Mathematical Biosciences and Engineering

MBE, 19(12): 12353–12370.
DOI: 10.3934/mbe.2022576
Received: 29 June 2022
Revised: 09 August 2022
Accepted: 14 August 2022
Published: 24 August 2022

http://www.aimspress.com/journal/MBE

Research article

The construction of transcriptional risk scores for breast cancer based on lightGBM and multiple omics data

Jianqiao Pan[1,2], Baoshan Ma[2,*], Xiaoyu Hou[2], Chongyang Li[2], Tong Xiong[2], Yi Gong[2] and Fengju Song[1,*]

[1] Department of Epidemiology and Biostatistics, Key Laboratory of Molecular Cancer Epidemiology, Tianjin, National Clinical Research Center of Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China
[2] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

* **Correspondence:** Email: mabaoshan@dlmu.edu.cn, songfengju@163.com.

Abstract: *Background:* Polygenic risk score (PRS) can evaluate the individual-level genetic risk of breast cancer. However, standalone single nucleotide polymorphisms (SNP) data used for PRS may not provide satisfactory prediction accuracy. Additionally, current PRS models based on linear regression have insufficient power to leverage non-linear effects from thousands of associated SNPs. Here, we proposed a transcriptional risk score (TRS) based on multiple omics data to estimate the risk of breast cancer. *Methods:* The multiple omics data and clinical data of breast invasive carcinoma (BRCA) were collected from the cancer genome atlas (TCGA) and the gene expression omnibus (GEO). First, we developed a novel TRS model for BRCA utilizing single omic data and LightGBM algorithm. Subsequently, we built a combination model of TRS derived from each omic data to further improve the prediction accuracy. Finally, we performed association analysis and prognosis prediction to evaluate the utility of the TRS generated by our method. *Results:* The proposed TRS model achieved better predictive performance than the linear models and other ML methods in single omic dataset. An independent validation dataset also verified the effectiveness of our model. Moreover, the combination of the TRS can efficiently strengthen prediction accuracy. The analysis of prevalence and the associations of the TRS with phenotypes including case-control and cancer stage indicated that the risk of breast cancer increases with the increases of TRS. The survival analysis also suggested that TRS for the cancer stage is an effective prognostic metric of breast cancer patients. *Conclusions:* Our proposed TRS model expanded the current definition of PRS from standalone SNP data to multiple omics data and outperformed the linear models, which may provide a powerful tool for diagnostic and prognostic

prediction of breast cancer.

Keywords: breast cancer; transcriptional risk scores; multiple omics data; lightGBM; diagnosis; prognosis

1. Introduction

Breast cancer is the most frequently diagnosed cancer in women worldwide [1]. In 2020, there were over 2 million new cases reported [2]. Although morbidity and mortality have declined in recent years due to changes in risk factors and improvements in screening tests and treatments, breast cancer survival rates vary widely across countries and regions. The establishment of effective prevention and treatment measures is essential to prevent breast cancer occurrence and reduce breast cancer mortality. Although carriers of BRCA1 and BRCA2 gene mutations confer a high risk of breast cancer, these gene mutations can be found in only a small part of breast cancer patients [3]. In recent years, genome-wide association study (GWAS) identified multiple high frequency and low penetrance susceptibility variants of breast cancer [4]. The accumulation effects of these susceptibility variants can be summarized as a polygenic risk score (PRS). Researchers have developed several PRS models for breast cancer by using a large amount of single nucleotide polymorphisms (SNPs) data [5]. These studies maintained the PRS to be an effective and reliable predictor of breast cancer risk that may provide screening and prevention strategies [6–8].

However, the PRS calculated using SNPs data can only assess the genetic risk of an individual, while ignoring the influence of the external environmental exposure on gene expression. With the development of high-throughput omics technology, a large number of related studies based on genomics and transcriptomics emerged [9,10]. Omics data have been widely used for cancer classification based on identified gene signatures [11], gene pathways [12], and protein-protein interaction networks, etc. [13,14]. For example, Zhang Y et al. [15] proposed a novel approach to predict prognosis in glioblastoma multiforme (GBM) by integrating histopathological images and multi-omics data. Zhang X et al. [16] used XGBoost to upgrade a previously developed cancer-related lncRNA classifier to improve the prediction accuracy of lncRNA-cancer associations, which was expected to contribute to the functional annotation of lncRNAs and guide cancer treatment. Tong D et al. [17] collected the clinical, DNA methylation and miRNA expression data of colon cancer from TCGA and proposed a predictive model based on the integration of clinical data and multi-omics data, and the results showed better predictive outcomes. These high-throughput molecular markers can dynamically reflect the comprehensive effects of genetic background, environmental exposure and lifestyle habits individually [18,19]. The analyses of multiple omics data may lead to new insights into diagnosis and prognosis of breast cancer [20]. In addition, in the standard approach of PRS, the effect sizes of the genetic variants are usually estimated in linear statistical models [21]. However, linear statistical model has some limitations and only be applied when specific requirements are satisfied [22]. Advanced machine learning (ML) models [23,24] such as LightGBM can account for non-linear relationships among large-scale variables and have an increasing trend on the applications for breast cancer research. Using these ML models may further improve the prediction accuracy of breast cancer.

Here, we used multiple omics data and LightGBM model to construct a novel transcriptional risk score (TRS) for breast cancer. The results illustrated that our proposed method outperforms traditional linear models and other ML models and can effectively predict individual risk of breast cancer.

2. Materials and methods

2.1. Data collection

The datasets in this study were downloaded from the cancer genome atlas (TCGA) project. Now, all TCGA data are accessible without limitations in publications or presentations according to the posted announcement from the TCGA website [25]. We collected four kinds of omics datasets on breast invasive carcinoma (BRCA), including DNA methylation data (Illumina Infinium Human DNA Methylation 450 K; Level 3) measured from 782 tumor tissues and 96 normal tissues (Paracancerous tissue), miRNA-seq data (IlluminaHiSeq_miRNASeq; Level 3) measured from 1078 tumor tissues and 104 normal tissues, mRNA expression (Illumina mRNA-seq; Level 3) measured from 1102 tumor tissues and 113 normal tissues, lncRNA expression (Illumina lncRNA-seq; Level 3) measured from 1102 tumor tissues and 113 normal tissues. We also collected the stage of tumor for the BRCA patients, including stages I–IV. According to the literature [26], the annotation of stages I and II were labelled as early-stage, stages III and IV as late-stage. For BRCA patients, most of the individuals are white, and a small number of individuals are black or African American and Asian. The ages of volunteers used in our study range from 26 to 90. Tables 1 and 2 show the sample sizes and clinical data of patients in BRCA datasets, respectively. An independent dataset (GSE66695) from the gene expression omnibus (GEO) measured by the Illumina Infinium 450 k Human DNA methylation Beadchip was used to validate the predictive performance of the proposed method. This dataset includes 80 BRCA tumor tissues and 40 normal tissues, and all volunteers are from Detroit, USA.

Table 1. The description of BRCA datasets from TCGA.

Omic type	Total of early-stage and late-stage tumor samples		Total of tumor samples	Total of normal samples	Total of biological variables
DNA methylation	Early-stage	562	782	96	14,797
	Late-stage	209			
miRNA	Early-stage	790	1078	104	360
	Late-stage	264			
mRNA	Early-stage	800	1102	113	16,499
	Late-stage	267			
lncRNA	Early-stage	800	1102	113	5382
	Late-stage	267			

The total of tumor samples is not equal to the sum of the early-stage and the late-stage samples, because some tumor samples have unknown breast cancer stage.

Table 2. Statistics of clinical data of patients in BRCA datasets from TCGA.

	Clinical data	Sample size	Percentage (%)
Race	White	575	62.8
	Black or African American	183	20.0
	Asian–	96	10.5
	Not available	61	6.7
Age	Average	58	—
	Range	2690	—
Survival status	Alive	947	88.7
	Dead	149	11.3

2.2. Data pre-processing

For DNA methylation, we retained the CpG sites that most negatively correlated with gene expression according to Firehose [27] and removed CpG sites with missing value to ensure the quality of the datasets [28]. For miRNA, mRNA, and lncRNA, two steps were performed to deal with the missing values in the datasets [29]. First, the probes were excluded if there is missing value in more than 20% of samples. Second, all data were normalized by Min-Max scaling to map the range from 0 to 1. For convenience, CpG sites of DNA methylation and probes of miRNA, mRNA and lncRNA are collectively referred to as biological variables. Table 1 also shows the summary of biological variables.

2.3. Construction of transcriptional risk scores

2.3.1. Overview of TRS model

According to the different phenotypes, we proposed to utilize multiple omics data and breast cancer status to construct two kinds of TRS models. The first phenotype only contains the normal samples (control) and tumor samples (case), which were labelled 0 and 1, respectively. The second phenotype contains the normal samples, early-stage and late-stage tumor samples, which were labelled 0, 1 and 2, respectively. We defined the above-mentioned two TRS models as TRS for case-control status and TRS for cancer stage status. The TRS can evaluate the individual risk of breast cancer and may improve the diagnosis of breast cancer. Moreover, since recent studies found the stage of cancer is highly associated with the prognosis [30], accurate construction of TRS for cancer stage status may facilitate the prediction of breast cancer prognosis. The framework of this study is shown in Figure 1. We provided an executable python program, which is available for downloading from the GitHub website (https://github.com/lab319/TRS_BRCA_omics_LightGBM).

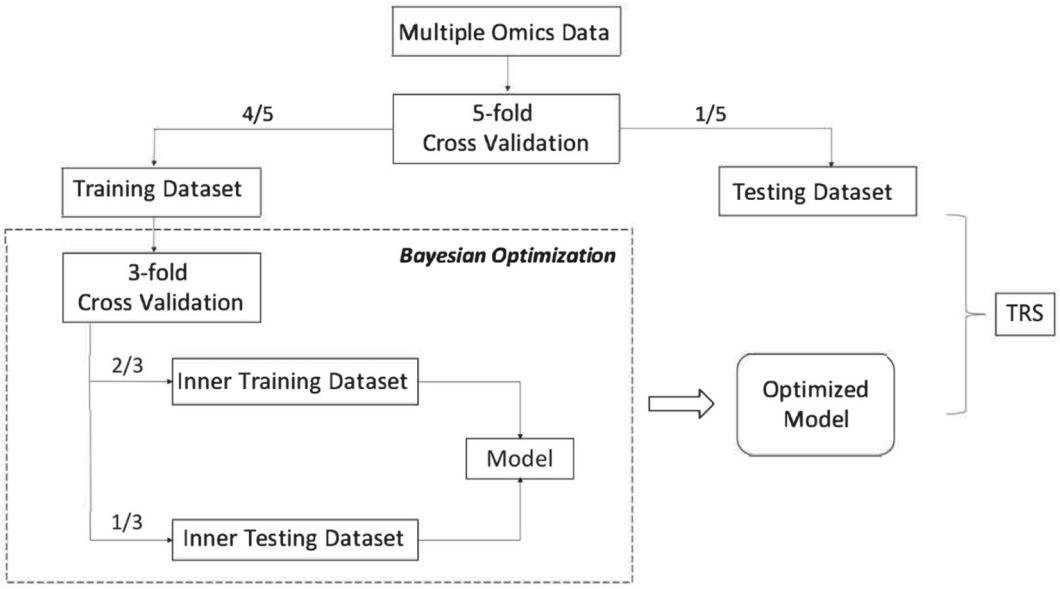

Figure 1. Schematic overview of the framework for constructing TRS model based on multiple omics data.

The dataset of BRCA was split into two groups as training dataset and testing dataset based on five-fold cross validation. We constructed TRS model by using MCP, LASSO, elastic net, SVM and LightGBM based on training dataset. The hyper-parameters of five models were optimized by using bayesian optimization and three-fold cross validation. The TRS of testing dataset was predicted by optimized model. The predictive performance of final models was evaluated with R^2.

2.3.2. TRS based on LightGBM

LightGBM is an ensemble model of classification and regression trees (CART) [31], in which each step generates a basic CART model and adds it to the overall model. The TRS models based on LightGBM were built using a training dataset to predict TRS in a testing dataset. We defined each omic dataset $D_1 = \{(\mathbf{X}_i, y_i)\}(|\, D_1 \,| = n_1, \mathbf{X}_i \in \mathbf{R}^m, y_i \in \mathbf{R})$ as training dataset, where $\mathrm{x_i}$ represents a matrix containing n_1 samples and m biological variables, y_i is the corresponding outcome (phenotype). Let $\hat{y}_i$ be the prediction of $\hat{y}_i$. $D_2 = \{(\mathbf{X}_i^*, y_i^*)\}(|\, D_2 \,| = n_2, \mathbf{X}_i^* \in \mathbf{R}^m, y_i^* \in \mathbf{R})$ was the testing dataset, where X_i^* represents a matrix containing n_2 samples and m biological variables, y_i^* denotes the TRS. We used T additive CART models to predict the TRS in the training dataset.

$$\hat{y}_i = \sum_{t=1}^{T} f_t(\mathbf{X}_i), f_t \in F \quad (1)$$

where $f_t(\mathbf{X}_i)$ corresponds to an independent CART model and F is the space of CART models. To learn the set of CART used in the TRS model, we minimize the following objective function.

$$L(f_t) = \sum_{i=1}^{n_1} l(y_i, \hat{y}_i) + \gamma K + \frac{1}{2}\lambda \sum_{k=1}^{K} w_k^2 \quad (2)$$

here $l(y_i, \hat{y}_i)$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and true phenotype y_i. The K and w_k respectively represent the number and value of leaf nodes in each CART model, γ and λ are constant coefficients. In general setting, the second-order approximation can be utilized to quickly optimize the objective function.

$$L(f_t) = \sum_{k=1}^{K} [(\sum_{i \in I_k} g_i) w_k + \frac{1}{2}(\sum_{i \in I_k} h_i + \lambda) w_k^2] + \gamma K \quad (3)$$

where g_i and h_i are the first and second-order gradient statistics of the loss function. $I_k = \{i \mid q(\mathbf{X}_i) = k\}$ was defined as the instance set of leaf nodes. LightGBM used two techniques including gradient-based one-side sampling and exclusive feature bundling to estimate the information gain in a high speed [23]. The structure and value of each CART model can be determined by the information gain. Thus, we generated the TRS model consisting of T additive CART models. For the samples in a testing dataset, TRS y_i^* can be calculated by applying $\mathbf{x}_i^*$ to the TRS model.

2.3.3. TRS based on linear model and other ML model

To evaluate the predictive performance of LightGBM objectively, we applied the linear model and other ML models to construct TRS. The traditional linear model contains minimax concave penalty (MCP) [32], least absolute shrinkage and selection operator (LASSO)[33] and elastic net [34]. The ML model contains support vector regression (SVR) [35]. Here, we compared the TRS methods that only utilize omics data, without considering the methods that use GWAS summary statistics, such as LDpred [36], Lassosum [37] and so on. Similar to the TRS method based on LightGBM model, we used each omic dataset as the input of these models, and the corresponding phenotypes as the output.

2.3.4. Model training and evaluation

To ensure the robustness and stability of the model, we trained and evaluated the proposed TRS model by five-fold cross validation. The five-fold cross-validation method used in this study is shown in Figure 2. This procedure divided each omic dataset into five subsets. In each fold, one of the five subsets was used as the testing dataset and the other four subsets were put together to form a training dataset. We applied bayesian optimization and 3-fold inner cross validation to optimize the hyper-

parameters of the TRS model in each training dataset. Specifically, for LASSO, we optimized the parameter "alpha". For MCP, we adjusted regularization parameter "lambda". For elastic net, the parameter "alpha" and "l1_ratio" were optimized. For SVR, we choose "rbf kernel" and optimized the regularization parameter "C". For LightGBM, the optimized parameters were "num_leaves", "n_estimators", "learning_rate", "max_depth", "max_bin", "min_split_gain", "subsample", "subsample_freq", "colsample_bytree", "min_child_sample", "min_child_weight", "reg_alpha", "reg_lambda". Finally, we obtained the TRS of each testing dataset which was predicted by the model with the optimized parameters. Each TRS was standardized based on its mean and standard deviation. The predictive performance of TRS model was evaluated by square of the Pearson correlation coefficient (R^2).

$$R^2 = \left(\frac{Cov(Y,\hat{Y})}{\sqrt{Var(Y)Var(\hat{Y})}}\right)^2 \tag{4}$$

where $Cov(Y,\hat{Y})$ represents the covariation of true phenotype and predicted TRS, $Var(Y)$ is the variance of true phenotype, and $Var(\hat{Y})$ is the variance of predicted TRS. In addition, for case-control status, we can also evaluate the predictive performance by the area under the receiver operating characteristic curve (AUC).

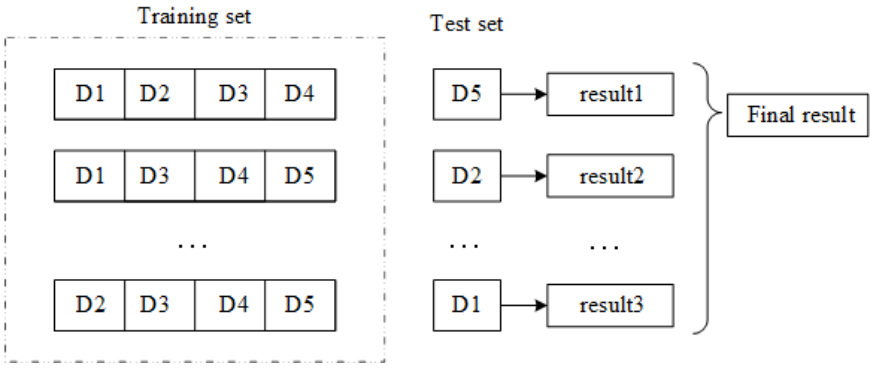

Figure 2. Five-fold cross-validation block diagram.

The procedure divided the dataset into five folds of approximate equal size. Each fold was used as a test set separately, and the other four folds were utilized as the training set. The performance of the prediction algorithm was estimated by averaging the accuracy on five test sets.

2.4. Combination model of TRS

To further improve the predictive performance of TRS, we utilized the TRS based on each omic dataset to construct a new combination model [38,39]. We first matched a common dataset from the four kinds of omics datasets for BRCA. In the common dataset of case-control status, there are 786 tumor samples and 75 normal samples. In the common dataset of cancer stage status, there are 553 early-stage samples, 205 late-stage samples and 75 normal samples. Next, we used the TRS based on four kinds of omics datasets as new feature variables for the combination model. Then, we built the

combination model using the LightGBM model. Bayesian optimization was applied to adjust hyper-parameters and five-fold cross validation was used to evaluate the overall predictive performance. To evaluate the performance of combination model of TRS, we standardized the TRS based on its mean and standard deviation. The framework of the combination model is shown in Figure 3.

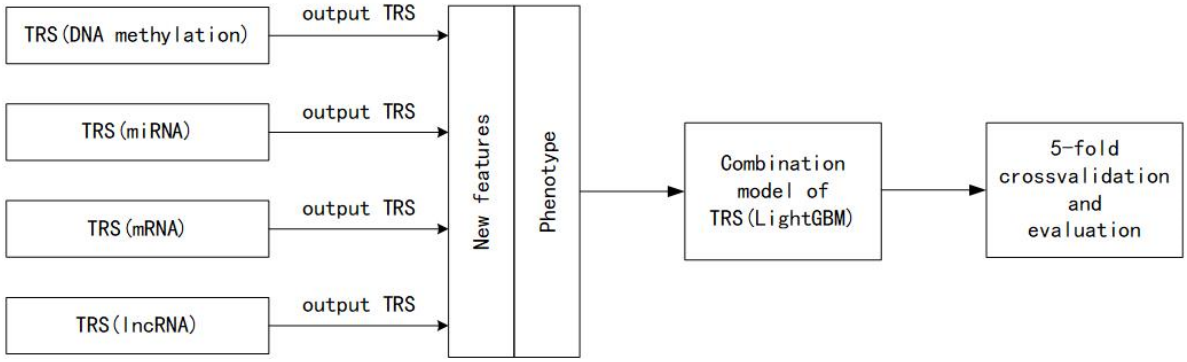

Figure 3. Schematic overview of the framework for constructing combination model of TRS.

We utilized proposed TRS model based on each omic data (DNA methylation, miRNA, mRNA and lncRNA) as new feature variables. The combination model was constructed by LightGBM model. The hyper-parameters of combination model were optimized by using bayesian optimization and 3-fold cross validation. The predictive performance was evaluated by five-fold cross validation.

3. Results

3.1. Predictive performance of TRS based on multiple omics data

We first compared our prediction model to the linear methods and other ML methods for case-control status. Figure 4.a shows the results of these methods on four kinds of omics datasets from TCGA. We observed that elastic net achieves the best performance in traditional linear models. The R^2 of SVR is 3.3, 7.7 and 0.5% higher than elastic net on DNA methylation, miRNA and lncRNA datasets and 3.1% lower than elastic net on mRNA dataset. The R^2 of our proposed model improved by 8.3, 14.8, 5.1 and 7.2% than elastic net on four kinds of omics datasets. Overall, our model outperformed other models and mRNA data exhibited better performance than other omics data.

Next, we applied our proposed model and other linear and ML methods for cancer stage status. Compared with the case-control status, this phenotype contains normal and two stage statuses of breast cancer. Thus, the predictive performance of TRS for cancer stage status is not as good. Nevertheless, the present results are consistent with the TRS for case-control status. According to the comparison results of our proposed model with other methods (Figure 4b), the LightGBM model performs the best predictive performance, outscoring other methods on four kinds of omics datasets from TCGA. The TRS based on LightGBM obtains the R^2 of 0.405, 0.371, 0.437 and 0.407, respectively. Compared with the elastic net with the highest prediction accuracy in the linear models, the R^2 of LightGBM improved by 12.8, 20.9, 9.8 and 12.4%, respectively. Compared with SVR, the R^2 of our proposed model improved by 10.4, 14.4, 10.6 and 3.3%, respectively. Moreover, the results showed mRNA data obtained better results than other omics data. Table 3 shows more detailed results about predictive performance of TRS based on multiple omics data.

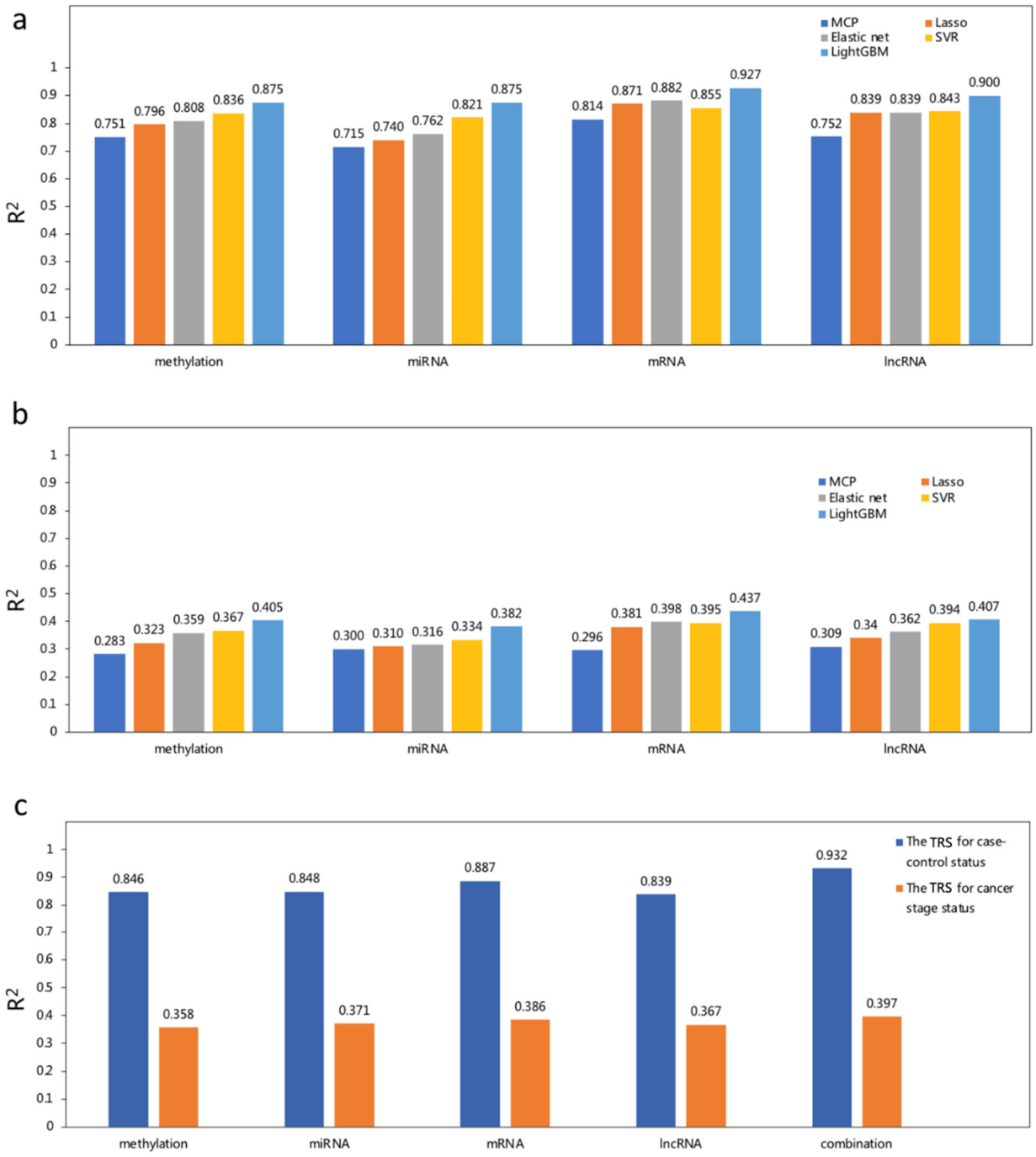

Figure 4. Predictive performance of MCP, LASSO, elastic net, SVR and LightGBM in four kinds of omics datasets. (A) Comparison results of multiple omics datasets for case-control status. (B) Comparison results of multiple omics datasets for cancer stage status. (C) Comparison results of multiple omics datasets and combination model in the common samples for case-control and cancer stage status.

Table 3. Predictive performance of MCP, LASSO, elastic net, SVR and LightGBM in four kinds of omics datasets.

	Omic type	MCP	Lasso	Elastic net	SVR	LightGBM
case-control status	DNA methylation	0.751	0.796	0.808	0.836	**0.875**
	miRNA	0.715	0.740	0.762	0.821	**0.875**
	mRNA	0.814	0.871	0.882	0.855	**0.927**
	lncRNA	0.752	0.839	0.839	0.843	**0.900**
cancer stage status	DNA methylation	0.283	0.323	0.359	0.367	**0.405**
	miRNA	0.300	0.310	0.316	0.334	**0.382**
	mRNA	0.296	0.381	0.398	0.395	**0.437**
	lncRNA	0.309	0.340	0.362	0.394	**0.407**

3.2. Predictive performance of TRS in independent dataset

To further validate the predictive performance of the proposed TRS model, we utilized the GSE66695 as an independent validation dataset. We first applied the LightGBM algorithm to construct the TRS model for case-control status using DNA methylation dataset from TCGA. The hyper-parameters of LightGBM were optimized by bayesian optimization of 3-fold cross validation. Next, we predicted TRS of each sample of GSE66695 dataset and obtained the R^2 of 0.887. Compared to the BRCA dataset from TCGA, although the predictive performance of the independent validation dataset has slightly decreased, the proposed TRS model can still achieve satisfactory results.

3.3. Predictive performance of TRS based on combination model

In this part, we evaluated the performance of combination model of TRS. Figure 4c shows the results of TRS models based on four types of omics datasets and combination model in the common samples. For four kinds of omics datasets from TCGA, although the prediction accuracy in the common samples has decreased, we found that TRS based on mRNA still obtained the best prediction accuracy. For case-control and cancer stage status, the R^2 of combination model were 0.932 and 0.397, respectively. Compared with the TRS model based on mRNA dataset, the R^2 of combination model improved by 5.1 and 2.8%, respectively. Thus, the combination of four types of molecular data can achieve better results of TRS for case-control and cancer stage status. Table 4 shows more detailed results about predictive performance of TRS based on combination model.

Table 4. Comparison results of multiple omics datasets and combination model in the common samples for case-control and cancer stage status.

	DNA methylation	miRNA	mRNA	lncRNA	Combination
case-control status	0.846	0.848	0.887	0.839	0.932
cancer stage status	0.358	0.371	0.386	0.367	**0.397**

3.4. Prevalence of breast cancer

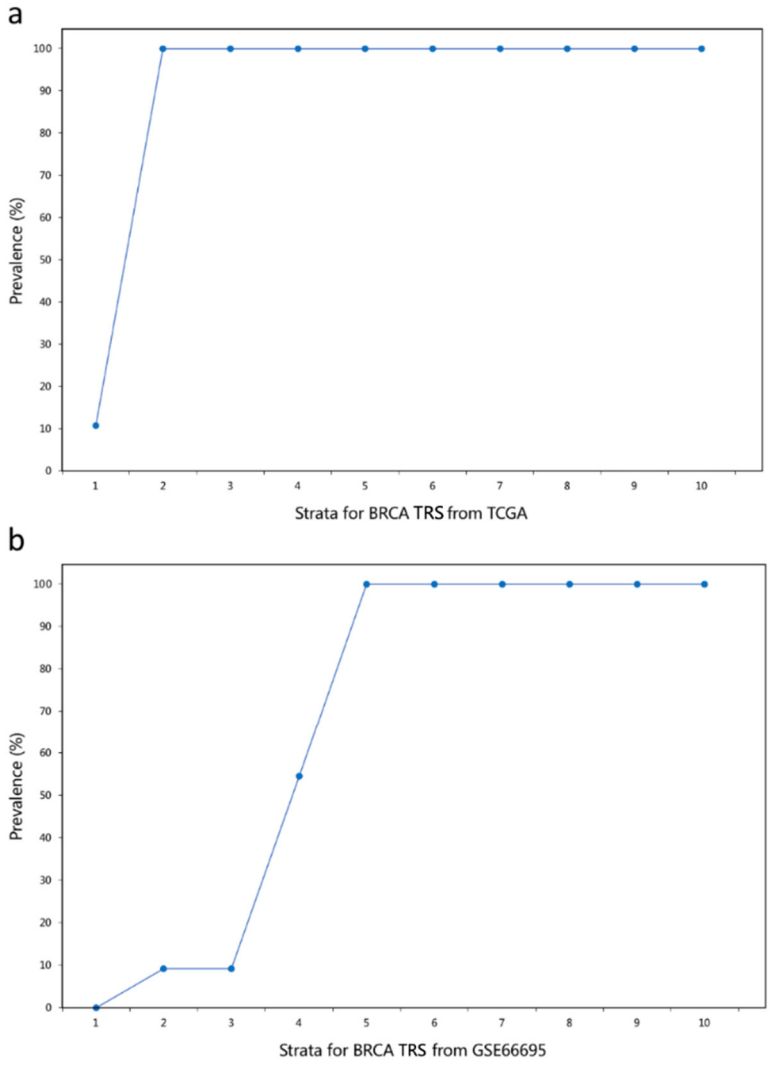

Figure 5. The prevalence curve of TRS for case-control status. (A) The prevalence strata plot of increasing TRS derived by TCGA. The 1st stratum can be regarded as a low-risk TRS stratum and the 2nd to 10th stratum as a high-risk stratum. (B) The prevalence strata plot of increasing TRS derived by GSE66695. The 1st to 3th stratum can be regarded as a low-risk TRS stratum and the 5th to 10th stratum as a high-risk stratum.

Exploring the prevalence of different TRS strata has a positive impact on the prevention and treatment of breast cancer [40]. The main goal of this part is to analyze the risk stratification of case-control status. Thus, we divided the common samples into 10 strata of increasing TRS from the combination model and calculated the prevalence of each stratum (Figure 5a). We defined the prevalence as the proportion of breast cancer patients in each stratum.

Across the common samples, we observed that the prevalence is about 10% in the first stratum then upgrades to 100% in the second stratum and remains steady afterward. The prevalence changes significantly at one stratum because our proposed method achieved relatively accurate prediction of breast cancer risk for case-control status. The trend plot of the prevalence also indicated that the individuals with high-TRS strata have greater breast cancer risk than the individuals with lower-TRS strata. In addition, we built the LightGBM model based on the DNA methylation dataset of TCGA, and then calculated the TRS using the GSE66695 dataset. We subsequently plotted the prevalence curve (Figure 5b) and obtained a similar result with the TCGA dataset.

3.5. Associations between TRS and breast cancer risk

We investigated the relationship of TRS with different phenotypes of breast cancer in this section (Table 5). For case-control status, the association of TRS was evaluated in predicted results from the combination model by logistic regression. The TRS generated by combination model and phenotype (case-control status) of breast cancer are used as the variable (x) and outcome (y) of the logistic regression model, respectively. We observed that TRS was associated with occurrence risk of breast cancer (odds ratio (OR) = 18.48; 95% confidence interval (CI): 9.60-35.55; $P = 2.46 \times 10^{-18}$), suggesting that per one standard deviation increase in TRS is associated with risk increase of breast cancer. In addition, we calculated the OR value by the association of TRS and outcome of breast cancer using the GSE66695 dataset (odds ratio (OR) = 58.19; 95%CI: 12.78-264.90; $P = 1.48 \times 10^{-7}$), and further verified the above conclusion. For cancer stage status, we performed a multinomial logistic regression model to evaluate the association of TRS and set the normal sample as the reference group. The TRS was associated with early-stage breast cancer risk (OR = 21.05; 95%CI: 10.26-43.19; $P = 9.63 \times 10^{-17}$) and late-stage breast cancer risk (OR = 46.62; 95%CI: 19.72-110.25; $P = 2.14 \times 10^{-18}$). The results indicated higher TRS is associated with a significantly increased risk for early-stage and late-stage breast cancer.

Table 5. Associations between TRS and breast cancer risk.

Phenotype		OR	95%CI	P-value
The TRS for case-control status (TCGA)		18.48	9.60–35.55	2.46×10^{-18}
The TRS for case-control status (GSE66695)		58.19	12.78–264.90	1.48×10^{-7}
The TRS for cancer stage status (TCGA)	Early-stage	21.05	10.26–43.19	9.63×10^{-17}
	Late-stage	46.62	19.72–110.25	2.14×10^{-18}

3.6. Prognosis prediction of breast cancer

We explored whether the TRS for cancer stage status can effectively assess the prognosis of patients. According to the predicted results of tumor samples using combination model, we firstly divided 758 patients of breast cancer into high-risk and low-risk groups based on the 50th percentile of TRS. Next, we utilized the survival time and the status at the end of their survival time for each patient to generate Kaplan-Meier curves (KM curve). We observed that high-risk patients had statistically significantly worse prognosis (Figure 6). The results showed the TRS for cancer stage may provide an effective prognostic tool of breast cancer patients.

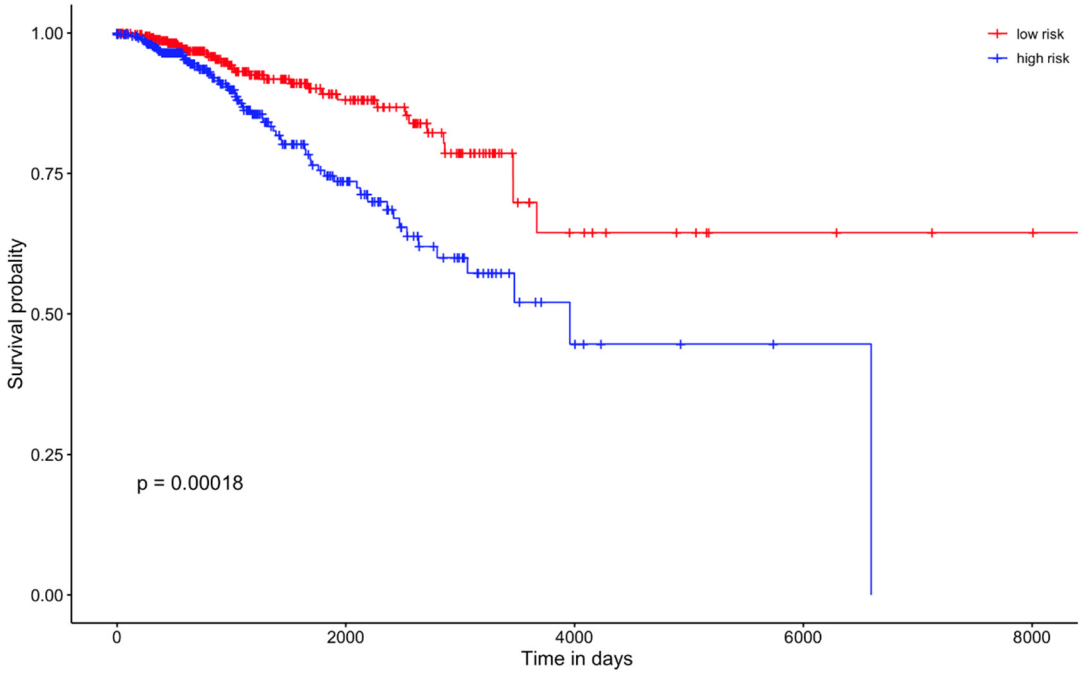

Figure 6. The KM survival curve of BRCA patients in the high-risk and low-risk groups.

We divided patients into high-risk and low-risk groups based on the 50th TRS. The patients with low-risk group have better prognosis than those with high-risk group.

4. Discussion

In our study, first of all, we have developed a novel TRS method for breast cancer using multiple omics data and LightGBM model. For case-control and cancer stage status, we showed that the proposed method had better prediction performance than linear models and other ML models using multiple omics data. Meanwhile, the prediction results of five-fold cross validation demonstrated the robustness and reliability of our proposed method. Second, the combination of TRS further improved the predictive performance for breast cancer. Finally, by analyzing the trend of prevalence and associations between TRS and breast cancer risk, the results bolstered the clinical understanding and application for breast cancer TRS. In addition, we also found that our TRS models for cancer stage status can improve the prognosis prediction of breast cancer patients.

Most of the previous PRS studies focused on the analysis of individual-level genotype data (SNPs)

using linear models. For example, Mavaddat et al. [15] utilized PRS derived from 313 SNPs in 69 studies of the Breast Cancer Association Consortium (BCAC) to predict the breast cancer risk and the AUC was 0.63. Khera et al. [14] derived a PRS based on 5218 SNPs in the UK Biobank and the AUC was 0.68. Although these studies obtain individual-level genetic risk of breast cancer, the current prediction accuracy still maintains at low level. In the independent validation dataset (GSE66695), our TRS model obtained the AUC of 0.98. Thus, the TRS based on multiple omics data and LightGBM not only improved the risk of predicting breast cancer, but also expanded the current definition of PRS from SNP data to genomics and transcriptomics data. In addition, some studies used gene expression data and clinical data to establish prognostic models to assess disease risk. For example, wang et al. [41] established an immune-related prognostic score in 22 breast cancer cohorts with a total of 6415 samples. Yang et al. [42] undertook a study of tumor infiltrating lymphocytes in a large group of ovarian cancer patients and found that high expression levels of the immune-related genes were associated with good prognosis in high-grade serous carcinomas. Distinct with these studies our investigation utilized the LightGBM algorithm and multi-omics data to build a transcriptional risk score (TRS) model and estimate the risk of breast cancer.

Our proposed method has the following advantages. First, the LightGBM model we used exploits gradient boosted trees to fits all biological variables simultaneously, especially high-dimensional data such as multiple omics data [23]. In addition, the LightGBM model takes advantage of ensemble learning, which helps to minimize the main causes of error in ML model such as noise, bias and variance than a single model [43]. Second, it is not easy to obtain SNP data because of ethical and legal constraints. With the development of high-throughput omics technology, related studies accumulate vast amounts of genomic and transcriptomic data which can be downloaded from many public databases. Third, as representative of genomics data, DNA methylation can be modulated by physiological and environmental exposures and provide biomarkers for diagnosis and prognosis for cancer [44–46]. Transcriptomics data including miRNA, mRNA, and lncRNA reveals the transcription and regulation mechanism of large-scale genes, which play an important role in determining the mechanism and treatment of cancer [47,48]. Compared to the individual-level genotype data, using multiple omics data to construct breast cancer TRS considered the interaction of genetic and environmental factors, and thus can provide higher prediction accuracy.

Although our TRS methods provide good predictive performance, they have some limitations. First, the LightGBM model has more hyper-parameters than traditional linear models such as MCP, LASSO and elastic net. Thus, we need more time to train the proposed model. We applied multithreading technology to effectively utilize computing resources and correspondingly reduced some running time. Second, the sample size of breast cancer from TCGA is relatively small compared to large-scale genome-wide association studies data. In addition, there are significantly more tumor samples than normal samples in our study. Imbalanced datasets significantly compromise the performance of most standard learning algorithms, because these models assume the balanced class distributions. Third, the results we obtained on the independent validation set are lower than the results on the TCGA breast cancer dataset. The reasons are as the following, the two datasets are collected from different studies. Besides, the sample size of the independent validation (GEO) set is much smaller than that of TCGA. In the future, one of our tasks is to collect more omics data and clinical information from public datasets (TCGA, GEO, METABRIC, et al.), more cancer datasets are needed to improve and validate the proposed TRS model for breast cancer.

5. Conclusions

We proposed a novel TRS model for two kinds of breast cancer phenotypes by using multiple omics data and LightGBM. The results demonstrated that our model improved the prediction accuracy of current PRS methods indeed and may provide an effective diagnosis and prognosis tool for breast cancer.

Acknowledgements

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All TCGA data are accessible without limitations in publications or presentations according to the posted announcement from the TCGA website.

Conflict of interest

The authors have no conflicts of interest to declare.

References

1. K. L. Britt, J. Cuzick, K. Phillips, Key steps for effective breast cancer prevention, *Nat. Rev. Cancer*, **20** (2020), 417–436. https://doi.org/10.1038/s41568-020-0266-x
2. C. Wild, E. Weiderpass, B. Stewart, World cancer report: cancer research for cancer prevention, *Lyon: Int. Agency Res. Cancer*, **1** (2020), 23–33. https://www.paho.org/en/node/69005
3. D. Thompson, D. Easton, The genetic epidemiology of breast cancer genes, *J. Mammary Gland Biol. Neoplasia*, **9** (2004), 221–236. https://doi.org/10.1023/B:JOMG.0000048770.90334.3b
4. L. Wu, W. Shi, J. Long, X. Guo, K. Michailidou, J. Beesley, et al., A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer, *Nat. Genet.*, **50** (2018), 968–978. https://doi.org/10.1038/s41588-018-0132-x
5. P. Maas, M. Barrdahl, A. D. Joshi, P. L. Auer, M. M. Gaudet, R. L. Milne, et al., Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States, *JAMA Oncol.*, **2** (2016), 1295–1302. https://doi.org/10.1001/jamaoncol.2016.1025
6. N. Mavaddat, P. D. Pharoah, K. Michailidou, J. Tyrer, M. N. Brook, M. K. Bolla, et al., Prediction of breast cancer risk based on profiling with common genetic variants, *J. Nat. Cancer Inst.*, **107** (2015), djv036. https://doi.org/10.1093/jnci/djv036
7. A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, et al., Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, *Nat. Genet.*, **50** (2018), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z
8. N. Mavaddat, K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, et al., Polygenic risk scores for prediction of breast cancer and breast cancer subtypes, *Am. J. Hum. Genet.*, **104** (2019), 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002
9. Y. Dor, H. Cedar, Principles of DNA methylation and their implications for biology and medicine, *Lancet*, **392** (2018), 777–786. https://doi.org/10.1016/S0140-6736(18)31268-6

10. R. Lowe, N. Shirley, M. Bleackley, S. Dolan, T. Shafee, Transcriptomics technologies, *PLoS Comput. Biol.*, **13** (2017), e1005457. https://doi.org/10.1371/journal.pcbi.1005457
11. Y. C. Chen, Y. C. Chang, W. C. Ke, H. W. Chiu, Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer, *J. Biomed. Inf.*, **56** (2015), 1–7. https://doi.org/10.1016/j.jbi.2015.05.006
12. H. Jin, H. C. Lee, S. S. Park, Y. S. Jeong, S. Y. Kim, Serum cancer biomarker discovery through analysis of gene expression data sets across multiple tumor and normal tissues, *J. Biomed. Inf.*, **44** (2011), 1076–85. https://doi.org/10.1016/j.jbi.2011.08.010
13. L. P. Zhao, H. Bolouri, Object-oriented regression for building predictive models with high dimensional omics data from translational studies, *J. Biomed. Inf.*, **60** (2016), 431–445. https://doi.org/10.1016/j.jbi.2016.03.001
14. S. Joe, H. Nam, Prognostic factor analysis for breast cancer using gene expression profiles, *BMC Med. Inf. Decis. Making*, **16** (2016), 56. https://doi.org/10.1186/s12911-016-0292-5
15. Y. Zhang, A. Li, J. He, M. Wang, A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data, *IEEE J. Biomed. Health. Inf.*, **24** (2020), 171–179. https://doi.org/10.1109/JBHI.2019.2898471
16. X. Zhang, T. Li, J. Wang, J. Li, L. Chen, C. Liu, Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy, *Front. Genet.*, **10** (2019), 735. https://doi.org/10.3389/fgene.2019.00735
17. D. Tong, Y. Tian, T. Zhou, Q. Ye, J. Li, K. Ding, et al., Improving prediction performance of colon cancer prognosis based on the integration of clinical and multi-omics data, *BMC Med. Inf. Decis. Making*, **20** (2020), 22. https://doi.org/10.1186/s12911-020-1043-1
18. J. A. Alegría-Torres, A. Baccarelli, V. Bollati, Epigenetics and lifestyle, *Epigenomics*, **3** (2011), 267–277. https://doi.org/10.2217/epi.11.22
19. C. P. Wild, The exposome: from concept to utility, *Int. J. Epidemiol.*, **41** (2012), 24–32. https://doi.org/10.1093/ije/dyr236
20. Y. V. Sun, Y. J. Hu, Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases, *Adv. Genet.*, **93** (2016), 147–190. https://doi.org/10.1016/bs.adgen.2015.11.004
21. S. W. Choi, T. S. Mak, P. F. O'Reilly, Tutorial: a guide to performing polygenic risk score analyses, *Nat. Protoc.*, **15** (2020), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1
22. J. Erenpreisa, A. Giuliani, Resolution of complex issues in genome regulation and cancer requires non-linear and network-based thermodynamics, *Int. J. Mol. Sci.*, **21** (2019), 240. https://doi.org/10.3390/ijms21010240
23. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 3146–3154. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/
24. E. Zhu, F. Jiang, C. Liu, J. Xu, Partition independent set and reduction-based approach for partition coloring problem, *IEEE Trans. Cybern.*, **52** (2022), 4960–4969. https://doi.org/10.1109/TCYB.2020.3025819
25. K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemp. Oncol.*, **19** (2015), A68–77. https://doi.org/10.5114/wo.2014.47136

26. A. Rahimi, M. Gönen, Discriminating early-and late-stage cancers using multiple kernel learning on gene sets, *Bioinformatics*, **34** (2018), i412–i421. https://doi.org/10.1093/bioinformatics/bty239
27. Y. Yuan, E. M. V. Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, et al., Assessing the clinical utility of cancer genomic and proteomic data across tumor types, *Nat. Biotechnol.*, **32** (2014), 644–652. https://doi.org/10.1038/nbt.2940
28. B. Liu, Y. Liu, X. Pan, M. Li, S. Yang, S. C. Li, DNA methylation markers for pan-cancer prediction by deep learning, *Genes*, **10** (2019) 778. https://doi.org/10.3390/genes10100778
29. B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, F. Song, Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data, *Comput. Biol. Med.*, **121** (2020), 103761. https://doi.org/10.1016/j.compbiomed.2020.103761
30. A. Weiss, M. Chavez-MacGregor, D. Y. Lichtensztajn, M. Yi, A. Tadros, G. N. Hortobagyi, et al., Validation study of the American joint committee on cancer eighth edition prognostic stage compared with the anatomic stage in breast cancer, *JAMA Oncol.*, **4** (2018), 203–209. https://doi.org/10.1001/jamaoncol.2017.4298
31. G. De'ath, K. E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, *Ecology*, **81** (2000), 3178–3192. https://doi.org/10.2307/177409
32. J. Liu, K. Wang, S. Ma, J. Huang, Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method, *Stat. Interface*, **6** (2013), 99–115. https://doi.org/10.4310/SII.2013.v6.n1.a10
33. R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, *J. R. Stat. Soc.: Ser. B*, **73** (2011), 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x
34. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc.: Ser. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
35. A. J. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.*, **14** (2004), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
36. B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, et al., Modeling linkage disequilibrium increases accuracy of polygenic risk scores, *Am. J. Hum. Genet.*, **97** (2015), 576–592. https://doi.org/ 10.1016/j.ajhg.2015.09.001
37. T. S. Mak, R. M. Porsch, S. W. Choi, X. Zhou, P. C. Sham, Polygenic scores via penalized regression on summary statistics, *Genet. Epidemiol.*, **41** (2017), 469–480. https://doi.org/10.1002/gepi.22050
38. A. Alves, Stacking machine learning classifiers to identify Higgs bosons at the LHC, *J. Instrum.*, **12** (2017), T05005. https://doi.org/10.1088/1748-0221/12/05/T05005
39. B. Pavlyshenko, Using stacking approaches for machine learning models, in *2018 IEEE Second International Conference on Data Stream Mining & Processing*, (2018), 255–258, https://doi.org/10.1109/DSMP.2018.8478522
40. J. J. Barendregt, S. A. Doi, Y. Y. Lee, R. E. Norman, T. Vos, Meta-analysis of prevalence, *J. Epidemiol. Commun. Health*, **67** (2013), 974–978. https://doi.org/10.1136/jech-2013-203104
41. S. Wang, Q. Zhang, C. Yu, Y. Cao, Y. Zuo, L. Yang, Immune cell infiltration-based signature for prognosis and immunogenomic analysis in breast cancer, *Briefings Bioinf.*, **22** (2021), 2020–2031. https://doi.org/10.1093/bib/bbaa026
42. L. Yang, S. Wang, Q. Zhang, Y. Pan, Y. Lv, X. Chen, et al., Clinical significance of the immune microenvironment in ovarian cancer patients, *Mol. Omics*, **14** (2018), 341–351. https://doi.org/10.1039/c8mo00128f

43. C. Zhang, Y. Ma, Ensemble machine learning || ensemble learning, *Chapter*, **1** (2012), 1–34. https://doi.org/10.1007/978-1-4419-9326-7.
44. Y. Pan, G. Liu, F. Zhou, B. Su, Y. Li, DNA methylation profiles in cancer diagnosis and therapeutics, *Clin. Exp. Med.*, **18** (2018), 1–14. https://doi.org/10.1007/s10238-017-0467-0
45. T. Hou, H. Chang, H. Jiang, P. Wang, N. Li, Y. Song, et al., Smartphone based microfluidic lab-on-chip device for real-time detection, counting and sizing of living algae, *Measurement*, **187** (2022), 0263–2241. https://doi.org/10.1016/j.measurement.2021.110304
46. Y. Cheng, C. He, M. Wang, X. Ma, F. Mo, S. Yang, et al., Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials, *Signal Transduction Targeted Ther.*, **4** (2019), 62. https://doi.org/10.1038/s41392-019-0095-0
47. J. Fan, K. Slowikowski, F. Zhang, Single-cell transcriptomics in cancer: computational challenges and opportunities, *Exp. Mol. Med.*, **52** (2020), 1452–1465. https://doi.org/10.1038/s12276-020-0422-0
48. J, Rodon, J. C. Soria, R. Berger, W. H. Miller, E. Rubin, A. Kugel, et al., Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial, *Nat. Med.*, **25** (2019), 751–758. https://doi.org/10.1038/s41591-019-0424-4

AIMS Press