Identifying Bio-markers Using Support Vector Machine to Understand the Racial Disparity in Triple-Negative Breast Cancer.

Bikram Sahoo,[1,*] Zandra Pinnix,[2] Seth Sims,[1]
Alex Zelikovsky,[1,*]
[1] Department of Computer Science, Georgia State University
Atlanta, GA 30303, USA
[2]Department of Biology and Marine Biology, University of North Carolina at Wilmington
Wilmington, NC, 28403, USA
*To whom correspondence should be addressed;
E-mail: alexz@gsu.edu

Keywords: Triple-negative breast cancer, Racial disparity, SVM, RNA sequencing, Gene expression, Machine learning and Feature engineering

Abstract: With the properties of aggressive cancer progression and heterogeneous tumor biology, triple-negative breast cancer is a type of breast cancer known for its poor clinical outcome. The lack of estrogen, progesterone and human epidermal growth factor receptor in the tumors of TNBC leads to fewer treatment options in clinics. The incidence of TNBC is higher in African Americans compared to European American women with worse clinical outcomes. The significant factors responsible for the racial disparity in TNBC are socio-economic lifestyle and tumor biology.

The current study considered the open-source gene expression data of triple-negative breast cancer samples' racial information. We implemented a state-of-the-art classification Support Vector Machine (SVM) method with a recurrent feature elimination approach to the gene expression data to identify significant biomarkers deregulated in African American women and European American women. We also included spearman's rho and ward's linkage method in our feature selection workflow. Our proposed method generates 24 features/genes that can classify the AA and EA samples 98% accurately. We also performed the Kaplan-Meier analysis and log-rank test on the 24 features/genes. We only discussed the correlation between deregulated expression and cancer progression with a poor survival rate of two genes, *KLK10* and *LRRC37A2*, out of 24 genes. We believe that further improvement of our method with a higher number of RNA-seq gene expression data will more accurately provide insight into racial disparity in TNBC.

1 Introduction

Breast cancer is the most frequent and second major cause of cancer-related death in women in the US and heterogeneous cancer with diverse biological subtypes Hendrick *et al.* (2021). Breast cancer tumor growth relies on estrogen, progesterone hormones, and Her2 (growth factor) protein. The molecular classification of breast cancer is based on the expression of three biomarkers: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (Her2). Triple-negative breast cancer lacks the expression of estrogen, progesterone, and HER2 receptors Cho *et al.* (2021); Moss *et al.* (2020). It can grow in the absence of estrogen and progesterone hormone, and Her2 protein. Therefore, standard breast cancer treatment options such as hormonal therapies and targetable drugs fail to cure TNBC Prakash *et al.* (2020).

Furthermore, TNBC has highly heterogeneous tumor biology with a solid metastatic potential leading to poor clinical outcomes compared to other breast cancer subtypes Cho *et al.* (2021). The rate of TNBC diagnosis in the US is 10-20% compared to different invasive breast cancer subtypes Prakash *et al.* (2020); Dietze *et al.* (2015); Lehrberg *et al.* (2021). In addition to that epidemiological and clinical studies, the incidence rate of TNBC is twice in African Americans compared to European American women Chen and Li (2015). The current finding suggests that biological and socioeconomic factors significantly contribute to poor clinical outcomes in AA women having TNBC Newman and Kaljee (2017). Consequently, the recent breast cancer research needs to provide the underlying biological mechanism of TNBC, especially biomolecules responsible for racial disparity in TNBC Newman and Kaljee (2017); Siddharth and Sharma (2018); Sturtz *et al.* (2014). In the era of machine learning and deep learning, we need robust computational models to build on biomolecular, clinical, and epidemiological data to find novel therapeutic targets that can be used in clinics to improve the treatment option and survival of African American TNBC patients.

Cancer researchers generated an unprecedented amount of next-generation sequencing data to study the molecular mechanism behind cancer progression and metastasis. This immense amount of NGS data used for different scientific studies is freely available on the internet Lachmann *et al.* (2018); Esposito *et al.* (2019). The dependency on NGS data to address various cancer-related questions are in progress at a higher rate than in earlier studies because of its cost reduction in data generation with upgradation in sequencing technology Rondel *et al.* (2021). With the help of NGS technology, a cancer researcher can efficiently study the whole transcriptome and genome of a cancer sample to get a broader biological perspective. RNA sequencing is one of the sequencing technologies from the NGS arsenal that can help study the whole transcriptome of a cancer sample. RNA-seq provides information about a tumor's transcripts and gene expression; it can also check gene fusion, immune cell expression, somatic mutations, splicing variants, intron-retention, pathway analysis, and gene interaction networks Rondel *et al.* (2021); Lachmann *et al.* (2018). Online databases such as ArrayExpress EMBL-EBI (2019) and NCBI-GEO geo (2019) store all forms of sequencing data, including RNA-seq, with raw, processed, and clinical information for induvial studies considered NGS to address their biological questions. Also, databases like TCGA and ICGC consortiums provide free access to the massive amount of sequencing data for different cancer types Institute (2019). Computational biologists take advantage of these databases to obtain reliable sequencing data for their studies to understand cancer. This study also considered the open-source RNA-sequencing gene expression data of TNBC patients to identify potential biomarkers expressed in African American women compared to European American women.

To identify highly discriminant biomarkers in African American and European American TNBC samples gene expression data, we considered it a problem to classify AA and EA samples with a selected number of genes. We applied the popular classification algorithm, Support Vector Machine (SVM), with feature/gene selection techniques on RNA-seq gene expression data to achieve our goal Guyon *et al.* (2002); Platt *et al.* (1999); Vanitha *et al.* (2015); Brown *et al.* (2000); Das *et al.* (2020). Here, feature/gene selection is an essential step because clinical/biological science always relies on the top-ranking genes from the computational data analysis. In addition to feature selection using the SVM method, we showed the enrichment/expression of selected features/genes in different groups. Therefore, the proposed method satisfies quantitative and qualitative analysis of features/genes that can accurately classify AA and EA TNBC samples. We further used a micro-array dataset of TNBC gene expression data with survival information to validate our genes and predict pathological outcomes. We considered the micro-array dataset because, as per our knowledge, that is the only dataset that satisfies our study's requirement. The pathological role of our genes is accessed by performing Kaplan-Meier and log-rank tests on the microarray expression and survival data.

Our proposed SVM-based feature/gene selection with the recurrence feature elimination method identifies 24 genes that can accurately classify AA and EA samples. We further validated the pathological significance of those genes with Kalan-Meier survival and log-rank test on a microarray dataset having survival information. Finally, we found that 12 genes have a significant pathological role in TNBC samples, and we also cherry-picked two genes, *KLK10* and *LRRC37A2*, from the 12 genes and discussed their biological role in TNBC. The high expression of *KLK10* is known for poor prognosis in TNBC and other cancers. Our methods also show that *KLK10* has an increased expression in AA samples compared to EA samples and can be a potential biomarker in this racial disparity study Rückert *et al.* (2008); Alexopoulou *et al.* (2013); Kioulafa *et al.* (2009); Yousef *et al.* (2005); Lin *et al.* (2020). The de-regulation of *LRRC37A2* is known for cancer progression and is a critical member of various gene-signature studies to assess the cancer prognosis Wisnieski *et al.* (2021); Xu *et al.* (2022); Wu *et al.* (2018); Feng *et al.* (2020). The role of *LRRC37A2* in TNBC progression is known Stewart *et al.* (2013); we also found low expression of the *LRRC37A2* gene in AA samples compared to EA samples, and it has a significant survival outcome in our validation analysis. Considering the above results, we can conclude that our SVM-based feature selection method identified essential genes significantly contributing to poor outcomes in AA TNBC women.

The current study is organized as follows. In section 2, we detail the RNA-seq gene expression data collection and SVM-based feature selection method. We also specify the feature/gene enrichment method. Finally, we provided the details of feature validation methods. In section 3, we provided the final feature list and discussed the biological significance of two features/genes from the final feature list such as *KLK10* and *LRRC37A2*. In section 4 , we discussed our approach's merit in the TNBC racial disparity study to identify potential biomarkers and contributions of our research.

2 Data and Methods

In this section, we report our RNA-seq gene expression dataset details in Sec. 2.1. We discuss our SVM-based feature/gene selection method used to study the racial disparity in TNBC in Sec. 2.2. The feature estimation method is discussed in Sec. 2.3. The validation method for our features using Kaplan-Meier survival analysis, log-rank test, and the differential gene expression analysis using DESeq2 discussed in Sec. 2.4 and 2.5 respectively.

2.1 RNA-seq gene expression data

Open source RNA sequencing gene expression data of breast cancer patients were considered for this study. The processed RNA-seq gene expression data with log2 median-centered was downloaded from geo (2019); Institute (2019). According to hormone-receptor status, breast cancer has various subtypes and can be accessed by immunohistochemistry tests. We considered the negative immunohistochemistry status for ER, PR, and HER2 receptors to collect the triple-negative breast cancer samples. The FISH status was considered for the samples with equivocal immunohistochemistry status for HER2 receptor. The immunohistochemistry status filtering criteria generated 145 triple-negative breast cancer samples from the breast cancer dataset. This study focuses on discovering molecular-level racial disparity in EA and AA women having TNBC. Therefore, we further reduced our sample size from 145 to 128, considering the race category information. For our final analysis, we considered 128 TNBC samples, where 87 were from European Americans, and 41 were from African American women.

2.2 Construction of SVM based model for feature selection

2.2.1 Initial feature selection using SVM classifier:

This current research addresses the classification problem of EA and AA women having TNBC to understand the racial disparity using RNA-seq gene expression data. As an initial pre-processing step, we dropped the features having few unique gene expression counts across the dataset. Next, we standardize the gene expression data to Z-scores in Eq. 1.

$$Z - score = \frac{(x - \mu)}{\sigma} \tag{1}$$

Where: $x =$ the original gene expression value

$\mu =$ the mean of the gene expression values

$\sigma =$ the standard deviation of the gene expression values

After that, the support vector machine (SVM) classifier with a linear kernel fitted to the Z-score normalized data. Finally, L1 regularization was considered to generate a classifier for the groups with a sparse feature set to select the initial feature set.

2.2.2 Feature reduction:

We calculated the pairwise correlation between the features using Spearman's rho non-parametric test Eq.2.

$$\rho = 1 - \frac{6 \sum d_i^2}{n_f(n_f^2 - 1)} \tag{2}$$

Where: $\rho =$ the spearman's rho coefficient

$d_i =$ the difference between the two ranks of each features

$n_f =$ number of features

Then we applied the ward's linkage method for hierarchical cluster analysis to find the highly correlated features by minimizing the increase in ESS (error sum squares).

$$ESS(X_f) = \sum_{i=1}^{Nx_f} \left| Xf_i - \frac{1}{Nx_f} \sum_{j=1}^{Nx_f} Xf_j \right|^2 \tag{3}$$

Then, we randomly selected one feature from each cluster to represent each. For Further analysis, the clusters were selected that have a linkage-distance score of 0.65. The value 0.65 was chosen by performing a grid search in five-fold cross-validation. The grid search was performed, including the value between 0.25 and 1.75 with a step size of 0.1. A linear SVM with standard L2 loss was fit to each fold, and the threshold was set to the most significant value, which maintained complete separation in all folds. Figure 1 plots the accuracy of each fold during the grid search.

2.2.3 Further feature reduction using recursive feature:

The recursive feature elimination with cross-validation was performed to reduce the features further Guyon *et al.* (2002). This method has been used successfully to find classification features for other cancers. We continued to use a linear kernel as the core classifier for the feature selection.

2.2.4 Permutation test on final features:

A permutation test was performed on the final features. The EA/AA labels are randomly permuted and fed to the feature selection pipeline. The number of times the magnitude of a feature coefficient from the permuted data exceeded the features selected on the original data was counted. This was repeated with 5,000 permutations.

2.2.5 Computation of Influence of each feature:

The influence of each feature was then assessed with Shapley Additive Explanation (SHAP) values Lundberg and Lee (2017). Bootstrapping was used to assess the distribution of variation of SHAP values for each feature. Each iteration was performed on a randomly selected test-train split of the data. A linear SVM was fit with the Platt scaling estimation of class probabilities Platt *et al.* (1999). The SHAP KernelExplainer method was used to assess the influence of each item on the test set. The mean of the SHAP value magnitudes was collected and plotted in Figure 2.

2.2.6 Feature evaluation:

Finally, each feature was evaluated using Student's T-test for the independence of the mean with false discovery rate (FDR) control using the Benjamini-Hochberg procedure. Features with FDR rates > 0.05 were suppressed. All feature selection was performed in a Jupyter notebook Kluyver *et al.* (2016), using SciKit-learn Pedregosa *et al.* (2011), SciPy Virtanen *et al.* (2020), Pandas, Seaborn Waskom (2021), and Matplotlib Hunter (2007).

2.3 Estimation of feature/Genes

The dataset consists of gene expression values for 128 samples. Here we used the term "features" to represent "genes," and the analysis is conducted in m-dimensional feature space. While formulating the problem, we restricted the analysis to a two-class classification problem. The classes represent with the symbols (+) and (-). A sample set considered $\{a_1, a_2, ..., a_k, ..., a_n\}$ having the true class labels $\{b_1, b_2, ..., b_k, ..., b_n\}$. Given that b_k belongs to (-1,+1). The training algorithm will build a scalar separation function D(a). Further, the separation function will use to classify the new samples.

$$D(\mathbf{a}) > 0 \Rightarrow \mathbf{a} \in \text{class } (+)$$

$$D(\mathbf{a}) < 0 \Rightarrow \mathbf{a} \in \text{class } (\text{-})$$

$$D(\mathbf{a}) = 0, \text{ decision boundary}$$

The separation function is:

$$D(\mathbf{a}) = \mathbf{w} \cdot \mathbf{a} + \mathbf{b}$$

Where: $w =$ the weight vector, $b =$ the bias

The sign of the separation function D(a) can be interpreted as a difference in gene expression. The positive sign represents the direction of the positive class, and the negative sign represents the direction of the negative class. To understand the pathological role of genes, we considered the signs (+,-) as up and down-regulation of gene expression and performed the survival analysis.

2.4 Feature selection using the kaplan-Meier and log-rank tests

The reduced feature/gene set generated by our SVM-based approach was further considered for Kaplan-Meir and log-rank tests. The Kaplan-Meier and log-rank tests will help to assess the pathogenicity of the selected features/genes in the poor survival outcome of TNBC in African American women. We considered a micro-array dataset of breast cancer women with survival data to perform the analysis. We collected the triple-negative breast cancer samples from the breast cancer dataset, considering the samples having negative ER, PR, and Her2 receptor status. We considered a cut-off value for each feature/gene and performed the Kaplan-Meier and log-rank tests. The feature with a p-value less than 0.09 was considered for further discussion and included in building a signature biomarker list to understand the racial disparity in TNBC.

2.5 Differential gene expression analysis using DESeq2

We performed the differential gene expression analysis to check the expression status and significance of the gene set generated by our SVM-based feature selection method. The idea behind the differential gene expression analysis is to validate the results generated from our SVM-based approach. We considered the DESeq2 Love *et al.* (2014) tool to perform differential expression analysis between EA and AA women having TNBC. Furthermore, we selected the genes having log2foldchange greater and less than +1.5 and -1.5 with a p-adjusted value less than 0.05 to get the up and down-regulated genes. Finally, we mark the status of the genes generated from our SVM-based model and report the results.

3 Results and Discussion

In this section, we discussed our results; in Sec. 3.1, we discussed the step-wise approach for selecting 24 genes by our SVM method. In Sec. 3.2, we discussed our validation method's results, including the Shapley Additive Explanation scores (SHAP). We reported the survival analysis results for the selected genes in Sec. 3.4. The gene expression status and statistics are reported in Sec. 3.3. Finally, we discussed the association of gene expression and survival outcome in 3.5.

3.1 Selection of final features.

We considered the RNA-seq gene expression data of 128 women's tumor samples having triple-negative breast cancer to study the bio-molecular mechanism behind the racial disparity between European American and African American women. Our dataset has gene expression counts for 87 European American and 41 African American women samples. Our study's sample/data point has 20,530 gene expression read-count values. When we performed the primary analysis to separate the classes into EA and AA considering each gene/feature, we found that no single gene/feature can divide the data into two categories. After that, we selected 95 genes/features using an L1 regularized classifier that can separate the gene expression data into two classes. We performed Spearman's rho correlation analysis to compute the pairwise correlation among the features. Then, we applied Ward's method with a specific threshold to generate 45 clusters using the pairwise correlation values. We selected one feature from each cluster and continued the recursive feature elimination procedure with cross-validation. Finally, we selected 24 genes/features with a final mean accuracy score of 98% Figure 3.

3.2 Validation of features

The final 24 features showed approximately a perfect separation of two classes, EA and AA. The five-fold cross-validation of the final 24 genes/features is represented in Table 1. We observed that the 1st, 4th, and 5th cross-validation have a similar value of balanced accuracy, ROC AUC, and weighted F1 score. The effect of the features is evaluated with Shapley Additive Explanation (SHAP) values. The bootstrap SHAP distribution for 24 features is illustrated in Figure 2. The SHAP values show that *CROCCP2* and *POLR1A* genes strongly influence the classification model. The permutation test results only generated four out of 24 features at a 5% level. The four features generated by the permutation test are *POLR1A, CROCCP2, KBF2*, and *SULTIE1*.

3.3 Gene expression status

The gene expression status and up and down-regulation of our final 24 genes are reported in Figure 4 and Table 2.The status of each gene in the proposed method is computed in the feature estimation step. The gene has a (+) sign is considered an up-regulated gene, and a (-) sign is considered a down-regulated gene. We reported and discussed the expression status of *KLK10* and *LRRC37A2* from our final 24 features de-regulated in EA and AA samples significantly. We validated our method by comparing the results with a popular differential expression analysis tool, DESeq2. The significant gene generated by DESeq2 is reported in 3. The table has information about *KLK10* expression and p-adjusted values.

3.4 Survival analysis

We performed the Kaplan-Meier survival analysis and log-rank test to understand the pathogenicity of the 24 genes generated by our SVM-based model. The survival analysis on gene expression status (low vs. high) helps us understand each gene's role in TNBC. We found that 12 out of 24 genes have significant p-value (<0.05) from the Kaplan-Meier analysis and log-rank test. The Hazard ratio [HR] and log-rank test the p-value for the 12 genes are reported in 4. Furthermore, we considered the gene expression (low or high) in European Americans (EA) and African Americans (AA) to correlate the pathogenic role and racial disparity. In this study, we discussed the pathogenic role of *KLK10* and *LRRC37A2* with expression in EA and AA samples.

3.5 Association of gene expression with survival outcome

In this section, we discussed the role of *KLK10* and *LRRC37A2* in cancer progression, especially in triple-negative breast cancer. We try to speculate the correlation of poor survival outcome with deregulated gene expression in African American women compared to European American women having TNBC.

3.5.1 *KLK10*

Our SVM-based feature selection and expression analysis show that *KLK10* has high expression in African Americans compared to European American TNBC women. *KLK10* is one of the non-classical family members of the kallikrein-related peptidases (KLKs). The KLKs are well-known for playing a significant role in cancer progression. The KLK10's expression is found in normal mammary epithelial cells. Therefore, KLK10 is also referred to as the Normal Epithelial cell-specific 1 (NSE1) gene. KLK10 is reported as a potential biomarker for cancer Borgoño and Diamandis (2004); Kioulafa *et al.* (2009). *KLK10* is a significant biomarker with CA125 for understanding the diagnosis and prognosis of ovarian cancer Dong *et al.* (2013); El Sherbini *et al.* (2018); White *et al.* (2010); Geng *et al.* (2017); Batra *et al.* (2010). High expression of *KLK10* mRNA plays an important role in colorectal and pancreatic cancer Rückert *et al.* (2008);

Alexopoulou *et al.* (2013). Low expression of *KLK10* is reported in testicular and breast cancer Kioulafa *et al.* (2009); Yousef *et al.* (2005); Lin *et al.* (2020). In TNBC, the study [2] reported that *KLK10*'s high expression promoted tumor growth and cancer progression. In the log-rank test, we found that increased expression of *KLK10* shows poor survival in triple-negative breast cancer (Hazard ratio [HR] = 2.41, $p < 0.08$) Figure 5. The KM- curve drags attention to the up-regulation of *KLK10* in African Americans compared to European American women may be the reason for the poor survival outcome and can be considered a biomarker for this racial disparity study. However, we believe more in-depth analysis will reveal more about it.

3.5.2 *LRRC37A2*

In this study, the proposed gene selection method shows that *LRRC37A2* has low expression in African Americans compared to European American women. The gene *LRRC37A2* belongs to the LRR37 family of genes. The name *LRRC37A2* comes from a leucine-rich repeat-containing 37-member A2 Giannuzzi *et al.* (2012); Wisnieski *et al.* (2021). *LRRC37A2* regulates the protein and ligand interaction. It is associated with non-cancer diseases, Parkinson's disease, and epilepsy Yao *et al.* (2021). High expression of *LRRC37A2* shows poor survival and is considered a member of the gene signature for predicting Lymphoma Xu *et al.* (2022). *LRRC37A2* is associated with ovarian and gastric cancer progression Wisnieski *et al.* (2021). It is also reported in several breast cancer-related studies Feng *et al.* (2020); Wu *et al.* (2018). One of the studies specifically related to finding de-regulated genes in AA and EA women in TNBC reported *LRRC37A2* Stewart *et al.* (2013). In our kM and log-rank test, low expression of *LRRC37A2* shows poor survival in triple-negative breast cancer samples (Hazard ratio [HR] = 0.36, p-value < 0.05). From KM-curve Figure 6, we can speculate that the down-regulation of *LRRC37A2* in African Americans compared to European American women may be the reason for the poor survival outcome and can be considered a biomarker for this racial disparity study. However, we believe more in-depth analysis will provide more information about the role of *LRRC37A2*.

4 Conclusion

Triple-negative breast cancer is a highly heterogeneous subtype among the other subtypes of breast cancer with aggressive tumor progression and poor survival outcomes. As a consequence of this, TNBC has limited treatment options in clinics. The current vital challenge for the breast cancer community is to discover targetable bio-marker playing a significant role in TNBC progression that can be useful in clinics. Epidemiological and clinical studies reported that TNBC has different consequences in terms of race. African descent women have a higher mortality rate compared to European descent women. Unfortunately, this adds a racial disparity challenge to the current puzzle of TNBC research. The clinical and epidemiological studies evidence reveals that the environmental and genetic elements are responsible for the aggressive progression of tumors and poor survival in African American women. Breast cancer research is going on several fronts to discover some biomarkers accountable for this racial disparity in TNBC to give better treatment options to African American women. Integrative genomics and proteomics studies can be a vital option for this study to discover a comprehensive underlying bio-molecular mechanism.

With the help of next-generation sequencing technology, the triple-negative breast cancer community generated an unprecedented amount of whole transcriptome and genome data for various cell lines and patient samples to study cancer at the nucleic acid level. In the NGS data processing journey, state-of-the-art computational algorithms and machine learning models aided the research community in understanding the molecular heterogeneity of TNBC. As a result of that, TNBC is divided into various sub-types considering the tumor parameters such as gene, protein, immune cell, and pathway expression. However, the current approach to NGS data processing has limitations and vagueness in selecting the methods to process the data.

This research aims to develop a method to identify critical biomarkers responsible for the racial disparity in TNBC among EA and AA women and, unfortunately, a poor outcome in AA women. We proposed an SVM-based method with recursive feature elimination to select the biomarkers for this study. We also computed the pairwise correlation using Spearman's rho method and clustered the features using Ward's. Finally, we found 24 features/genes that can classify African American and European American TNBC samples with 98% accuracy. The pathological role of the 24 features/gene was further validated using a micro-array dataset having survival information. The Kaplan-Meier and log-rank test shows that 12 out of our final 24 genes have a significant pathological role in TNBC.

In this study, we cherry-picked two genes, *KLK10* and *LRRC37A2*, and reported the consequence of their deregulated expression leading to poor survival in African American women. Besides our survival analysis results, cancer research studies report that high expression of *KLK10* is known for metastasis and cancer progression, and low expression of *LRRC37A2* aids in cancer procession. *KLK10* and *LRRC37A2* were also reported as essential bio-marker in triple-negative breast cancer research. Our proposed SVM-based gene selection model not only captured these two genes but also clearly showed the expression status that is high expression of *KLK10* and low expression of *LRRC37A2* in African Americans compared to European American TNBC samples. In traditional gene expression analysis, getting significant genes directly from an analysis is complex and requires prior biological and statistical knowledge. However, we also agree that further improvement of our model and wet lab experiments requires getting a significant bio-marker list that will accurately use as a target in clinics to provide better treatment options to African American TNBC women.

Availability of data and materials Data sets and Jupyter notebook hosted at Github: https://github.com/xzy3/SVM-TNBC-racial-disparity

Contributions BS prepared the data set, validated and analysed ML results, and wrote the manuscript. ZP analysed the results and wrote the manuscript. SS performed data ML analysis, and wrote the manuscript. AZ supervised the project.

Acknowledgements

BS was partially supported by a GSU Molecular Basis of Disease Fellowship. AZ was partially supported from NSF Grant CCF-2212508, and NIH grant 1R21CA241044-01A1.

References

Alexopoulou, D.K., Papadopoulos, I.N., and Scorilas, A. 2013. Clinical significance of kallikrein-related peptidase (klk10) mrna expression in colorectal cancer. *Clinical Biochemistry* 46, 1453–1461.

Batra, J., Tan, O.L., O'Mara, T., *et al.* 2010. Kallikrein-related peptidase 10 (klk10) expression and single nucleotide polymorphisms in ovarian cancer survival. *International Journal of Gynecological Cancer* 20, 529–536.

Borgoño, C.A. and Diamandis, E.P. 2004. The emerging roles of human tissue kallikreins in cancer. *Nature Reviews Cancer* 4, 876–890.

Brown, M.P.S., Grundy, W.N., Lin, D., *et al.* 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* 97, 262–267.

Chen, L. and Li, C.I. 2015. Racial Disparities in Breast Cancer Diagnosis and Treatment by Hormone Receptor and HER2 Status. *Cancer Epidemiology, Biomarkers & Prevention* 24, 11, 1666–1672.

Cho, B., Han, Y., Lian, M., *et al.* 2021. Evaluation of Racial/Ethnic Differences in Treatment and Mortality Among Women With Triple-Negative Breast Cancer. *JAMA Oncology* 7, 7, 1016–1023.

Das, P., Roychowdhury, A., Das, S., *et al.* 2020. sigfeature: Novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. *Frontiers in Genetics* 11.

Dietze, E.C., Sistrunk, C., Miranda-Carboni, G., *et al.* 2015. Triple-negative breast cancer in african-american women: Disparities versus biology. *Nature Reviews Cancer* 15, 4, 248–254.

Dong, Y., Loessner, D., Irving-Rodgers, H., *et al.* 2013. Metastasis of ovarian cancer is mediated by kallikrein related peptidases. *Clinical and Experimental Metastasis* 31, 135–147.

El Sherbini, M.A., Mansour, A.A., Sallam, M.M., *et al.* 2018. Klk10 exon 3 unmethylated pcr product concentration: a new potential early diagnostic marker in ovarian cancer? - a pilot study. *Journal of Ovarian Research* 11.

EMBL-EBI. 2019. Arrayexpress ¡ embl-ebi.

Esposito, D., Weile, J., Shendure, J., *et al.* 2019. Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology* 20.

Feng, H., Gusev, A., Pasaniuc, B., *et al.* 2020. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genetic Epidemiology* 44, 442–468.

Geng, X., Liu, Y., Diersch, S., *et al.* 2017. Clinical relevance of kallikrein-related peptidase 9, 10, 11, and 15 mrna expression in advanced high-grade serous ovarian cancer. *PLOS ONE* 12, e0186847.

geo. 2019. Home - geo - ncbi.

Giannuzzi, G., Siswara, P., Malig, M., *et al.* 2012. Evolutionary dynamism of the primate ¡i¿lrrc37¡/i¿ gene family. *Genome Research* 23, 46–59.

Guyon, I., Weston, J., Barnhill, S., *et al.* 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1, 389–422.

Hendrick, R.E., Helvie, M.A., and Monticciolo, D.L. 2021. Breast cancer mortality rates have stopped declining in u.s. women younger than 40 years. *Radiology* 299, 143–149.

Hunter, J.D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 9, 3, 90–95.

Institute, N.C. 2019. The cancer genome atlas program.

Kioulafa, M., Kaklamanis, L., Stathopoulos, E., *et al.* 2009. Kallikrein 10 (klk10) methylation as a novel prognostic biomarker in early breast cancer. *Annals of Oncology* 20, 1020–1025.

Kluyver, T., Ragan-Kelley, B., Pérez, F., *et al.* 2016. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas.* IOS Press. pages 87 – 90.

Lachmann, A., Torre, D., Keenan, A.B., *et al.* 2018. Massive mining of publicly available rna-seq data from human and mouse. *Nature Communications* 9.

Lehrberg, A., Davis, M.B., Baidoun, F., *et al.* 2021. Outcome of african-american compared to white-american patients with early-stage breast cancer, stratified by phenotype. *The Breast Journal* 27, 7, 573–580.

Lin, C.L., Ying, T.H., Yang, S.F., *et al.* 2020. Transcriptional suppression of mir-7 by mta2 induces sp1-mediated klk10 expression and metastasis of cervical cancer. *Molecular Therapy - Nucleic Acids* 20, 699–710.

Love, M.I., Huber, W., and Anders, S. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology* 15.

Lundberg, S.M. and Lee, S.I. 2017. A unified approach to interpreting model predictions. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30.* Curran Associates, Inc.. pages 4765–4774.

Moss, J.L., Tatalovich, Z., Zhu, L., *et al.* 2020. Triple-negative breast cancer incidence in the united states: ecological correlations with area-level sociodemographics, healthcare, and health behaviors. *Breast Cancer* 28, 82–91.

Newman, L.A. and Kaljee, L.M. 2017. Health Disparities and Triple-Negative Breast Cancer in African American Women: A Review. *JAMA Surgery* 152, 5, 485–493.

Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12, Oct, 2825–2830.

Platt, J. *et al.* 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3, 61–74.

Prakash, O., Hossain, F., Danos, D., *et al.* 2020. Racial disparities in triple negative breast cancer: A review of the role of biologic and non-biologic factors. *Frontiers in Public Health* 8.

Rondel, F.M., Hosseini, R., Sahoo, B., *et al.* 2021. Pipeline for analyzing activity of metabolic pathways in planktonic communities using metatranscriptomic data. *Journal of Computational Biology* 28, 8, 842–855.

Rückert, F., Hennig, M., Petraki, C.D., *et al.* 2008. Co-expression of klk6 and klk10 as prognostic factors for survival in pancreatic ductal adenocarcinoma. *British Journal of Cancer* 99, 1484–1492.

Siddharth, S. and Sharma, D. 2018. Racial disparity and triple-negative breast cancer in african-american women: A multifaceted affair between obesity, biology, and socioeconomic determinants. *Cancers* 10, 12.

Stewart, P.A., Luks, J., Roycik, M.D., *et al.* 2013. Differentially expressed transcripts and dysregulated signaling pathways and networks in african american breast cancer. *PLoS ONE* 8, e82460.

Sturtz, L.A., Melley, J., Mamula, K., *et al.* 2014. Outcome disparities in african american women with triple negative breast cancer: a comparison of epidemiological and molecular factors between african american and caucasian women with triple negative breast cancer. *BMC Cancer* 14.

Vanitha, C.D.A., Devaraj, D., and Venkatesulu, M. 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. *Procedia Computer Science* 47, 13–21.

Virtanen, P., Gommers, R., Oliphant, T.E., *et al.* 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, 261–272.

Waskom, M.L. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60, 3021.

White, N.M.A., Chow, T.F.F., Mejia-Guerrero, S., *et al.* 2010. Three dysregulated mirnas control kallikrein 10 expression and cell proliferation in ovarian cancer. *British Journal of Cancer* 102, 1244–1253.

Wisnieski, F., Geraldis, J.C., Santos, L.C., *et al.* 2021. Differential regulation of ¡i¿lrrc37a2¡/i¿ in gastric cancer by dna methylation. *Epigenetics* 17, 110–116.

Wu, L., Shi, W., Long, J., *et al.* 2018. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature Genetics* 50, 968–978.

Xu, H., Li, Y., Jiang, Y., *et al.* 2022. A novel defined super-enhancer associated gene signature to predict prognosis in patients with diffuse large b-cell lymphoma. *Frontiers in Genetics* 13.

Yao, S., Zhang, X., Zou, S.C., *et al.* 2021. A transcriptome-wide association study identifies susceptibility genes for parkinson's disease. *npj Parkinson's Disease* 7.

Yousef, G.M., Obiezu, C.V., Luo, L., *et al.* 2005. Human tissue kallikreins: From gene structure to function and clinical applications. *Advances in Clinical Chemistry* , 11–79.

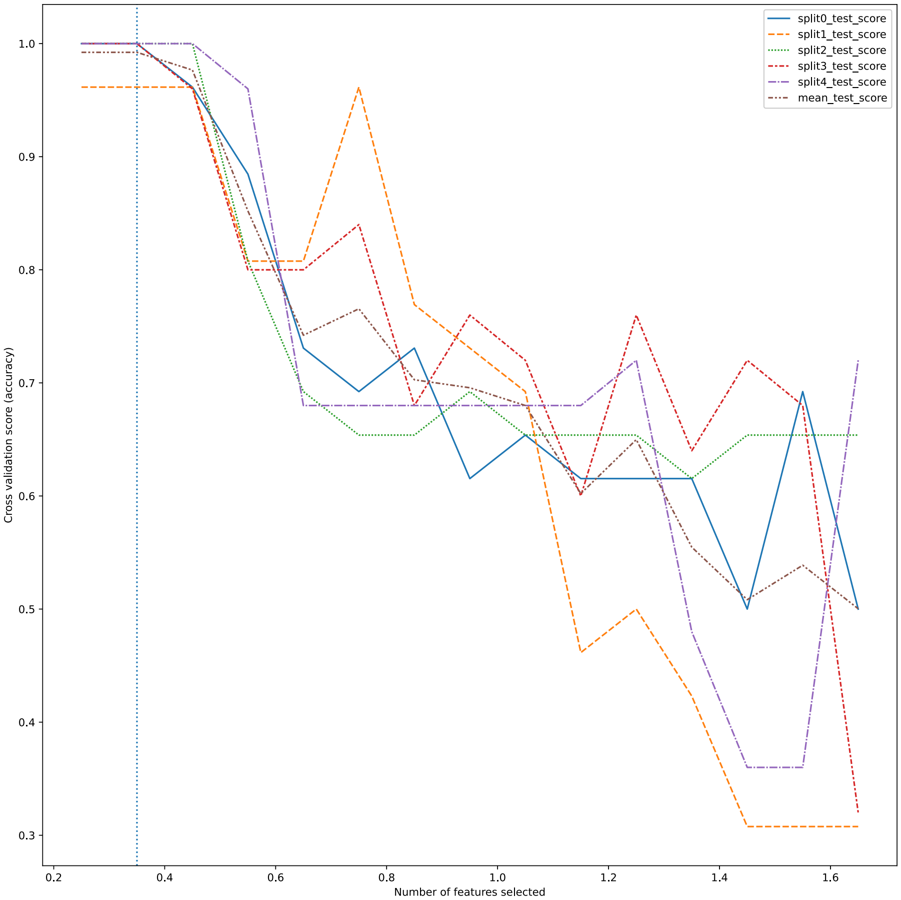

Figure 1: The figure shows the accuracy of SVM classification during grid search for collinearity threshold. The search bound values between 0.25 and 1.75 with an increasing step size of 0.1.

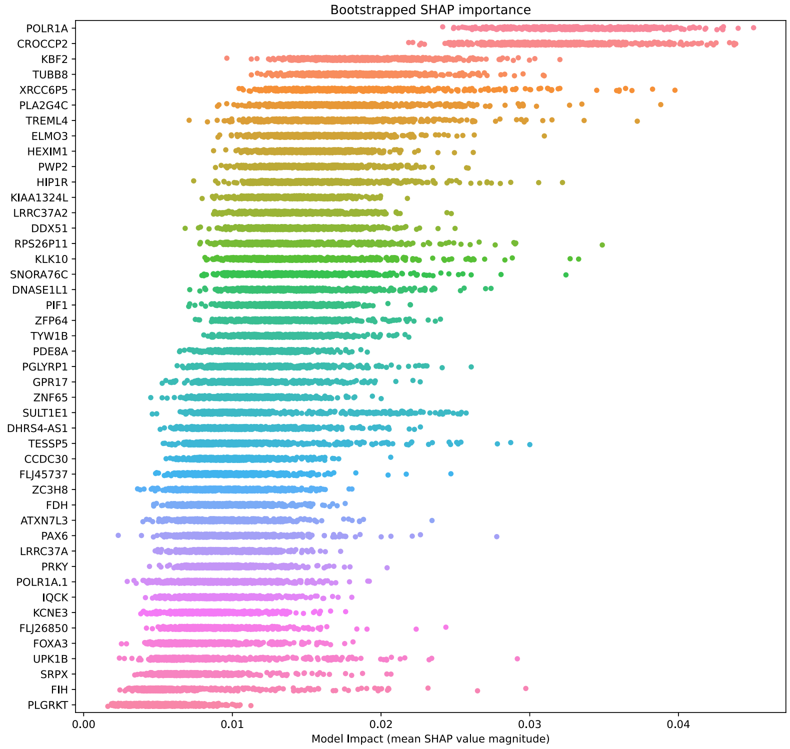

Figure 2: The above figure represents the bootstrapped distribution of the Shapley Additive Explanation (SHAP) feature importance for each feature/gene. The median value of each bootstrap orders them.

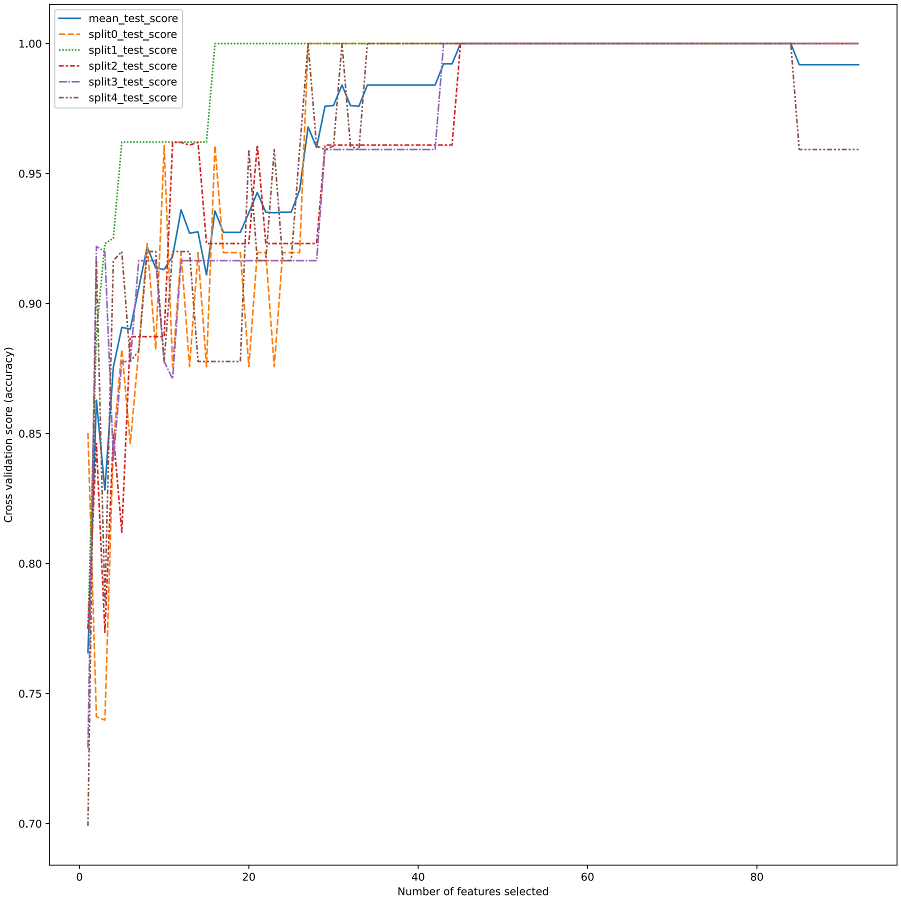

Figure 3: The figure shows the 98% mean accuracy of SVM classification with 24 features during recursive feature selection.

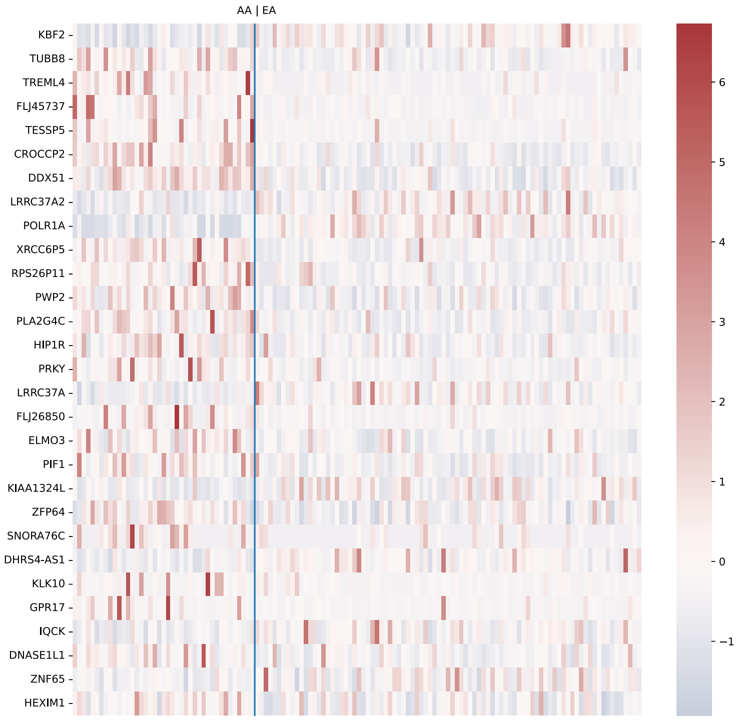

Figure 4: The above heatmap represents the expression of 24 features/genes selected by our SVM-based method. The blue line in the heatmap separates African American (AA) from European American (EA) women.

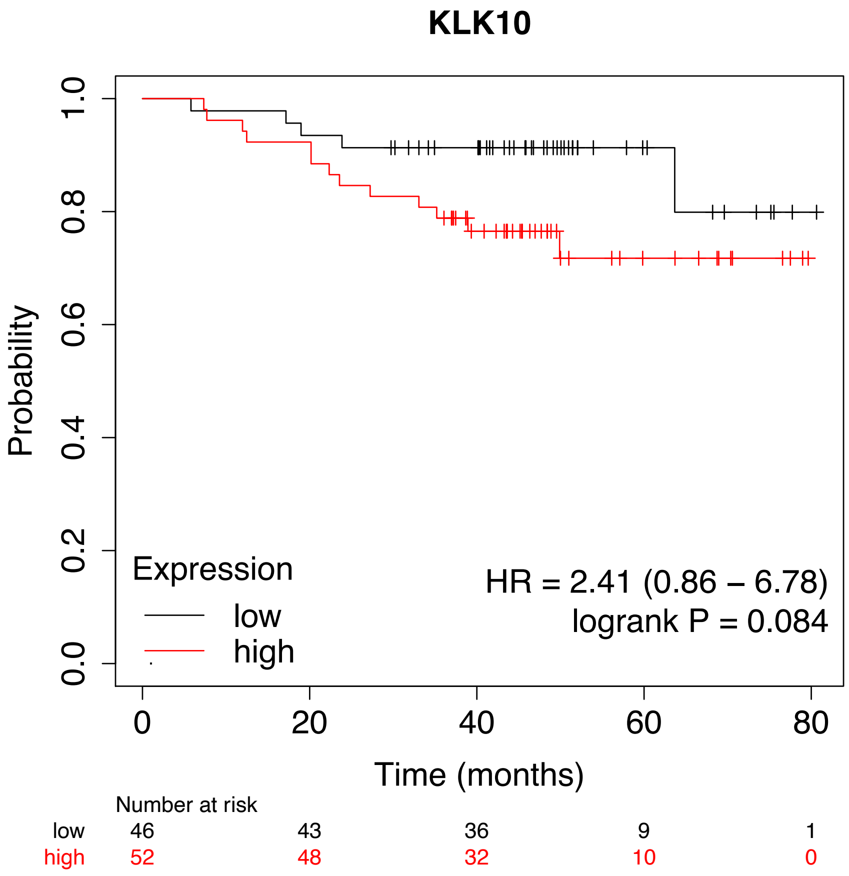

Figure 5: The figure shows the high expression of *KLK10* shows poor overall survival in the TNBC patients with p-value < 0.09 and Hazard ratio [HR] $=$ 2.41 (0.86-6.78). In the above plot, the red line represents the TNBC samples with high *KLK10* expression, and the black line represents the TNBC samples with low *KLK10* expression.

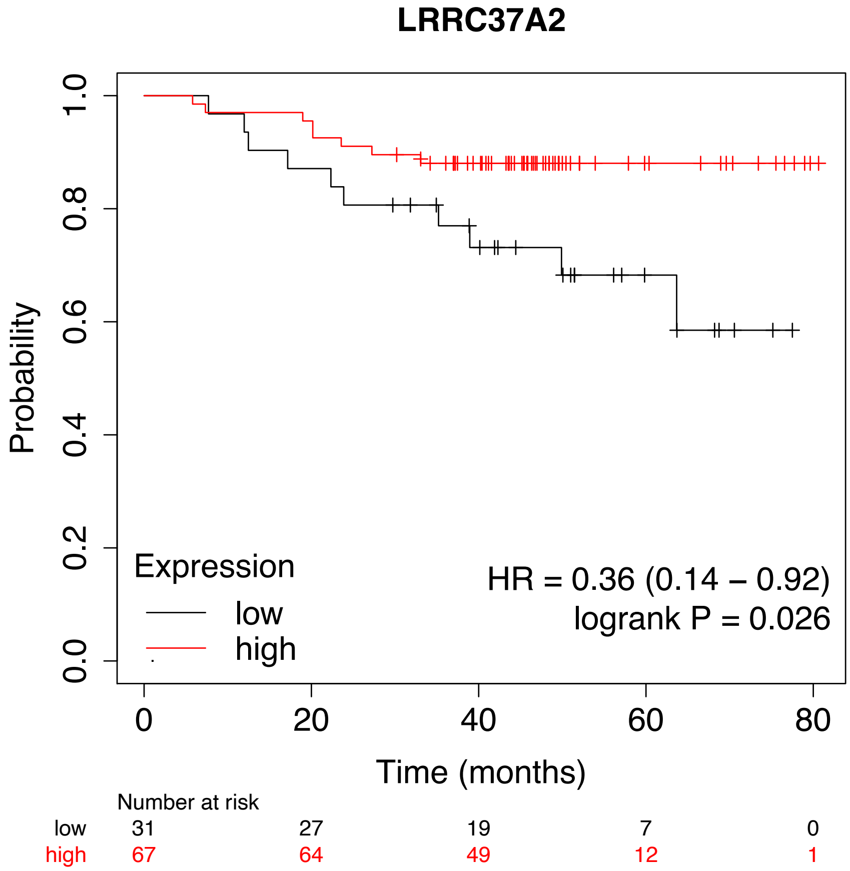

Figure 6: The figure shows the high expression of *LRRC37A2* shows poor overall survival in the TNBC patients with p-value < 0.09 and Hazard ratio [HR] $= 0.36$ (0.14-0.92). In the above plot, the red line represents the TNBC samples with high *LRRC37A2* expression, and the black line represents the TNBC samples with low *LRRC37A2* expression.

Fold	1st	2nd	3rd	4th	5th
Balanced Accuracy	1.0	0.94	0.96	1.0	1.0
ROC AUC	1.0	1.0	1.0	1.0	1.0
Weighted F1	1.0	0.96	0.96	1.0	1.0

Table 1: The above table presents the cross-validation of the final SVM classifier.

	EA	**AA**	**EA means**	**AA means**	**FDR**	**reject**
KBF2	H	L	746.9	525.83	<0.001	TRUE
TUBB8	L	H	11.01	23.6	<0.001	TRUE
TREML4	L	H	0.53	2.76	<0.001	TRUE
FLJ45737	L	H	13.35	51.86	<0.001	TRUE
TESSP5	L	H	1.99	10.31	<0.001	TRUE
CROCCP2	L	H	234.93	520.63	<0.001	TRUE
DDX51	L	H	411.13	721.8	<0.001	TRUE
LRRC37A2	H	L	125.21	61.89	<0.001	TRUE
POLR1A	H	L	407.11	161.32	<0.001	TRUE
XRCC6P5	L	H	0.74	1.98	<0.001	TRUE
RPS26P11	L	H	0.86	2.06	<0.001	TRUE
PWP2	L	H	1062.2	1619.12	<0.001	TRUE
PLA2G4C	L	H	132.7	279.76	<0.001	TRUE
HIP1R	L	H	932.62	1620.51	<0.001	TRUE
PRKY	L	H	2.26	5.93	0.001	TRUE
LRRC37A	H	L	320.96	201.34	0.001	TRUE
FLJ26850	L	H	0.95	3.64	0.001	TRUE
ELMO3	L	H	381.44	679.34	0.001	TRUE
PIF1	L	H	105.26	190.83	0.001	TRUE
KIAA1324L	H	L	242.47	112.44	0.002	TRUE
ZFP64	L	H	485.34	607.47	0.002	TRUE
SNORA76C	L	H	0.13	0.45	0.003	TRUE
DHRS4-AS1	H	L	649.67	388.31	0.003	TRUE
KLK10	L	H	735.21	2746.76	0.005	TRUE
GPR17	L	H	4.95	18.19	0.006	TRUE
IQCK	H	L	408.64	284.22	0.025	TRUE
DNASE1L1	L	H	472.65	661.69	0.028	TRUE
ZNF65	H	L	1014.11	782.45	0.029	TRUE
HEXIM1	L	H	782.62	987.46	0.029	TRUE
FDH	H	L	2414.35	1998.87	0.063	FALSE
KCNE3	H	L	154.85	112.82	0.063	FALSE
CCDC30	H	L	55.81	36.68	0.065	FALSE
PGLYRP1	L	H	0.17	0.56	0.112	FALSE
TYW1B	L	H	93.65	126.9	0.125	FALSE
PDE8A	H	L	561.4	455.48	0.148	FALSE
PAX6	L	H	133.19	298.26	0.148	FALSE
ATXN7L3	L	H	1471.25	1654.79	0.152	FALSE
FIH	L	H	0.35	1.8	0.202	FALSE
FOXA3	L	H	8.36	19.11	0.22	FALSE
SULT1E1	L	H	9.34	39.47	0.22	FALSE
UPK1B	L	H	11.14	39.22	0.297	FALSE
PLGRKT	H	L	433.78	363.53	0.362	FALSE
ZC3H8	L	H	182.43	208.53	0.362	FALSE
POLR1A.1	L	H	1079.08	1224.81	0.362	FALSE

Table 2: The above table represents the gene expression status in EA and AA women. Student's T-test investigated the expression difference for each gene for Independence of means with FDR correction by the Benjamini-Hochberg procedure.

Gene Name	**log2FoldChange**	**p-value**	**p-adj**
LOC387860	3.853183	6.02E-09	3.09E-06
CLD	3.563536	0.000612	0.010813
SCGB1B2P	3.292141	4.01E-17	3.71E-13
SP7	3.272564	1.13E-07	2.95E-05
RETN	2.994998	2.72E-09	1.93E-06
KLK14	2.953535	1.61E-09	1.42E-06
KRT34	2.699778	0.000285	0.006826
MPZ	2.487386	1.03E-09	9.53E-07
NFJ	2.465433	5.53E-08	1.71E-05
PRB2	2.419292	0.003462	0.032299
TESSP5	2.40138	8.15E-10	7.93E-07
MGC34772	2.397453	0.004814	0.039604
LRRC14B	2.359004	2.01E-06	0.000242
R3HDML	2.340992	0.00097	0.014333
TREML4	2.332817	3.91E-09	2.52E-06
HPR	2.295351	0.002648	0.027265
PGLYRP3	2.188858	0.001763	0.021088
HS3ST6	2.126105	0.000336	0.007547
AZU1	2.11035	0.000115	0.003854
HAPLN1	2.105996	9.60E-07	0.000146
LOC93086	2.103133	0.000197	0.005493
FLJ26850	2.041965	1.60E-07	3.90E-05
UPK3B	2.038656	9.19E-08	2.58E-05
MYEOV	2.027528	2.93E-05	0.001569
JSRP1	2.026862	4.26E-06	0.00042
LOC100131650	2.017893	3.35E-06	0.000353
SYCE1	2.012454	5.85E-05	0.002499
TCTE1	1.991721	7.62E-07	0.000123
CMD1D	1.967604	2.81E-06	0.000312
FLJ45737	1.964475	3.29E-12	1.22E-08
RGR	1.944269	0.000231	0.006042
KRT38	1.926021	0.001432	0.018548
KRT3	1.912656	0.000856	0.013151
KLK10	**1.895774**	**1.08E-05**	**0.000807**
VGR1	1.891378	2.64E-10	3.05E-07
LOC115824	1.885044	5.44E-07	9.50E-05
FLJ41941	1.884628	9.56E-05	0.003443
CCL3L3	1.882881	1.91E-06	0.000236
NKX2-3	1.878537	0.000778	0.012487
SPDYC	1.867796	0.001711	0.020739
KLK11	1.862105	0.000801	0.01267
KRT8P41	1.854303	1.82E-09	1.53E-06
SOX15	1.840919	5.19E-06	0.000471
NACA2	1.826642	3.00E-15	1.39E-11
VIL1	1.776616	0.006256	0.046661
C1QL2	1.739977	0.000109	0.003725
ART3	1.737035	0.000356	0.007821
LAIR2	1.674734	3.99E-06	0.000397
MYO7B	1.658655	1.46E-05	0.00097
LEFTY1	1.65524	9.72E-07	0.000146
ACOXL	1.633604	4.50E-07	8.17E-05
PPP1R14A	1.63021	3.88E-07	7.41E-05
FGF17	1.627988	9.73E-06	0.000738

Gene name	**HR**	**p-value**
RPS26P11	0.27(0.09-0.82)	0.014
PWP2	0.32(0.12-0.85)	0.016
LRRC37A2	0.36(0.14-0.92)	0.02
PIF1	0.34(0.13-0.89)	0.021
PRKY	0.29(0.1-0.9)	0.022
TREML4	0.34(0.13-0.92)	0.026
LRRC37A	3.3(1.09-10.03)	0.026
XRCC6P5	0.29(0.08-1.01)	0.038
GPR17	0.27(0.06-1.19)	0.063
TUBB8	3.65(0.84-15.87)	0.065
POLR1A	3.4(0.78-14.78)	0.083
KLK10	2.41(0.86-6.78)	0.084

Table 4: The above table represents the Hazard ratio [HR] and log-rank test p-value for the 12 genes (Out of 24 genes selected by our method, only 12 genes have significant survival differences with p-value < 0.05).