



Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods

Zongzhen He¹, Junying Zhang^{1*}, Xiguo Yuan¹ and Yuanyuan Zhang²

¹ School of Computer Science and Technology, Xidian University, Xi'an, China, ² School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China

OPEN ACCESS

Edited by:

Wei Lan,
Guangxi University, China

Reviewed by:

Ergude Bao,
Beijing Jiaotong University, China
Zhenhua Yu,
Ningxia University, China

*Correspondence:

Junying Zhang
jyzhang@mail.xidian.edu.cn

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 24 November 2020

Accepted: 30 December 2020

Published: 18 January 2021

Citation:

He Z, Zhang J, Yuan X and
Zhang Y (2021) Integrating Somatic
Mutations for Breast Cancer Survival
Prediction Using Machine Learning
Methods. *Front. Genet.* 11:632901.
doi: 10.3389/fgene.2020.632901

Breast cancer is the most common malignancy in women, and because it has a high mortality rate, it is urgent to develop computational methods to increase the accuracy of breast cancer survival predictive models. Although multi-omics data such as gene expression have been extensively used in recent studies, the accurate prognosis of breast cancer remains a challenge. Somatic mutations are another important and promising data source for studying cancer development, and its effect on the prognosis of breast cancer remains to be further explored. Meanwhile, these omics datasets are high-dimensional and redundant. Therefore, we adopted multiple kernel learning (MKL) to efficiently integrate somatic mutation to currently molecular data including gene expression, copy number variation (CNV), methylation, and protein expression data for the prediction of breast cancer survival. Before integration, the maximum relevance minimum redundancy (mRMR) feature selection method was utilized to select features that present high relevance to survival and low redundancy among themselves for each type of data. The experimental results demonstrated that the proposed method achieved the most optimal performance and there was a remarkable improvement in the prediction performance when somatic mutations were included, indicating that somatic mutations are critical for improving breast cancer survival predictions. Moreover, mRMR was superior to other feature selection methods used in previous studies. Furthermore, MKL outperformed the other traditional classifiers in multi-omics data integration. Our analysis indicated that through employing promising omics data such as somatic mutations and harnessing the power of proper feature selection methods and effective integration frameworks, the breast cancer survival predictive accuracy can be further increased, thereby providing a more optimal clinical diagnosis and more effective treatment for breast cancer patients.

Keywords: breast cancer, multi-omics, survival prediction, somatic mutation, mRMR, MKL

INTRODUCTION

Breast cancer is the most common malignant tumor in women. Although there are millions of breast cancer survivors in the United States, breast cancer is the main cause of cancer-related deaths worldwide because of its high mortality rate (Ferlay et al., 2010). Thus, it is urgent to design highly accurate methods to predict the survival of breast cancer patients. Accordingly, effective survival

predictors could finally contribute to the reduction of the overall mortality of breast cancer and could further improve the life quality and increase the lifespan of breast cancer patients.

Recently, the Cox regression model (Yuan et al., 2014; Xu et al., 2016) and traditional machine learning classification methods, such as support vector machine (SVM) (Xu et al., 2013), Bayes classifier (Gevaert et al., 2006), and random forest (RF) (Nguyen et al., 2013), have been widely deployed to identify breast cancer prognostic biomarkers. Multiple survival prediction models have been mainly developed based on gene expression data. The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2013; Brennan et al., 2014) provides multiple types of molecular data such as gene expression (Exp), copy number variation (CNV), methylation (Methy), protein expression (Protein), and somatic mutation (SM) data for various cancers, including breast cancer. Moreover, the advancement of machine learning technologies enables various data types to be combined within a model (Chen et al., 2019; Lan et al., 2020), which may increase the accuracy of predictive models.

One of the biggest challenges in breast cancer research involves the effective combination of heterogeneous data sources into survival prediction models, making the selection of a proper integration method essential. In previous studies (Seoane et al., 2014; Zhang et al., 2016; Sun et al., 2018; Zhang A. et al., 2019; Zhang Y. et al., 2019), multiple kernel learning (MKL) (Lanckriet et al., 2004; Rakotomamonjy et al., 2008; Kloft et al., 2011) was successfully used to integrate different types of data into a universal model to distinguish short-term and long-term cancers survivors. MKL uses different kernels for different types of data, and then trains the weight of each kernel to select the best combination of kernel functions for classification. These studies have demonstrated that models that were obtained using integrated data improved the performance of survival prediction compared to models that used only one single data type.

A previous study (Sun et al., 2018) showed that MKL outperformed Cox-based regression models for breast cancer survival prediction. However, omics data, such as Exp, CNV, and methylation data, are usually extremely high-dimensional and redundant (Dey et al., 1990). In the previous study (Sun et al., 2018), information gain ratio (IGR) was utilized to select survival relevant features from multi-omics data, but the redundancy of dataset features was not considered. Despite the promising performance of the above MKL-based studies for breast cancer prognosis, somatic mutations are rarely considered for breast cancer survival prediction due to their complexity and heterogeneity in serious disease. Therefore, there is still much room to increase the accuracy of breast cancer survival models by incorporating somatic mutations into the MKL model.

Currently, somatic mutations are strongly correlated with the clinical symptoms of breast cancer (Griffith et al., 2018), and they have been successfully adopted for the classification of primary cancer sites (Chen et al., 2015) and identification of survival-related cancer subtypes (Hofree et al., 2013; He et al., 2017; Ronen et al., 2018; Arslanturk et al., 2020). Somatic mutations are sparse but common mutations of that offer less accuracy in the prediction of cancer survival (Zhang et al., 2018; Ye et al., 2019). Previous studies (Haricharan et al., 2014;

Griffith et al., 2018; Zhang et al., 2018; Ye et al., 2019) have reported that mutations enriched in specific pathways have shown potential for breast cancer survival prediction. The authors of a previous study (Griffith et al., 2018) stated that uncommon recurrent somatic mutations should be further explored to explain breast cancer survival outcomes. In the present study, the effect of somatic mutations on the integrated prognosis of breast cancer is explored.

In the present study, we applied the state-of-the-art MKL method in the integration of somatic mutation datasets with previously used omics data, including Exp, CNV, Methy, and Protein, to train and test an integrated breast cancer survival prediction model. The maximum relevance minimum redundancy (mRMR) algorithm (Ding and Peng, 2005; Radovic et al., 2017) was used to alleviate the redundancy of the data, by simultaneously selecting highly predictive but non-redundant features from each type of molecular data. Then, selected features from multiple data type were integrated into the MKL classification.

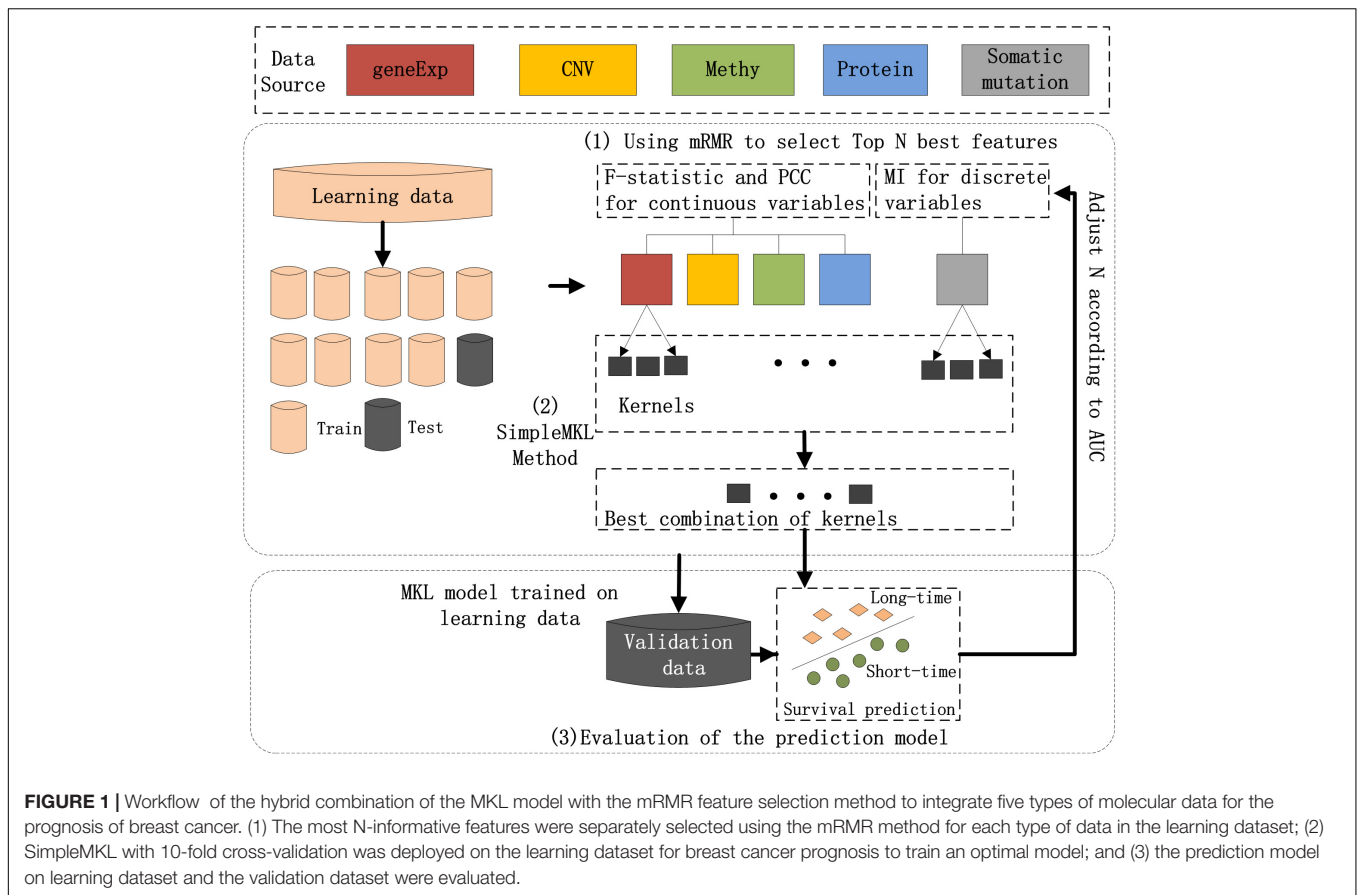
In order to gauge the performance of our method, first, the newly introduced method was compared with different single data types and integrated datasets to verify the effectiveness of somatic mutations, and the results indicated that there was a remarkable improvement in the prediction performance when somatic mutations were included. Different feature selection algorithms were then studied, and the experimental results demonstrated that mRMR was the most optional among them. Furthermore, the MKL classification method was compared with other traditional classifiers, and the experimental results proved the superiority of MKL in data integration. Finally, the newly introduced model was validated in an independent validation dataset and achieved a promising high accuracy in survival prediction. According to the results, the most optimal performance was achieved by our method, which demonstrated the feasibility of integrating somatic mutations in the prognostic models and the usefulness of mRMR and MKL in breast cancer prognosis.

The reminder of this article is organized as follows. A workflow of our proposed method and related methods are described. Next, comparative studies were carried out to evaluate the performance of the proposed methods and their comparison methods, as well as to analyze the most informative features discovered by our model. Then, we applied our model on the validation dataset. Finally, the proposed method is discussed, and it is expected to undergo the improvement in future studies.

MATERIALS AND METHODS

Workflow of the Proposed Method

The workflow chart of the proposed method is shown in **Figure 1**. Preprocessing of the input dataset initially occurred, during which entire datasets were randomly divided into a learning dataset (80% of the entire dataset) and validation dataset (20%). Then, three main steps were carried out to realize the prediction of breast cancer prognosis.



The three main steps include: (1) The most N-informative features were separately selected using the mRMR method for each type of data in the learning dataset; (2) SimpleMKL with 10-fold cross-validation was deployed on the learning dataset for breast cancer prognosis to train an optimal model; and (3) the prediction model on learning dataset and the validation dataset were evaluated for their ability to learn data. A detailed description of each of the steps is listed below.

Data Input and Preprocessing

The Cancer Genome Atlas provides multiple types of biomolecular data. High-level molecular data for breast cancer were retrieved from TCGA, including gene expression, gene CNV, gene methylation, protein expression, and somatic mutation along with clinical features from the University of California Santa Cruz (UCSC) cancer browser website¹ (Mary et al., 2014). The downloaded dataset consisted of five types of data, including different numbers of samples, and the original data matrixes were structured with rows denoting patient samples and columns denoting features. A total of 139 true normal, seven metastatic, and 13 male patients' samples were removed, and regarding somatic mutations, samples with less than 10 mutations were removed (Hofree et al., 2013; He et al., 2017). We finally obtained 488 primary breast tumors

¹<https://xenabrowser.net/datapages/>

together with survival time, and all samples of them included all of the five aforementioned genomic data types. The details of our dataset are illustrated in **Table 1**. The median age at diagnosis was 57.37, and the median survival time was 42.43 months, which is in agreement with the previous research (Sun et al., 2018).

We followed the protocol from our previously published studies (He et al., 2017, 2019), and we first removed the genes with missing values in more than 10% of samples for gene expression, CNV, gene methylation, protein expression, and somatic mutations. After that, flat variables that had the same values in more than 80% of the samples (non-informative) were discarded except in the case of somatic mutations (Yuan et al., 2014; He et al., 2019). According to the previous study (He et al., 2019), the RNA-Seq gene expression level 3 transcription was log₂ transformed and RSEM-normalized (Li and Dewey, 2011). Regarding the CNV features, we directly utilized the gene-level

TABLE 1 | The detailed information in our breast cancer dataset.

Properties	Number
Total population of primary cancer	488
Long-term survivors	119
Short-term survivors	369
Mean age at diagnosis (years old)	57.37
Median survival (months)	42.43

copy number values that were estimated using the GISTIC2 method (Mermel et al., 2011; Yuan et al., 2017; Yuan et al., 2019, 2020a,b). For gene methylation and protein expression, we directly used the original data with z-score normalization. For somatic mutation, we also directly utilized the original binary data, and in addition, genes that were mutated in more than one sample were reserved for further analysis. The gene expression, CNV, gene methylation, and somatic mutations contained 18,000, 25,000, 22,000, and 14,000 features, respectively, after data filtering, and the properties of these datasets are shown in **Table 2**.

In the present study, the survival prediction for breast cancer was defined as a binary classification problem with a threshold of 5 years as conducted in previous studies (Seoane et al., 2014; Zhang et al., 2016; Sun et al., 2018; Zhang A. et al., 2019). Of the total, 369 out of the final 488 patients with survival shorter than 5 years were considered as short-term survivors, and 119 patients with survival longer than 5 years were considered as long-term survivors. Moreover, the long-term patients were labeled as 1, while short-term patients were labeled as 0. After the initial data preprocessing, the entire dataset was randomly divided into the learning dataset (80%) and validation dataset (20%). For each type of data, we initially conducted the following feature selection on the learning dataset containing 390 breast cancer patients, and trained and tested the integrated MKL model on it to obtain the optimal parameters. Then, we applied the optimal model on the validation dataset that included 98 patients.

mRMR Feature Selection

Five different types of genomic data were used in the present study, as described above, and the number of variables for most types of genomic data exceeded 10,000 after feature preprocessing. However, this large number of features may cause poor performance due to dimensionality and high redundancy (Jain and Zongker, 1997; Jie et al., 2015). Therefore, according to our previous study (He et al., 2019), mRMR was adopted in the present study to select the most useful features for the prognostic model.

The mRMR is a feature selection method that aims to select a subset of features that are highly related to the output classes and have low redundancy between them (Radovic et al., 2017). In the present work, mRMR was deployed to select features from five types of molecular data that are the most highly relevant with respect to survival and the least correlated among themselves. Then, the most relevant features for each molecular dataset were combined to form a candidate feature set to be used for classification. A feature of one type of genomic dataset for the i th

variable with N individuals is denoted as $v_i \in R^M, i = 1, \dots, M$, and the survival prediction labels with N individuals as $l \in R$. For label l , mRMR aims to search a feature subset S with k features $\{v_i\}$, which collectively have the maximal relevance (Max-Relevance) $Rel(S, l)$ on the target label l and the minimal redundancy (Min-Redundancy) $Red(S)$.

The F-statistic (F) was used to calculate the relevance between feature variables with binary survival terms and the Pearson correlation coefficient (PCC) was used to measure the redundancy for the continuous feature variables of the gene expression, CNV, gene methylation, and protein datasets. Max-Relevance is defined in Eq. 1, where relevance $Rel(S, l)$ is calculated using the mean value of all F-statistic values F of the individual variables v_i with the label l . In parallel, the Min-Redundancy $Red(S)$ constraint was adopted to select irrelevant features, and is shown as Eq. 2.

$$\max Rel(S, l), \quad Rel = \frac{1}{|S|} \sum_{v_i \in S} F(v_i; l), \quad (1)$$

$$\min Red(S), \quad Red = \frac{1}{|S|^2} \sum_{v_i, v_j \in S} PCC(v_i; v_j) \quad (2)$$

For binary discrete feature variables of somatic mutation data, the mutual information (MI) was used to calculate both the relevance between feature variables and survival terms, and the redundancy between mutations. Max-Relevance is used to select features satisfying Eq. 3, where relevance $Rel(S, l)$ is obtained by the mean value of all MI values of individual variable v_i with label l . The Min-Redundancy constraint $Red(S)$ is used to select irrelevant features, and is shown as Eq. 4.

$$\max Rel(S, l), \quad Rel = \frac{1}{|S|} \sum_{v_i \in S} MI(v_i; l), \quad (3)$$

$$\min Red(S), \quad Red = \frac{1}{|S|^2} \sum_{v_i, v_j \in S} MI(v_i; v_j) \quad (4)$$

Finally, as shown in Eq. 5, the operator $\phi(Rel, Red)$ was deployed to simultaneously optimize the two constraints “Max-Relevance” and “Min-Redundancy” based on the MI quotient (MIQ) criterion (Radovic et al., 2017; He et al., 2019) to obtain the best feature subsets, as shown in Eq. 5:

$$\max_{v_k} \phi(Rel, Red), \quad \phi = Rel/Red \quad (5)$$

The area under the curve (AUC) value is used as a metric to evaluate the performance and the most optimal number of the most relevant and non-redundant features k for each data type was determined by comparing the AUC valued for the models. After the mRMR features were selected for each type of genomic data, the most informative features were combined and used as the input feature set for the classification problems.

Multiple Kernel Learning

In our study, we aimed to integrate multiple types of genomics data, with a focus on somatic mutations. Although the fusion

TABLE 2 | The properties of five types of genomic data for our breast cancer prediction.

Data types	Feature number
Gene expression	18624
CNV	24774
Gene methylation	21136
Protein expression	170
Somatic mutations	13602

of multiple types of data into one model is one of the most widely used methods for classification, this is not feasible due to the fact that different types of molecular data present different feature representations (Khademi and Nedialkov, 2016). MKL has become a natural method to enhance the interpretability of models and to address the data integration problem. The optimal function can be obtained by constructing a linear weighted combination of predefined M kernels. The optimal combination of kernels is given as Eqs 6 and 7:

$$K(x_i, x_j) = \sum_{m=1}^M d_m K_m(x_i, x_j), \tag{6}$$

$$s.t. \ d_m \geq 0, \text{ and } \sum_{m=1}^M d_m = 1, \tag{7}$$

where d_m denotes the weight of the m th different kernel $K_m(x_i, x_j)$.

Some methods based on MKL have been proposed and many of them outperformed uni-MKL (Rakotomamonjy et al., 2008; Gönen and Alpaydin, 2011; Kloft et al., 2011). However, most of the weights d_m of the kernels were 0 and thus non-contributory to the MKL model (Ikononov et al., 2013). In the present work, SimpleMKL (Zhang et al., 2016), which is based on a weighted L2-norm regularization and is more powerful than other methods (Yan et al., 2009), was adopted as our classification model. It employs dual kernels in the of classic kernel optimization problem, which can be presented as Eq. 8:

$$f(x) = \sum_{i=1}^l \alpha_i^* K(x_j, x_i) + b^* \tag{8}$$

The decision function is given as:

$$\begin{aligned} \min_{f,b,\varepsilon} \quad & \frac{1}{2} f_H^2 + C \sum_i \varepsilon_i \\ s.t. \quad & y_i(f(x_i) + b) \geq 1 - \varepsilon_i \quad \forall_i, \\ & \varepsilon_i \geq 0 \quad \forall_i \end{aligned} \tag{9}$$

where $\|f\|_H$ denotes a kernel in Hilbert space related to a kernel K_m . The overall kernel can be divided into different kernels, and we replace $\|f\|_H$ with $\sum_m \|f_m\|_{HM}$ to obtain:

$$\begin{aligned} \min_{f_m,b,\varepsilon,d} \quad & \frac{1}{2} \sum_m \|f_m\|_{HM}^2 + C \sum_i \varepsilon_i \\ s.t. \quad & y_i \sum_m f_m(x_i) + y_i b \geq 1 - \varepsilon_i \quad \forall_i, \\ & \varepsilon_i \geq 0 \quad \forall_i \\ & \sum_m d_m = 1, \ d_m \geq 0 \quad \forall_m \end{aligned} \tag{10}$$

Optimization matter is performed using the convex optimization mathematical algorithm (Rakotomamonjy et al., 2008). Using multiple kernels increases the decision the power of the decision function and also increases the prediction performance compared to using one single kernel. In the present study, SimpleMKL was deployed to integrate five different types of molecular data including gene expression, CNV, gene methylation, protein expression, and somatic mutation.

Considering the number of data types used in our study, five different kernels were independently built and further integrated into a generic model. Each kernel corresponds to each individual data type (gene expression, CNV, gene methylation, protein expression, and somatic mutation). The ‘‘Poly’’ (Eq. 11) polynomial base kernel with a search range of degrees of freedom $d \in \{1, 2, 3\}$ (Seoane et al., 2014) and the ‘‘Gaussian’’ (Eq. 12) kernel with a search range of the parameter $\delta \in \{0.25, 0.5, 1, 2, 5, 7, 10, 12, 15, 17, 20\}$ (Zhang et al., 2016; Sun et al., 2018) were used as kernel types.

$$K(x_i, x_j) = (x_i^T x_j + 1)^d, \tag{11}$$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right) \tag{12}$$

In summary, the SimpleMKL directly addressed a multiple kernel SVM optimization problem and greatly reduced computation costs when compared to the use of learning kernel combinations from individual kernels.

Evaluation

The dataset used in our study was randomly divided into learning and validating sets in order to assess the performance of the proposed method. For the learning set, we used mRMR to select the most optimal features and to determine the model through 10-fold cross-validation experiments. Then, the pre-trained MKL model and its optimal parameters were used to predict the validation set. Because the validation dataset was not used in the cross-validation process, the model derived from the learning dataset was tested on an independent validation dataset.

To assess the performance of our model, AUC, the most widespread evaluation metric for classification problems, was used to assess the performance of the proposed model. AUC is defined as the area under the receiver operating characteristic (ROC) curve, and it is used to quantify the overall performance of a classification model. Specifically, AUC = 1 denotes perfect performance, and 0.5 denotes random guessing. Pre (precision, Eq. 13), Sn (sensitivity, Eq. 14), Sp (Specificity, Eq. 15), and Acc (Accuracy, Eq. 16) were also employed in addition to AUC as classification performance metrics for breast cancer prognosis. The definitions of those metrics are provided below:

$$Pre = \frac{TP}{TP + FP}, \tag{13}$$

$$Sn = \frac{TP}{TP + FN}, \tag{14}$$

$$Sp = \frac{TN}{TN + FP}, \tag{15}$$

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{16}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

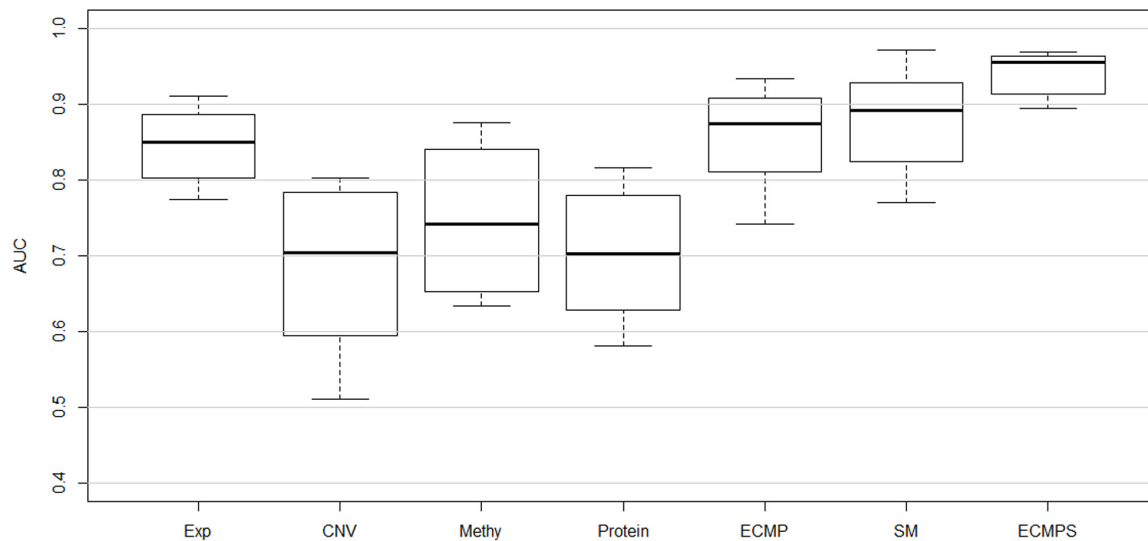


FIGURE 2 | Performance of classifying long-term and short-term survivors from a breast cancer dataset using different types of data based on the proposed hybrid combination of mRMR feature selection and MKL classification methods.

RESULTS

Comparison Studies on Learning Datasets

The proposed method was compared with other methods in three different applications: (1) comparison of the results of the models with different datasets based on the same method; (2) comparison of the results of different feature selection methods under the same datasets; and (3) comparison of the integration results of classification methods, under the same integrated datasets. AUC was used as an evaluation metric when comparing different methods and 10-fold cross-validation was applied for all methods.

Comparison of ECMPS and Other Data Types

Seven different MKL-based models were built using five single types of molecular data [gene expression (Exp), CNV, gene methylation (Methy), protein expression (Protein), and somatic mutations (SM)] and two integrated datasets with and without somatic mutation data in order to evaluate the role of somatic mutations in breast cancer survival prediction. The dataset integrating gene expression, CNV, gene methylation, and protein expression is abbreviated as “ECMP,” and the dataset integrating all five molecular datasets including somatic mutations is denoted as “ECMPS.”

The corresponding mean of the AUC value of 10-fold cross-validation (CVmean_AUC) for each of the seven models, using the mRMR feature selection and the MKL classification method, was calculated to compare the predictive performance of breast cancer survival models. The results are displayed in **Figure 2**, with the mean values of the boxplots corresponding to the red line in **Figure 3**. As shown in **Figure 3**, the ECMPS model consistently exhibited significantly more optimal performances than all the other models for all three feature selection methods. The two

integrated models present obvious improvements compared to the single data type model results, suggesting that integrated models are more optimal than single data type ones, which is consistent with previous studies (Zhang et al., 2016; Sun et al., 2018).

In **Figure 2**, the mean value of the AUC for the multi-data ECMP model without somatic data is 0.8854, and the corresponding value for the ECMPS model increased to 0.9421 when incorporating somatic mutation. In addition, among the single data type models, the AUC of the somatic mutation model was higher than that of the model using the other four single data types and ECMP. Thus, our experimental results indicated that the somatic mutation data is able to increase the accuracy of the survival prediction for breast cancer patients.

The Pre, Sn, Sp, and Acc values for each dataset model were calculated in addition to the AUC based on the proposed method, and the results are presented in **Figure 4A**. **Figure 4A** shows that the integrative models combining different types of data, including somatic mutations, overcome the models using single data types for classification. The experimental results indicated that the proposed integrated model can successfully predict the survival time for breast cancer patients and somatic mutations can improve predictive accuracy.

Comparison of mRMR With Different Feature Selection Methods

We used mRMR to select the variables for each of the five types of molecular data. Then, the features with the largest relevance to the survival and lowest redundancy among themselves were selected, and they were combined as integrated features using the MKL classification model. The most optimal number of selected non-redundant features k for each molecular data type was determined by comparing the AUC values in the prediction results. According to the number of features reported in the

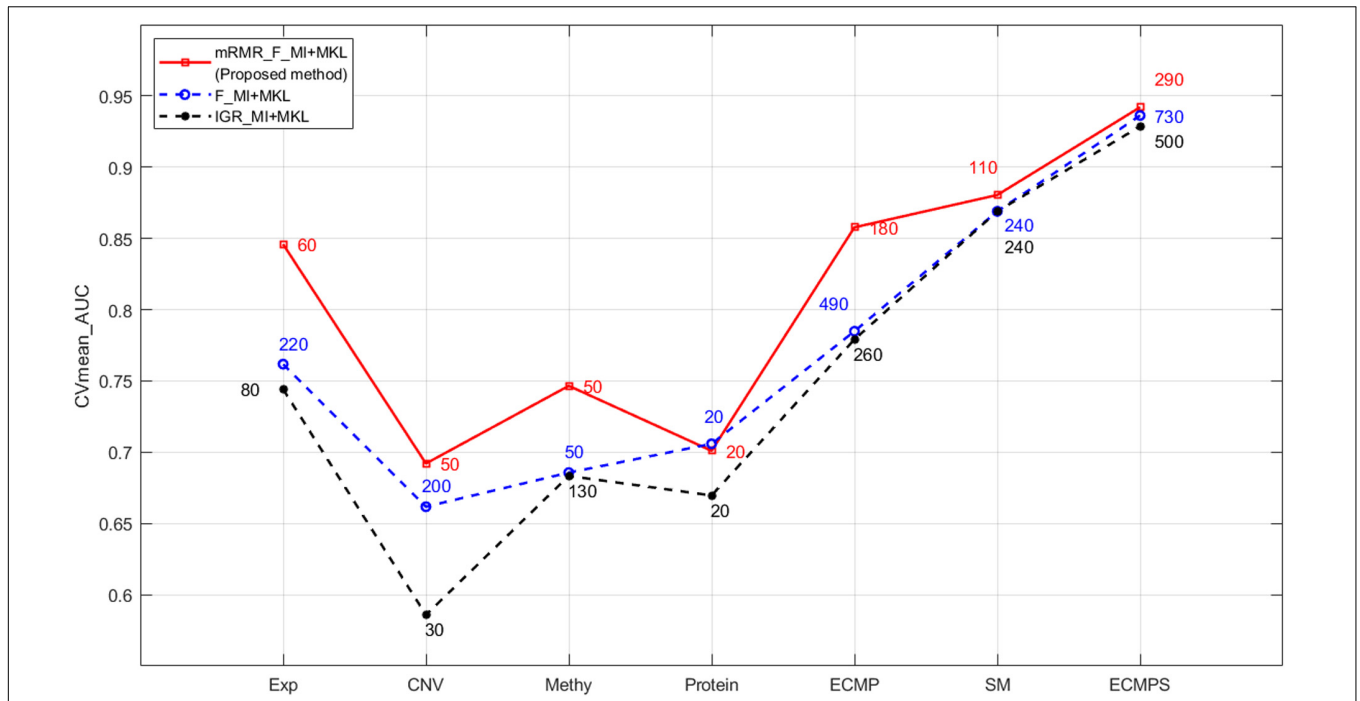


FIGURE 3 | Performance comparison of mRMR and the two k-best methods based on MKL under different data types. The numbers in different colors on the lines indicate the number of optimal features selected by the corresponding method.

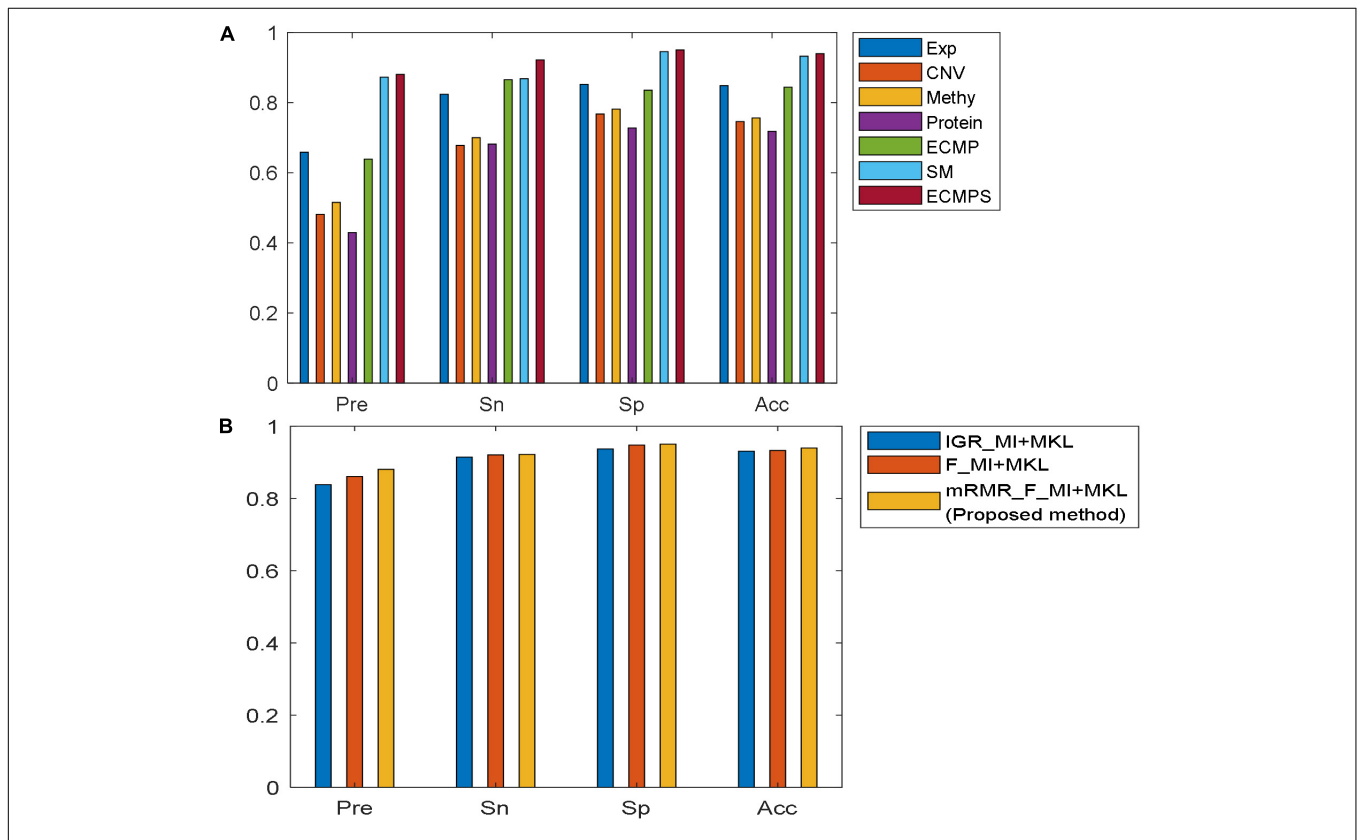


FIGURE 4 | Comparison of performances of the models using different evaluation metrics: Pre, Sn, Sp, and Acc. **(A)** Performance of the proposed method in seven datasets. **(B)** Performance of various feature selection methods based on MKL under the same data type “ECMPs.”

previous study (Sun et al., 2018), we set $k = [10, 20, \dots, 300]$ in our work and chose the optimal parameter k as the final parameter for each data type in our study based on the prediction result.

The classification outcomes of the five data types under different parameters are presented in **Supplementary File S1**. The optimal feature number was selected based on the position of the maximum AUC value as the final parameter for a model of further integration. Take gene expression for example, as shown in **Figure 5**, the optimal number of features in the gene expression model using the proposed method is 60, which achieves the largest mean value of AUC with 10-fold cross-validation. Finally, we chose $k = [60, 50, 50, 20, 110]$ as the optimal parameters for the five types of molecular data (Exp, CNV, Methy, Protein, and SM), respectively, for further integration analysis, and the total 290 features were obtained for our integrated ECMPS model.

The F-statistic (F) and PCC were used for the mRMR feature selection method to calculate the relevance and redundancy (Radovic et al., 2017), respectively, for four continuous data types, including Exp, CNV, Methy, and Protein, in order to maintain the original information for different types of data. MI was used to calculate both the relevance and redundancy of somatic mutation features, and is short for “mRMR_F_MI.” In all cases, the selected features were integrated using MKL classification. To assess the performance of the mRMR feature selection method in the selection of features for our breast cancer survival prediction model, the proposed mRMR feature selection was compared with two commonly used k-best methods, which only consider relevance with the output, based on the same datasets and classification method MKL: (1) F-MI. Compared to the proposed method, it only uses the F-statistic and MI to select

the most optimal k-best features for four continuous molecular datasets and discrete somatic mutation. (2) IGR-MI. It adopts a recently used feature selection method, the IGR (Sun et al., 2018), for four continuous molecular datasets and MI for discrete somatic mutation.

The proposed mRMR method outperformed both k-best feature selection methods F-statistics and IGR for four continuous molecular data types and their integration ECMPS model according to the results shown in **Figure 3**. For instance, 260 features were selected by IGR based on the ECMPS model and the AUC value was 0.7791, which was consistent with previous studies (Sun et al., 2018). Next, 180 features were selected using mRMR and AUC was 0.8578 showing that mRMR can achieve higher predictive accuracy using fewer features. The mRMR method also outperformed MI for discrete somatic mutation returning a smaller number of features. The most optimal result was obtained by mRMR and the total integration model ECMPS. The metrics Pre, Sn, Sp, and Acc were calculated in addition to the AUC for each dataset model, with a more optimal performance by mRMR as compared to the other the two k-best methods (**Figure 4B**). Our findings indicated that the use of proper feature selection methods is crucial to the classification process.

As the red line shows in **Figure 3**, for the integrated ECMPS model, 290 features were selected as more relevant to survival and non-redundant features in the integrated ECMPS mode consisting of 60 Exp, 50 CNV, 50 Methy, 20 Protein, and 110 SM using mRMR, and the most optimal AUC (0.9421) in the present study was achieved. Next, mRMR was applied again for the set of 290 features, which is

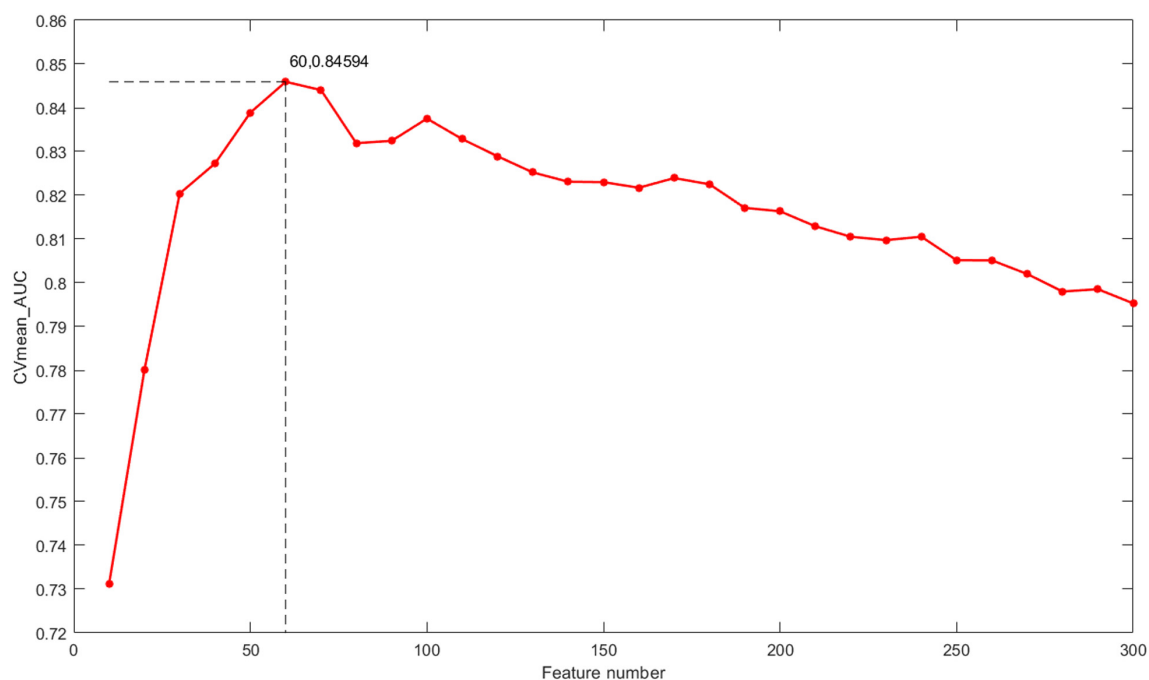


FIGURE 5 | The mean value of the AUC for 10-fold cross-validation (CVmean_AUC) under the feature numbers ranging from 10 to 300 for the model based on gene expression.

TABLE 3 | Comparison of mRMR and 2-mRMR on survival prediction power and feature numbers.

	AUC(ECMPS)	Number of features					
		ECMPS	Exp	CNV	Methy	Protein	SM
mRMR	0.9421 ± 0.0281	290	60	50	50	20	110
2-mRMR	0.9439 ± 0.0264	220	49	22	29	10	110

TABLE 4 | Comparative results of the proposed MKL method and existing traditional classifiers using AUC values under two mRMR selected integrated data models.

	ECMP	ECMPS
RF	0.7135 ± 0.054	0.7916 ± 0.027
SVM	0.8325 ± 0.037	0.9086 ± 0.058
MKL	0.8578 ± 0.049	0.9421 ± 0.028

termed as “2-mRMR,” to resolve the redundancy that exists in the selected features of different data types. However, as shown in **Table 3**, the number of SM remained at 110 among the new 220 features, and the AUC value was only marginally improved. These results showed that there is a large internal redundancy within one type of data, while the redundancy between different types of data is small. It further indicated that the importance of somatic mutations to the prognosis is relatively stable. Finally, we retained the integrated 290 features originally selected by mRMR and used them for further classification, considering the stable high performance and simpler simple computational complexity of mRMR. We observed that mRMR outperformed k-best methods, and integrating somatic mutations achieved the most accurate prognosis.

Comparison of MKL With Traditional Classification Methods

The proposed method achieves a stronger performance by integrating somatic mutations compared with those methods incorporating single data types and integrated datasets without somatic mutations. The MKL classification method was compared with two widely used classifiers, SVM and RF, to further verify its ability to combine different types of data. Experiments were conducted in two integrated datasets: ECMP and ECMPS, which were selected by mRMR. The AUC value (mean value and standard error) was used to assess the performance of different methods and the results are provided in **Table 4**. **Table 4** shows that a more optimal performance was obtained from MKL for both integrated datasets compared to other classifiers, and this finding indicated the superiority of MKL in data integration.

In addition, the performances of all the classifiers were improved when employing ECMPS compared with ECMP, which further suggested that somatic mutations can provide adequate supplementary information for survival prediction of breast cancer. Finally, our method achieves the most optimal performance due to its ability to integrate multiple molecular data types, including somatic mutations, and MKL was quite efficient in integrating the data from distinct sources in breast cancer survival prediction.

Analysis of the Most Desirable Features From Somatic Mutation and Gene Expression Data

The top 10 features ranked by mRMR for each molecular data type were further analyzed by conducting a simple analysis on their association with breast cancer.

TABLE 5 | Genes previously associated with breast cancer.

Genes	Reports	References
HCN4	HCN4 was highly correlated with lower survival rates of breast cancer.	Phan et al., 2017
RGPD3	30 most enriched new HOXB7 binding sites on breast cancer cell chromatin for which an annotated nearest gene exists: RGPD3, PIK3R1, etc.	Heinonen et al., 2015
EFCAB13	Variants that induce premature stop codons were identified in the DENND2D, EFCAB13, and TICRR genes.	Määttä et al., 2016
NFATC1	NFATC1 overexpression results in oncogenic BMI1 transcriptional upregulation. Co-expression of FUNDC1 and BMI1 in BC patients predicted worse prognosis.	Wu et al., 2019
VAC14	VAC14 selectively prevents rapid degradation of Sac3.	Ikonomov et al., 2013
PRB2	A novel six-gene (TMEM252, PRB2, SMCO1, IVL, SMR3B, and COL9A3) signature was significantly associated with prognosis as an independent prognostic signature.	Lv et al., 2019
HIPK1	The deletion of the miR-200c/141 cluster resulted in increased tumor metastasis and inhibited tumor growth by directly upregulating the target gene HIPK1.	Liu et al., 2018
IRF2	Interferon regulatory factor 1 (IRF-1) and IRF-2 expression in breast cancer tissue microarrays.	Connett et al., 2005
HMGB2	Promotion of breast cancer progression by HMGB2.	Fu et al., 2018
FRMPD1	Rat Mcs5a is associated with breast cancer risk. Mcs5a1 is located within the ubiquitin ligase Fbxo10, whereas Mcs5a2 includes the 5' portion of FRMPD1.	Samuelson et al., 2007
RPS27	The best ranked cancer immunotherapy proteins related to BC were RPS27, SUPT4H1, and CLPSL2.	López-Cortés et al., 2020
PTPRR	PTPRR and myocyte enhancer factor 2C (MEF2C) genes were upregulated in the classical MAPK and p38 MAPK pathways.	Motaghdet et al., 2014

Only features from somatic mutations and gene expression datasets were explored to further assess the effectiveness of our method. The results of this analysis showed that it was previously reported that some of the genes are associated with breast cancer survival. These genes and their references are listed in **Table 5**. It has previously been reported in the literature that seven of the top 10 ranked gene names from the somatic mutation features play critical roles in breast cancer prognosis. For example, the HCN4 gene is highly correlated with lower survival rates of breast cancer (Phan et al., 2017), and the gene PRB2 is significantly related to prognosis as an independent prognostic marker (Lv et al., 2019). On the other hand, five of the top 10 genes selected from gene expression datasets have also been found to be associated with breast cancer. For instance, the expression of IRF2 has been found to be related to breast cancer (Connett et al., 2005), and it has been reported that HMGB2 directly and significantly promotes breast cancer progression (Fu et al., 2018). Thus, the top ranked features were shown to be important for breast cancer prognosis.

Validation

Optimization techniques have been previously applied (Zhang et al., 2016; Zhang A. et al., 2019) to select the most optimal feature subsets in a wrapper feature selection framework. Therefore, experiments were performed on an independent validation dataset to further evaluate our proposed method. Our model was initially trained and tested on a learning dataset containing 390 breast cancer patients, and then, to predict patient survival, it was applied to a 98-patient validation dataset that was not involved in training or testing. The survival of most of the 98 breast cancer patients was correctly classified, and the accuracy of the proposed method on the validation dataset was 0.9808.

DISCUSSION

We integrated somatic mutations and previously used data types, including Exp, CNV, Methy, and protein, using MKL to predict breast cancer patient survival. Applying mRMR-selected features and MKL classification, we found that the integration of somatic mutations enriched the diversity of features and was conducive to the improvement of the prediction model. In all, integrating promising data sources such as somatic mutations and harnessing the powerful feature selection method mRMR and the effective data fusion method MKL can increase the prediction accuracy of breast cancer patient survival.

REFERENCES

- Arslanturk, S., Draghici, S., and Nguyen, T. (2020). Integrated Cancer subtyping using heterogeneous genome-scale molecular datasets. *Pac. Symp. Biocomput.* 25, 551–562.
- Brennan, C. W., Verhaak, R. G. W., Mckenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2014). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477.
- Cancer Genome Atlas Research Network (2013). Comprehensive genomic characterization defines human glioblastoma genes

and core pathways. *Nature* 494, 506–506. doi: 10.1038/nature11903

Chen, Q., Lai, D., He, L., Yan, Y., Li, E., Liu, Y., et al. (2019). “ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion” in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Piscataway, NJ: IEEE.

Chen, Y., Sun, J., Huang, L.-C., Xu, H., and Zhao, Z. (2015). Classification of cancer primary sites using machine learning and somatic mutations. *Biomed. Res. Int.* 2015, 1–9. doi: 10.1155/2015/491502

Although our method is effective and can accurately predict the survival of breast cancer patients, some limitations remain in the prognosis of breast cancer. For instance, there may be more effective methods that can be used to construct kernels for an improved multi-kernel learning method in the future that will further improve the performance in multi-omics data fusing. In addition, our available sample size was limited by the intersection of multiple types of molecular data samples. Thus, the performance of our method could be promoted when a larger population of samples becomes available in the future. Furthermore, somatic mutations are highly heterogeneous among patients, and therefore, further understanding of the mechanism of somatic mutation in cancer may lead to a more accurate prognostic model for breast cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZH and JZ participated in the design of algorithms and experiments and participated in the design of the whole framework of prediction of breast cancer survival. JZ directed the whole work. YZ participated in the analysis of the performance of the proposed method. JZ and XY conceived of the study and helped edit the manuscript. All authors read the final manuscript and approved the submission.

FUNDING

This research was funded by the Natural Science Foundation of China under Grants 61571341 and 61902430 and the Fundamental Research Funds for the Central Universities under Grant JB190302.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.632901/full#supplementary-material>

- Connett, J. M., Badri, L., Giordano, T. J., Connett, W. C., and Doherty, G. M. (2005). Interferon regulatory factor 1 (IRF-1) and IRF-2 expression in breast cancer tissue microarrays. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res.* 25, 587–594. doi: 10.1089/jir.2005.25.587
- Dey, S., Gupta, R., Steinbach, M., and Kumar, V. (1990). *Integration of Clinical and Genomic Data: A Methodological Survey*. Minneapolis, MN: University of Minnesota Digital Conservancy.
- Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/s0219720005001004
- Ferlay, J., Héry, C., Autier, P., and Sankaranarayanan, R. (2010). *Global Burden of Breast Cancer*. New York, NY: Springer.
- Fu, D., Li, J., Wei, J., Zhang, Z., Luo, Y., Tan, H., et al. (2018). HMGB2 is associated with malignancy and regulates Warburg effect by targeting LDHB and FBP1 in breast cancer. *Cell Commun. Signal.* 16:8.
- Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, A. B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, e184–e190.
- Gönen, M., and Alpaydin, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.
- Griffith, O. L., Spies, N. C., Anurag, M., Griffith, M., Luo, J., Tu, D., et al. (2018). The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat. Commun.* 9:3476.
- Haricharan, S., Bainbridge, M. N., Scheet, P., and Brown, P. H. (2014). Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. *Breast Cancer Res. Treat.* 146, 211–220. doi: 10.1007/s10549-014-2991-x
- He, Z., Zhang, J., Yuan, X., Liu, Z., Liu, B., Tuo, S., et al. (2017). Network based stratification of major cancers by integrating somatic mutation and gene expression data. *PLoS One* 12:e0177662. doi: 10.1371/journal.pone.0177662
- He, Z., Zhang, J., Yuan, X., Xi, J., Liu, Z., and Zhang, Y. (2019). Stratification of breast cancer by integrating gene expression data and clinical variables. *Molecules* 24:631. doi: 10.3390/molecules24030631
- Heinonen, H., Lepikhova, T., Sahu, B., Pehkonen, H., Pihlajamaa, P. I., Louhimo, R., et al. (2015). Identification of several potential chromatin binding sites of HOXB7 and its downstream target genes in breast cancer. *Int. J. Cancer J. Int. Cancer* 137, 2374–2383. doi: 10.1002/ijc.29616
- Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651
- Ikonomov, O. C., Filios, C., Sbrissa, D., Chen, X., and Shisheva, A. (2013). The PKfyve-ArPKfyve-Sac3 triad in human breast cancer: functional link between elevated Sac3 phosphatase and enhanced proliferation of triple negative cell lines. *Other* 440, 342–347. doi: 10.1016/j.bbrc.2013.09.080
- Jain, A., and Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 153–158. doi: 10.1109/34.574797
- Jie, T., Hammond, J. H., Hogan, D. A., and Greene, C. S. (2015). ADAGE analysis of publicly available gene expression data collections illuminates *Pseudomonas aeruginosa*-host interactions. *mSystems* 1:e00025-15. doi: 10.1128/mSystems.00025-15
- Khademi, M., and Nediakov, N. S. (2016). “Probabilistic graphical models and deep belief networks for prognosis of breast cancer,” in *Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL.
- Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2011). lp-norm multiple kernel learning. *J. Mach. Learn. Res.* 12, 953–997.
- Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). “LDICDL: LncRNA-disease association identification based on collaborative deep learning,” in *Proceedings of the IEEE/ACM Trans Comput Biol Bioinform*, Piscataway, NJ: IEEE.
- Lanckriet, G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.* 5, 27–72.
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323
- Liu, B., Du, R., Zhou, L., Xu, J., Chen, S., Chen, J., et al. (2018). miR-200c/141 regulates breast cancer stem cell heterogeneity via Targeting HIPK1/β-Catenin Axis. *Theranostics* 8, 5801–5813. doi: 10.7150/thno.29380
- López-Cortés, A. L., Cabrera-Andrade, A., ázquez-Naya, J. M. V., Pazos, A., and Munteanu, C. R. (2020). Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks. *Entific Rep.* 10:8515.
- Lv, X., He, M., Zhao, Y., Zhang, L., and Wei, M. (2019). Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer. *Cancer Cell Int.* 19:172.
- Määttä, K., Rantaperi, T., Lindström, A., Nykter, M., Kankuri-Tammilehto, M., Laasanen, S. L., et al. (2016). Whole-exome sequencing of Finnish hereditary breast cancer families. *Eur. J. Hum. Genet. Ejhg* 25, 85–93. doi: 10.1038/ejhg.2016.141
- Mary, G., Brian, C., Teresa, S., Melissa, C., Olena, M., Mark, D., et al. (2014). The UCSC cancer genomics browser: update 2015. *Nucleic Acids Res.* 43, D812–D817.
- Mermel, C. H., Schumacher, S. E., Hill, B., and Meyerson, M. L. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, R41–R41.
- Motaghed, M., Al-Hassan, F. M., and Hamid, S. S. (2014). Thymoquinone regulates gene expression levels in the estrogen metabolic and interferon pathways in MCF7 breast cancer cells. *Int. J. Mol. Med.* 33, 8–16. doi: 10.3892/ijmm.2013.1563
- Nguyen, C., Yong, W., and Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Ence Eng.* 06, 551–560. doi: 10.4236/jbise.2013.65070
- Phan, N. N., Huynh, T. T., and Lin, Y. C. (2017). Hyperpolarization-activated cyclic nucleotide-gated gene signatures and poor clinical outcome of cancer patient. *Transl. Cancer Res.* 6, 698–708. doi: 10.21037/tcr.2017.07.22
- Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* 18:9. doi: 10.1186/s12859-016-1423-9
- Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). SimpleMKL. *J. Mach. Learn. Res.* 9, 2491–2521.
- Ronen, J., Hayat, S., and Akalin, A. (2018). Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* 2:e201900517. doi: 10.26508/lsa.201900517
- Samuelson, D. J., Hesselton, S. E., Aperavich, B. A., Zan, Y., Haag, J. D., Trentham-Dietz, A., et al. (2007). Rat Mcs5a is a compound quantitative trait locus with orthologous human loci that associate with breast cancer risk. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6299–6304. doi: 10.1073/pnas.0701687104
- Seoane, J. A., Day, I. N. M., Gaunt, T. R., and Colin, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* 30, 838–845. doi: 10.1093/bioinformatics/btt610
- Sun, D., Li, A., Tang, B., and Wang, M. (2018). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Comput. Methods Progr. Biomed.* 161, 45–53. doi: 10.1016/j.cmpb.2018.04.008
- Wu, L., Zhang, D., Zhou, L., Pei, Y., Zhuang, Y., Cui, W., et al. (2019). FUN14 domain-containing 1 promotes breast cancer proliferation and migration by activating calcium-NFATC1-BMI1 axis. *Ebiomedicine* 41, 384–394. doi: 10.1016/j.ebiom.2019.02.032
- Xu, X., Huang, L., Chan, C. H., Yu, T., Miao, R., and Liu, C. (2016). Assessing the clinical utility of genomic expression data across human cancers. *Oncotarget* 7, 45926–45936. doi: 10.18632/oncotarget.10002
- Xu, X., Zhang, Y., Liang, Z., Wang, M., and Ao, L. (2013). “A gene signature for breast cancer prognosis using support vector machine, biomedical engineering and informatics (BMEI),” in *Proceedings of the 2012 5th International Conference on BioMedical Engineering and Informatics*, Chongqing: IEEE.
- Yan, F., Kittler, J., Mikolajczyk, K., and Tahir, M. A. (2009). “Non-sparse multiple kernel learning for fisher discriminant analysis,” in *Proceedings of the IEEE International Conference on ICDM*, Miami, FL.
- Ye, Z. L., Guan, W. L., Tang, T., Wang, F., and He, C. Y. (2019). Gene mutation profiling in chinese colorectal cancers patients and its association with clinicopathological characteristics and prognosis. *Ssrn Electron. J.* 9, 745–756. doi: 10.1002/cam4.2727
- Yuan, X., Bai, J., Zhang, J., Yang, L., Duan, J., Li, Y., et al. (2020a). CONDEL: detecting copy number variation and genotyping deletion zygosity from single

- tumor samples using sequence data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1141–1153.
- Yuan, X., Gao, M., Bai, J., and Duan, J. (2020b). SVSR: a program to simulate structural variations and generate sequencing reads for multiple platforms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1082–1091. doi: 10.1109/tcbb.2018.2876527
- Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., et al. (2019). “CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data,” in *Proceedings of the IEEE/ACM Trans Comput Biol Bioinform*, Piscataway, NJ: IEEE.
- Yuan, X., Zhang, J., and Yang, L. (2017). IntSIM: an integrated simulator of next-generation sequencing data. *IEEE Trans. Biomed. Eng.* 64, 441–451. doi: 10.1109/tbme.2016.2560939
- Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* 32, 644–652. doi: 10.1038/nbt.2940
- Zhang, A., Li, A., He, J., and Wang, M. (2019). LSCDFS-MKL: a multiple kernel based method for lung squamous cell carcinomas disease-free survival prediction with pathological and genomic data. *J. Biomed. Inform.* 94:103194. doi: 10.1016/j.jbi.2019.103194
- Zhang, Y., Li, A., He, J., Wang, M., and Novel, A. (2019). MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data. *IEEE J. Biomed. Health Inform.* 24, 171–179. doi: 10.1109/jbhi.2019.2898471
- Zhang, Y., Li, A., Peng, C., and Wang, M. (2016). Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 825–835. doi: 10.1109/tcbb.2016.2551745
- Zhang, Y., Yang, W., Dan, L., Yang, J. Y., Guan, R., and Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genomics* 11(Suppl. 5):104. doi: 10.1186/s12920-018-0419-x
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2021 He, Zhang, Yuan and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*