Discovering Potential Biomarkers for Uterine and Cervical Cancers with Machine Learning

Genevieve Ferguson
Knight Foundation School of Computing and Information Sciences - MS in Data Science and AI

Florida International University
Miami, Florida, USA
gferg020@fiu.edu

Abstract

Genetic information holds the power to give insight into the human condition- even revealing potential for illness in the future. Oncogenes, or cancer-causing genes, and tumor suppressing genes have been found to play a major role in the development of cancer. Though the scientific and medical communities are aware of several biomarkers for cervical and uterine cancers, this does not indicate the impossibility of others existing. In this study, three datasets of copy number alterations (CNA), gene expression, and microRNA (miRNA) were used, as well as one combined dataset consisting of overlapping participants in all three datasets. These data were used to find previously discovered biomarkers as well as new potential biomarkers via the feature importance of several different types of classification models. The important features were determined by SHapley Additive exPlanations, or SHAP values. The values were indicative of the overall impact of each feature for the model.

Overall, it was found that among the three datasets, the most promising evidence was found in that of the miRNA, HtSeq, and Combined datasets. hsa-mir-224, ENSG00000269899.1, ENSG00000274501.1, ENSG00000215267.7, ENSG00000215030.5, ENSG00000225131.2, and ENSG00000128228.4 were found as potential candidates for biomarkers. While these are currently biomarkers for either other cancers or no cancers at all, further research may be able to reveal their impact on the cancers evaluated in this study, uterine and cervical.

Keywords— biomarker, feature importance, explainable AI, SHAP, CNA, gene expression

I. Introduction

Though there have been a number of breakthroughs in recent years regarding cancer treatment, research must continue in order to find more effective methods of treatment and eventually a cure. While oncogenes were discovered over 50 years ago [36], they still remain a vital subject of interest in the search for better cancer treatments and prevention. A particular method of study involves the use of genetic information in order to determine a person's predisposition for developing specific types of cancer. This predisposition for disease was coined as a 'biomarker'. In machine learning, classification models can be used to represent the difference between classes in a dataset.

Machine learning models are able to recognize patterns in data that would not have been discoverable with the simple naked eye. They can then use this large amount of data to make predictions. Features refer to the columns in a dataset, or the categories of data of a sample. Though all of the features are considered when training a machine learning model, some contribute to the model in different strengths. Therefore by determining the feature importance of each feature in these genetic datasets, potential biomarkers for each cancer are revealed. These feature importances were determined using SHAP values, an explainable AI approach to feature importance [28]. These values are calculated using a game theory technique that calculates the contribution of each feature to the final prediction. This technique is particularly valuable in that it is model-agnostic [28], meaning that it can be used to interpret any machine learning model, regardless of the type. Though many models are black-box, SHAP is able to give insight into what is actually going on behind the scenes.

In order to properly explain all biological elements included, the introduction will be approached in a basic to detailed manner. This ensures that the reader, regardless of familiarity with the subject, will have learned something of value. First, the basic components of genetic material will be overviewed. Next, the types of genetic material used in this study will be described. Then, the types of cancer will be outlined. Finally, the aims, motivation, and significance of the work will be discussed.

To begin, some of the most basic components of genetics include nucleic acids named DNA and RNA. DNA is a molecule found in the nucleus of most cells that contains the genetic information for all of an organism's functioning. A gene is a specific sequence of DNA that holds instructions to make some sort of molecule, whether that is a specific protein or an RNA molecule. RNA is another component that is quite essential for most biological functions. The process of producing a strand of RNA from DNA is called transcription. This RNA is either able to perform a certain function itself or forms instructions for the development of proteins, which is called translation. When an RNA molecule has instructions that do not strictly create proteins, the RNA would be called non-coding RNA. This sequence of events where DNA is transcribed to RNA and RNA creates protein is an integral part of biology called The Central Dogma.

For the purpose of the study, a type of non-coding RNA called miRNA was used for analysis. These are small RNA that can alter gene expression, or the extent to which DNA is represented throughout the body, which will be discussed more thoroughly in a subsequent paragraph. For the RNA that do help create proteins, they would have the title of messenger RNA (mRNA), as it carries the message of how to manufacture proteins. This process is called translation. Proteins are especially important because they help facilitate cellular functions that regulate the human body's tissues and organs. Though there are other types of RNA, the ones that are most important to this study's data types have already been discussed or fall under one of the two categories.

Copy number alterations (CNAs) occur when the number of genes are altered in some way in a genome, which can result in adverse effects on cellular functioning [6]. Copy number duplication is when a gene or a part of a gene is repeated and the copy number is increased, while copy number deletion is when a gene or a part of a gene is removed and the copy number is decreased. When a gene's copy number is altered, the way that the gene is expressed is also changed. It can change when, where, and in what quantity proteins are expressed.

Gene expression refers to the way that the products of genetic information are converted into a protein that carries out a certain function in the body [9]. These are especially important in cancer, as over expressions and under expressions could result in specific genes being promoted and inhibited. A specific type of RNA, called miRNA, acts as oncogenes or tumor suppressor genes when they block mRNA from creating proteins by binding to them [4]. When oncogenes are promoted, cancer cells are encouraged to continue growing throughout the body. Additionally, when there are tumor suppressor genes that are inhibited, the body has no defense against the growth of a tumor, thus resulting in uninhibited cancer cell growth. This can cause several complications in functions at the cellular level or, depending on the scale of the growth, across tissues and organs. For this reason, finding the specific genes that control these processes becomes especially important to successfully predict and hopefully prevent cancer.

Cervical cancer (CESC) and endometrial uterine cancer (UCEC) are two cancers that negatively impact the lives of women across the globe. As with most cancers, a diagnosis such as this results in a complete upheaval of one's life. From surgery, to radiation, to even chemotherapy, the detriment to a person's well-being is impossible to deny. In the US alone, there were 11542 new cases of cervical cancer and 4272 deaths from 2016 to 2020. Within that same time frame, 54744 people were diagnosed with and 11995 died of uterine cancer. Learning more information about cancers affecting the female reproductive system could lead to a mitigation of death by early detection or total prevention via biomarkers. Because these cancers have high rates of metastasis — the spreading of cancer to another part of the body — with each other due to proximity, they were chosen as the subjects of study for this experiment.

II. Materials and Methods

A. *Data Collection*

Three types of data were used in this study. Gene expression data, copy number alteration data (CNA), & microRNA (miRNA) data, were all sourced

via the XenaBrowser, which collected the data from The Cancer Genome Atlas (TCGA) that was published in 2019 [33]. The gene expression data was in the form of high-throughput sequencing data, or Htseq [18]. There were 892 samples and 60483 features. The features in this dataset represented the coding, non-coding, and pseudogenes. The CNA data was in the form of binary GISTIC values where '-1' represented significant evidence of a copy number deletion, '0' represented no evidence of a copy number alteration, and '1' represented significant evidence of a copy number amplification. There were 845 samples and 19729 columns. The columns represented the number of only protein-coding genes that we have in our genome as humans. The miRNA values were quantified by signal intensity of the stem loop expression. There were 887 samples and 1881 columns. The miRNA dataset was quantified by the signal intensity of miRNA only, with no other non-coding genes. Finally, the last dataset consisted of the 'combined' data that existed in all three categories. There were a total of 824 samples and 82093 concatenated features of all datasets in this final dataset.

B. Data Cleaning and Preparation

In order to accomplish the goal of finding the biomarkers, several processing steps were implemented before beginning. Originally, the datasets were separated by cancer type. The uterine cancer dataset and cervical cancer datasets were originally separate, but were merged by adding a target feature label column to each sample in the datasets. The cervical cancer (CESC) target variable was labeled with a '1' & the uterine cancer (UCEC) target variable was labeled with a '0'. The two datasets were then merged to represent both cancer types. This ensured binary simplicity for the classification task. After undergoing this process, the datasets were ready to begin being manipulated for the study.

To start, each of the features were required to be standardized in order for the feature importance values to be used. This ensures that the model is not being influenced solely by values that have a large range. By giving the values of each feature a position on the same scale, we ensure that these can be justly compared down the line. This way, the feature importance is not overly influenced by large values.

C. Methods

Binary classification is essentially creating a model that can take in new data and predict in which class - either one or the other - the sample belongs. In order to build this model, one requires a set of data that this model can learn from in order to predict correctly when given new, unseen data. Therefore, it is pertinent to split the study's data into two groups, one to train the model and one to evaluate the predictions. Usually this is done in an 80/20 fashion, where 80% of the data goes towards training and 20% goes into testing.

In order to accurately represent the two cancer types, it was necessary to incorporate an alternate method of sampling. The two datasets did not have the same amount of samples per cancer type, but instead were slightly imbalanced for all genomic data types. For the GISTIC dataset, there were 548 cervical cancer samples and 297 uterine cancer samples. In the HtSeq dataset, there were 583 cervical cancer samples and 309 uterine cancer samples. In the miRNA dataset, there were 575 cervical cancer samples and 312 uterine cancer samples. Finally, the Combined dataset had 530 cervical cancer samples and 294 uterine cancer samples. In order to account for the slightly imbalanced data, stratified sampling was employed, where each target variable was equally sampled .

The next step involved the use of a feature selection method called LASSO in order to reduce the number of features in the dataset. As the datasets given by TCGA contained thousands- and in some cases tens of thousands- of features, it was necessary to reduce that number in order to prevent the models from overfitting. Overfitting refers to a model becoming too tuned to the data that it has been trained on, which results in the model being unable to accurately predict results when given new, unseen data. LASSO reduces the number of features by identifying strongly correlation features that are redundant or irrelevant, without a significant loss of information when they are removed. This is done via linear regression that uses shrinkage, or the reduction of data values towards the mean (in this case, zero) in order to reduce the impact of unimportant features to nothing. In the case of this study, LASSO resulted in a feature reduction that left 131 out of 1881 features for the miRNA dataset, 77 out of 19729 features for the GISTIC dataset, 118 out of 60483 features for the Htseq dataset, and 152 out of 82093 features for the combined dataset. This significant reduction was intended to contribute to a lessening of overfitting for all models evaluated.

Another control for overfitting was cross validation. In this study, a ten-fold cross validation was employed. Ten fold cross validation is a resampling method that takes ten different splits of the training data and testing data. The data is split into ten equal groups, or folds, where each fold takes a turn of being the one testing data fold, while the other nine folds are used as the training data. This is repeated until every fold has had an opportunity to be used as the sole testing data fold. The model is evaluated for each of the ten splits and the final accuracy is the average accuracy of all of the splits. This was repeated for each algorithm in the study, which are listed below.

D. Algorithms and Packages

For this study, the coding language used was Python, specifically in Jupyter Notebook format. Almost all models were imported from Scikit-learn except XGBoost, which has its own library. The classifiers included in the study are listed below in Table I., with the exception of the Artificial Neural Network (ANN). This classifier was created manually using Keras Sequential. The total number of classifiers used were 18.

Packages included Pandas for DataFrame construction, Numpy for mathematical functions and arrays, Matplotlib.pyplot for visualizations, Scikit-Learn for the classification models and data splitting, Tensorflow for the ANN, & SHAP for feature importance values. SHAP values were calculated by the KernelExplainer for all classification models except XGBoost, which necessitated the TreeExplainer. The mean absolute value SHAP values were calculated and stored in a pandas DataFrame named shap_importance.

To construct the neural network (ANN), Keras was imported and utilized. A Sequential model was created with three Dense layers. The input layer was built with the number of features for each individual dataset and the ReLU activation function. The second and hidden layer was built to further define the features and had the same structure as the input layer. Finally, the output layer was a single neuron with the Sigmoid activation function, as the problem had a binary output.

III. Experimental Result

A. Evaluation of Algorithms

To simplify the names of the algorithms, abbreviations were used to save space in the formatting. The table below shows the algorithms and their corresponding abbreviation. Though the means of accomplishing the classification differ, the overall result remains the same - classifying a sample into one class or the other via a set of rules determined by fitting the model to the training data.

Table I. Algorithm Abbreviations

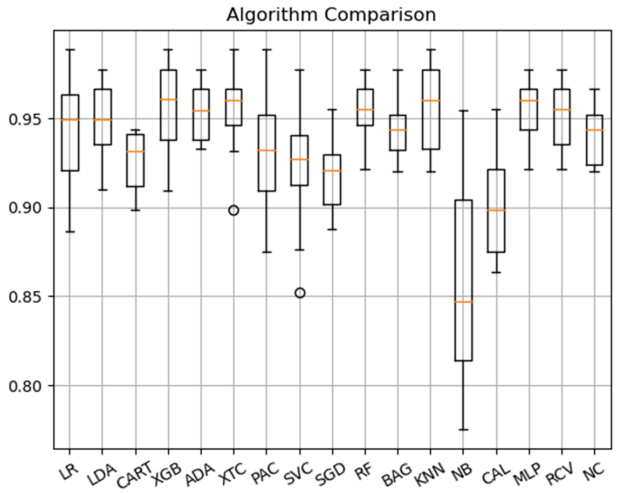
Algorithm	Abbreviation
Logistic Regression	‘LR’
Linear Discriminant Analysis	‘LDA’
Classification and Regression Tree	‘CART’
eXtreme Gradient Boosting Classifier	‘XGB’
Adaptive Boosting Classifier	‘ADA’
Extra Trees Classifier	‘XTC’
Passive Aggressive Classifier	‘PAC’
C-Support Vector Classification	‘SVC’
Stochastic Gradient Descent Classifier	‘SGD’
Random Forest Classifier	‘RF’
Bagging Classifier	‘BAG’
K Neighbors Classifier	‘KNN’
Gaussian Naive Bayes Classifier	‘NB’
Calibrated Classifier	‘CAL’
Multi-layer Perceptron Classifier	‘MLP’
Ridge Classifier	‘RCV’
Nearest Centroid Classifier	‘NC’

Each dataset was fit to the models with the stratified training set and evaluated using the testing set. The accuracies of each fold in the 10 fold cross validation were recorded and evaluated. By looking at the average accuracies for the ten folds, it was clear to see the overall effectiveness of the algorithm in predicting cancer types. For the HtSeq dataset, all algorithms performed either with perfect accuracy or near perfect accuracy. The classification algorithms in the GISTIC dataset had average accuracies in the 70s, with some dropping all the way to the high 60s. The miRNA dataset had average accuracies around the mid 90s for all classification algorithms. Finally, the Combined dataset was similar to the HtSeq dataset in that it had near perfect accuracy.

By looking at the boxplots, it was clear to see the underperforming algorithms. Though all of the classification models for the HtSeq and Combined datasets performed at around the same level, the miRNA and GISTIC datasets had underperformers. For the miRNA dataset, the Naive Bayes (NB) had a median accuracy of about 0.85, while several others were performing at least in the 90s. For this reason, it was determined that the value would be omitted for the remainder of the study. For the GISTIC dataset, the underperformer was the Passive Aggressive Classifier, or PAC. For almost all of the classification tasks, the model performed under 50% accuracy, which would be worse than even just random guessing.

Because of the variation in accuracies in the GISTIC and miRNA datasets, only the top performer was chosen to be analyzed for feature importance for the remainder of the study. On the other hand, because of the consistency in accuracy for the HtSeq and Combined datasets, it was determined that the feature importance would be decided by taking the average feature importance values for all of the classification algorithms.

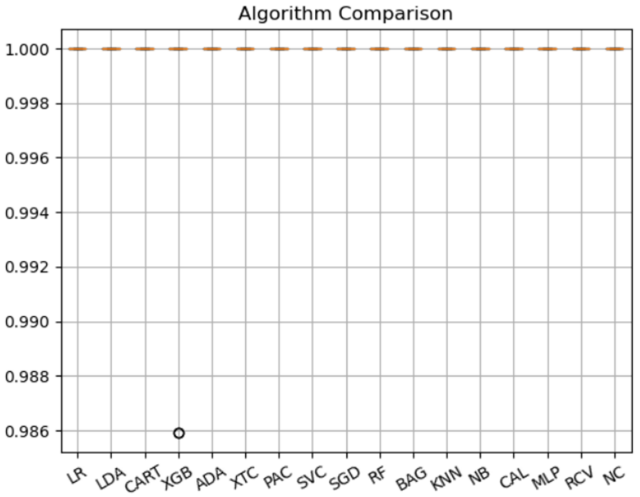
Figure A. Boxplot of 10-Fold Cross Validation Accuracies for the miRNA dataset



X-axis: Classification Algorithm Name
Y-axis: Accuracy

Description: Most classification algorithms had accuracies in the lower to mid nineties, with the exception of NB.

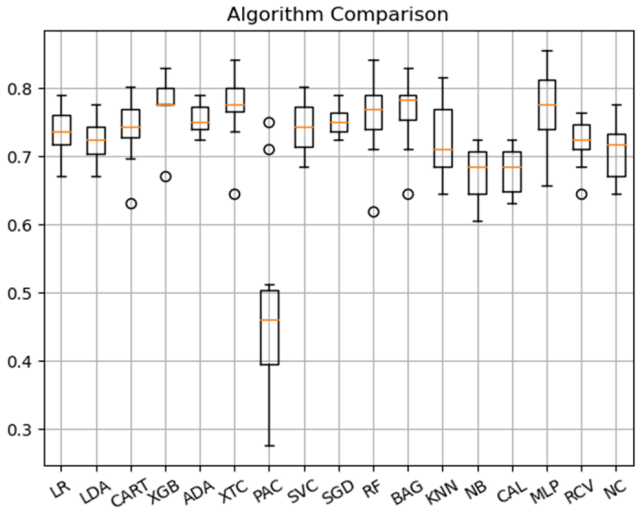
Figure B. Boxplot of 10-Fold Cross Validation Accuracies for the HtSeq dataset



X-axis: Classification Algorithm Name
Y-axis: Accuracy

Description: All classification algorithms had the majority of accuracies close to 1.0.

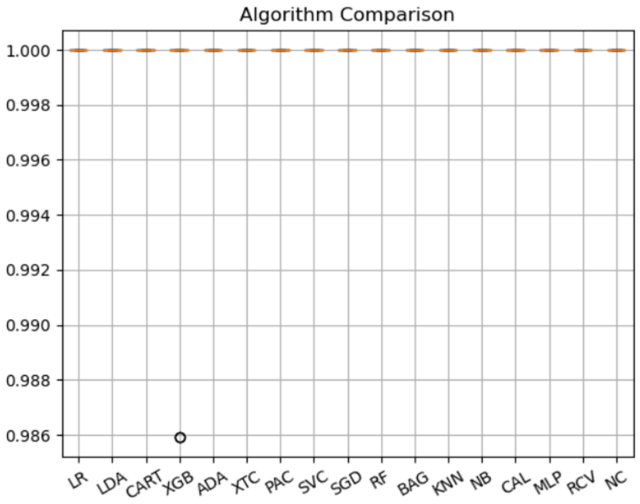
Figure C. Boxplot of 10-Fold Cross Validation Accuracies for the GISTIC dataset



X-axis: Classification Algorithm Name
Y-axis: Accuracy

Description: Most classification algorithms had accuracies between 0.7 and 0.8, with the exception of PAC, which severely underperformed.

Figure D. Boxplot of 10-Fold Cross Validation Accuracies for each algorithm for the Combined dataset



X-axis: Classification Algorithm Name
Y-axis: Accuracy

Description: All classification algorithms had the majority of accuracies close to 1.0.

Table II. Average Accuracy for All Classification Algorithms using HtSeq Dataset

HtSeq Dataset	
Algorithm	Average Accuracy
LR	1.0
LDA	1.0
CART	1.0
XGB	0.999
ADA	1.0
XTC	1.0
PAC	1.0
SVC	1.0
SGD	1.0
RF	1.0
BAG	1.0
KNN	1.0
NB	1.0
CAL	1.0
MLP	1.0
RCV	1.0
NC	1.0
ANN	1.0

Table III. Average Accuracy for All Classification Algorithms using GISTIC Dataset

GISTIC Dataset	
Algorithm	Average Accuracy
LR	0.728
LDA	0.717
CART	0.724
XGB	0.767
ADA	0.741
XTC	0.755
SVC	0.728
SGD	0.722
RF	0.759
BAG	0.753
KNN	0.724
NB	0.671
CAL	0.667
MLP	0.763
RCV	0.714
NC	0.704
ANN	0.711

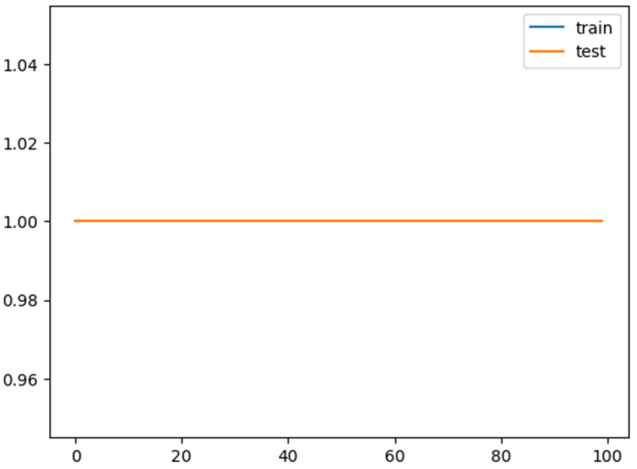
Table IV. Average Accuracy for All Classification Algorithms using miRNA Dataset

miRNA Dataset	
Algorithm	Average Accuracy
LR	0.942
LDA	0.950
CART	0.926
XGB	0.957
ADA	0.955
XTC	0.955
PAC	0.933
SVC	0.923
SGD	0.919
RF	0.955
BAG	0.942
KNN	0.955
CAL	0.901
MLP	0.956
RCV	0.953
NC	0.940
ANN	0.937

Table V. Average Accuracy for All Classification Algorithms using Combined Dataset

Combined Dataset	
Algorithm	Average Accuracy
LR	1.0
LDA	1.0
CART	0.999
XGB	0.999
ADA	1.0
XTC	1.0
PAC	1.0
SVC	1.0
SGD	1.0
RF	1.0
BAG	1.0
KNN	1.0
NB	1.0
CAL	1.0
MLP	1.0
RCV	1.0
NC	1.0
ANN	1.0

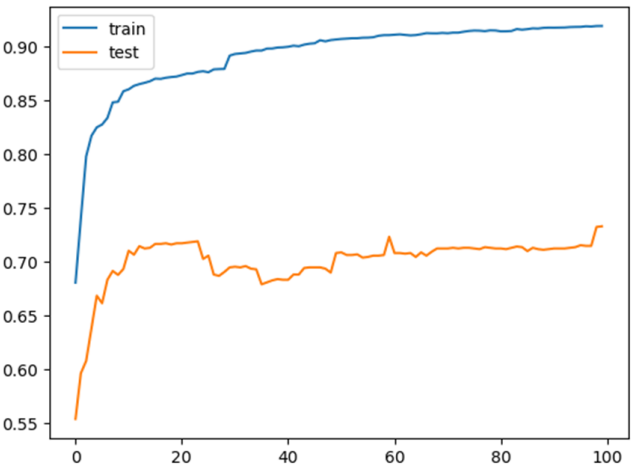
Figure E. ROC AUC Curve Across 100 Epochs for HtSeq Dataset - XGBoost



X-axis: Number of Epochs
Y-axis: Accuracy

Description: Perfect AUC across all epochs.

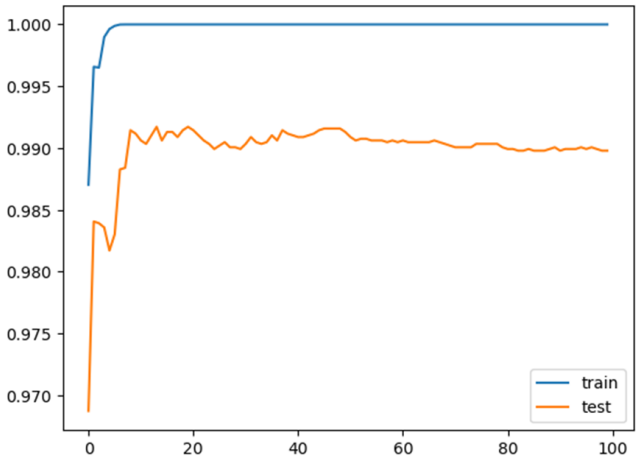
Figure F. ROC AUC Curve Across 100 Epochs for GISTIC Dataset - XGBoost



X-axis: Number of Epochs
Y-axis: Accuracy

Description: Evident overfitting in the model, as the training accuracy is very high (~90%) and the testing accuracy is around 75%.

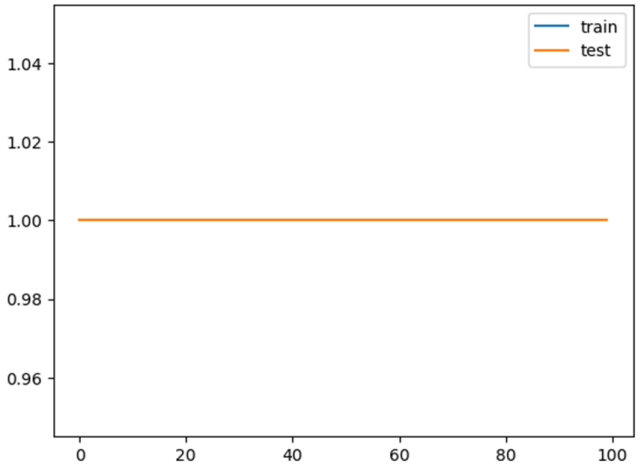
Figure G. ROC AUC Curve Across 100 Epochs for miRNA Dataset - XGBoost



X-axis: Number of Epochs
Y-axis: Accuracy

Description: Slight overfitting but with exceptional performance at around 99% accuracy for the testing data.

Figure H. ROC AUC Curve Across 100 Epochs for Combined Dataset - XGBoost



X-axis: Number of Epochs
Y-axis: Accuracy

Description: Perfect accuracy across all epochs.

B. Feature Importance via SHAP

Ultimately, because of the performance of certain datasets, two routes were taken. For the GISTIC and miRNA datasets, the XGBoost classifiers were clearly the top performers according to the average accuracy tables. This is important because a classification model without high accuracy is not likely to represent the data well, which means that their feature importance may not have much significance for either uterine or cervical cancer. Therefore, only the average absolute value SHAP scores for XGBoost were used to find the feature importance of these two datasets.

Another route was taken for the HtSeq and Combined data because of the high accuracy across all of the algorithms. Because all of the classification models performed exceptionally, it was determined that an average of all of the average absolute value SHAP scores across all algorithms would give a better indication of the true feature importance of the datasets. In the tables below, the top 5 features for each dataset and their respective SHAP scores were listed. In addition, supplementary research was performed to understand the previously discovered biomarkers for each of the genes, which are listed under 'Cancer Specificity'.

For the GISTIC Data, there were several already identified biomarkers for Endometrial cancer and Cervical cancer that were found in the top 5 most important genes. The fourth most influential gene for this dataset was detected in all cancers. The first and fifth most influential genes were previously discovered biomarkers for endometrial uterine cancer (UCEC) and the second most influential gene was a biomarker for cervical cancer (CESC). Interestingly, the third most influential gene was not a biomarker for any cancer or disease in general, though it held a SHAP score at just under the second most influential gene.

For the HtSeq data, there were also previously discovered biomarkers for endometrial uterine cancer and cervical cancer. However, the only biomarkers for these two were the first and second most important, respectively. The next three did not have any cancer specificity with either cancers. It is interesting to note that the third most influential feature has a SHAP value only slightly below the value of the second most influential feature.

The miRNA data also shows several important features that have been researched and found as biomarkers for UCEC and CESC. The top

three were found to be biomarkers for both, while the fifth is only a biomarker for uterine cancer. The fourth gene was found to be a biomarker for several other cancers, but not the targets of this study. There is a significant jump down from the SHAP values of the top three and the bottom two, suggesting a bigger impact on the model for the former.

The last dataset, the Combined, showed some results that were unlike all of the other datasets. All of the top four genes were not biomarkers for either of the cancers in the study. However, the last gene was found to be a biomarker for endometrial cancer. The accuracy given for this dataset makes one question if these genes do have influence for these types of cancer, or if there are any confounding variables that are interacting with the data.

Table VI. GISTIC - Average SHAP Values for XGBoost

Gene Name	Cancer Specificity	Average SHAP value
ENSG0000 0141905.16	Endometrial cancer	0.093
ENSG0000 0073282.11	Cervical, head and neck, lung, urothelial	0.063
ENSG0000 0176009.3	None	0.062
ENSG0000 0028310.16	Detected in all	0.055
ENSG0000 0121858.9	Prostate cancer, Myeloma, Endometrial cancer, Breast cancer	0.054

Table VII. HtSeq - Average SHAP Values Across All Algorithms

Gene Name	Cancer Specificity	Average SHAP value
ENSG0000 0206630.1	Endometrial cancer	0.061
ENSG0000 0182117.5	Diffuse large B cell lymphoma (DLBCL), bladder, & cervical	0.060
ENSG0000 0269899.1	Ovarian cancer	0.057
ENSG0000 0274501.1	Acute myeloid leukemia (LAML) & thymoma	0.044
ENSG0000 0215267.7	None	0.030

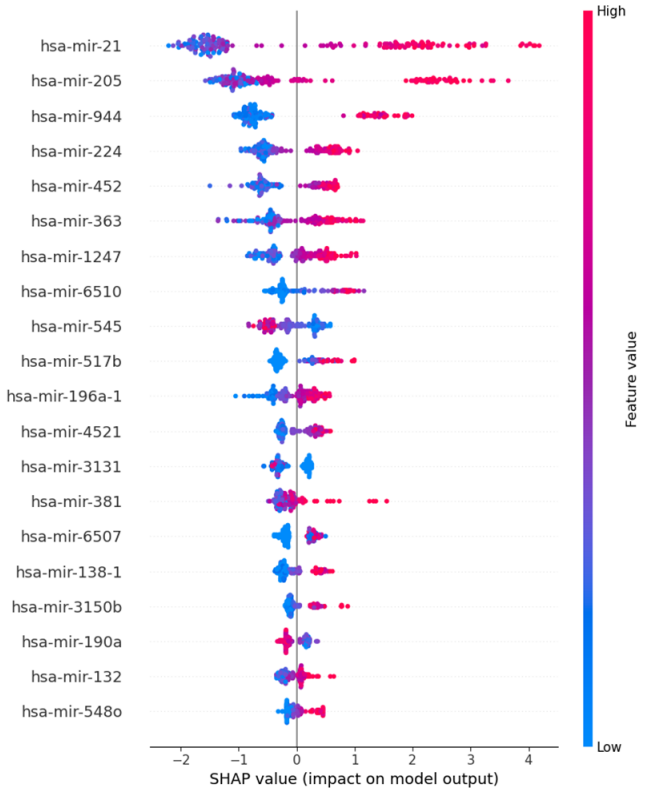
Table VIII. miRNA - Average SHAP Values for XGBoost

Gene Name	Cancer Specificity	Average SHAP value
hsa-mir-21	Many cancers, incl. endometrial & cervical	0.154
hsa-mir-205	Endometrial, cervical, squamous cell carcinoma, colon cancer	0.110
hsa-mir-944	Endometrial, cervical, & breast cancers	0.082
hsa-mir-224	Hepatocellular carcinoma (HCC), Pancreatic ductal adenocarcinoma (PDAC), & Non-small cell lung cancer (NSCLC)	0.049
hsa-mir-452	Bladder & uterine cancer	0.048

Table IX. Combined - Average SHAP Values Across All Algorithms

Gene Name	Cancer Specificity	Average SHAP value
ENSG0000 0280231.1	Thymoma	0.064
ENSG0000 0215030.5	Ovarian and bladder cancer	0.061
ENSG0000 0225131.2	Diffuse large B cell lymphoma (DLBCL) and Glioblastoma	0.058
ENSG0000 0128228.4	Diffuse large B cell lymphoma (DLBCL), Uterine, Bladder cancer	0.058
ENSG0000 0244268.1	endometrial cancer	0.034

Figure I. Shap Directionality - miRNA data - XGBoost



C. Directionality

By examining the global impact of each model, it becomes possible to understand the directionality of the impact of the SHAP values for each feature in a dataset. Each point on the graph represents the SHAP score for one feature of one sample [32]. The example above demonstrates a number of insights into several features. For features that do not have such a clear divergence, no conclusions would be possible to make. However, it is clear to see that some features in the visualization have a distinct separation between the blue and red points on the plot. For instance, the most impactful feature for the XGBoost model was hsa-mir-21. Across that feature, the color coding shows a concentration of red points to the right of the y axis at the SHAP value of zero.

On the opposite side, there is a concentration of blue points. The two target values of both uterine and cervical cancers were set to 0 and 1, respectively, so this chart indicates that lower SHAP values for this feature contributed to a UCEC classification, while high SHAP values for this feature contributed to a CESC classification. The reverse could be said for the feature hsa-mir-190a. The high SHAP values indicated a classification of UCEC and the low SHAP values indicated a classification of CESC. By following this process for all features in each beeswarm plot, it becomes feasible to determine the directionality of each one of the features. It is also interesting to note that SHAP values of zero indicate that the model is ignoring the value of that feature. A feature with all values at zero would be considered unimportant to the model.

D. Supplementary Data: Rate of Cancers

A supplementary visualization was incorporated into this study to demonstrate the impact of personalized medicine, which refers to the emerging medical practice that acknowledges the impact of interpersonal differences in the treatment of illness. As seen in Figure J., the rate of new cervical cancers in the US seems to be heavily concentrated in the southern region. Conversely, the rate of new uterine cancers in the US shows high numbers in the northern half in Figure K.. These figures show that there may be differences not only between individuals, but also regions. This shows the ever-importance of choosing samples that will represent the population to which you are looking to generalize the results of a study.

Figure J. Supplementary Visualization: Rate of New Cervical Cancers in the US from 2016-2020 [5].

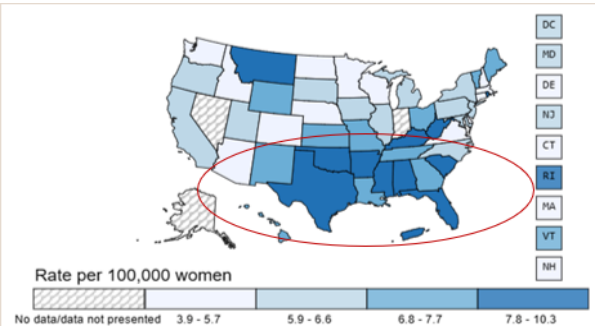
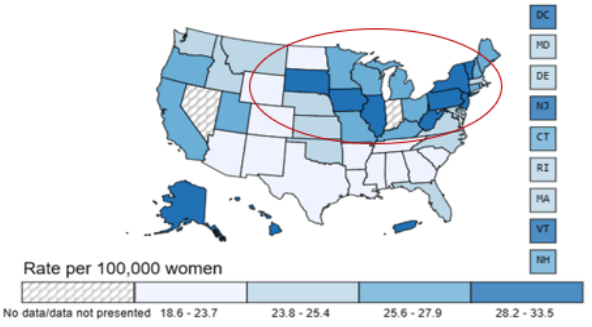


Figure K. Supplementary Visualization: Rate of New Uterine Cancers in the US from 2016-2020 [5].



IV. Discussion

Though there were several genes that were found as influential in the study, how does one decide which ones are actually impacting the cancers? It first starts with the evaluation. As displayed in the results, the accuracies for the GISTIC dataset were far from perfect and very clearly underperformed despite all of the cautions taken in order to minimize overfitting. Therefore, it would be improper to assume that the feature importances discovered from this dataset actually reflect the copy number alteration data in cancer patients. Though it is not ideal to have to ignore a full dataset, it is unlikely that there is valuable information to be gathered in that area regarding biomarkers. In the future, a larger dataset with more samples may lead to better results, even without any modification to the code.

Inversely, the other datasets did not have this problem. The highest-performing datasets in terms of accuracy were the HtSeq and Combined dataset, with perfect accuracies in nearly every model in all the cross validation folds. The miRNA dataset had relatively similar, but slightly lower accuracies around 99%. Given this performance, there is potential for all of the genes that have not been determined as biomarkers for UCEC or CESC to have some sort of impact on the illnesses. The most prominent in terms of future research would be hsa-mir-224, ENSG00000269899.1, ENSG00000274501.1, ENSG00000215267.7, ENSG00000215030.5, ENSG00000225131.2, and ENSG00000128228.4 due to their high SHAP values.

In the future, experiments should focus on these genes and perhaps how valid these results are via a replication study. Knowing that these have potential to be biomarkers, looking into how models perform without the influence of these genes may be able to demonstrate their significance to the classification of both cancers. An improvement of the GISTIC dataset may lead to better insights regarding copy number alterations. As TCGA updates, it would be interesting to see if all of the samples can be combined. Low sample size and high computational cost of SHAP were some of the largest obstacles that were put on the path during the course of this study. However, the biggest asset to this form of study is that the code is easily replicable for any other two types of cancers, meaning that any two available datasets could be exchanged with the UCEC and CESC data and still give just as equally valuable results.

V. Conclusion

Though the common thought may be that all cancer research is done with test tubes and Erlenmeyer flasks in a lab, Data Science methods have proven time and time again to also have value in the biological research community. Machine learning is a valuable resource when it comes to processing large quantities of data, as it's able to learn all of the minute details and patterns that would take humans a very long time to process by hand. By leveraging these classification algorithms, researchers are able to reveal insights into the human body to combat illness. Biomarkers have the potential to save countless lives via early detection, as many do not realize they have cancer because the symptoms are not severe. Therefore, there is less of an opportunity to fight the cancer while it is still at a low stage or grade. By discovering biomarkers, the future of illness may lie in a simple genetic sequencing, some of which are relatively low-cost, especially in comparison to cancer treatments.

References

[1] AKR1C7P Gene - Aldo-Keto Reductase Family 1 Member C7, Pseudogene. (n.d.). GeneCards. Retrieved December 4, 2023, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=AKR1C7P

[2] ASCL3 protein expression summary. (n.d.). The Human Protein Atlas. Retrieved December 4, 2023, from https://www.proteinatlas.org/ENSG00000176009-ASCL3

[3] Biomarkers in endometrial cancer: Possible clinical applications (Review). (n.d.). NCBI. Retrieved December 4, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392571/

[4] Bushell, M. (n.d.). *How do microRNAs regulate gene expression?* PubMed. Retrieved December 4, 2023, from https://pubmed.ncbi.nlm.nih.gov/19021530/

[5] CDC.gov. (n.d.). USCS Data Visualizations. Retrieved December 4, 2023, from https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/

[6] Chauhan, T. (2020, August 24). *What is Copy Number Variation and How to Detect it? – Genetic Education*. Genetic Education. Retrieved December 4, 2023, from https://geneticeducation.co.in/what-is-copy-number-variation-and-how-to-detect-it/

[7] Doyle, M., Phipson, B., & Dashnow, H. (n.d.). *1: RNA-Seq reads to counts*. Galaxy Training! Retrieved December 4, 2023, from https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-reads-to-counts/tutorial.html#map-reads-to-reference-genome

[8] Expression of BRD9 in cancer - Summary. (n.d.). The Human Protein Atlas. Retrieved December 4, 2023, from https://www.proteinatlas.org/ENSG00000028310-BRD9/pathology

[9] Gene Expression. (2023, November 30). National Human Genome Research Institute. Retrieved December 4, 2023, from https://www.genome.gov/genetics-glossary/Gene-Expression

[10] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000182117

[11] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000269899

[12] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000274501

[13] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000280231

[14] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000215030.5

[15] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000225131.2

[16] GEPIA. (n.d.). Gepia. Retrieved December 4, 2023, from http://gepia.cancer-pku.cn/detail.php?gene=ENSG00000128228.4

[17] Jenike, A. E., & Halushka, M. K. (2021, March 12). *miR-21: a non-specific biomarker of all maladies - Biomarker Research*. Biomarker Research. Retrieved December 4, 2023, from https://biomarkerres.biomedcentral.com/articles/10.1186/s40364-021-00272-1

[18] Kellis, M. (2021, March 17). *15.2: Methods for Measuring Gene Expression*. Biology LibreTexts. Retrieved December 4, 2023, from https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/15%3A_Gene_Regulation_I_-_Gene_Expression_Clustering/15.02%3A_Methods_for_Measuring_Gene_Expression

[19] Li, Y. (2020, October 27). *Upregulation of miR-205 induces CHN1 expression, which is associated with the aggressive behaviour of cervical cancer cells and correlated with lymph node metastasis - BMC Cancer*. BMC Cancer. Retrieved December 4, 2023, from https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07478-w

[20] Mapping the Chromatin Landscape of Human Cancers. (2018, October 26). HHMI. Retrieved December 4, 2023, from https://www.hhmi.org/news/mapping-the-chromatin-landscape-of-human-cancers

[21] microRNA-944 overexpression is a biomarker for poor prognosis of advanced cervical cancer. (2019, May 6). PubMed. Retrieved December 4, 2023, from https://pubmed.ncbi.nlm.nih.gov/31060525/

[22] miR-205: A Potential Biomedicine for Cancer Therapy. (n.d.). NCBI. Retrieved December 4, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7564275/

[23] MIR452 Gene. (n.d.). GeneCards. Retrieved December 4, 2023, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR452

[24] miR-944 acts as a prognostic marker and promotes the tumor progression in endometrial cancer. (n.d.). PubMed. Retrieved December 4, 2023, from https://pubmed.ncbi.nlm.nih.gov/28178620/

[25] Momeni-Boroujeni, A., & Nguyen, B. (2023, June 16). *Genomic landscape of endometrial carcinomas of no specific molecular profile*. YouTube. Retrieved December 4, 2023, from https://www.nature.com/articles/s41379-022-01066-y

[26] NFIC protein expression summary. (n.d.). The Human Protein Atlas. Retrieved December 4, 2023, from https://www.proteinatlas.org/ENSG00000141905-NFIC

[27] Prior, I. A., Lewis, P. D., & Mattos, C. (n.d.). A comprehensive survey of Ras mutations in cancer. NCBI. Retrieved December 4, 2023,

from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3354961/

[28] SHAP. (n.d.). Welcome to the SHAP documentation — SHAP latest documentation. Retrieved December 4, 2023, from https://shap.readthedocs.io/en/latest/

[29] *SNORD60 promotes the tumorigenesis and progression of endometrial cancer through binding PIK3CA and regulating PI3K/AKT/mTOR signaling pathway*. (2022, December 23). PubMed. Retrieved December 4, 2023, from https://pubmed.ncbi.nlm.nih.gov/36562475/

[30] *TNFSF10 protein expression summary*. (n.d.). The Human Protein Atlas. Retrieved December 4, 2023, from https://www.proteinatlas.org/ENSG00000121858-TNFSF10

[31] *TP63 protein expression summary*. (n.d.). The Human Protein Atlas. Retrieved December 4, 2023, from https://www.proteinatlas.org/ENSG00000073282-TP63

[32] Trevisan, V. (2022, January 17). *Using SHAP Values to Explain How Your Machine Learning Model Works*. Towards Data Science. Retrieved December 4, 2023, from https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

[33] *UCSC Xena*. (n.d.). UCSC Xena. Retrieved December 4, 2023, from https://xenabrowser.net/datapages/?dataset=TCGA-UCEC.mirna.tsv&host=https%3A%2F%2Fgdc.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

[34] Wang, X. (2021, August 19). *How to interpret and explain your machine learning models using SHAP values*. Mage. Retrieved December 4, 2023, from https://m.mage.ai/how-to-interpret-and-explain-your-machine-learning-models-using-shap-values-471c2635b78e

[35] Wang, Y., & Lee, A. T.C. (2023, June 16). *Profiling MicroRNA Expression in Hepatocellular Carcinoma Reveals MicroRNA-224 Up-regulation and Apoptosis Inhibitor-5 as a MicroRNA-224-specific Target**. YouTube. Retrieved December 4, 2023, from https://www.jbc.org/article/S0021-9258(20)59775-5/fulltext

[36] *Bister, K. (2015, December 7). Discovery of oncogenes: The advent of molecular cancer research. NCBI. Retrieved December 4, 2023, from* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687565/