UNIVERSITY OF CALIFORNIA,
IRVINE

Breast Cancer Prediction from Genome Segments with Machine Learning

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Biomedical Engineering

by

Xinhan Tong

Thesis Committee:
Associate Professor James Brody, Chair
Assistant Professor Jered Haun
Professor Frithjof Kruggel

2018

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLES

ACKNOWLEDGMENTS

I'd like to express my great appreciation to the chair of my thesis committee, Dr. Brody. His meticulous attitude and devoted passion in research and scholarship inspired me a lot. His genius ideas can also help me rethink and get out of the struggle. Without his generous help and continuous guidance, it's impossible for me to finish this thesis.

I also want to thank the committee members of the thesis, Dr. Kruggel and Dr. Haun. Their knowledge in statistics and biology is a substantial backup. I really appreciated their helpful support in this thesis.

ABSTRACT OF THE THESIS

Breast Cancer Prediction from Genome Segments with Machine Learning

By

Xinhan Tong

Master of Science in Biomedical Engineering

University of California, Irvine, 2018

Associate Professor James Brody, Chair

Breast cancer is the most common diagnosed cancer for the worldwide women. Due to the multiformity of the clinical behaviors, it is difficult to predict and diagnosed only with clinical information. In order to find out a better solution to make some prediction in the early stage, the genome wide analysis is introduced. In this paper, we got the dataset from The Cancer Genome Atlas (TCGA) database to find a best predictive machine learning model. Since the copy number variations (CNVs) is highly related with the breast cancer, CNVs is used as a fundamental indicator of each genome segmentation in the study. Based on the start and the end positions, the datasets can be sorted and reorganized into five grouping sets. We tested the predictive power of the Gradient Boosting Machine, Distributed Random Forest, XGBoost and Deep Learning Neural Network. With the different genome segmentation grouping dataset and different machine learning models, we finally found the Gradient Boost Machine is the most powerful model for this problem. It can finally reach AUC of 0.756799 after 15-fold cross validation trained with “merged” grouping dataset.

INTRODUCTION

Breast cancer is always a nightmare for the female. It's the most common cancer diagnosed among U.S. woman. There are about 252,710 cases diagnosed in 2017, and it is estimated 266,120 new cases expected to be diagnosed in 2018, which means 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer (from the report of U.S. Breast Cancer Statistics: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.pdf). To make things worse, there is currently no effective way to completely cure breast cancer except for breast removal. The best option it the surgery before its metastasis.

According to the estimates of the fraction of cases of breast cancer, approximate 47% of breast cancer cases and 41% of the pathological in the total U.S. population can be ascribed to well-established risk factors. More specifically, later age at first birth and nulliparity accounts for 29.5% (95% confidence interval [CI] = 5.6%–53.3%); higher income contributed 18.9% (95% CI = −4.3% to 42.1%), and family history of breast cancer accounted for 9.1% (95% CI = 3.0%–15.2%)(Madigan, Ziegler, Benichou, Byrne, & Hoover, 1995). The high proportion indicates that the breast cancer should be attached importance to, and can be predicted with high risk factors, which may provide significant results.

However, breast cancer patients with the stage of the disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumors

according to their clinical behavior(van 't Veer et al., 2002). The prediction of the breast cancer cannot merely rely on the clinical behaviors and the risk factors.

The genetic mutation is the root of the breast cancer. The prediction with genome information can help focus the research in a narrower area as well as make quicker diagnose in clinical. There is a previous study about prediction test, but it can only test BRCA1 mutations. In reality, less than 10% breast cancer have BRCA1 mutations.

According to related genome-wide studies of patients carrying mutations in Breast Cancer, there're strong associations between the copy number variations (CNVs) and the cancer risk, besides the single-nucleotide polymorphisms (SNPs) and risk. The detection of CNVs can explained at least 40% of the predicted common variants (Walker et al., 2017).

Genome studies of the breast cancer involves a great range of the genome pieces. It's very hard to establish a traditional statistic model to fit and predict. The machine learning methods can extract the genome features and make a predictive model with these features. It's perfectly suitable in this situation. The area under the receiver operating characters (ROC) curve (AUC) can be used as a metric to measure the performance of the learners.

According to a recent study about the heritability of the breast cancer, the best predictive breast cancer tests incorporating multiple SNPs and family history have an AUC in the range 0.7 to 0.8 (Toh & Brody, 2018). That is good enough to help the further genetic research and diagnosis. In this thesis, instead studying the heritability, we will study and

predict the breast cancer in a single generation.

Materials and Methods

Data Source

The original data is collected by National Cancer Institute and National Human Genome Research Institute. The gene-sequence based genetic cancer data collection is called TCGA (National Cancer Genome Atlas) in the data portal (https://cancergenome.nih.gov). The data of the breast cancer is derived from all the female samples that has been diagnosed with breast cancer. Finally, there are 4,692 female breast cancer samples in the dataset. For each sample, there are 90 genetic segmentation tags that can be selected as features in the machine learning algorithms.

Platform

H2O Flow is a web-based notebook style interface of H2O, which is an open-source package for AI. It's like the Jupyter Notebook (http://jupyter.org) interface for iPython. The H2O provides multiple prototypes of the common machine learning models, which has been encapsulated to use. All manipulation of the data frame and the creation of the machine learning models can be implemented with several lines of codes. With H2O Flow user interface, the implementation can be accomplished with several clicks. It can save a lot of time to focus on tuning the hyper parameters and ~~the~~ algorithm improvement.

K-fold Cross Validation

Although there're thousands of samples in the dataset, according to the loss function curve,

the model is far from being well trained with only these samples. Meanwhile, when we used more proportion of the samples to train the model, the predictive power of the model will strongly increase. Due to the limitation of the medical data records, there's few ways to get more data due to the privacy of the medical data. If the sources of the data are collected from different medical instruments or different institutes, it will be hard to sort these data into a unified format. Cross validation can be used as a tricky and powerful method to "augment" the dataset and train a better learner. K-fold cross validation can reduce the variance while increasing the bias with moderate k-values (10-20). As k decreases (2-5) and the sample gets smaller, the variance will increase due to the instability go the training sets (Kohavi, 1995).

Machine Learning Models

Gradient Boosting Machine

Boosting a quiet fundamental and naive strategy in machine learning. Some weak learners (base models) are bagged together as an ensemble learner. The weight of each weak learner is adjusted in each iteration according to the performance of the learning. Gradually, the ensemble learner will get better and better performance on the training datasets and validation datasets.

Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (weak learner) to current "pseudo"-residuals by least squares at each iteration (Jerome H. Friedman, 2018). In other words, the loss function of the model is always descending in the gradient direction during the training process. Gradient boosting

can be applied to many different differentiable loss functions. Using different loss functions, most common problems such as regression, classification and even ranking can be handled.

It seems to be the most brute-force learning method, but it is really efficient. The breast cancer prediction is a classical regression problem in supervised learning. The encapsulated Gradient Boosting Machine can be easily implemented to work as a "base-line" model.

Distributed Random Forest

Classification by Random Forests has been a kind of popular ensemble learning since 2001, when Breiman first proposed this algorithm. Instead of contracting each tree with a different bootstrap dataset, the random forest changes how the classification and regression trees are constructed (Liaw Merck, Liaw, & Wiener, 2002).

The random forest algorithm requires only two parameters: the number ***t*** of decision trees to be constructed, and the number ***m*** of input features to be considered when each node of the decision tree is split.

Each single decision tree can be constructed in the following way:

1. Let ***N*** be the number of training examples. Then the number of input samples for a single decision tree is ***N***. There are ***N*** training samples randomly retrieved from the training set.
2. Let the number of input features of the training examples be ***M***, and the cut ***m*** is much smaller than ***M***. Then we split the ***m*** input features from each of the ***M*** input features

randomly at each node of each decision tree, and then select one of the ***m*** input features to split the best one. ***m*** does not change during the construction of the decision tree.
3. Each tree has been split until this time, until all the training examples for that node belong to the same category. No pruning is required.

According to the steps mentioned above, the random decision trees can be fast constructed compared to the normal decision trees, because there is no extra computation of loss function when splitting the feature. After constructed ***t*** random trees, all these decision trees can be bagged together to create a strong learner. For each new test case, the classification results of multiple decision trees are combined as a random forest classification or regression result. When the target feature is continuous, the output takes the average of ***t*** decision trees as the classification result. When the target feature is categorical, the minority follows the majority, and the category with the highest tree classification result is used as the entire random forest classification result.

The distributed random forest is actually not innovative at the algorithmic level. However, it optimizes the implementation of the random forests to accurate the calculation of the whole process. In random forests, cross-validation is not needed to evaluate the accuracy of the classification. Random forests have an out-of-bag (OOB) error estimation, which means in the construction of a single decision tree, ***N*** samples are randomly selected to be put back, so the accuracy of the classification of the decision tree can be tested using samples that have not been extracted. These samples account for approximately one third of the total sample size (Liaw Merck et al., 2002). Luckily, the OOB error estimation was proven to

be unbiased.

For the breast cancer detection problem, the probability of the cancer should be calculated with regression random forest. Some or all of the 90 gene segmentation tags can be selected as the features of the random trees. There is only the number of the decision trees to be decided when constructing the forest.

eXtreme Gradient Boosting

XGBoost (short for eXtreme Gradient Boosting) is a kind of tree boosting method, which has been widely considered as a highly efficient machine learning method. XGBoost is an improved and scalable system to process ~~the~~ more sparse data, and it can be fit into a wide range of scenarios.

The basic unit of the Boosting Tree is called Regression Tree. It assigns input to each leaf node based on the input attributes, and each leaf node will have a real score on it. More specifically, the score below the leaf indicates the probabilities of the expected result happening.

Sometimes, the regression tree is too simple and limited to increase the accuracy. Since a single regression tree cannot provide useful prediction results, a more powerful model called Ensemble Tree is introduced.

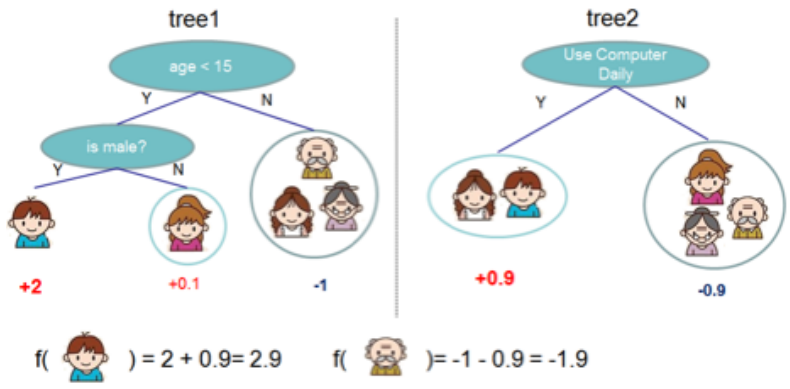

Figure 1(Chen & Guestrin, 2016).

Two regression trees are used in the figure above to predict whether someone likes video games. The results from both trees are summed up to get a new vector of the prediction results. In the process of the ensemble, more and more regression trees are added to the model with appropriate weights. Then, final results of the prediction are summarized after the whole process. It is very similar to the Random Forest. However, it differs a lot in the hyper-parameter adjustment and model construction.

The scalability of XGBoost is due to several important systems and algorithmic optimizations. These innovations include: a novel tree learning algorithm is for handling sparse data; a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning(Chen & Guestrin, 2016). Through a series of mathematical interpretation, the complexity of each regression tree can be decided by the structure. Once the structure of the regression tree is determined, the corresponding objective function can be calculated. The problem of the generating a regression tree can also be treated as the optimization of the structure of a regression tree, which has the

minimum value in objective function.

Theoretically speaking, the XGBoosting provides a better optimization than most boosting tree algorithms. Technically, it also applies weighted column subsampling to reduce over fitting and computation workload, which draws the features by the importance during growing each tree (Gao et al., 2017).

Breast cancer prediction problem is a classical supervised learning problem. This algorithm is an improved gradient boosting method implemented with the regression tree. Hence, it should have the advantages from gradient boost machine and regression tree forest, which means it can have a higher regression accuracy as well as a quite flashing training rate along the gradient direction.

Deep Learning Neural Network

Deep learning is a special machine learning algorithm with long history. Due to the limitation of the computation power, the accuracy of the deep learning was always low. It took several months to train a network that can be used in the industry in the past. Thanks to the development of GPU, the training speed is boosting nowadays.

Deep Learning also is known as the artificial neural network, which is the base of the artificial intelligence. The structure and single components are inspired by the structure of biological neuronal networks. Each single unit in the neural network is introduced as perceptron, which has the same logic characteristic as the biological neuron.

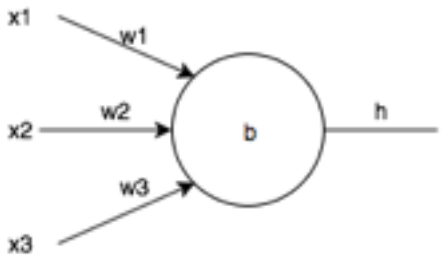

Figure 2.

The weight is like the synapse, and the bias of the function can be analogies as the threshold voltage. Then the output of the perception and the biological neuron are surprisingly similar. After the complicated connections between perceptions are established, the neural network can work as a dynamic system to simulate part of functions of the cerebral cortex. The system can be nonlinear, non-limited, qualitative and non-convex. After the hierarchy structure and the connections has been decided, data must be fed into the input layer of the neural net and result obtained from the output layer. The efficiency and the accuracy is totally determined by the structure and the hyper parameters such as learning rates, the output activation function and the number of the hidden layer.

As for the breast cancer prediction problem, the genetic series can be seen as features in a gene sequence. Neural network work well in feature detection, especially in image, audio records and natural language processing. It may be work also great in genetic feature detection. Hence, it is worth a try.

The whole neural network can be treated as any another machine learning model. That means we only need to define the input layer and output layer instead of decide the details about the connections among the neurons in the neural network. The genetic series is fed into the input layer, and we can get the prediction results on the output layer. After comparison, the neural net can be adjusted and tuned until the accuracy of the results and the predictive power of the whole model cannot be changed any more.

Copy Number Variation & Segmented Mean Value

Sometimes, a section of the genome can be duplicated and deleted, and the number of the duplication can differ among individuals. This phenomenon is called Copy Number Variation (CNVs). It is a common type of structural variation. Another genetic variation associated to disease susceptibility is, namely, single-nucleotide polymorphisms (SNPs). SNPs and CNVs captured 83.6% and 17.7% of the total detected genetic variation in gene expression, respectively, but the signals from the two types of variation had little overlap (Stranger et al., 2007). Breast cancer is highly associated with copy number variation. Segmented Mean Value is the number to measure how many copy number variations in the certain region of a chromosome. According to the database, there are three types of features in the original dataset.

The first kind of feature provides a starting position. It will calculate the Segmented Mean Value from the "start" index to the dead end of the chromosome. The 30 features that only provides start index are sorted in "BRCA_startgrouping_30.csv".

Similarly, the second one provides an end position. It measures the Segmented Mean Value from the very beginning of the chromosome to the "end" position. The 30 features that only

provides end index are sorted in “BRCA_endgrouping.csv”.

The other kind of the data provides both “start” and “end” index, which means it corresponds to a specific chromosome segments. These 30 features are sorted in “BRCA_start-endgrouping.csv”.

There is also a dataset called “BRCA_merged.csv” to merge the 60 features from the “start grouping” and “end grouping”. The original dataset is marked and stored as “BRCA_totalmerge.csv”.

In order to find out which way is the best for the learner model to predict the model, all of them will be tested to compare with each other for different machine learning algorithms.

Results

The Area Under the ROC Curve (AUC) is selected as the metric to measure the prediction power of the models. Although the AUC is usually used to evaluate the performance of a classifier, it can also be applied into the regression problem, especially cancer detection. Because a threshold will be calculated according to the list of predicted probabilities to decide whether or not a patient will be diagnosed as cancer, the regression problem will finally also be defined as a classification. However, in this process no classifier is used.

The default setting of the models can help to estimate the ability of the algorithm. From the table below, the deep learning neural network seems to be the least powerful but most time-consuming method. The other three methods seem to have the same power and spend much less time. However, neural network can have more hyper-parameters to tune, there is a chance for it to “win back”.

Feature	Gradient Boost Machine	Distributed Random Forest	eXtreme Gradient Boosting	Deep Learning Neural Network
Default Hyper-Parameters Setting	Number of trees: 50 Maximum tree depth: 5	Number of trees: 50 Maximum tree depth: 20	Number of trees: 50 Maximum tree depth:6	Hidden layer sizes: 200, 200 Epochs: 10
AUC	0.713596	0.717411	0.714054	0.645654
Running Time	00:18.497	00:16.205	00:39.467	01:03.671

Table 1.

The “total merged” dataset is used to find out the best parameters for each algorithm. The data is split into training dataset and validation dataset with the proportion of 0.75:0.25. According to the conclusion made in (Kohavi, 1995), the variance will decrease while the bias increase when the ***k*** of the k-fold cross validation is between 10-20. The ***k*** is set as 15 during the experiments.

AutoML Project

It is a packaged methodology provided by H2O platform, which can run a series of the

machine learning models automatically and list all the results. After making some appropriate settings in the AutoML Project, the project will run all the tests and help find out the best parameters for the model within limited time.

AUC	**Gradient Boost Machine**	**Distributed Random Forest**	**eXtreme Gradient Boosting**	**Deep Learning Neural Network**
Start Grouping	0.754568	0.715357	**0.729618**	0.687892
End Grouping	0.733892	0.694433	0.711557	0.683110
Start-End Grouping	0.702409	0.665750	0.683130	0.642165
Merged Grouping	**0.756799**	0.719647	0.721330	0.548369
Total Merged	0.756587	**0.733054**	0.725867	**0.703651**

Table 2.

For each machine learning algorithm, there is a most “suitable” copy number variation grouping method. Meanwhile, the “total merged” grouping have the best overall

performance. For machine learning method, the “merged” grouping have a best AUC. For the distributed random forest and the deep learning network, the “total merged” grouping have the best performance. For XGBoost, the AUC reach 0.729618 when using the “start” grouping. The Neural Network took a few hours to search for the optimization of the hyper-parameters, while it took several minutes to find out the best model for the other three methods.

In summary, the most powerful prediction model of the breast cancer is Gradient Boost Machine with “merged” grouping as the input. The hyper-parameters and the prediction results of this model is shown as below.

Hyper Parameters	**Best Value**
number_of_trees	54
number_of_internal_trees	54
model_size_in_bytes	32208
min_depth	2
max_depth	9
mean_depth	8.8704
min_leaves	3
max_leaves	58
mean_leaves	42.4259

Table 3.

When the number of the trees is 44 and the depth is 10, the Gradient Boost Machine has a greatest AUC value. The scoring history of Log Loss during training and the ROC Curve can be shown as below.

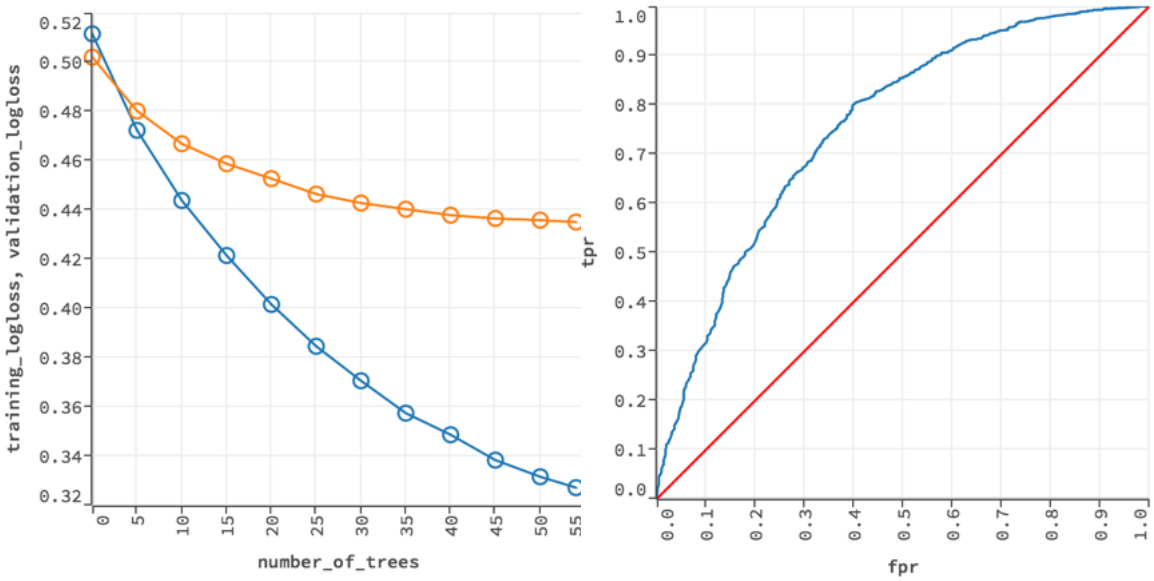

Figure 3.

With this model, the ranking of the importance of the features can also be plotted as below. The histogram indicates how important the feature is for diagnosing the breast cancer. From this chart, it can be deduced which section of the genome has the largest influence on breast cancer. The CNVs on these sections provides a much narrower area to be focused on for further clinical diagnosis and genetic research.

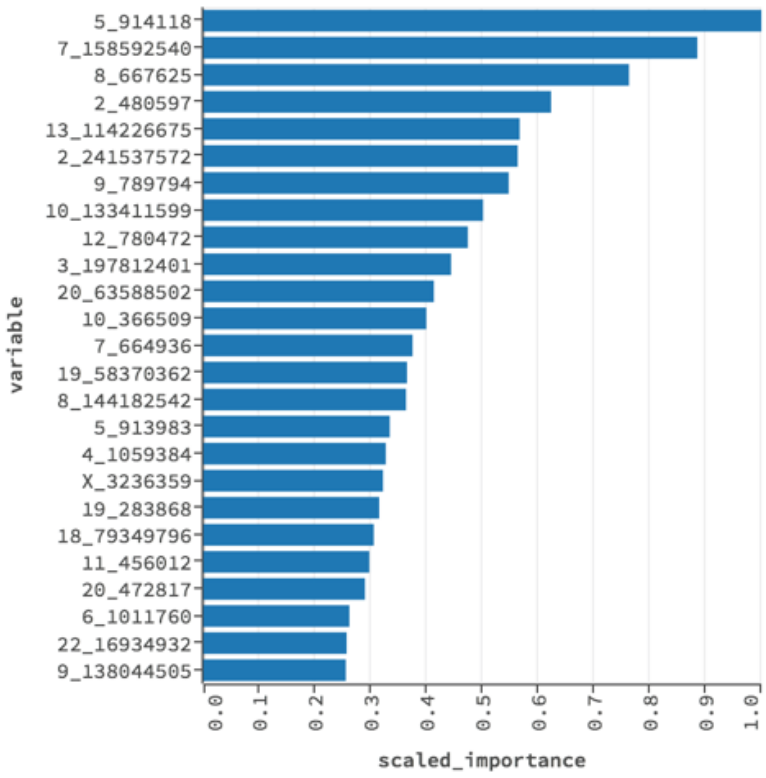

Figure 4.

There are 23 chromosome pairs in the human genome. The first 22 is indexed with 1 to 22. The 23rd is marked as XY (male) or XX (female). The name of each feature is composed by two parts, and these parts are connected with the underline. The first part is the index of the chromosome. The second part is the "start" index, "end" index or "start-end" index. According to different grouping method, it corresponds to different genetic segmentation. However, the format of the feature is fixed to make sure the corresponded genome section can be found. The easiest way to find out the exact correspond genome section is through the Genome Browser by USCS (https://genome.ucsc.edu/).

Discussion

The Deep Learning Neural Network turns out to be the worst model, which takes hours but

makes the least powerful prediction. Although it has a strong ability to detect the feature, it performs badly with sparse data. The original data is actually pre-processed to select the least sparse 30 features in “start” grouping, 30 in “end” grouping and 30 in “start-end” grouping.

In order to improve the performance of the Neural Network, the primary target is to encode the data into dense form. It may need other machine learning method to do more pre-processing on the genetic data. Then these processed dense data can be fed into the neural network to get better results. This topic is left as future work.

In this research, the best method is to use the Gradient Boost Machine with the “merged” grouping dataset. The AUC of the model is 0.756799. According to the figure of the importance of the features, the CNVs in the genome segment on 5th chromosome from index 914118 to the tail may serve as a strong cause for breast cancer.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, 785–794. https://doi.org/10.1145/2939672.2939785

Gao, X., Fan, S., Li, X., Guo, Z., Zhang, H., Peng, Y., & Diao, X. (2017). An improved XGBoost based on weighted column subsampling for object classification. In *2017 4th International Conference on Systems and Informatics (ICSAI)* (pp. 1557–1562). IEEE. https://doi.org/10.1109/ICSAI.2017.8248532

Jerome H. Friedman. (2018). Greedy Function Approximation : A Gradient Boosting Machine, *29*(5), 1189–1232. https://doi.org/10.1214/009053606000001389

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and

Model Selection. Retrieved from http://robotics.stanford.edu/~ronnyk

Liaw Merck, A., Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest, *23*. Retrieved from https://www.researchgate.net/publication/228451484

Madigan, M. P., Ziegler, R. G., Benichou, J., Byrne, C., & Hoover, R. N. (1995). Proportion of Breast Cancer Cases in the United States Explained by Well-Established Risk Factors. *JNCI Journal of the National Cancer Institute*, *87*(22), 1681–1685. https://doi.org/10.1093/jnci/87.22.1681

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, *315*(5813), 848–853. https://doi.org/10.1126/science.1136678

Toh, A. C., & Brody, J. P. (2018). Analysis of copy number variation from germline DNA can predict individual cancer risk, 1–14.

van ’t Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536. https://doi.org/10.1038/415530a

Walker, L. C., Marquart, L., Pearson, J. F., Wiggins, G. A. R., O’Mara, T. A., Parsons, M. T., … Spurdle. (2017). Evaluation of copy-number variants as modifiers of breast and ovarian cancer risk for BRCA1 pathogenic variant carriers. *European Journal of Human Genetics : EJHG*, *25*(4), 432–438. https://doi.org/10.1038/ejhg.2016.203