but the most used is a biopsy, where doctors collect a tissue sample from the body, and test it in a laboratory.

A potential new method for diagnosing cancer is looking at circulating RNAs in the patient's blood, which there has been a lot of ongoing research in the past decade

1.2 Context

MicroRNAs (miRNAs) are endogenous non-coding functional RNAs that regulate gene expression.[4] They are essential to biological processes such as cell division and cell differentiation, cell death, and DNA repair.[5] But in the context of cancer, miRNAs could act as tumor suppressors and promoters, but also could be an indicative factor of the presence of cancer.

Failure to regulate miRNAs is a possible cause of non-small cell lung cancer (NSCLC), gastric cancer, and esophageal cancer.[5] If miRNAs are aberrantly expressed, they could inhibit the pathway that regulates cancer, leading to tumors forming. When miRNAs are overexpressed, they could inhibit processes in the cells that regulate tumor suppression. When miRNAs are underexpressed, there could be overly high amounts of protein produced, and this could also lead to cancer.

1.3 Goals

The major goals of this project were to improve understanding of gene expression in relation to disease and cancer. By using data on miRNA blood levels that are publicly available from real patient blood samples, we can then begin to use machine learning to identify the most prominent cancer-inducing miRNAs. We can then analyze the shared targets between these miRNAs and find novel pathways. A primary goal is to be able to produce a small simple model that maintains high diagnostic accuracy, a potential development for the therapeutic field.

2. Materials and Methods

2.1 Data Collection

Data Collection began with sourcing datasets from GEO datasets (https://www.ncbi.nlm.nih.gov/gds). The criteria that we looked for was that the dataset would contain at least 500+ samples and contain both cancer and control patients. If there was more than one malignancy in a dataset, we only focused on the primary cancer.

We chose datasets based on a few criteria. First of all, we looked for large amounts of data points for both negative-controls and cancer-positive patients. To increase the accuracy of training, we created balanced sets of each dataset, where the number of control patients was the same as cancer patients, this was done randomly so as not to bias. We chose to look at a binary outcome rather than a multivariate outcome by focusing on one cancer per model, to make our analysis less convoluted. All three of the datasets come from New Frontiers Research Laboratory in Kanagawa, Japan. This laboratory used reagents and an array chip developed by Toray Industries, meaning that the same miRNAs were screened for each site. The reason for this is taking all three datasets from the same provider allows us to ensure standardized testing methods and accurate results.

The process begins with data processing. This includes organizing and transposing the matrix and splitting it into frames for training. Then we import the data onto Jupyter notebook for the model.

First Dataset:[1] GSE122497 Esophageal cancer, Homo Sapiens, 5531 samples, 566 Esophageal cancer patients, 4965 non-cancer controls. Data was obtained using Serum miRNA profiling

Second Dataset:[2] GSE137140 Lung cancer, Homo Sapiens, 3924 samples, 1566 preoperative Lung cancer patients, 180 postoperative Lung cancer patients 1774 non-cancer controls. Data was obtained using Serum miRNA profiling

Third Dataset:[3] GSE164174 Gastric cancer, Homo Sapiens, 2940 samples, 1423 Gastric cancer patients, 50 Esophageal patients, 50 Colorectal cancer patients, 1417 non-cancer controls. Data was obtained using Serum miRNA profiling

2.2 Model development

We began model development by using the Google Collab notebook:

Pycaret was initially used as the platform for machine learning, and we had to install packages including pandas, NumPy, warnings, and Pycaret. These packages are the basis of all other codes.

The next step was to import and set up the dataset, the term binary refers to the column that shows 1 if the patient has cancer, or 0 if they are a control. For the training, 80% of the dataset was used (the training set) and the other 20% was used for verifying and refining the model (test set).

After the dataset was set up we proceeded to explore different machine learning algorithms to compare their levels of accuracy, the value that was used for overall model accuracy was the 'AUC', which is the

amount of area under the ROC curve. It is the measure of how well the classifier can distinguish different classes, and the higher the value of the AUC, the better the performance of the model.

The next step was model evaluation and predictions to finally get the feature importance plots and partial dependence plots for the miRNAs However, a bug on the developer's side resulted in an error. This was when we switched the platform to another machine learning platform called H2O.

Same as before while using Pycaret, installing and importing packages was the first step of new code. Then we imported the dataset into Jupyter.

Just as before, the dataset was split into an 80% train set and a 20% test set

After that, instead of training on Pycaret, we trained on H2OAutoML. After training was completed, the explain function revealed the feature importance plots, the confusion matrix, and the partial dependence plots.

Creating a feature importance plot helped to identify the most predominant miRNAs and to be able to perform pathway analysis on their gene targets. The confusion matrix shows the accuracy of the model, and the partial dependence plots reveal the levels of how much these miRNAs are expressed to yield certain results (cancer or cancer-free).

After the training and the plots were generated, we consolidated each model by taking the top 2 miRNAs and training a model using a dataset using only the top 2. This allowed us to create a model that used only 2 features to build the model versus the many more in the original models.

2.3 Pathway Analysis

We found the gene targets of the most prevalent miRNAs using the model and used mirBase to find the gene targets. Using Excel, we were able to deduce the shared gene targets between the different

miRNAs. Then, we used pathway analysis to find protein targets, and deduce what effect miRNAs had on tumor development/suppression.

Using Reactome (the program used for pathway analysis, https://reactome.org/), we identified potential pathways by identifying pathways with a high "entity found" value, meaning a high number of gene targets in a pathway. After locating novel pathways, we were able to find pathway diagrams of the entire web of signaling targets that a pathway has, then we could begin to create pathway plots based on specific entities present within the web of signaling targets.

2.4 Therapeutic proposal

Using this model, we hope to be able to accurately identify cases of lung, esophageal, and gastric cancer among patients at a very early stage. This could potentially allow early intervention and insight into potential targeted therapy. For example, several targeted therapies are approved for lung cancer, [14] such as immunotherapy. In development are antisense oligonucleotides, an experimental strategy to approach cancer. The overall concept of this is that antisense oligonucleotides bind to the miRNA which then prevents miRNA from binding to mRNA, hereby potentially able to limit protein generation and the pathogenesis of many diseases.[13] When the model identifies oncogenic pathways, antisense oligonucleotides could be a future option, as long as delivery systems are properly applied.

Another potential method of treatment that can be applied to early-diagnosed cancer patients is the bifunctional molecule RIBOTAC, a chimera molecule that targets miRNAs by recruiting a ribonuclease.[8] This was designed to work against breast cancer by exterminating breast cancer oncogenic miRNAs to cause rapid malignant cell death, but could be repurposed to target other miRNAs.[11]

2.5 Obstacles

During the project, we encountered a few major obstacles that set back progress. First of all, there was a Pycaret update that caused a bug in the software, which led to the original code based on Pycaret not being able to run without error. After a long debugging process and eventually realizing the issue was not on our end of the code but rather a bug that was not yet fixed, we switched to an alternative called H2O. During the process of model creation on H2O, we also got many errors during the training process due to incompatibilities with syntax inside the imported matrices, we fixed this by standardizing the datasets so that they could be compatible with the functions.

Apart from the coding process, sorting through the noise and the messiness of the pathway analysis website (Reactome) proved to be a very tedious task.

3. Results

The final results consist of a few components: ROC plots, variable importance plots, partial dependence plots, and potential novel pathway plots.

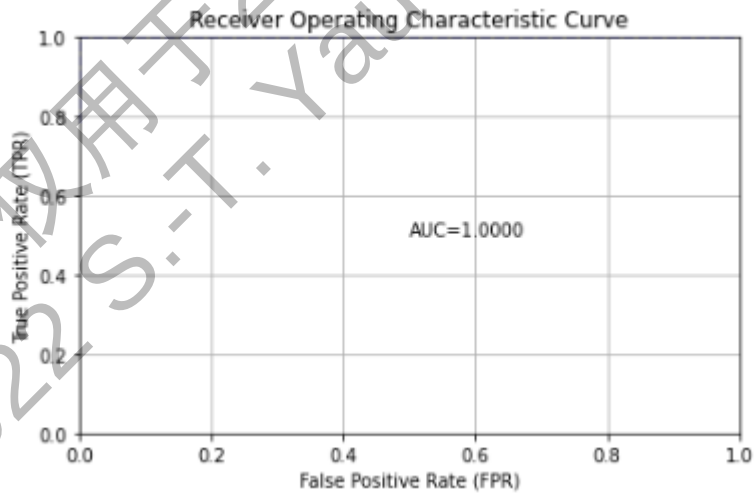


Figure 3.1 The ROC Curve above is representative of all 3 models, this shows an extremely high model accuracy rate for the datasets we trained on

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.5777988994718207:

	Esophageal_cancer	Non-cancer	Error	Rate
Esophageal_cancer	110.0	0.0	0.0	(0.0/110.0)
Non-cancer	0.0	117.0	0.0	(0.0/117.0)
Total	110.0	117.0	0.0	(0.0/227.0)

Figure 3.2 The confusion matrix for the esophageal cancer model. The confusion matrix depicts the rate of the inaccuracy of the models. From this example, there was no inaccuracy when the model used the test set. There could, however, be a rating that is lower if we used a larger dataset (10,000+ data points)

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.9993215949255042:

	Lung cancer	Non-cancer	Error	Rate
Lung cancer	317.0	0.0	0.0	(0.0/317.0)
Non-cancer	0.0	325.0	0.0	(0.0/325.0)
Total	317.0	325.0	0.0	(0.0/642.0)

Figure 3.3 Confusion matrix for the Lung Cancer model

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.9373178332975879:

	Control	Gastric Cancer	Error	Rate
Control	293.0	0.0	0.0	(0.0/293.0)
Gastric Cancer	0.0	284.0	0.0	(0.0/284.0)
Total	293.0	284.0	0.0	(0.0/577.0)

Figure 3.4 Confusion matrix for the Gastric Cancer model

From the above images, we can get an overall impression on the high level of accuracy the model possesses. To put this into perspective, the plot below demonstrates the AUC curve of random miRNAs taken from a dataset.

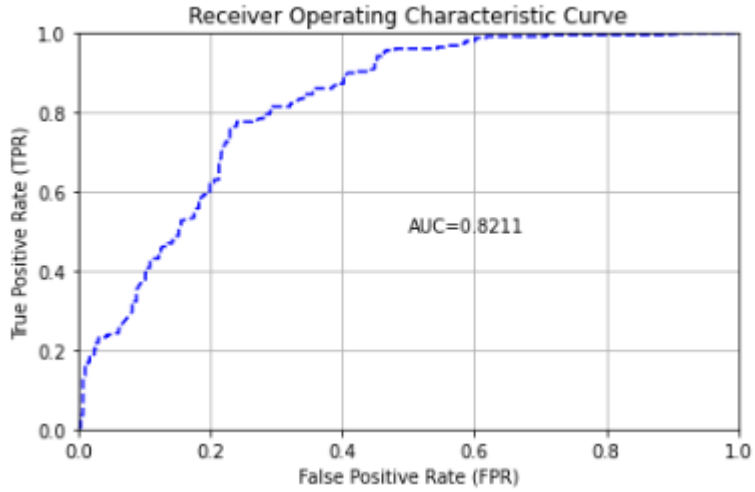


Figure 3.5 Control AUC curve to provide information on the default levels of AUC

On top of using a control AUC curve to compare how accurate the AUC curves we obtained were, we took further measures to ensure unbiased results by taking the top 2 miRNAs from each cancer, and running them in a separate dataset. If this still yielded high AUC levels then it shows that the top 2 miRNAs are in fact prevalent in the cancer.

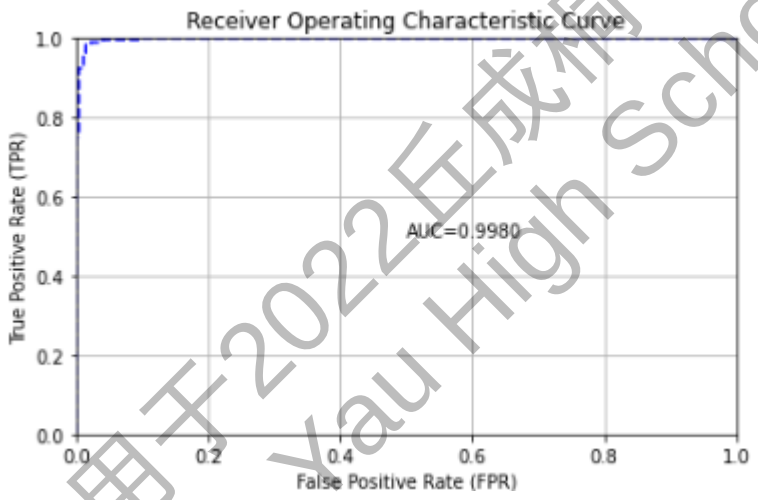


Figure 3.6 AUC curve for the top 2 miRNAs from Gastric cancer. A high AUC remains

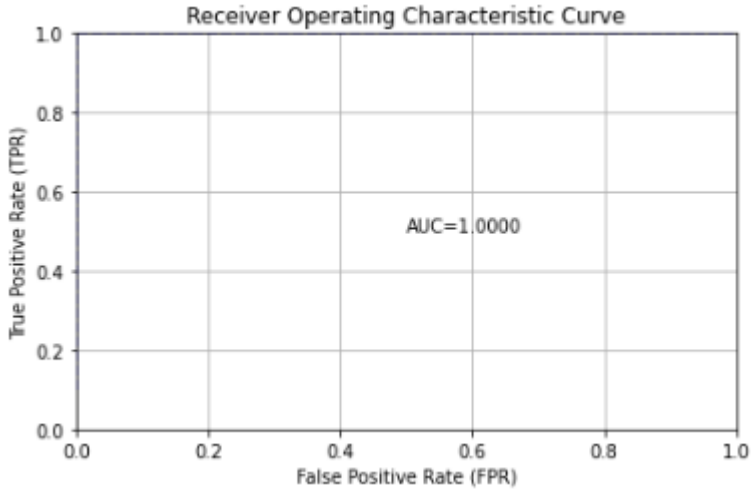


Figure 3.7 AUC curve for both lung and esophageal top 2 miRNAs (they were the same)

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.42189956390750893:

	Control	Gastric Cancer	Error	Rate
Control	761.0	360.0	0.3211	(360.0/1121.0)
Gastric Cancer	105.0	1052.0	0.0908	(105.0/1157.0)
Total	866.0	1412.0	0.2041	(465.0/2278.0)

Figure 3.8 Confusion matrix for the Gastric cancer consolidative model

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.07191671406388075:

	Lung cancer	Non-cancer	Error	Rate
Lung cancer	310.0	2.0	0.0064	(2.0/312.0)
Non-cancer	0.0	304.0	0.0	(0.0/304.0)
Total	310.0	306.0	0.0032	(2.0/616.0)

Figure 3.9 Confusion matrix for the Lung and Esophageal cancer mixed consolidative model

After the consolidation and verification process, we moved on to the feature importance plots from the models.

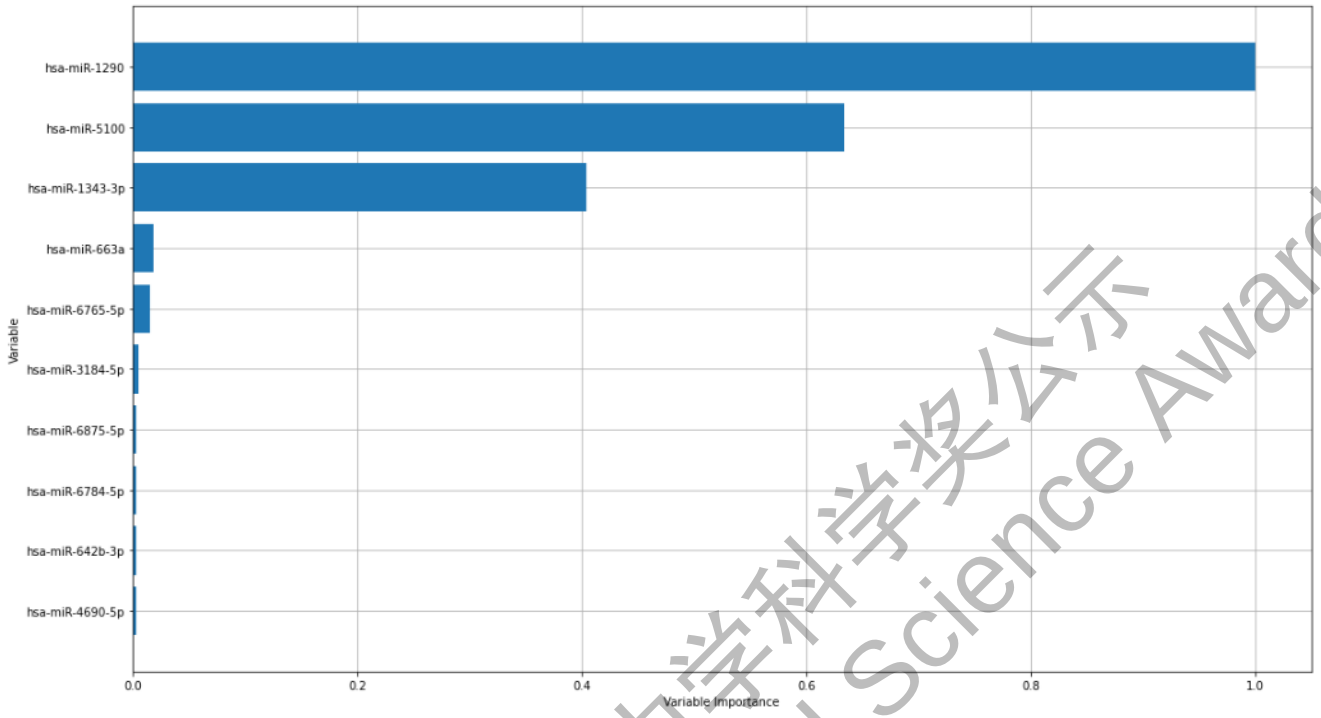


Figure 3.10 Variable importance plot generated from lung cancer model. The top 2 miRNAs were used for pathway analysis

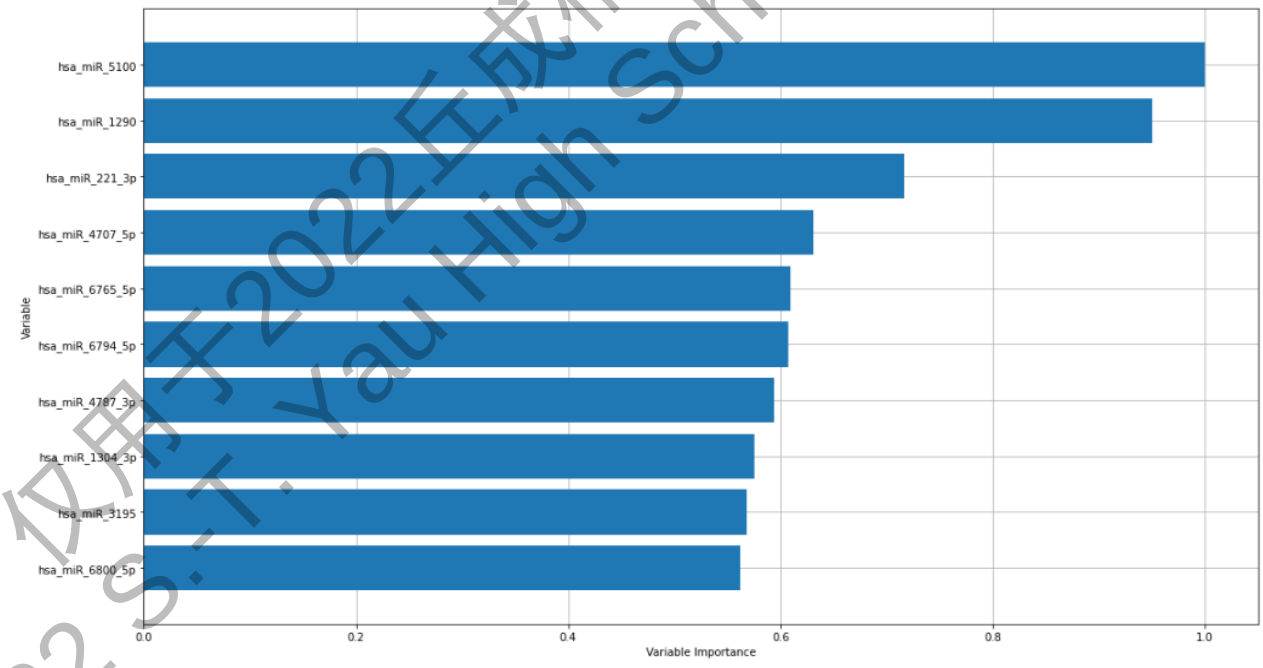


Figure 3.11 Variable importance plot generated from esophageal cancer model

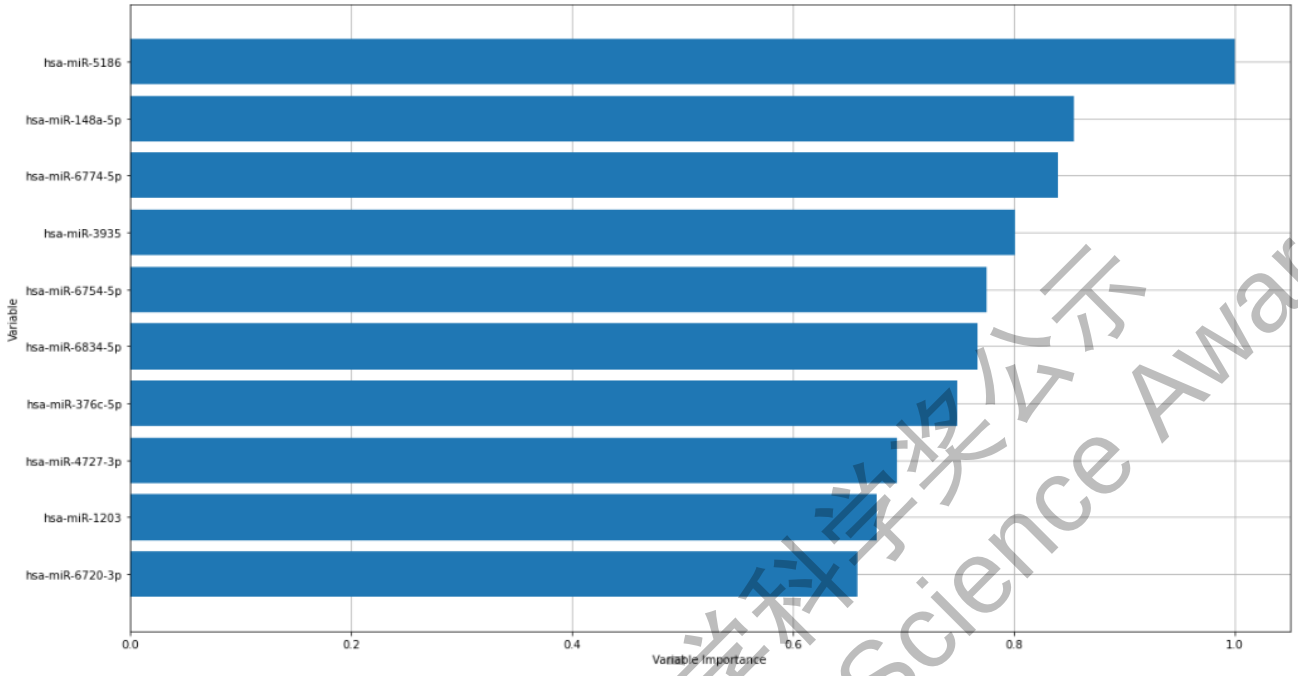


Figure 3.12 Variable importance plot generated from gastric cancer model

The variable importance plots allowed us to identify the important miRNAs, and right after that we also obtained partial dependence plots. Variable importance plots were used for pathway analysis, which we performed on the top 2 miRNAs for the 3 different models.

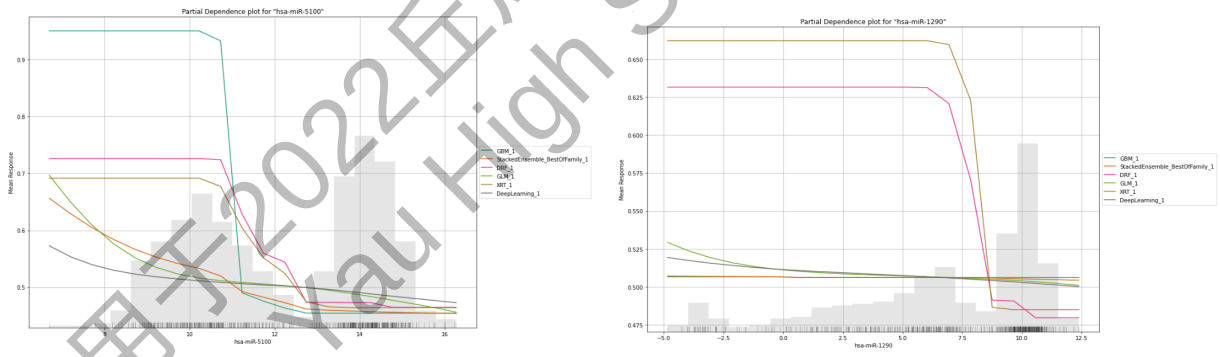


Figure 3.13 Partial dependence plots for top 2 miRNAs in the lung cancer variable importance plot. Almost all of the different algorithms (different colored lines) have the same general trend, which further solidifies the plots. (whether they are overexpressed or underexpressed). In this plot, the higher expression of the miRNA leads to a lower mean response (y-axis), which means a higher frequency of cancer patients. The miRNA is underexpressed in cancer.

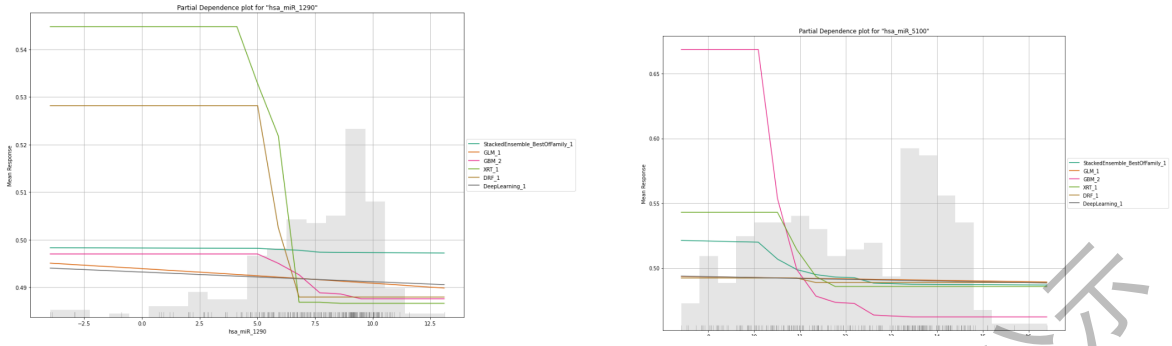


Figure 3.14 Partial dependence plots for top 2 miRNAs in the esophageal cancer variable importance plot. A lower mean response (y-axis) means a higher frequency of cancer patients.

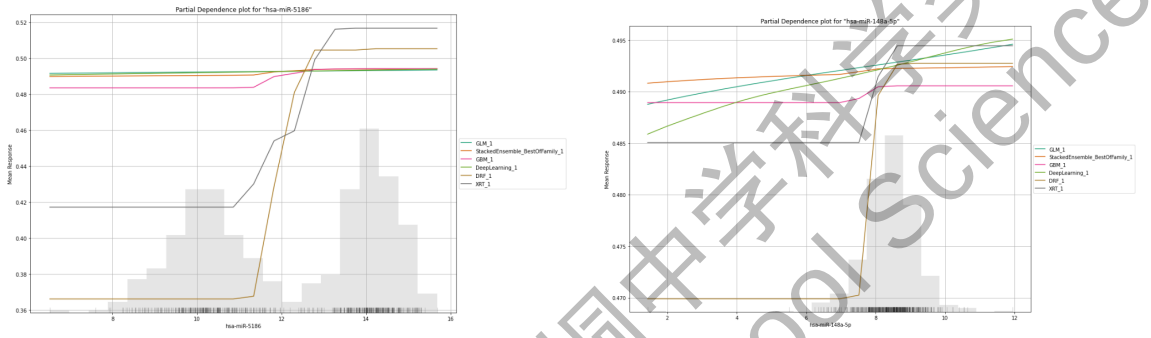


Figure 3.15 Partial dependence plots for top 2 miRNAs in the gastric cancer variable importance plot. In contrary to the other miRNA's plots, the mean response for the gastric cancer model reversed, meaning that as the expression of the miRNA increased, the frequency of cancer patients got higher.

After this, taking the top miRNAs, putting them into mirBase, and getting gene targets, we then found overlapping gene targets within the top 2 miRNAs and performed pathway analysis. Pathway analysis yielded certain high-entity pathways, which allowed us to generate these pathway plots (below)

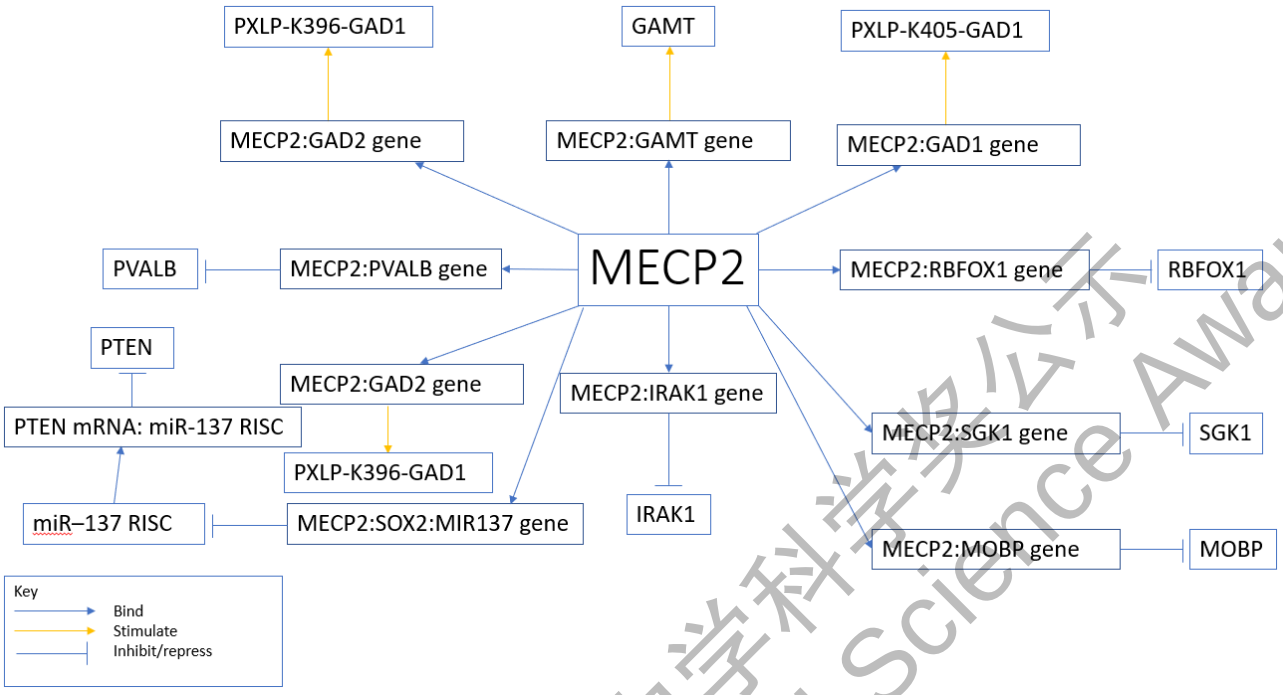


Figure 3.16 The MECP2 pathway appeared multiple times with high entity counts in pathway analysis of gastric cancer. The graph depicts the gene targets that the miRNA contains, and their connection to the MECP2 pathway.

4.1 Magnitude of impact

This model can be used side-by-side during health checkups because miRNA screening can take place as soon as blood samples are extracted. The impact of this could be significant for two main reasons. Firstly, in areas where access to healthcare is more limited, or inaccessible, a non-invasive and effective method will enable cancer to be diagnosed during health checks. Secondly, the early diagnosis of cancer is crucial to prevent the progression of symptoms, which would also in turn increase the survival rate. Although this method of testing still requires the patient to attend regular health checks, it is still a great increase in accessibility compared to those who do tests for cancer only. Waiting for symptoms to appear before seeking treatment is detrimental to chances of recovery, and recovering at later stages could cause permanent disabilities. Additionally, a less invasive method for cancer detection could motivate many who were deterred by the invasiveness of traditional cancer detection to partake in testing.

A single miRNA binds to many gene transcripts, therefore it is hard to elucidate the pathways that miRNAs regulate because of how many genes they regulate. Through our analysis, we were able to find miRNAs that have a shared relationship with cancer, in terms that they are linked to a similar cancer outcome. They may not both be over/underexpressed, but they are both linked to a cancer outcome, by then comparing the targets of these miRNAs, we can narrow down those which are in common. Performing a pathway analysis allows us to find novel pathways that are linked to these miRNAs.

4.2 Further study

In the future, a few immediate goals would be to implement a wider range of cancer types into the model. This would allow for wider coverage, as the current model is limited to esophageal, gastric, and lung cancer. While more cancers are included, diagnostic accuracy must remain the same, or even higher. Currently, the datasets are limited to below 5,500 data points for training, an increase in data would be beneficial to increasing diagnostic accuracy as well. On top of this, the model could potentially develop into a different field, also involving miRNAs, such as autoimmune diseases (when the immune system targets native body cells). Eventually, if the model reaches a high-enough point of accuracy, it could be implemented as a tool for doctors to use alongside their own diagnostic abilities, to further refine the process of the detection of cancer among the population.

4.3 Conclusion

Overall, this project serves as a hypothetical model for a future diagnostic method for cancer treatment. It demonstrates how machine learning can be integrated into hospitals for therapeutic usage, and how doctors can use this tool to diagnose cancer sufferers at an earlier stage, which could be greatly beneficial to increasing overall coverage of cancer treatment, and lowering treatment costs.

5. References

1.Sudo K, Kato K, Matsuzaki J, Boku N et al. Development and Validation of an Esophageal Squamous Cell Carcinoma Detection Model by Large-Scale MicroRNA Profiling. *JAMA Netw Open* 2019 May 3;2(5):e194573. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122497

2.Asakura K, Kadota T, Matsuzaki J, Yoshida Y et al. A miRNA-based diagnostic model predicts resectable lung cancer in humans with high accuracy. *Commun Biol* 2020 Mar 19;3(1):134 Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137140

3.Abe S, Matsuzaki J, Sudo K, Oda I et al. A novel combination of serum microRNAs for the detection of early gastric cancer. *Gastric Cancer* 2021 Jul;24(4):835-843 Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164174

4.Liu B, Shyr Y, Cai J, Liu Q. Interplay between miRNAs and host genes and their role in cancer. Brief Funct Genomics. 2018 Jul 22;18(4):255-266. doi: 10.1093/bfgp/elz002. PMID: 30785618; PMCID: PMC6609535. Available from: https://pubmed.ncbi.nlm.nih.gov/30785618/

5.Peng Y, Croce CM. The role of MicroRNAs in human cancer. Signal Transduct Target Ther. 2016 Jan 28;1:15004. doi: 10.1038/sigtrans.2015.4. PMID: 29263891; PMCID: PMC5661652. Available from: https://pubmed.ncbi.nlm.nih.gov/29263891/#:~:text=MiRNAs%20may%20function%20as%20either,and%20metastasis%2C%20and%20inducing%20angiogenesis.

6.Mukkamalla SKR, Recio-Boiles A, Babiker HM. Esophageal Cancer. [Updated 2022 Jul 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459267/

7.Mukkamalla SKR, Recio-Boiles A, Babiker HM. Gastric Cancer. [Updated 2022 Jul 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459142/

8.Kargbo, Robert B. "RIBOTACs: Small Molecules Selectively Destroy Cancer-Associated RNA." *ACS Publications*, 8 Nov. 2021, https://pubs.acs.org/doi/10.1021/acsmedchemlett.1c00576.

9."Lung Cancer: The World's Deadliest Cancer." *Roche*, https://www.roche.com/stories/about-lung-cancer.

10.Rawla, Prashanth, and Adam Barsouk. "Epidemiology of Gastric Cancer: Global Trends, Risk Factors and Prevention." *Przeglad Gastroenterologiczny*, Termedia Publishing House, 28 Nov. 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6444111/.

11.Satyanarayana, Megha. "Small-Molecule Selectively Destroys Cancer-Associated RNA." *Cen.acs.org*, 26 Aug. 2021, https://cen.acs.org/acs-news/acs-meeting-news/Small-molecule-selectively-destroys-cancer/99/i31#:~:text=The%20researchers%20found%20that%20an,also%20associated%20with%20miR%2D21.

12.Siddiqui, Abdul H. "Lung Cancer - StatPearls - NCBI Bookshelf." *National Library of Medicine*, 5 May 2022, https://www.ncbi.nlm.nih.gov/books/NBK482357/.

13.Stahel, Rolf A. “Antisense Oligonucleotides for Cancer Therapy—an Overview.” *Lung Cancer Journal*, 1 Aug. 2003, https://www.lungcancerjournal.info/.

14.“Non-Small Cell Lung Cancer Targeted Drug Therapy: Lung Cancer Drugs.” *American Cancer Society*, 15 Aug. 2022, https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/targeted-therapies.html#:~:text=Larotrectinib%20(Vitrakvi)%20and%20entrectinib%20(,pills%2C%20once%20or%20twice%20daily.

Thank you,

to my teacher, who aided me when I was getting bugs and errors during coding, and assisting me with paper-writing.