2023

Biomarker Identification for Breast Cancer Types Using Feature Selection and Explainable AI Methods

David E. La Rosa Giraud
University of Central Florida

BIOMARKER IDENTIFICATION FOR BREAST CANCER TYPES USING FEATURE SELECTION AND EXPLAINABLE AI METHODS

by

DAVID LA ROSA GIRAUD

A thesis submitted as partial fulfillment of the requirements
for the Honors in the Major Program in Computer Science
in the College of Engineering and Computer Science
and in the Burnett Honors College
at the University of Central Florida
Orlando, Florida

Fall Term, 2023

Thesis Chair: Liqiang Wang, Ph.D.

ABSTRACT

This paper investigates the impact the LASSO, mRMR, SHAP, and Reinforcement Feature Selection techniques on random forest models for the breast cancer subtypes markers ER, HER2, PR, and TN as well as identifying a small subset of biomarkers that could potentially cause the disease and explain them using explainable AI techniques. This is important because in areas such as healthcare understanding why the model makes a specific decision is important it is a diagnostic of an individual which requires reliable AI. Another contribution is using feature selection methods to identify a small subset of biomarkers capable of predicting if a specific RNA sequence will have one of the cancer labels positive. The study begins by obtaining baseline accuracy metric using a random forest model on The Cancer Genome Atlas's breast cancer database to then explore the effects of feature selection, selecting different numbers of features, significantly influencing model accuracy, and selecting a small number of potential biomarkers that may produce a specific type of breast cancer. Once the biomarkers were selected, the explainable AI techniques SHAP and LIME were applied to the models and provided insight into influential biomarkers and their impact on predictions. The main results are that there are some shared biomarkers between some of the subsets that had high influence over the model prediction, LASSO and Reinforcement Feature selection sets scoring the highest accuracy of all sets and obtaining some insight into how the models used the features by using existing explainable AI methods SHAP and LIME to understand how these selected features are affecting the model's prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TCGA – The Cancer Genome Atlas

ML – Machine Learning

AI - Artificial Intelligence

LASSO - least absolute shrinkage and selection operator

mRMR- Maximum Relevance Minimum Redundancy

RFS – Reinforcement feature selection

SHAP - SHapley Additive exPlanations

LIME - Local Interpretable Model-Agnostic Explanations

INTRODUCTION

This research will explore biomarker identification for breast cancer subtypes focusing on feature selection techniques the Random Forest, LASSO, Maximum Relevance Minimum Redundancy (mRMR), and reinforcement feature selection methods on the TCGA breast cancer dataset to then obtain an explainable result by applying Explainable AI methods SHAP and LIME.

This paper covers two important topics. The first one is biomarker selection using feature selection methods. This topic has been researched before, but here the focus will be on the different biomarker subsets selected by the algorithm and how they compare in a random forest model. Once the models are trained, explainable AI will be applied to them. This relatively new concept pursues a human interpretation of how the model manages the data.

Cancer

Cancer is a well-known disease that has been affecting humanity for years. It is an expression of genetic mutation with more than two hundred types [1]. The cancer organization organized the TCGA to make an atlas of cancer information to be used and studied, containing over five hundred test cases, and being used widely to research the properties of the disease. Breast cancer-specific can be divided into four subtypes depending on the affected receptor hormone. ER: receptor hormone, HER2 Receptor hormone, PR, and triple-negative TN [1]. In this paper, these are the labels the models will try to predict.

Biomarkers

The TCGA data contains many different data: DNA, DNA-meth, and RNA sequences [3]. This paper will focus on RNA sequences that represent the genetic sequences of each sample. Biomarker refers to specific RNA sequences related to the presence of something, the breast cancer type. The TCGA BRCA dataset has 90 thousand RNA sequences for each sample.

Machine Learning

Machine learning is using algorithms for machines to recognize patterns mimicking learning. It can be divided into supervised learning, where the machine learns with the data, and prediction and unsupervised learning, where the machine only has the data instead of learning a pattern, so it tries to group data into clusters. This paper will only go over the supervised approach.

Feature Selection

In machine learning, feature selection refers to finding an optimal subset of features, or in this case biomarkers, which will improve the accuracy and efficacy of a machine learning model. The approaches to feature selection important for this paper are LASSO or least absolute shrinkage and selection operator where coefficients are reduced and values of zero dropped, minimum Redundancy Maximum Relevance, which focus on feature correlation, and Reinforcement learning feature selection [2], which applies reinforcement learning concepts into the process of feature selection.

Explainable AI

Explainable AI refers to the process by which humans can interpret a model's reasoning and logic. The importance of explainable AI has surged over recent years. Explaining a model validates and provides reassurance on how it works, allowing it to be used in more sensitive topics like biomarker identification. Explainability or referred to as transparency in literature [6], specifically aims to explain how AI models work as they are black boxes; what exactly the model does with the data, and how each feature used to influence the prediction is unknown without explainable AI. For this paper, we will focus on two explainable AI techniques SHAP and LIME.

Starting with how features are used, we have the SHAP or SHapley Additive exPlanation[7]. SHAP is an algorithm that provides a value that shows how each feature impacts the prediction from class-wise impact or an overall subset of features. The next one is LIME, Local Interpretable Model-Agnostic [8]. It is a surrogate model evaluator. It calculates prediction probabilities and a class feature weight. This technique does not look into the model but learns from modifying the input and checking how the output behaves.

HYPOTHESIS

This section provides a hypothesis and goes over the study's objectives and the process that will be used. This study has two stages. The identification of biomarkers and the explainability of them.

Identification

Reinforcement Feature selection will select a subset of biomarkers that will produce a higher accuracy score than the baseline Random Forest score, LASSO, mRMR, and SHAP feature selection techniques to classify breast cancer subtypes.

Explainability

Applying SHAP and LIME techniques to the models from the identification set will provide an explanation that will connect some of the biomarkers selected by the different feature selection methods.

Contributions

- A comparison between LASSO, mRMR, and SHAP feature selection techniques against Reinforcement feature selection using health care data.
- Comparison of the accuracy of random forest model and different feature subsets using the different selected biomarkers.
- Explain reasons for model prediction in biomarker selection of TCGA breast cancer subtype using explainable AI techniques SHAP and LIME.
- Provide common biomarkers that have a high influence in prediction models.

METHODOLOGY

This study will be performed in a Python environment. The random forest and LASSO are the implementations defined in the scikit-learn library. The mRMR implementation will be the one used in the source paper [9]**.** The Reinforcement Feature Selection model will follow the one in the NN implementation from the source paper [2]**.** During the identification phase, SHAP will also be applied to the entire dataset to use its feature evaluation to select the top k features. The last model used will be a Pytorch neural network. On the explainable step, the SHAP and LIME implementation will be the ones from their respective libraries.

Of the 411 data samples available the data will be broken down into 70% (288 samples) into the training set, 21% (86 samples) into the test set, and 9% (37 samples) into the validation set. Because there can be multiple positive labels for a single sample there will be a model for each label, treating it as a binary classification problem.

For each feature selection technique, a new random forest model will be created, trained, and optimized using specific subsets of features. The main metric used to evaluate the model will be the accuracy score. Initial testing will be done with only the test set and the final comparison will be done based on the validation set.

Once the subset of features is selected for each model SHAP will be applied to obtain the average impact of the feature on the model output. Then the LIME algorithm[10] will be applied and used to compare with the SHAP table. Once both the LIME and SHAP table are obtained they

will be used to analyze the model: Look at common features between different models and which features have higher influence over the model.

Data Sources

The data used is the TCGA breast cancer dataset. This is the most TCGA dataset used for research. The RNA seq data of the TCGA can be found here [4]. The labels used are here [5].

RESULTS

Selected Subsets

The objective of these subsets is to get the highest accuracy possible by using the least number of biomarkers. Taking this into account, some of the feature selection algorithms automatically select some features, which was the case for LASSO, in which the algorithm selected 61 features for its final set for each of the labels.

The next model was mRMR. This algorithm requires the user to define the number of features to pick to be defined. The procedure was to pick one as a starting k, evaluate the accuracy, and then increase k by one until the accuracy stopped increasing. The accuracy fluctuated after passing the 10-feature mark. Even after using 61 features, it only slightly increased the accuracy compared to the ten feature sets for some labels. The feature set containing ten elements was the one chosen for mRMR. For the SHAP model, the same technique was applied, and its score kept increasing until the 60th feature.

Lastly, the RFS algorithm resulted in a set of 461 features. A high number of features compared to the other algorithms. This might be because of the nature of the algorithms. RFS focuses on selecting an optimal set of features to achieve the highest accuracy possible while tending to be computationally inefficient.

Test Set

The following table contains the accuracy score of a random forest model using the specific feature selection method on the test set. It also contains the accuracy score of a neural network to compare it to a black box neural network performance. Note that this neural network is taking the entire feature set.

Table 1: Accuracy percentage of models using different feature selection methods on the data labels for the test data.

LABELS	NO FS	LASSO	MRMR	SHAP	RFS	NN
ER	90.7	93.0	90.7	93.0	93.0	80.2
HER2	88.4	91.9	90.7	90.7	90.7	83.7
PR	83.7	86.0	60.5	86.0	88.4	70.9
TN	90.7	90.7	88.4	90.0	91.9	82.5

Starting with the baseline model using the full set of genetic sequences, it already got scores around 90%, which is surprisingly high considering the number of features against the number of samples. Another thing to note is that the labels are not uniform in terms of score. The ER label yielded the highest accuracy out of all the labels, with PR being the lowest. The Reinforcement feature selection set achieved the highest accuracy overall, only falling behind LASSO in the HER2 label by 1.2%. LASSO was the second-best performing set. Another notable fact is that the PR label accuracy using the mRMR set was extremely low compared to the others, which could mean that the nature of the mRMR algorithm goes against the pattern of the PR label. The worst-performing model is the neural network model, which is not surprising. The data have an extremely high dimensionality and a low sample size.

Validation Set

The following are the results using the validation set. Note that the tests using the validation set were run once.

Table 2: Accuracy percentage of models using different feature selection methods on the data labels for the validation data.

LABELS	NO FS	LASSO	MRMR	SHAP	RFS	NN
ER	94.7	94.7	92.1	94.7	94.7	92.1
HER2	89.5	94.7	89.5	89.5	92.1	89.5
PR	86.8	92.1	73.7	89.5	92.1	81.5
TN	92.1	89.4	89.5	92.1	94.7	94.7

On the validation set, the same label trends can be seen, but overall, the accuracy scores were higher than on the test set. Also, many of the sets got the same score. This could be due to the small set size of the validation set, only containing 37 samples. Some notable facts from this table are that the base random forest model with no feature selection matched the highest accuracy for the ER label. The Lasso set scored remarkably similar to the RFS set. Lastly, the neural network scored considerably higher than the test set score, matching the highest accuracy obtained for the TN label.

Exploring Selected Features

This section will go over the most important features selected for every label starting with the SHAP values for each feature and followed by an analysis of these graphs. Next will the LIME explanation of a prediction of single test sample with values of ER positive, HER2 negative, PR positive and TN negative.

Label ER

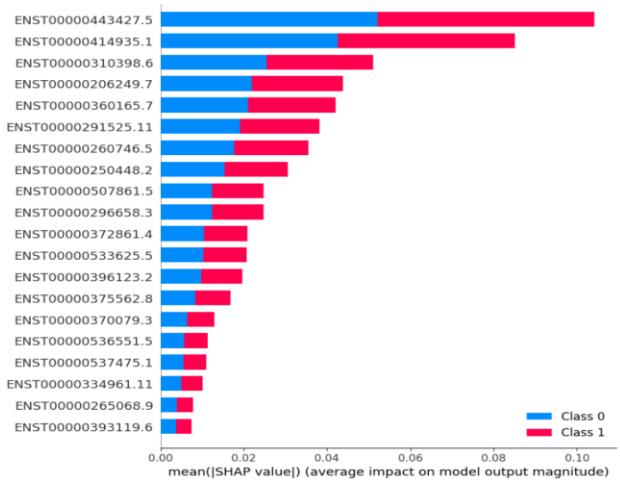

Figure 1: SHAP feature values for LASSO set on the ER label.

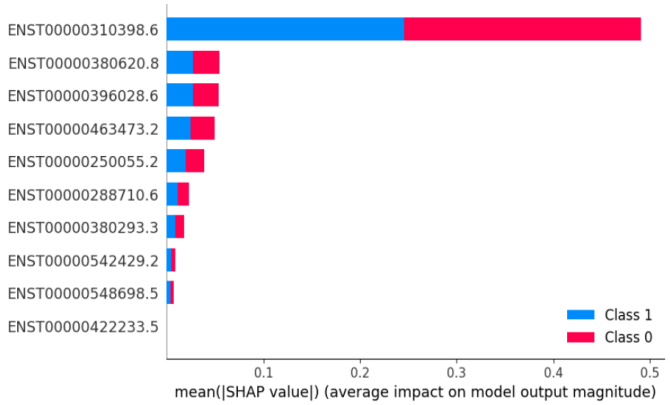

Figure 2: SHAP feature values for mRMR set on the ER label.

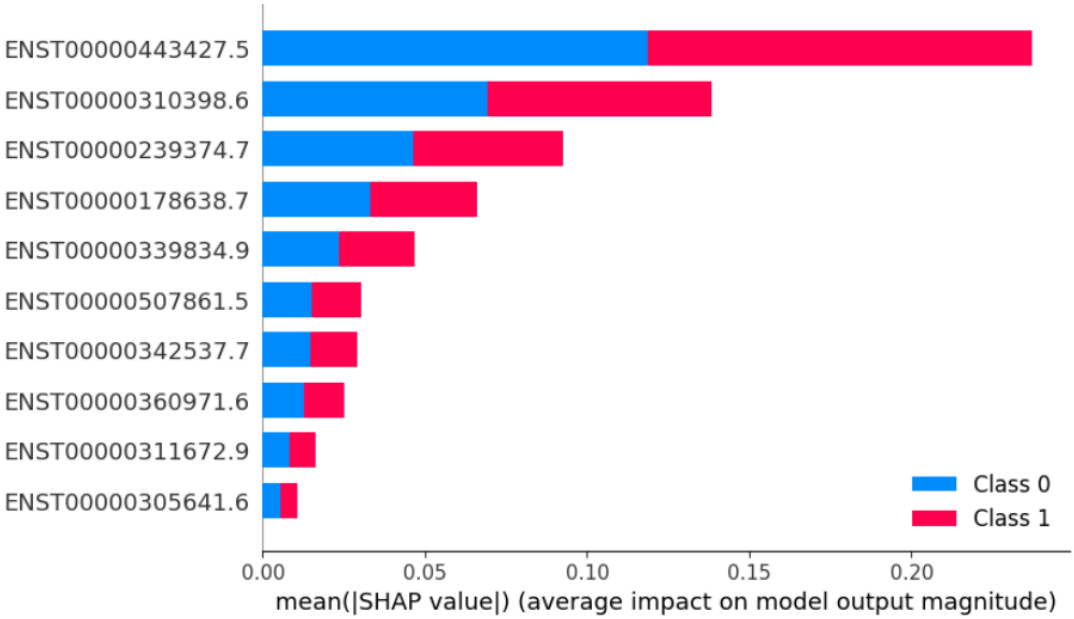

Figure 3: SHAP feature values for SHAP set on the ER label.

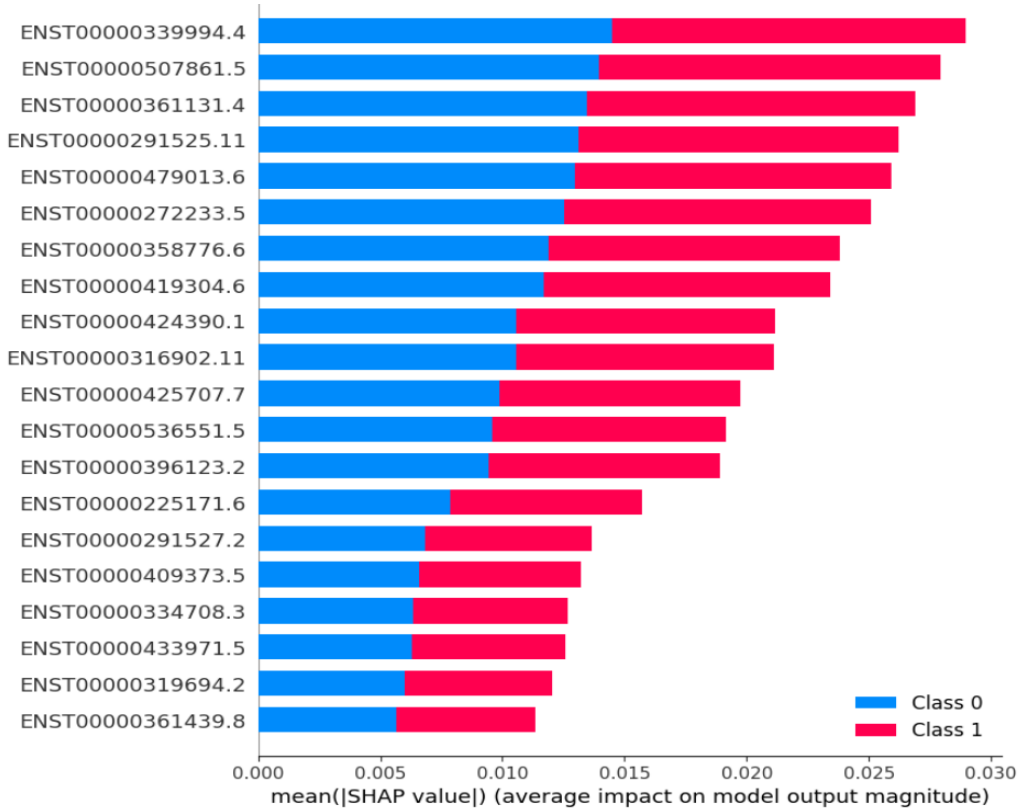

Figure 4: SHAP feature values for RFS set on the ER label.

These tables show the SHAP values of the top features to show the effect they have on the decision of the model. Some features are shared in the LASSO, MRMR, and SHAP. These are highly active biomarkers and have high values on all the models they appear on. The following table shows some of the most shared biomarkers between the different sets:

Table 3: Shared features between feature selection methods for ER label.

BIOMARKER	LASSO	MRMR	SHAP	RFS
ENST00000310398.6	X	X	X	
ENST00000443427.5	X		X	
ENST00000291525.11	X			X

These shared features were some of the most influential in their respective models. Overall, there were not many features shared between the sets, but Lasso, mRMR, and SHAP had a few in common, while RFS had only one high importance feature shared, and it was only contained in the LASSO set. The next set of tables are the LIME results of the ER label. The LIME representation is based on a sample-by-sample basis and the sample used for this is ER positive. The following are the LIME explanations of the same sample on the different feature sets.

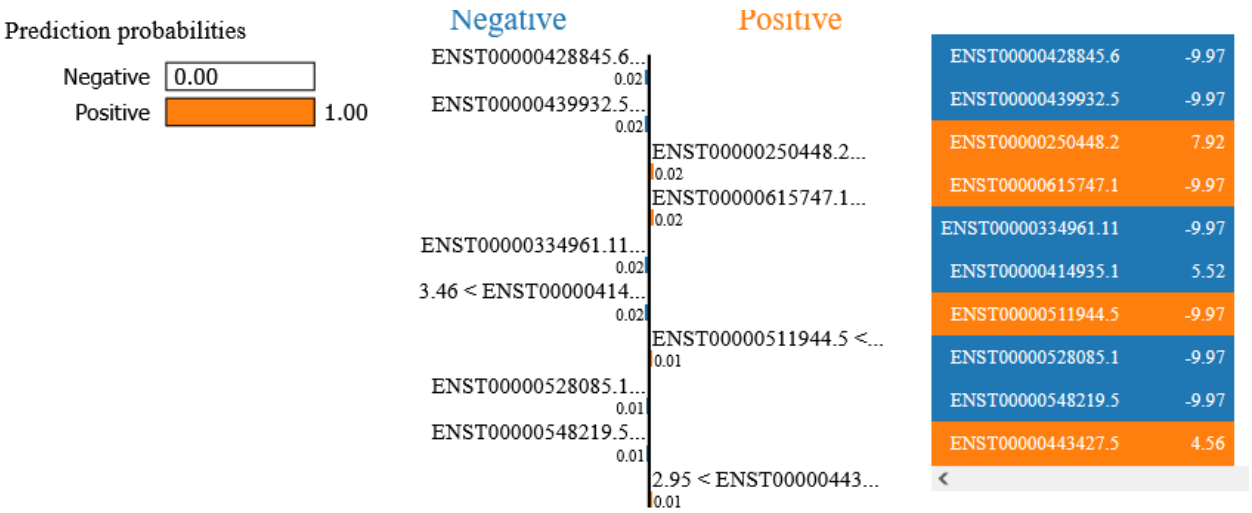

Figure 5: LIME table for Lasso set on ER label.

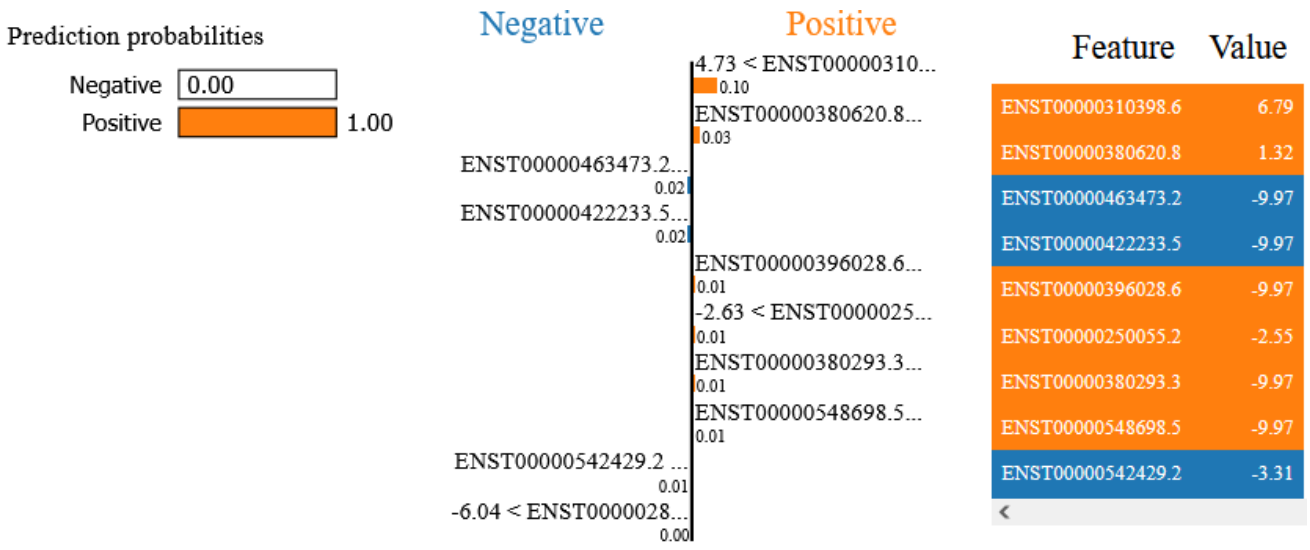

Figure 6: LIME table for mRMR ER label

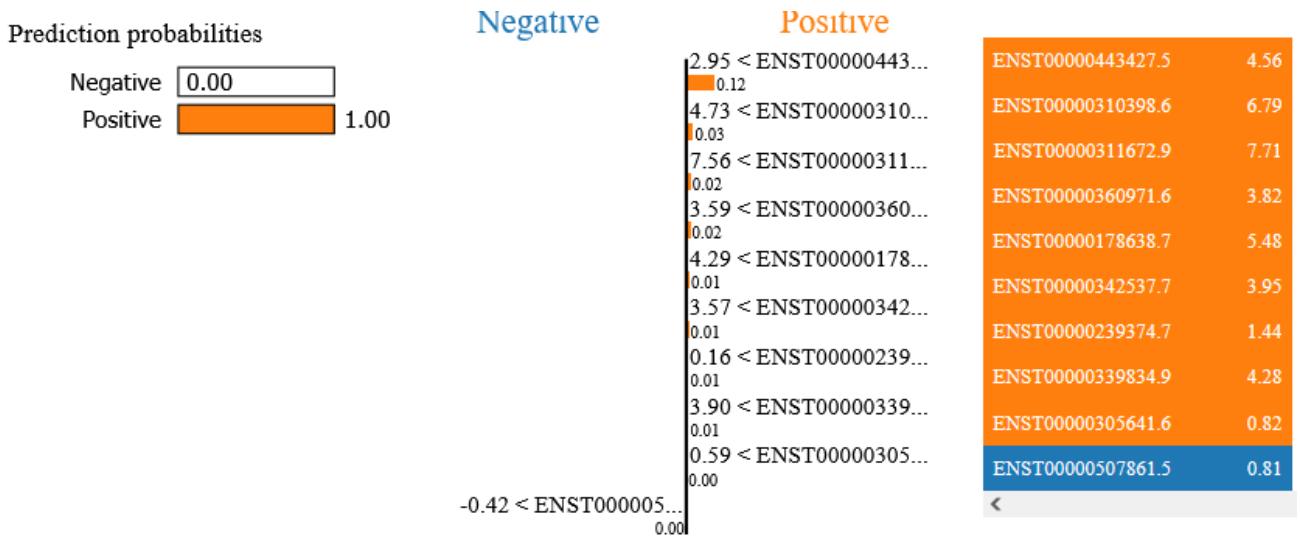

Figure 7: LIME table for SHAP ER label

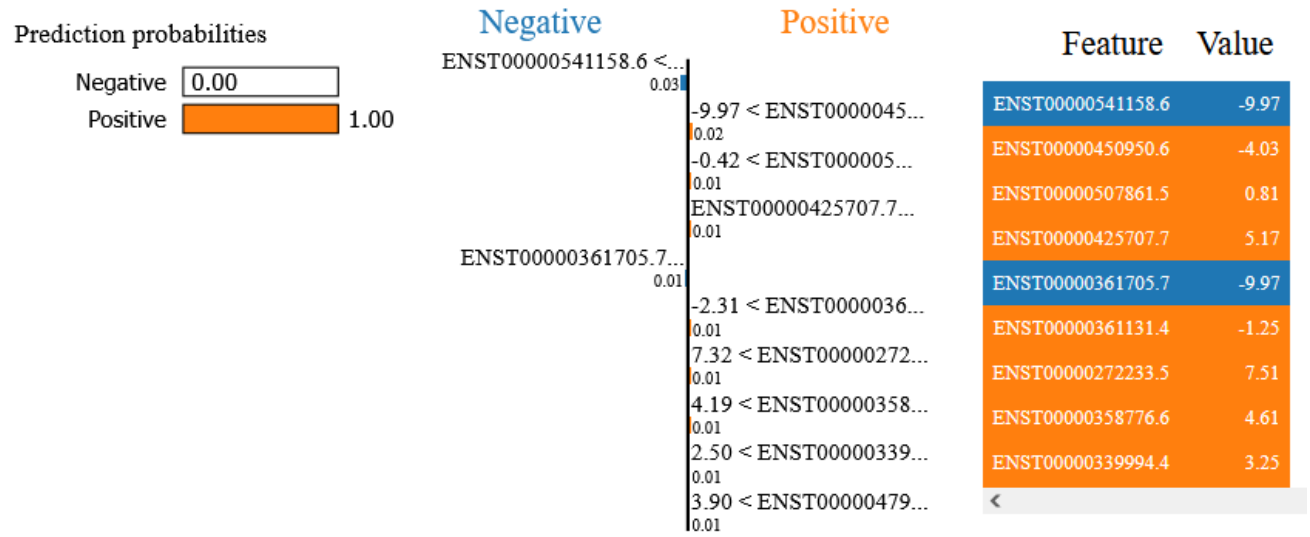

Figure 8: LIME table for RFS ER label.

The most interesting data from the LIME tables is the entire feature set prediction probability. For this sample, every model calculated that the ER label was 100% there. LIME also provides a different explanation for the features, showing if the value of a specific feature is making the model lean toward an optimistic prediction or a pessimistic prediction. In this case, for this sample, most of the features on all models contribute some value towards predicting a positive presence of the ER label.

Label HER2

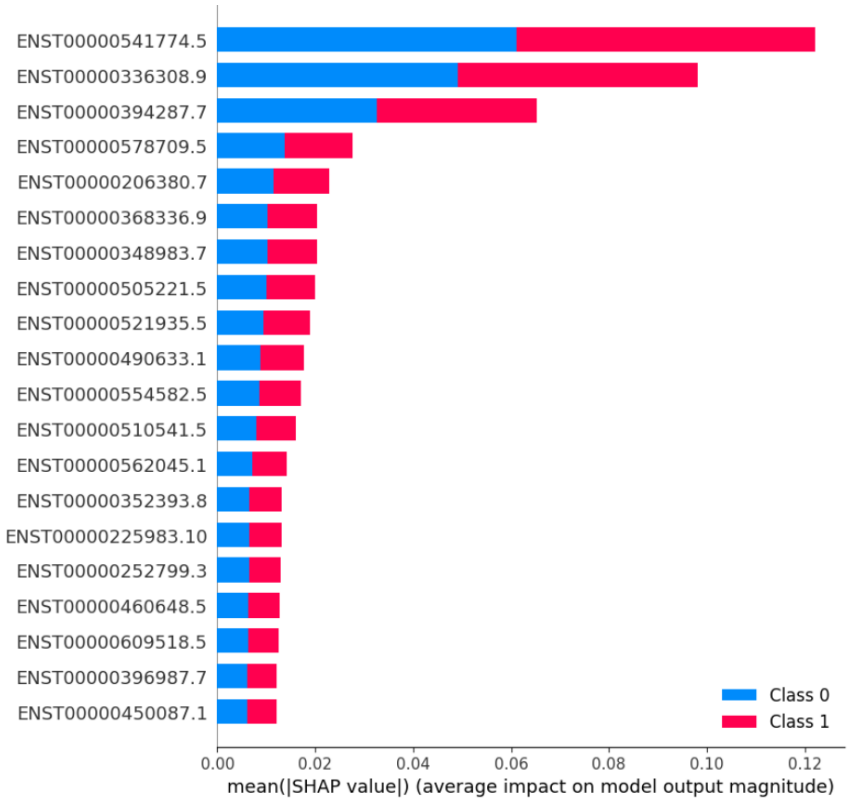

Figure 9: SHAP feature values for LASSO set on the HER2 label.

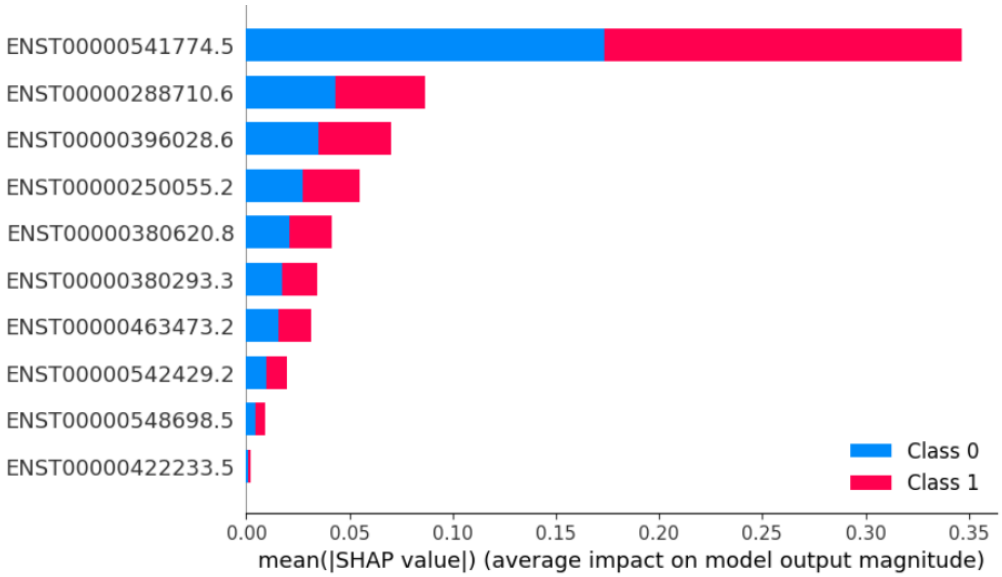

Figure 10: SHAP feature values for mRMR set on the HER2 label.

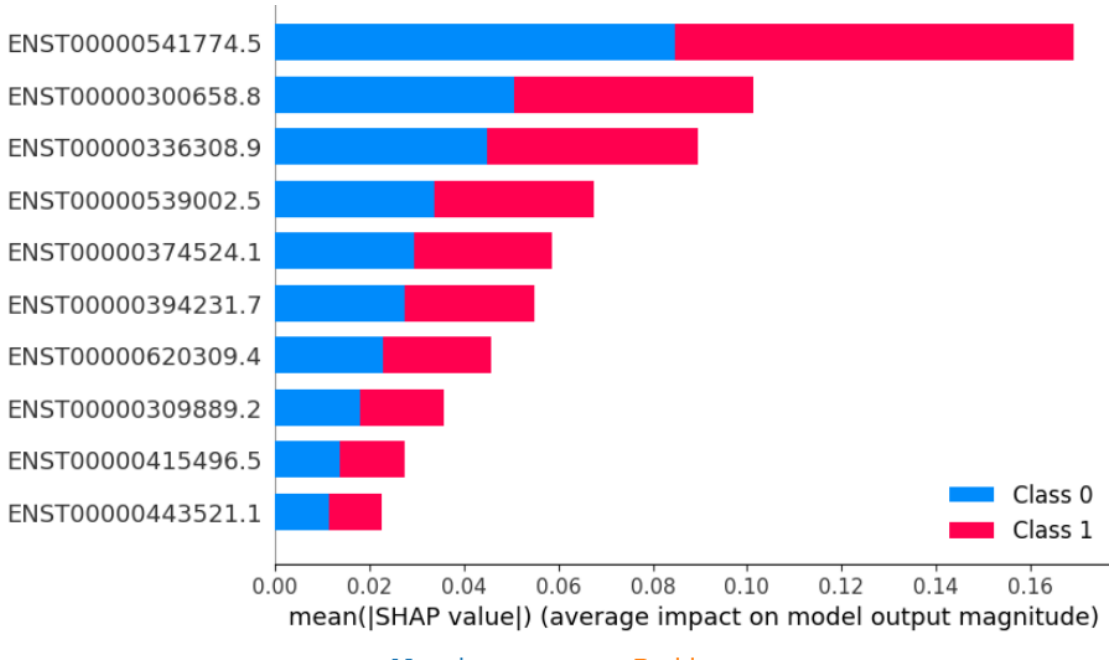

Figure 11: SHAP feature values for SHAP set on the HER2 label.

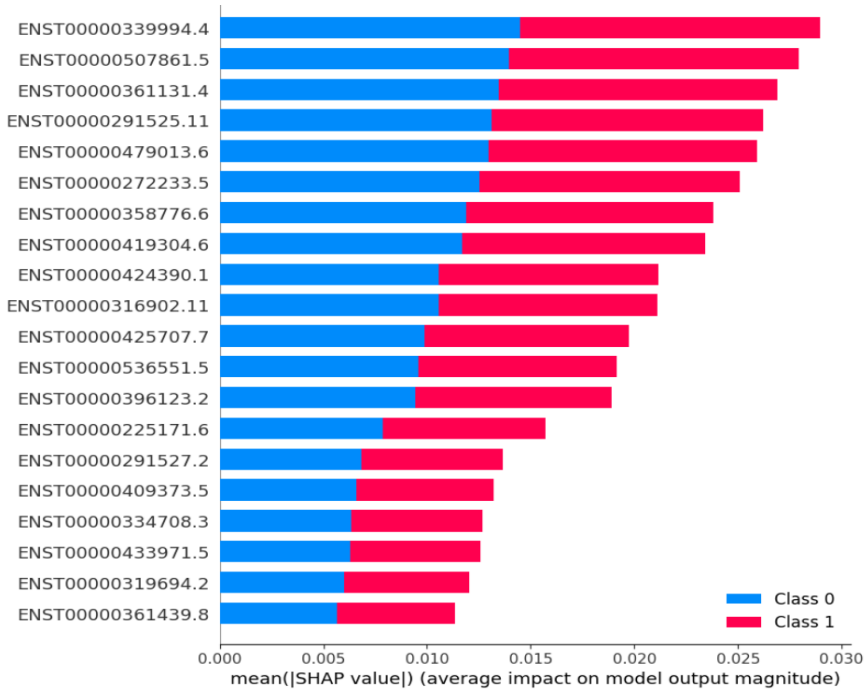

Figure 12: SHAP feature values for RFS set on the HER2 label.

Table 4: Shared features between feature selection methods for HER2 label.

BIOMARKER	LASSO	MRMR	SHAP	RFS
ENST00000541774.5	X	X	X	
ENST00000336308.9	X		X	

There were fewer features shared in this label, but the ones that are shared between the LASSO, mRMR, and SHAP sets were the most important out of all the others of the set. Also, note that for this label, and this is the label that has an extremely high importance score over the rest for the models that selected it.

The following are the LIME tables using same sample for the previous section. This sample is HER2 negative.

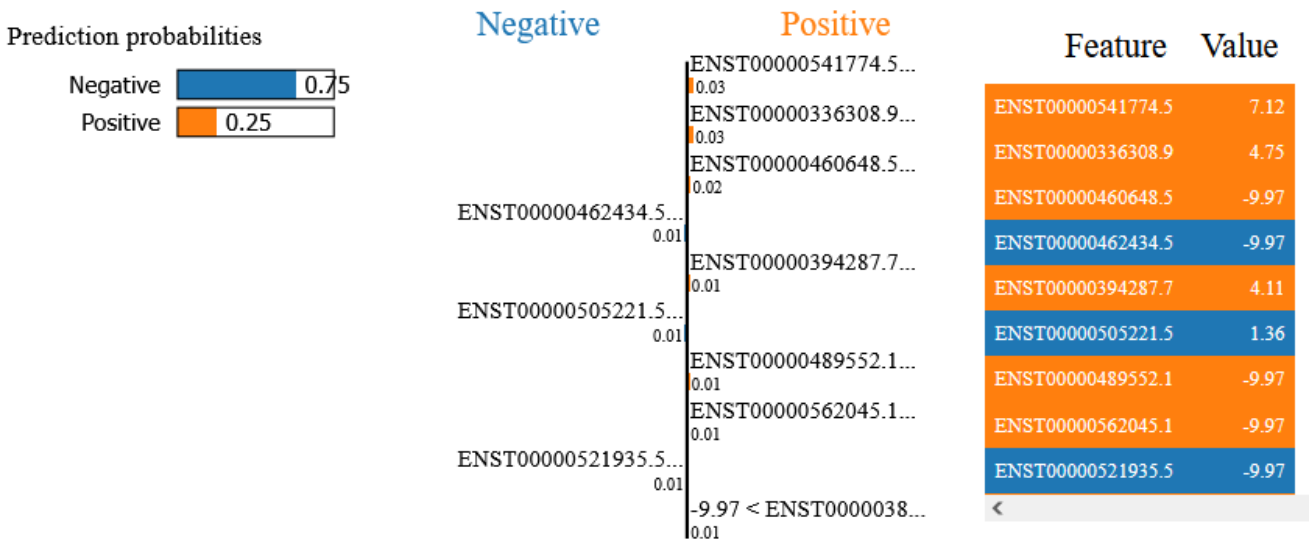

Figure 13: LIME table for Lasso set on HER2 label.

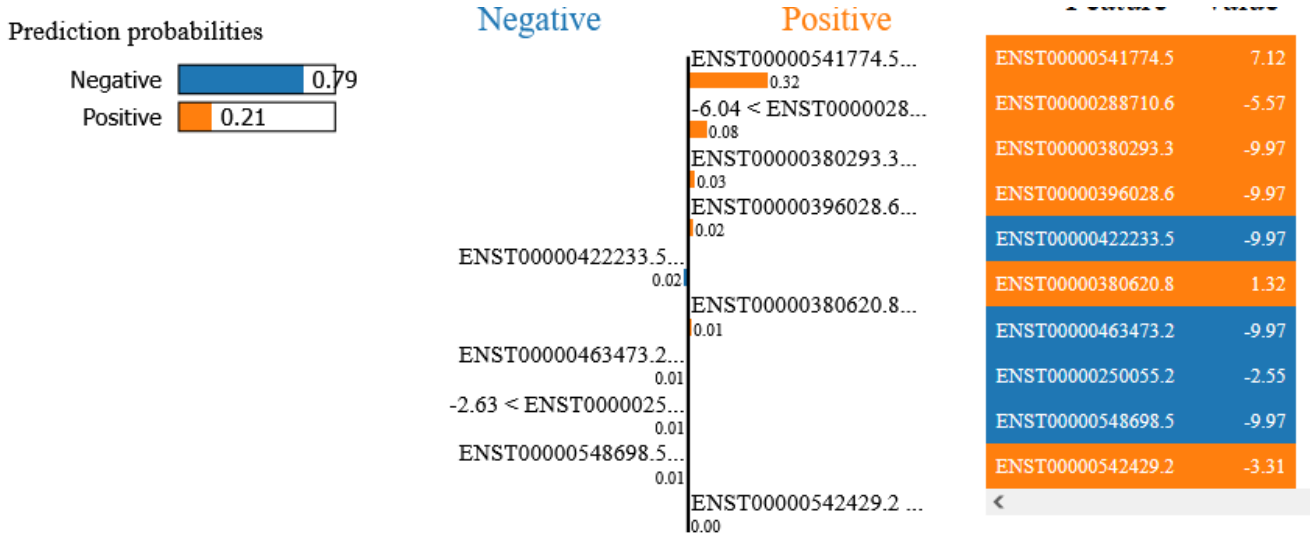

Figure 14: LIME table for mRMR HER2 label

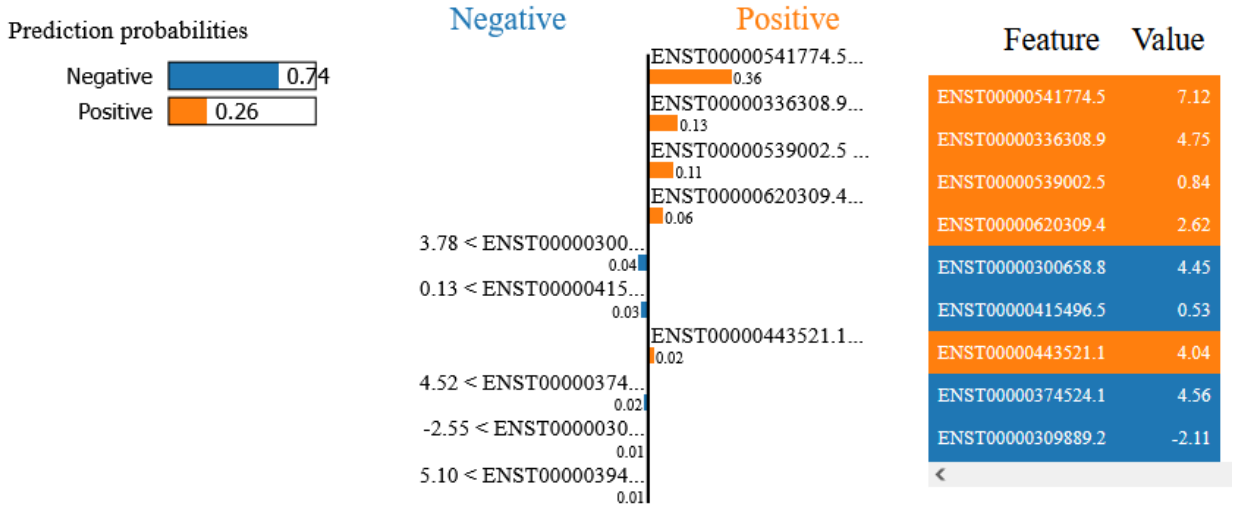

Figure 15: LIME table for SHAP HER2 label

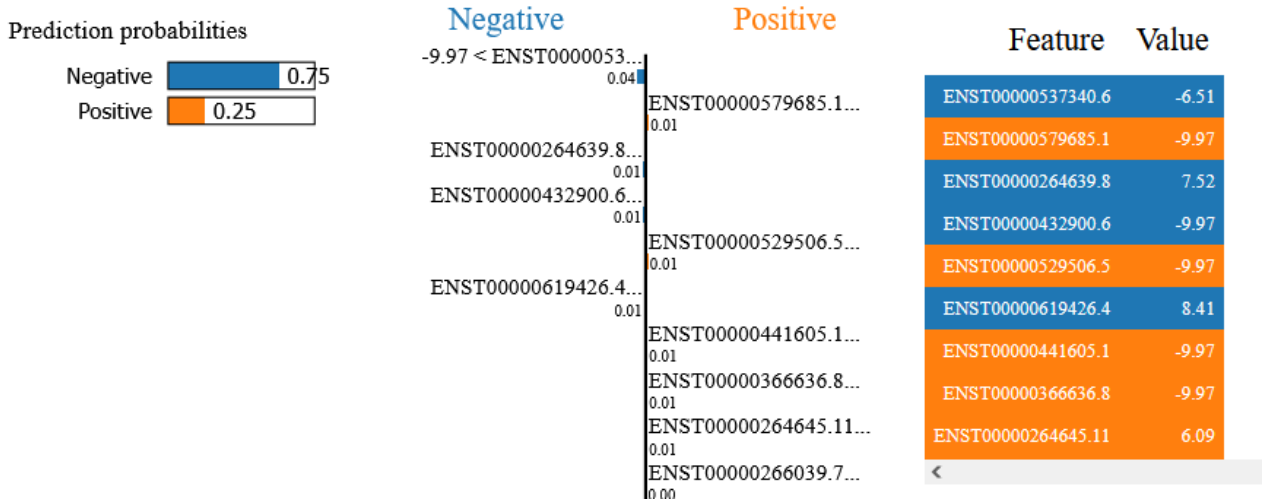

Figure 16: LIME table for RFS HER2 label

Each feature set has a different feature structure and feature values as they do not share many features however the final probabilities are similar for all the models.

PR label

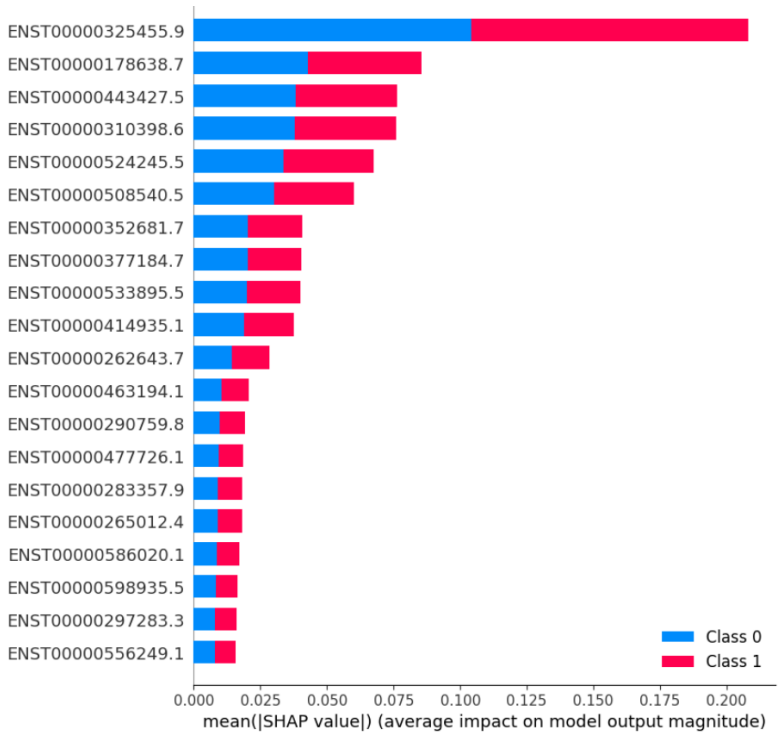

Figure 17: SHAP feature values for LASSO set on the PR label.

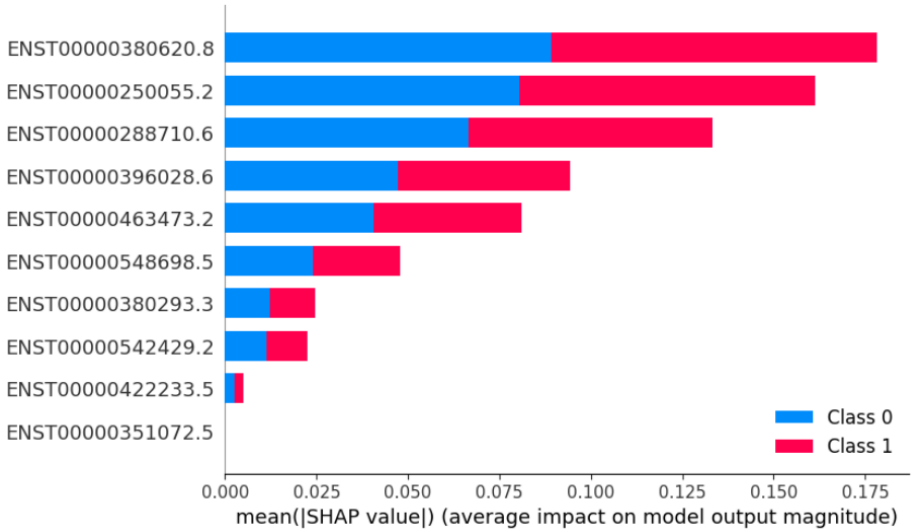

Figure 18: SHAP feature values for mRMR set on the PR label.

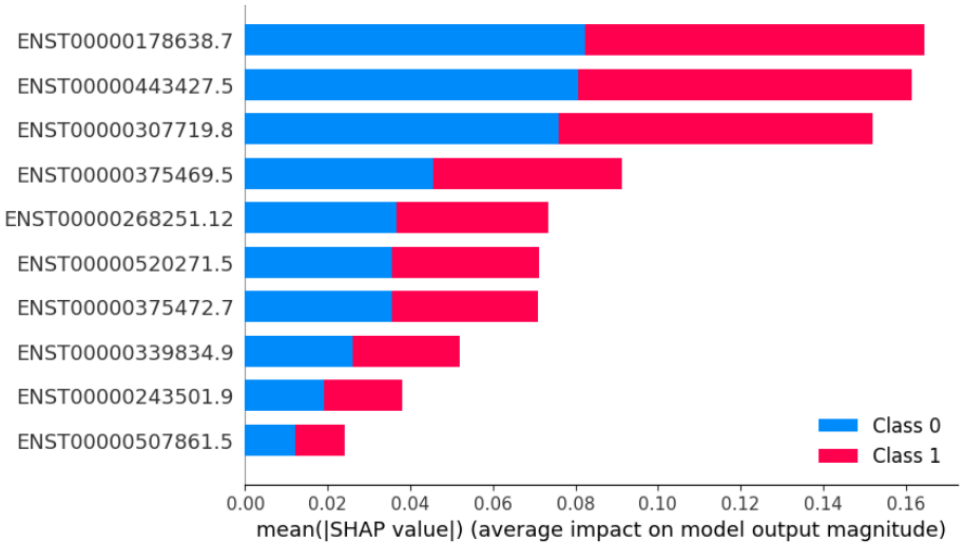

Figure 19: SHAP feature values for SHAP set on the PR label.

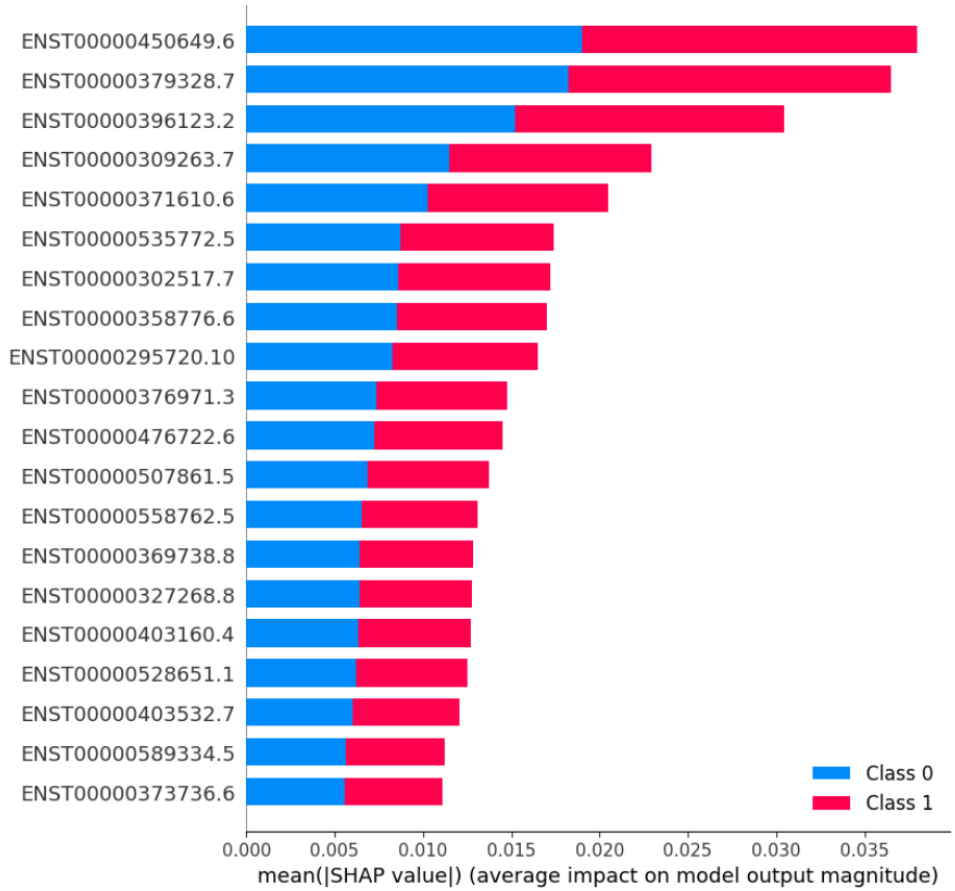

Figure 20: SHAP feature values for RFS set on the PR label.

The PR label has been the least stable of all the labels, having low and fluctuating accuracies. There is only a single feature shared feature between the mRMR sets and the SHAP for the PR label which are the lowest scoring feature sets of all, and it did not score a high importance value. Every model has different feature importance, from depending mainly on a single feature to three.

The following are the LIME tables using same sample for the previous section. This sample is PR positive.

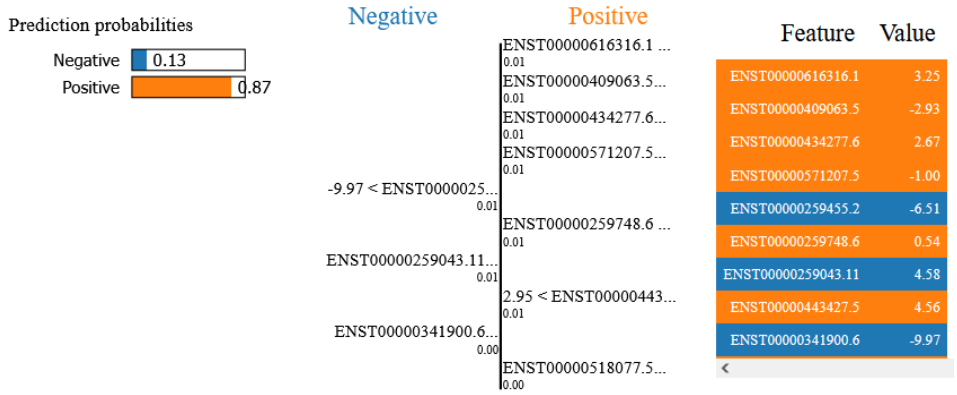

Figure 21: LIME table for Lasso set on PR label.

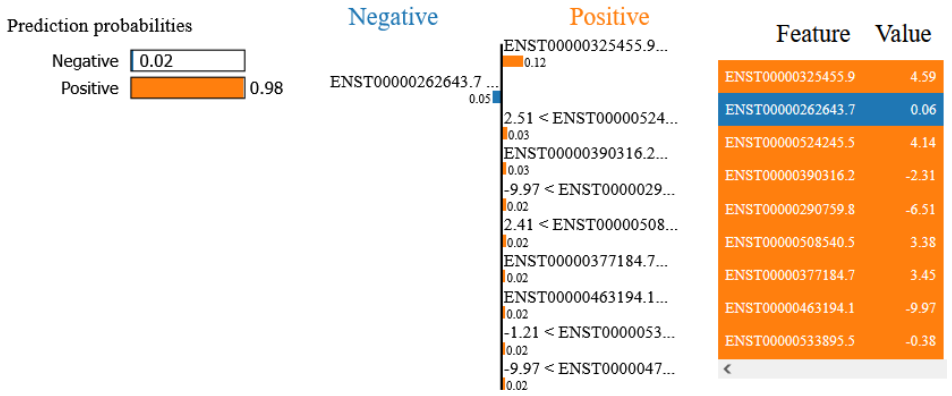

Figure 22: LIME table for mRMR PR label

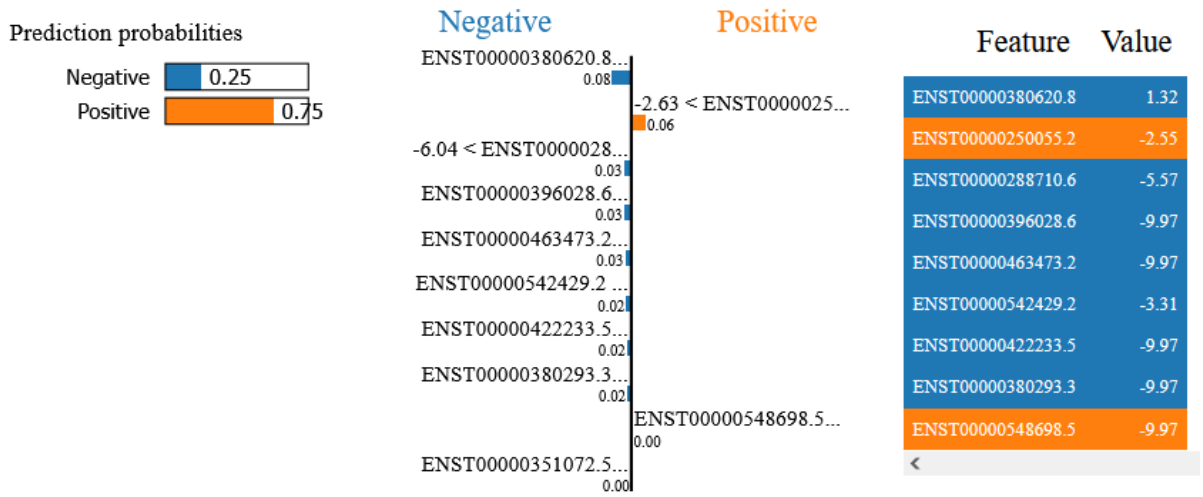

Figure 23: LIME table for SHAP PR label.

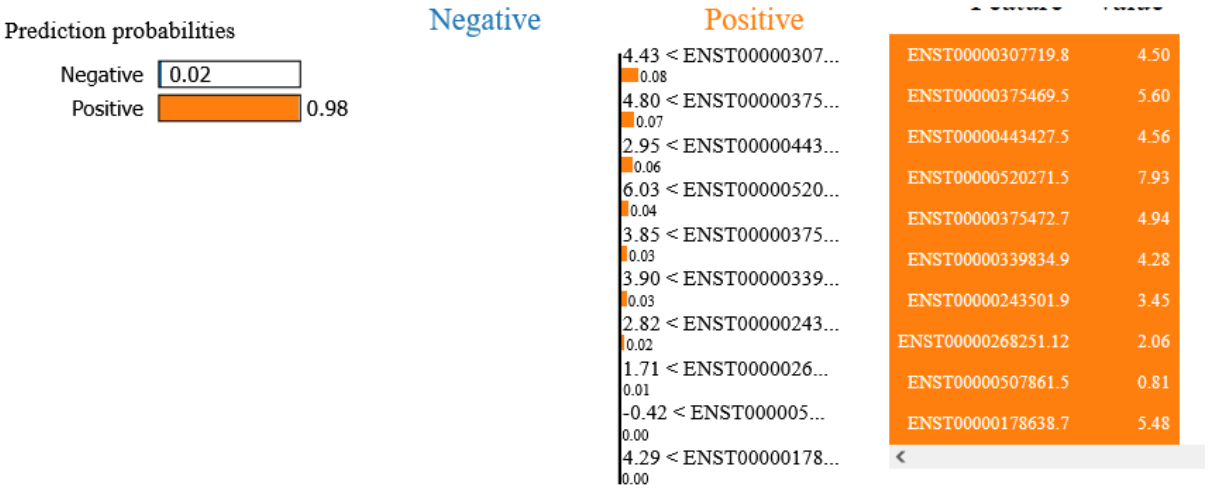

Figure 24: LIME table for RFS PR label

The effects of each set being a completely different feature set can be observed on the LIME tables. While every model produces a high probability for positive, the probability and the feature distribution are different for every feature set.

TN label

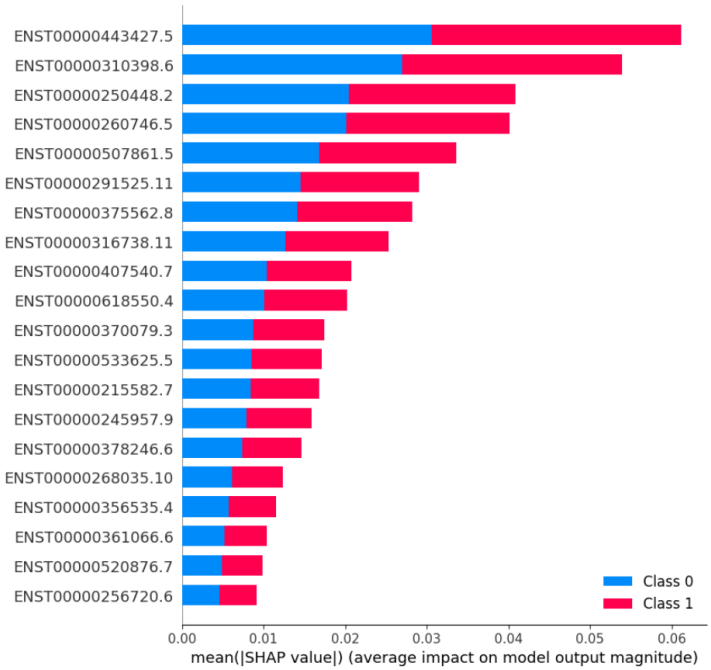

Figure 25: SHAP feature values for LASSO set on the TN label.

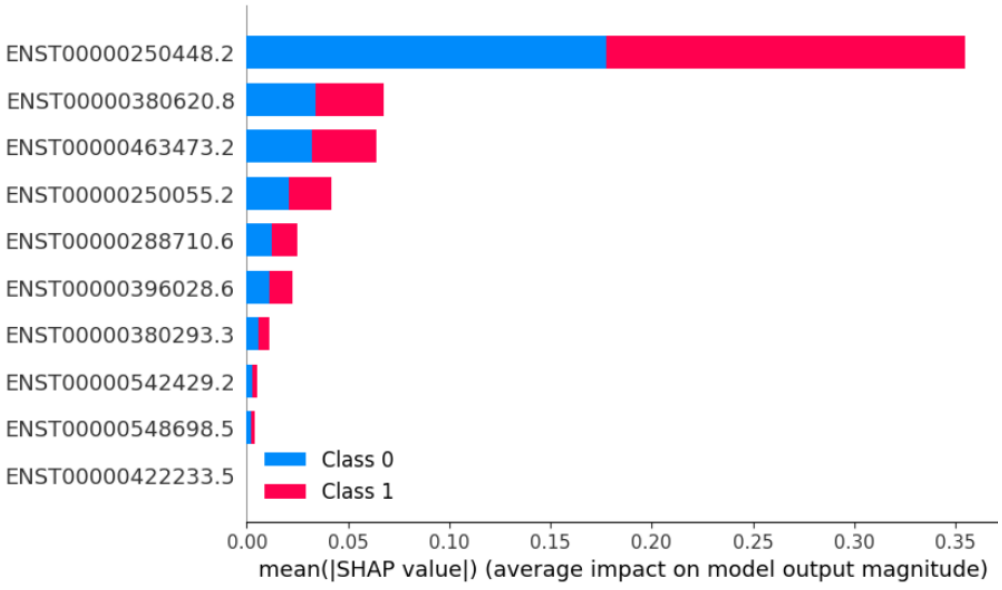

Figure 26: SHAP feature values for mRMR set on the TN label.

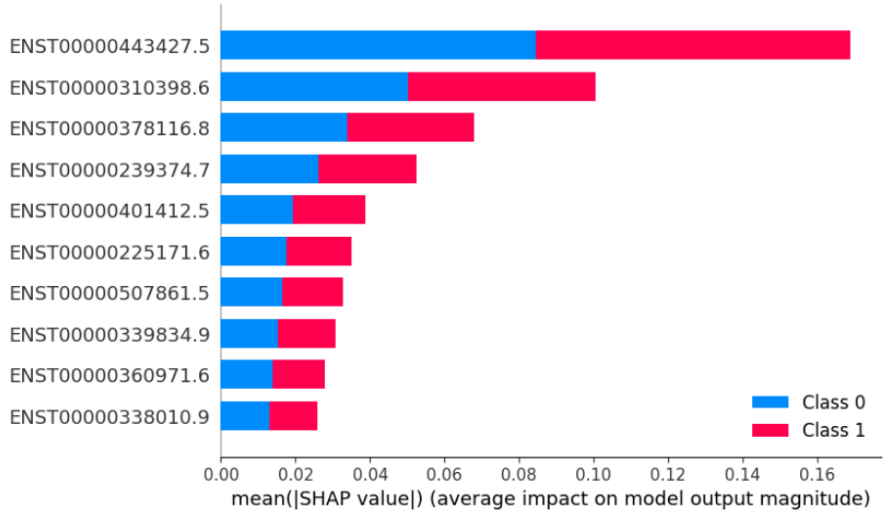

Figure 27: SHAP feature values for SHAP set on the TN label.

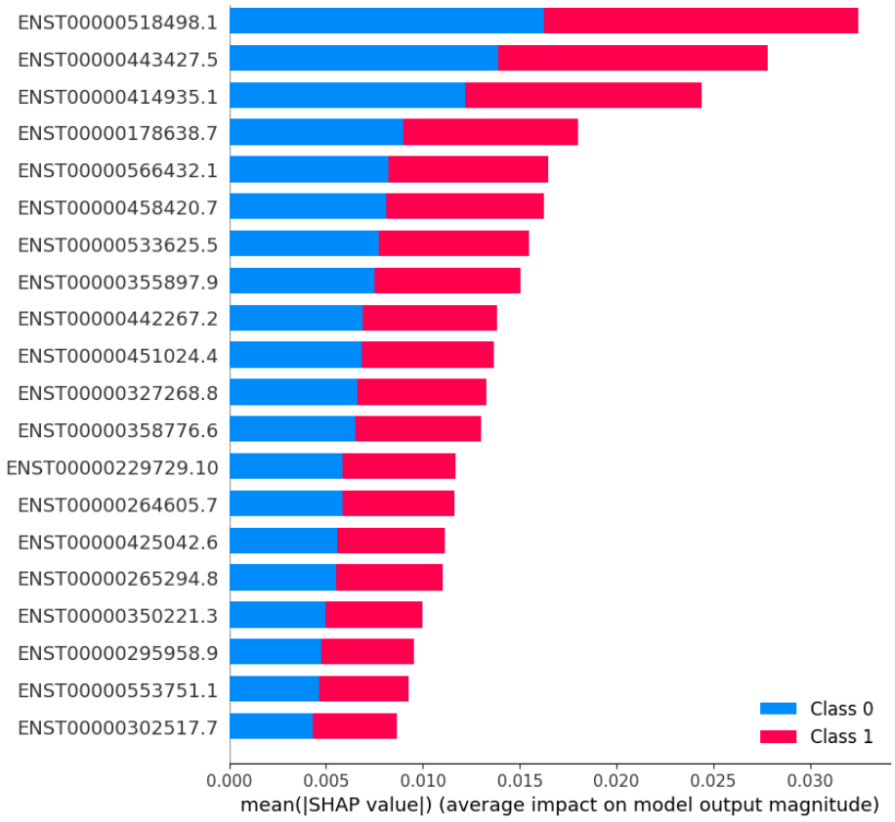

Figure 28: SHAP feature values for RFS set on the TN label.

Table 5: Shared features between feature selection methods for TN label.

BIOMARKER	LASSO	MRMR	SHAP	RFS
ENST00000443427.5	X		X	X
ENST00000310398.6	X		X	

For the first time there is a feature that is in the LASSO and SHAP set on top of also showing up in the RFS set. Another thing to note is that each model has a different feature distribution. mRMR is essentially relying on a single feature and that feature is not present in any of the other models.

The following are the LIME tables using same sample for the previous section. This sample is TN negative.

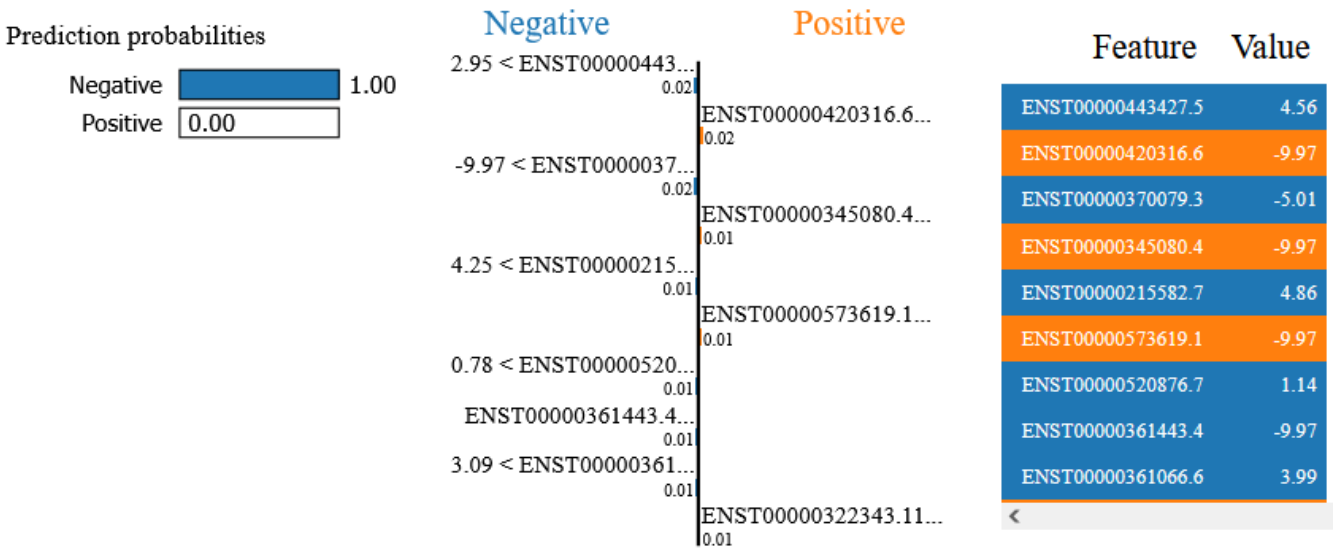

Figure 29: LIME table for Lasso set on TN label.

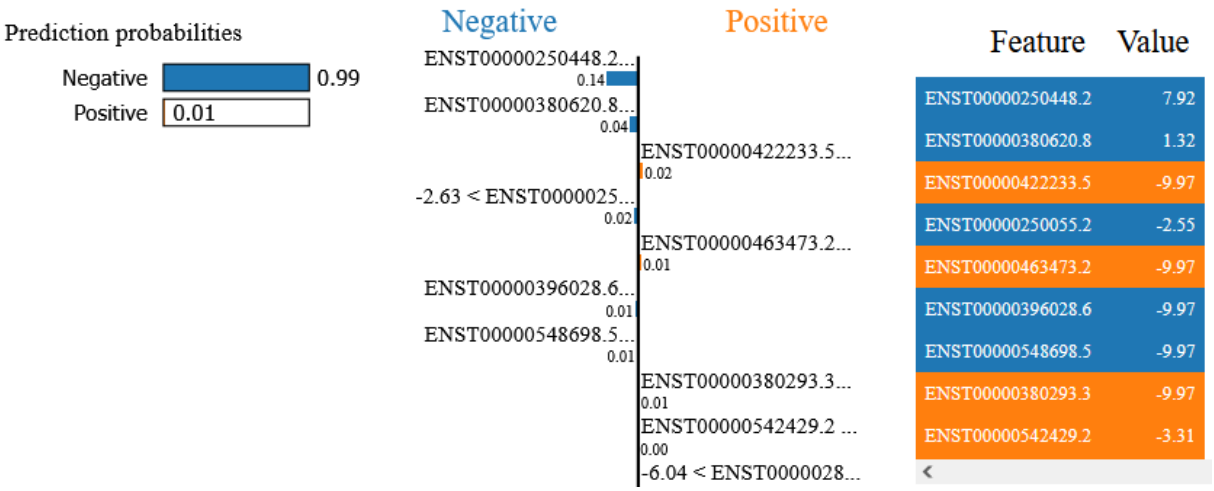

Figure 30: LIME table for mRMR TN label.

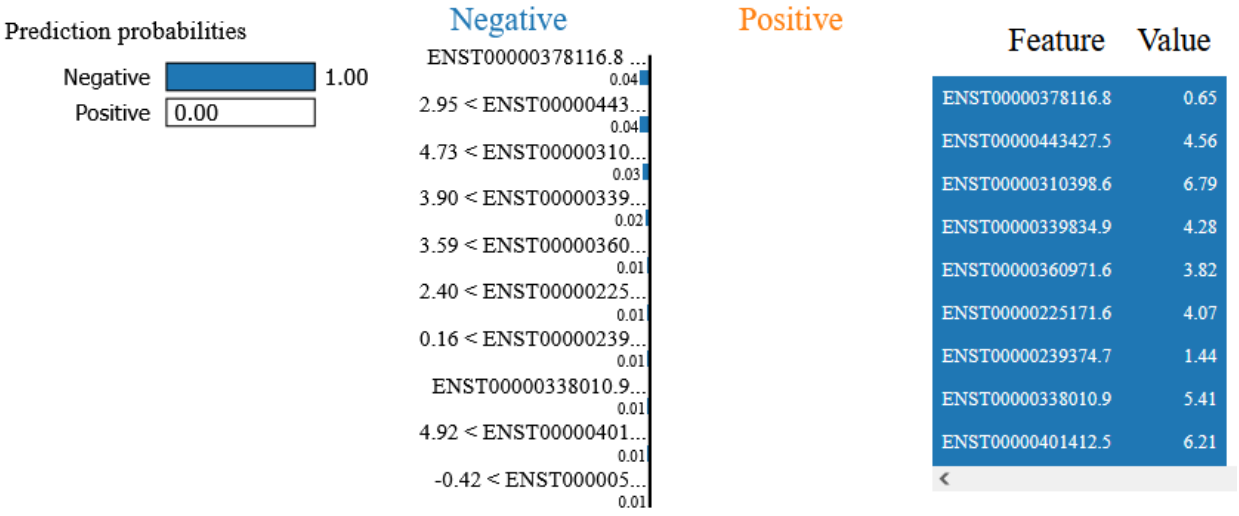

Figure 31: LIME table for SHAP TN label.

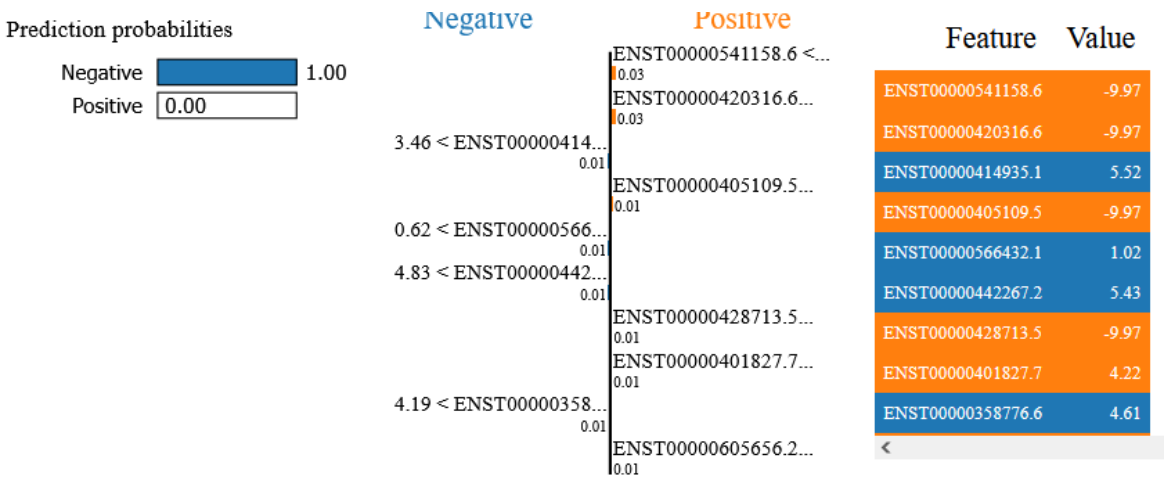

Figure 32: LIME table for RFS TN label

The LIME tables are similar except for the SHAP set, which considers all features to indicate a negative prediction. This label is a particular case because the TN label is used when all the other labels are negative. Because the same test sample was used for all explainable steps and at least one of the previous labels was predicted positive, it would make sense to expect the TN predictor to be negative. This is interesting as the different labels and models are entirely disconnected, yet the results obtained for this label are straightforward.

CONCLUSION

This paper went over application of LASSO, mrMR, SHAP, and RFS feature selection algorithms and their impact on predictive models for the breast cancer biomarkers ER, HER2, PR, and TN labels. Even with the high-dimensional feature space relative to the sample size, the baseline model achieved high accuracy in predicting labels. From the explored feature selection methods RFS scored the highest accuarcy at a high but still reduced number of features and while that helps when seeking the highest accuarcy it produced a large set. On the other hand, the LASSO set scored results close to the ones from RFS, It also did it with just 13% of the feature set size allowing for easier feature set analisys. Because not all labels yielded the same performance it's important to consider working with them individualy. The ER label outperformed others, while PR exhibited the lowest accuracy across different feature selection methods and the HER2 and TN labels scoring high accuracy around the 90% mark. The evaluations of the explainable techniques of SHAP values and LIME tables highlighted features with high influence and their impact on model predictions giving an idea of how each biomarker was being evaluated by the model and allowed to observe some similarities between sets. Shared influential features across different feature selection methods were identified for most labels making them a good target for future research.

Some future research areas to explore with this would be to test these feature sets to a diferent database to ensure reability, and use the lime tables on more samples to gain more information on possible patterns on a case by case basis.

LIST OF REFERENCES

[1] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 68-77. https://doi.org/10.5114/wo.2014.47136

[2] Zhao, X., Liu, K., Fan, W., Jiang, L., Zhao, X., Yin, M., & Fu, Y. (2020). Simplifying reinforced feature selection via restructured choice strategy of Single Agent. *2020 IEEE International Conference on Data Mining (ICDM)*. https://doi.org/10.1109/icdm50108.2020.00096

[3] Liñares-Blanco, J., Pazos, A., & Fernandez-Lozano, C. (2021). Machine learning analysis of TCGA cancer data. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/peerj-cs.584

[4] *GDC TCGA Breast Cancer (BRCA).* UCSC Xena. (n.d.). Retrieved April 7, 2023, from https://xenabrowser.net/datapages/

[5] *Breast Invasive Carcinoma (TCGA, Firehose Legacy)*. CBioPortal for Cancer Genomics. (n.d.). Retrieved April 7, 2023, from https://www.cbioportal.org/study/clinicalData?id=brca_tcga

[6] Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5). https://doi.org/10.1002/widm.1424

[7] Lundberg. M. Scott, Lee. Su-In . (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing System*

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

[9] Smazzanti. (n.d.). *Smazzanti/MRMR: MRMR (minimum-redundancy-maximum-relevance) for automatic feature selection at scale.* GitHub. Retrieved April 9, 2023, from https://github.com/smazzanti/mrmr

[10] Marcotcr. (n.d.). *MARCOTCR/Lime: Lime: Explaining the predictions of any machine learning classifier*. GitHub. Retrieved April 9, 2023, from https://github.com/marcotcr/lime.