Title:

Circulating tumor cells (CTCs) enumeration and machine-learning based diagnostic biomarkers for breast cancer detection

Authors

Chun-Yu Liu1,2,3, Yu-Hsiang Lin1, Yi-Fang Tsai1,2,4, Po-Yen Lu5, Ji-Lin Chen1,2, Yu-Hsuan Li5, Chi-Cheng Huang2,4,6, Yen-Shu Lin1,2,4, Ta-Chung Chao1,2,3, Chin-Jung Feng1,2,4, Chih-Yi Hsu1,7, Jen-Hwey Chiu2,4,8, Chyong-Mei Chen5,\*, Ling-Ming Tseng1,2,4,\*

Affiliations

1School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei 11217, Taiwan.

2Comprehensive Breast Health Center, Department of Surgery, Taipei Veterans General Hospital, Taipei 11217, Taiwan.

3Division of Medical Oncology, Department of Oncology, Taipei Veterans General Hospital, Taipei 11217, Taiwan.

4Division of Breast Surgery, Department of Surgery, Taipei Veterans General Hospital, Taipei 11217, Taiwan.

5Institute of Public Health, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan.

6Department of Public Health, College of Public Health, National Taiwan University, Taipei, Taiwan.

7Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei 11217, Taiwan.

8Institute of Traditional Medicine, School of Medicine, National Yang Ming Chiao Tung University, Taipei 11217, Taiwan.

|  |
| --- |
| This folder contains the R code for implementing the proposed method.  To use the new code, please conduct the following two steps.  Step 1: Implement the code “code\_SVM\_linear\_Model.r” to construct the svm-forest machine. In this file, readers need to determine the features by setting biomarker.name=c("Age", "CK18", "MGB", "WBC","Platelet") or biomarker.name=c("Age", "CK18", "MGB").  Step 2: Implement “code\_prediction for test data.r” to predict the test data, test\_data.csv. |

**Details**

The provided code package includes two scripts: **code\_SVM\_linear\_Model.r** and **code\_prediction\_for\_test\_data.r**.

* **code\_SVM\_linear\_Model.r** is used to construct the ensemble model and generates two output files: svm\_bestmodel.rds and imp.train.X\_1000.rds.
* **code\_prediction\_for\_test\_data.r** is used to evaluate the performance of the ensemble model on the independent test dataset.

The following provides a step-by-step walkthrough of the proposed method, demonstrating how the ensemble model is trained and subsequently applied for prediction.

**code\_SVM\_linear\_Model.r**

1. **Feature selection**  
   Users first specify the features by setting, for example,

|  |
| --- |
| # R  biomarker.name = c("Age", "CK18", "MGB", "WBC", "Platelet") |

or

|  |
| --- |
| # R  biomarker.name = c("Age", "CK18", "MGB") |

1. **Data import**  
   The dataset is imported using

|  |
| --- |
| # R  read.data = read.csv("read.data.csv", fileEncoding = "big5") |

In our analysis, the dataset contained a total of 398 subjects.

1. **Model training**
   * The required R packages are installed automatically.
   * Sampling was performed according to the number of cancer cases and non-cancer/healthy individuals to ensure that both the training and test datasets preserved the same structure as the original data.
   * The test dataset consisted of 48 subjects, as described in the main article.
   * The remaining 350 subjects were repeatedly partitioned using a Monte Carlo cross-validation (MCCV) scheme. In each iteration, the data were randomly split into a training set (75%, n=262) and a validation set (25%, n=88).
   * Within each split, machine learning models were trained on the training data using 10-fold cross-validation.
   * A support vector machine (SVM) with a linear kernel was trained for each split.
   * This procedure was repeated 1000 times, and the resulting “small” models were aggregated to form the ensemble classifier.

**code\_prediction\_for\_test\_data.r**

This script applies the ensemble classifier to the test dataset. The output includes the predicted probabilities, predicted disease status, and performance metrics including precision, Sensitivity, Specificity, F1\_score, the ROC curve and AUC with 95% CI.