Predicting BRCA Using Cancer Associated Microbiome

Maitree Patel

Many studies have shown the link between microbiome and cancer types. This association can be studied and utilized to predict or diagnose different types of cancers. A study by Poore *et al., 2020* correlated microbiome signatures with 33 different cancers with almost perfect accuracy. Faced with controversy due to invalidation of their results by another study by Giwahi *et a.l*, *2023*, the findings were proved erroneous. Stemming from this back-and-forth between research groups, the main question at hand still remains reliability and reproducibility of methods used. The main aim of this study was to use the data produced by Poore *et al.*, *2020* to devise a machine learning-based technique to predict breast cancer from normal samples. A random forest classifier was used for this and revealed better than average model accuracy. Another random forest classifier was used to discriminate breast cancer from other types of cancer to investigate if it was associated with a distinct microbiome signature. This model revealed a high accuracy suggesting a unique microbial signature associated with breast cancer. One Rule (OneR) machine learning model was also built to determine the one microbial genus most associated with breast cancer. The OneR model revealed the association of *Achromobacter* as the “one rule” in predicting breast cancer. The main aim underlying these models was to test the predicting power of the cancer microbiome dataset.

Introduction

The human microbiome is the entire repertoire of microbial species that colonize the human body. Changes in this population of microbes reflects diseases like cancer. Microbial populations in tumors have been shown by many studies to have distinct signatures associated with different cancers which can be exploited to predict them using patient data. Biologically this makes sense as metabolites secreted by microbes induce changes in the host's cells, directly affecting cancer biology. Therefore, patterns in changes of microbiome can be used to diagnose or predict cancer types.

Next Generation Sequencing (NGS) technologies have enabled us to sequence the microbiome of patients to help us decipher the above problem. And the vast volume of data from sequencing can be analysed using the advent of Artificial Intelligence (AI) and more specifically Machine Learning (ML) algorithms. A paper titled “Microbiome analyses of blood and tissues suggest cancer diagnostic approach” by Poore et al., 2020 filtered and used microbial counts to characterise the cancer microbiome for 33 types of cancer. Though the paper sheds light on the possibility of use of microbiome data in cancer diagnostic applications, the paper itself has been retracted due to analysis done by another paper titled “Major data analysis errors invalidate cancer microbiome findings” that invalidated its results since the original paper had extremely high accuracy of their ML models. This nudges researchers to ask questions about

reproducibility and replicability; the confidence in methodology used and the results interpreted from them.

The main purpose of this analysis was to use data produced by *Poore et al., 2020* to explore the potential of microbiome data in detecting and predicting breast cancer (BRCA). And to compare the results of this analysis to their retracted paper.

The specific aims of this analysis includes, using microbiome data to:

1. Discriminate between normal and BRCA samples.
2. Discriminate between BRCA and other cancers to investigate whether a microbial signature is present for BRCA.
3. Find one taxa (one rule) that can predict BRCA (i.e. is one taxa associated with BRCA)

Methodology

Data Description

Data from the supplemental tables of the Poore et al., 2020 paper were used for the analysis in this paper, they included the TableS8_T2T_KrakenUniq_BIO_Fullset.csv (raw microbial reads data matrix) and the TableS9_metadata_KrakenUniq_BIO_Fullset.csv (metadata table).

Data Preprocessing

First step in data preprocessing was merging the Table S8 and S9 using HashMaps in Java. The merged data was then manipulated, filtered and normalized in python using jupyter notebooks. Normalization of raw microbial reads was carried out by taking the log of raw counts divided by the sum of counts of that row. A small constant of 1e-10 was added to the division to prevent log of any zeros produced.

Data preprocessing output files that were generated for this analysis include:

1. **brca_vs_others.csv:** included raw microbial reads with disease_type column that was coded to have two categories, “brca” or “others” (for other cancer types).
2. **brca_vs_normal_raw.csv:** raw microbial reads of samples with “Breast Invasive Carcinoma” under disease_type and sample_type column which was coded to “brca” or normal.
3. **brca_vs_normal.csv**: the same data as brca_vs_normal.csv but with normalized counts.

Models

Three ML models were made to answer the three main aims of the analysis, these include:

1. A Random Forest classifier to distinguish BRCA from normal samples (using sklearn python)

2. A Random Forest classifier to distinguish between BRCA and other cancers (using sklearn python)
3. An One Rule classifier to identify a single microbial taxa that can predict BRCA from normal samples (using OneR package in R)

Details of input data, model metrics and evaluation are given below.

Random Forest: BRCA vs Normal Samples

To test the prediction of BRCA from normal samples using microbial reads data, the preprocessed brca_versus_normal.csv data was used as the input to compare between a random forest and a logistic regression model in python. The accuracy of predictions and ROC curves were obtained for comparison. Due to better performance and accuracy, the random forest model was chosen.

Leave-One-Group-Out Cross-Validation (LOGO-CV) was used to validate the model and to account for batch effects. The target class of sample type did not contain significant imbalance (as shown in *figure 1*) and therefore, it was not accounted for in this analysis.

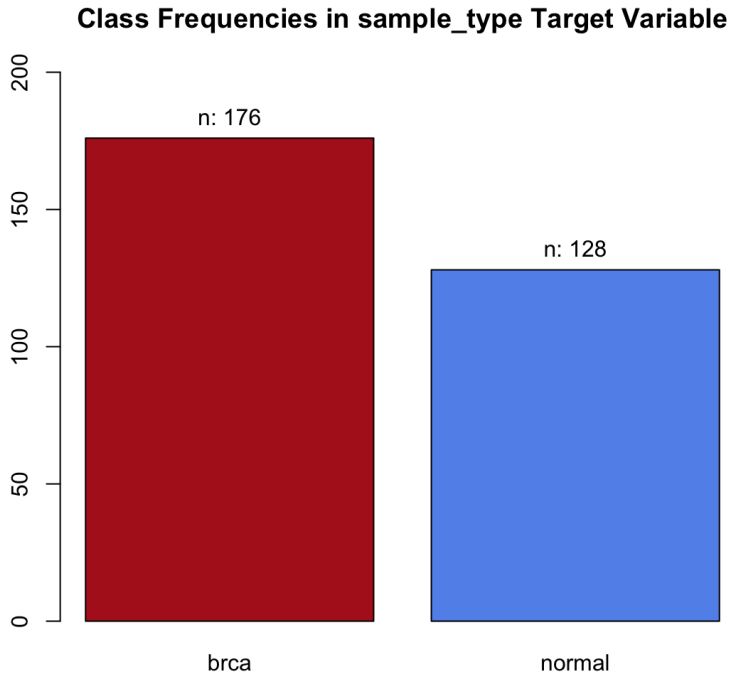


Figure 1: *BRCA and normal category representation in the target class sample type*

Random Forest: BRCA vs Other Cancers

To test whether BRCA has a microbial signature associated with it (BRCA versus other cancers) using microbial reads data, the preprocessed brca_versus_others.csv data was used as the input data to compare between a random forest and a logistic regression model in python. The accuracy of predictions and ROC curves were obtained for comparison. Due to better performance and accuracy, the random forest model was chosen.

The model was trained on 80% of the data and performance was tested on the remaining 20%. There was target class imbalance, i.e. the sample size of BRCA was 304 counts and other cancers were 7523 counts (as shown in *figure 2A)*. Therefore, oversampling was used to handle class imbalance. This was done using the *imblearn* package in Python as shown in *figure 2B*.

Table 1: *Class imbalance in disease type target class*

	"BRCA" count	**"Other" count**
Before oversampling	304	7523
After oversampling	6017	6017

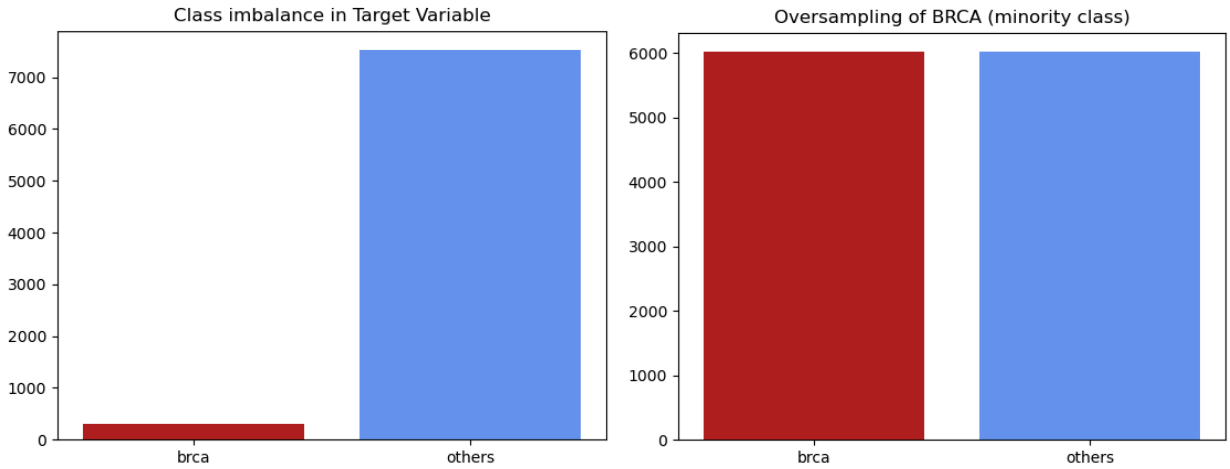


Figure 2: A) *Class imbalance in target variable disease_type,*
B) *class frequency after oversampling*

OneR: BRCA versus Normal Samples

The preprocessed data brca_vs_normal.csv was used as the input data for the OneR model. The model was built and predictions were made using the *OneR()* and *predict()* functions from the *OneR* package; and results visualized in R. Analysis of class frequency in sample_type target variable did not reveal class imbalance as shown in *figure 1*.

Results

Exploratory Data Analysis

The top genera in BRCA samples were Saccharomyces, Stenotrophomonas, Escherichia and Clostridium as shown in *figure 3.*

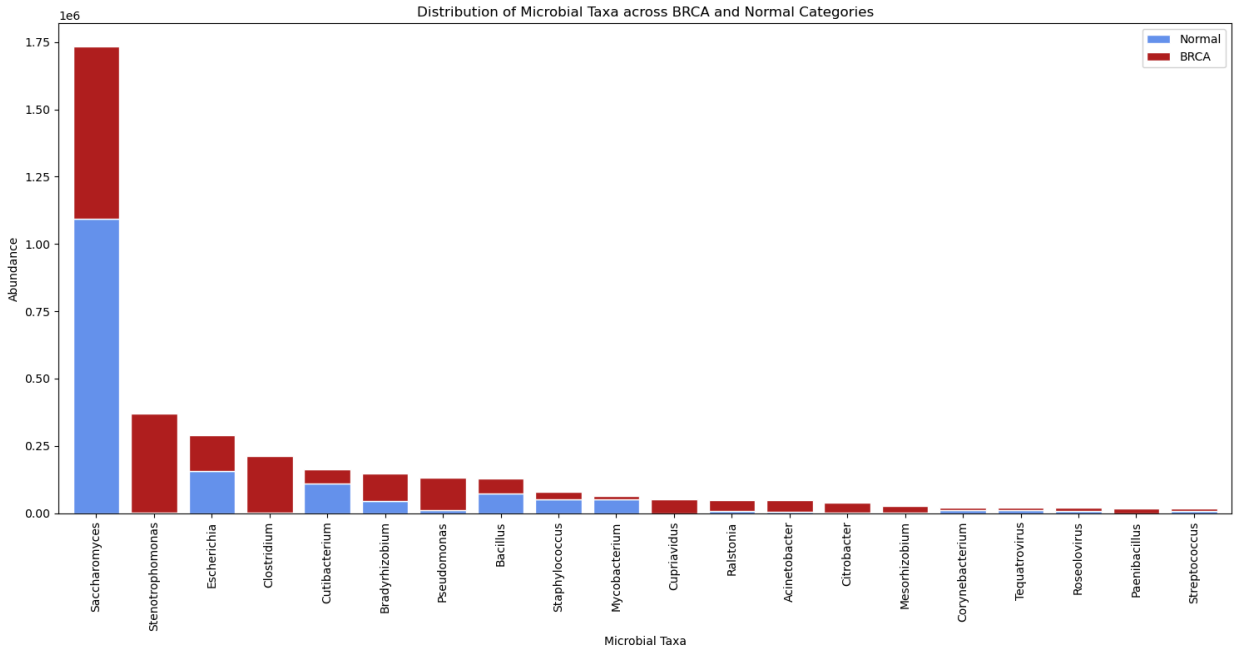


Figure 3: *Total number of reads of the top 20 most-abundant genera in BRCA samples*

Principal Component Analysis (PCA) of the entire normalized microbial reads data (*figure 4*) and of BRCA samples (*figure 5A and 5B*) was performed to visualize the highly-dimensional data. The PCA of the entire data was produced as a Graphical User Interface (GUI) as a part of this project in python (available at: PCAGUI). The GUI can be used to visualize the PCA of the log-normalized microbial reads with an option to color code it with any of the metadata columns. The mergedFullData.csv (provided in the GitHub repository) should be loaded into the GUI.

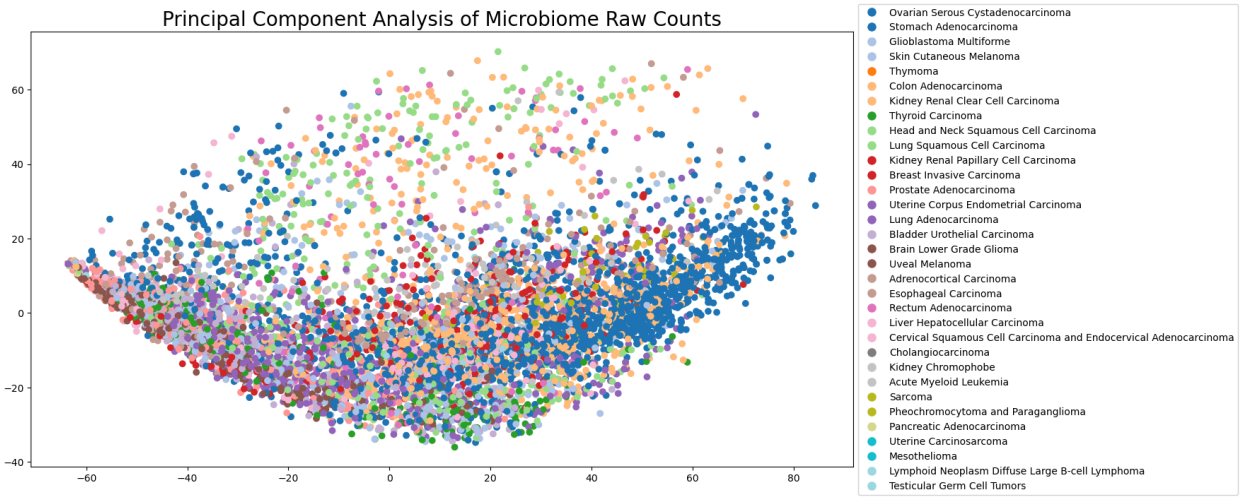
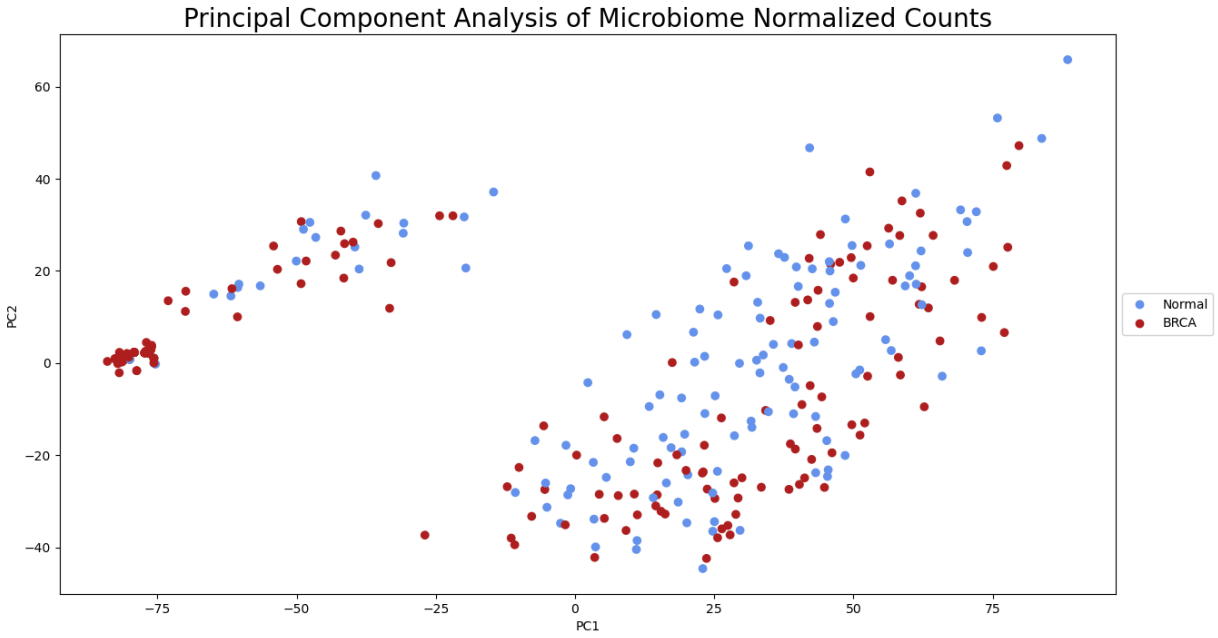


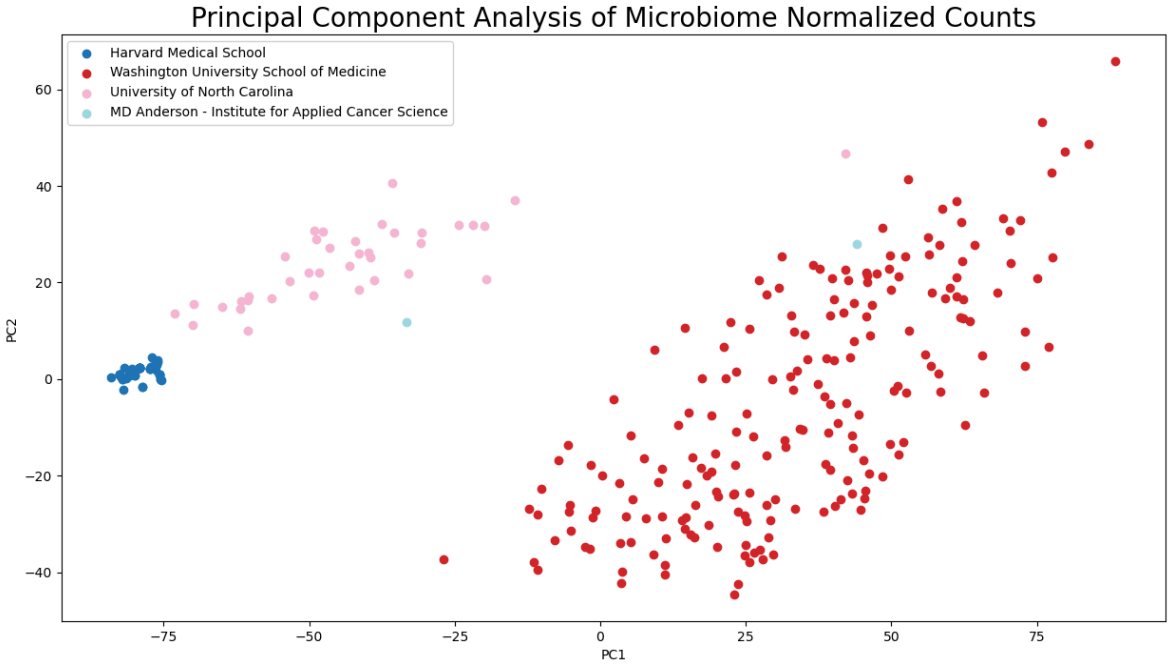
Figure 4: *PCA of the entire microbial reads data visualized in color by cancer type.*

Figure 5B shows that the separation in the data is due to the center from which data was sourced. Therefore, batch effects were accounted for while building the model using LOGO-CV.

The PCA also showed that MD Anderson had only 2 samples, both BRCA samples. Therefore, this center was left out entirely during model training using the LOGO-CV.



A)



B)

Figure 5: A) *PCA of BRCA samples visualized in color by the sample type,* ***B)*** *PCA of BRCA samples visualized in color by the center the data was obtained from.*

Figure 6 shows PCA using the brca_vs_others.csv output file and visualized using the disease_type target column as “brca” and “other”.It suggested class imbalance and therefore, this was handled before model training to prevent bias. Therefore, the PCA was key in decision making around model validation for the study.

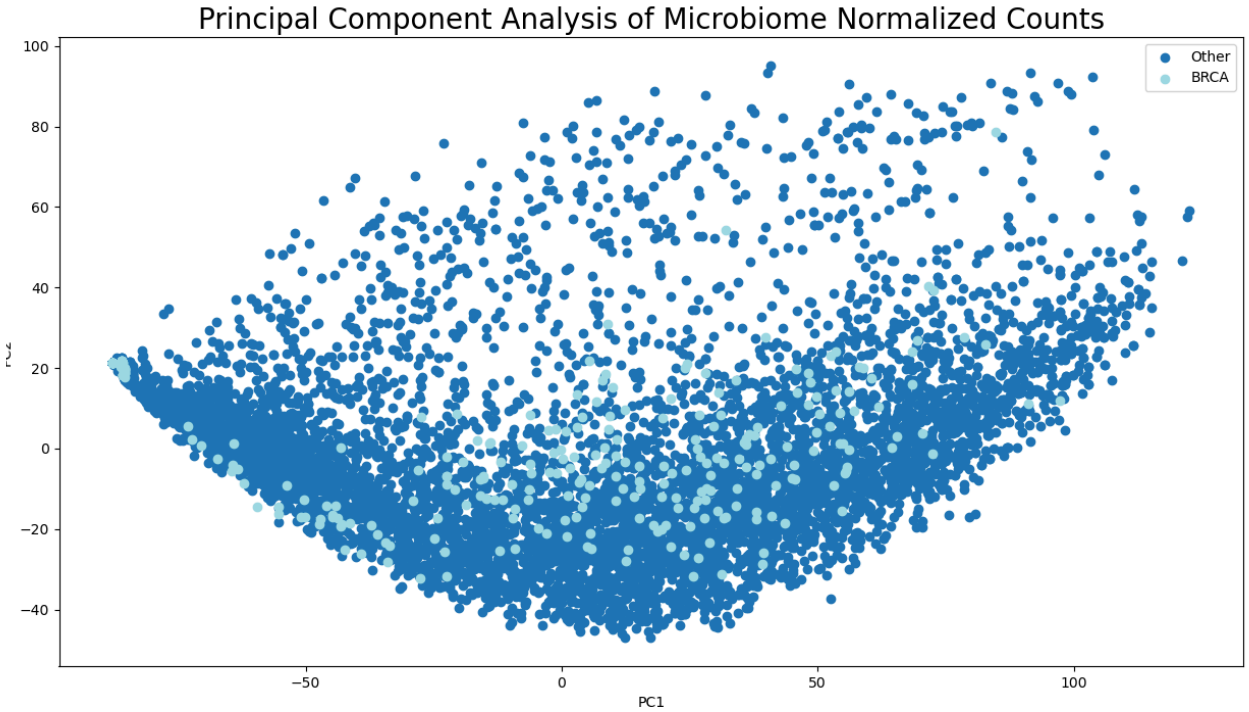


Figure 6: *PCA of the microbial reads data visualized in color by disease type.*

Model Performance

Random Forest: BRCA vs Normal Samples

The random forest model was used to predict BRCA from normal samples using LOGO-CV. The accuracy of testing the model on different centers is given in *table 2*, the average across the validations from all the centers was **0.62** and **AUC score** was **0.67**.

Table 2: *Model performance from LOGO-CV (MD Anderson - Institute for Applied Cancer Science was filtered out since it had only 2 samples)*

Center (on which model was trained)	Sample Size	Accuracy	AUC
University of North Carolina	65	0.49	0.50
Washington School of Medicine	199	0.73	0.77
Harvard Medical School	38	0.64	0.67

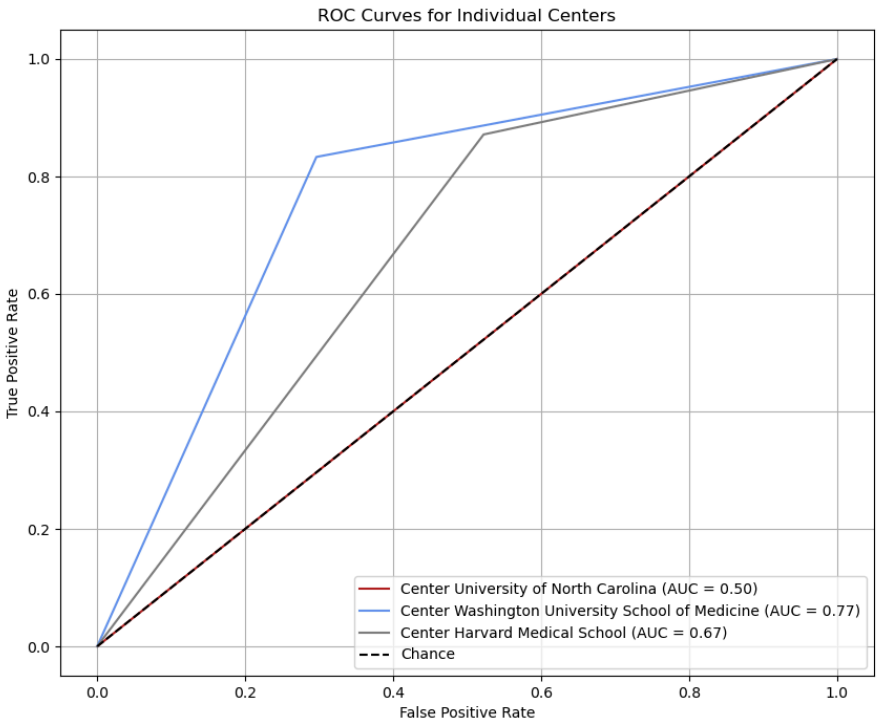


Figure 7: *ROC curves from LOGO-CV of the Random Forest model to predict BRCA from normal samples*

Random Forest: BRCA versus Other Cancers

To check whether BRCA has a distinct microbial signature, in other words, if BRCA can be distinguished from other cancers using microbial counts data, a random forest model was used. The **accuracy on 20%** of the data showed **0.99** and the **AUC score** was **0.76**.

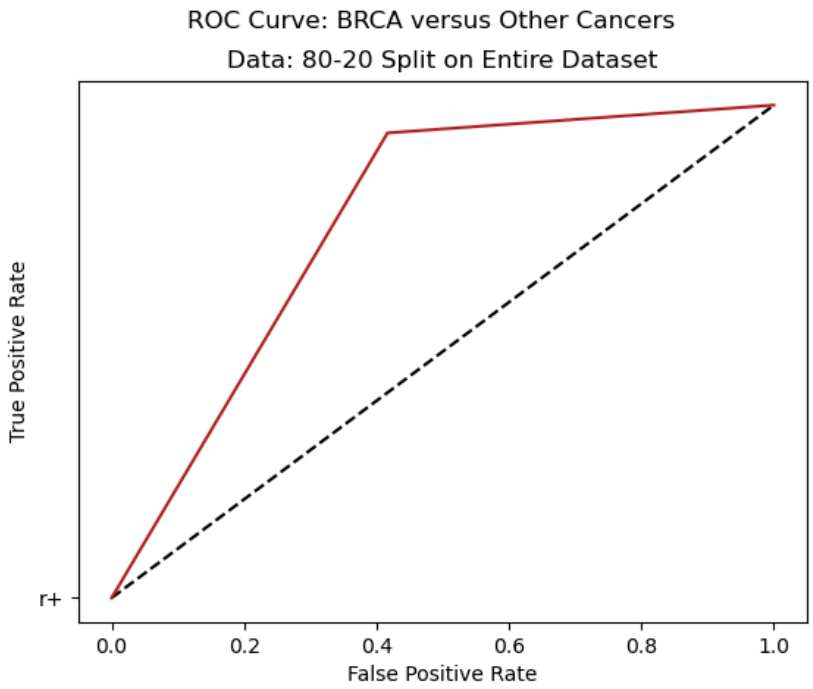


Figure 8: *ROC from the Random Forest model to predict BRCA from other cancers*

OneR

OneR model selected the *Achromobacter* as the feature with the least error, i.e. *Achromobacter* is the "one rule" which can make predictions about disease type target variable. The decision rules of the model include:

- If *Achromobacter* = low then sample_type = brca
- If *Achromobacter* = moderate then sample_type = brca
- If *Achromobacter* = high then sample_type = normal

This means the presence of low or moderate microbial counts of *Achromobacter* in a sample can help discriminate BRCA from normal.

The model uses Pearson's chi-squared test to assess association of *Achromobacter* reads (feature) with the target class (BRCA versus normal). The resulting **X-squared value** was **44.081** and **p-value = 2.679e-10** was obtained. The contingency table is as given below in *table 3*.

Table 3: *Contingency table for Burkholderia feature*

Disease Type	Low	Moderate	High	Sum
BRCA	*112	*27	37	**176**
Other	40	14	*74	**128**
Sum	**152**	**41**	**111**	**304**

Maximum in each column: ''*

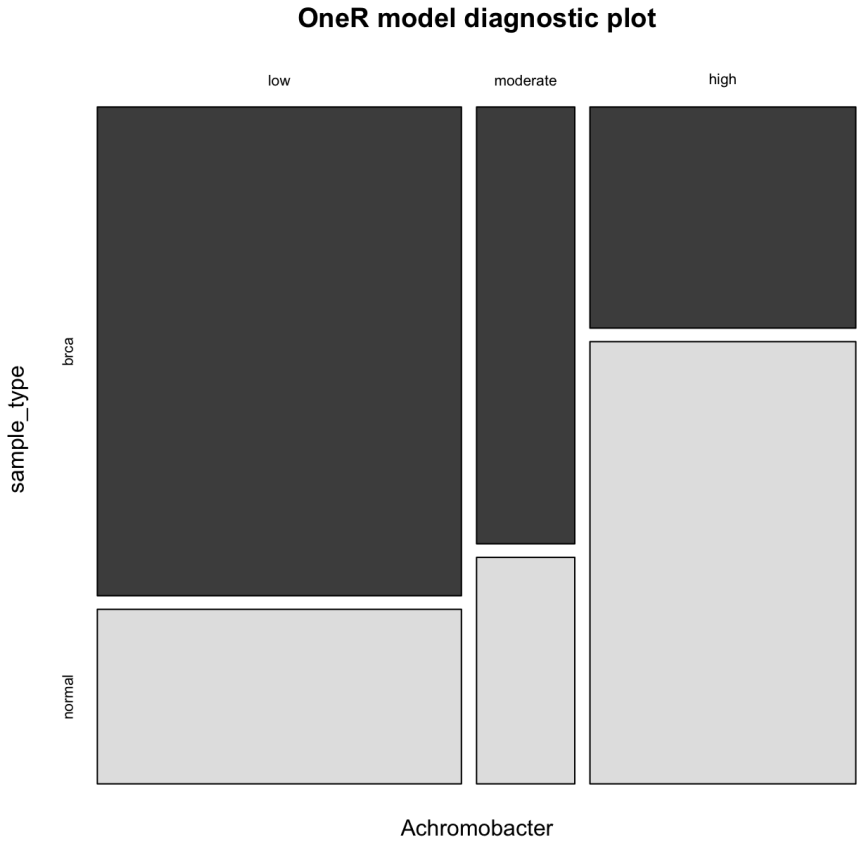


Figure 9: Mosaic plot of feature *Achromobacter* and sample type from OneR model

The resulting accuracy from the predictions on the data using the OneR model was 213 of 304 instances classified correctly (**70.07%**) with an error rate of **0.2993 (91/304).**

Discussion

Random forest model to predict BRCA from normal sample revealed an accuracy of 0.67, a little better than average, but still shows some biological signal that could help in predicting or diagnosing BRCA.

A random forest model showed a high accuracy in predicting BRCA from other cancers which suggests that BRCA is associated with a unique microbial signature which could be used in diagnosis as well. This result agrees with the authors of the Poore et al., 2020 paper in using this data for cancer-versus-cancer discriminations.

OneR model also showed a little more than average (0.70) accuracy in predicting cancer from *Achromobacter* microbial counts. Which also shows potential in diagnosing patients with a higher count of this taxa in their blood samples.

Achromobacter is a genus of gram-negative bacteria that are found to be opportunistic pathogens found in patients of cystic fibrosis. They have been associated with multiple infections in humans. They are also resistant to several antibiotics including cephalosporins. Wirtz, H.S. et al., 2013 shows that cephalosporins, along with penicillins, were the most prescribed antibiotic classes in breast cancer patients. This significantly highlights one of the limitations given below that the microbial reads could be a result of selection under drugs prescribed to the patients. There is no direct association of *Achromobacter* with breast cancer but they could be opportunistic in cancer immunosuppressed patients. This contradicts the OneR model, since low and moderate presence shows prediction for breast cancer.

These results come with limitations in the data and study. The analysis done by this paper showed that simply accounting for batch effects using a simple validation method resulted in much lower accuracies than as claimed by the source paper. This questions the methodology used by the authors.
A bigger sample size and standardized pipelines to filter human samples for microbial reads can help validate results from these studies. The pipeline also generated read identification at the genus level. In microbial genomes, we can see differences even at the strain level. These starin-level differences reflect microbial biology. Therefore diagnosis or predictions using them should be made with strain level data to better understand the diagnosing power of these models or even understand the underlying biology.
The blood sample data was also from cancer patients that could be assumed to undergo chemotherapy. The microbial reads could be a selection under certain drugs that might not reflect or be a result of cancer biology. Time series data in microbiome studies as these can help distinguish species associated with individuals before they develop a certain kind of cancer which can be even more valuable data in early diagnosis. Changes in microbial signatures in the presence of cancer or even healthy microbial signatures can help in therapeutic development.

References

Poore, G.D. et al. (2020) ‘Microbiome analyses of blood and tissues suggest cancer diagnostic approach’, Nature [Preprint]. Available at: https://doi.org/10.1038/s41586-020-2095-1.

Gihawi, A. et al. (2023) ‘Major data analysis errors invalidate cancer microbiome findings’, MBio, 14(5). Available at: https://doi.org/10.1128/mbio.01607-23.

Wirtz, H.S. et al. (2013) ‘Frequent Antibiotic Use and Second Breast Cancer Events’, Cancer Epidemiology, Biomarkers & Prevention, 22(9), pp. 1588–1599. Available at: https://doi.org/10.1158/1055-9965.epi-13-0454.

Isler, B. et al. (2020) ‘Achromobacter Infections and Treatment Options’, Antimicrobial Agents and Chemotherapy, 64(11). Available at: https://doi.org/10.1128/AAC.01025-20.