Supplementary Material of "Machine Learning Integrated Ensemble of Feature Selection Methods followed by Survival Analysis for Predicting Breast Cancer Subtype specific miRNA Biomarkers"

Jnanendra Prasad Sarkar[⋆,a,c], Indrajit Saha[⋆,b], Anasua Sarkar[c], Ujjwal Maulik[c]

[a]*Larsen & Toubro Infotech Ltd., Pune, India*
[b]*Department of Computer Science and Engineering, National Institute of Technical Teachers' Training & Research, Kolkata-700106, India*
[c]*Department of Computer Science and Engineering, Jadavpur University, Kolkata, India*

Abstract

Breast cancer is the second leading cancer type in female population among other different cancer types. In this regard, it is found that microRNAs play an important role by regulating the gene expression at the post-transcriptional phase. However, identification of most influencing miRNAs in breast cancer subtypes is a challenging task, while the recent advancement in Next Generation Sequencing techniques allows analyzing high throughput expression data of miRNAs. Thus, we have conducted this research with the help of NGS data of breast cancer in order to identify the most significant miRNA biomarkers which are highly associated with multiple breast cancer subtypes. For this purpose, a two-phase technique, called Machine Learning Integrated Ensemble of Feature Selection Methods followed by survival analysis, is proposed. In the first phase, we select the best machine learning technique among seven techniques based on classification accuracy using the entire set of features (in this case miRNAs). Subsequently, eight different feature selection methods are used separately in order to rank the features and validate each set of top features using the selected machine learning technique by considering a multi-class classification task of breast cancer subtypes. In the second phase, based on the classification accuracy the top features from each feature selection method are considered to make an ensemble to provide a further categorization of miRNAs as 8*, 7* up to 1*. The 8* miRNAs provide the highest average classification accuracy of 86% after 10-fold cross-validation. Thereafter, 27 miRNAs are identified from the list which is confined within 8* to 4* miRNAs based on their importance in survival for breast cancer subtypes using Cox regression based survival analysis. Moreover, expression analysis, regulatory network analysis, protein-protein interaction analysis, KEGG pathway and gene ontology enrichment analysis are performed in order to validate biological significance. Additionally, we have prepared a miRNA-protein-drug interaction network to identify possible drug for selected miRNAs. Thus, our findings may be considered during a clinical trial for the treatment of breast cancer patients.

Keywords: Breast Cancer, Cox Regression, Drug Repurposing, Feature Selection, Machine Learning, miRNA Sequencing.

1. Additional Experiment Results

Table S1: Additional experiment results

Sr. No.	URL for additional files
1.	List of miRNAs selected by eight feature selection methods
2.	Cox regression analysis used in table 4
3.	Box plot of expression values for selected 27 miRNAs
4.	Kaplan-Meier plot of selected 27 miRNAs
5.	Refined miNRA-Gene-TF details used in figure 4
6.	miNRA-Gene correlation score used in step 9 of figure 4
7.	Refined miNRA-Gene-TF with correlation
8.	Experimental data used to prepare network analysis
9.	KEGG Pathway analysis
10.	Biological Process (GO) analysis
11.	Molecular Function (GO) analysis
12.	Cellular Component (GO) analysis
13.	miRNA-Protein-Drug interaction network analysis

*Corresponding authors: correspondence should be addressed at indrajit@nitttrkol.ac.in

⋆Joint first authors and contributed equally