IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2021, accepted January 21, 2021, date of publication January 26, 2021, date of current version February 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054823

Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall

SHUAI LIU[1,2], HAN LI[3], QICHEN ZHENG[1,2], LU YANG[1,2], MEIYU DUAN[1,2], XIN FENG[4,5], FEI LI[1,2], LAN HUANG[1,2], AND FENGFENG ZHOU[1,2], (Senior Member, IEEE)

[1]Health Informatics Laboratory, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun 130012, China
[3]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[4]Department of Epidemiology and Biostatistics, School of Public Health, Jilin University, Changchun 130012, China
[5]Jilin Institute of Chemical Technology, Jilin 132022, China

Corresponding author: Fengfeng Zhou (fengfengzhou@gmail.com)

This work was supported in part by the Jilin Provincial Key Laboratory of Big Data Intelligent Computing under Grant 20180622002JC, in part by the Education Department of Jilin Province under Grant JJKH20180145KJ, in part by the Startup Grant of the Jilin University, in part by the Bioknow MedAI Institute under Grant BMCPP-2018-001, in part by the High Performance Computing Center of Jilin University, and in part by the Fundamental Research Funds for the Central Universities, JLU.

ABSTRACT Breast cancer is one of main causes of death for women. Most of the existing survival analyses focus on the features' associations with whether the patients may survive five years or not. The personalized question remains largely unresolved about how long a breast cancer patient will live. This study aims to predict the patient-specific survival time of breast cancer patients. It formulates the personalized question into two machine learning problems. The first problem is the binary classification of whether a patient will live longer than five years or not. The second one is to build a regression model to predict the patient's survival time within five years. The methylome of a breast cancer patient is used for the prediction. A new algorithm Crystall is presented to find the methylomic features for this regression model. Our models perform well in the above two problems, and achieve the mean absolute error (MAE) of about 1 month for predicting how long a breast cancer patient will live within five years. The detected biomarker genes demonstrate close connections with breast cancers.

INDEX TERMS Methylation, breast cancer, lifespan, prediction, feature selection.

I. INTRODUCTION

Breast cancer is a major cancer type for females and was ranked top two in both new cases and deaths among all the cancer types in major countries [1]–[3]. The rapid innovation and development of the modern high-throughput technologies facilitated the precision diagnosis of cancer types and personalized risk estimations [4]–[7]. And cancer patients become more concerned about the remaining life span and the life quality [8]–[10].

Most of the existing studies investigated the survival or relapse risks of a patient after a specific length of time, e.g., 1 or 5 years. Kolben, *et al.*, investigated the guideline-based clinical practice of uPA/PAI-1 treatment on the early breast cancers and observed that the five-year relapse-free survival in the intermediate-risk patients (N0, G2) achieved 99% even without chemotherapy [11]. Kaplan, *et al.*, also demonstrated that the adjuvant chemotherapy achieved 98% in the five-year relapse-free survival rate in the HR+/her2- group, while only 89% for the triple negative breast cancer patients [12]. Some acute cancer subtypes may need to investigate the one-year survival rate and Geerse, *et al.*, demonstrated that distress had a major impact on the one-year survival rate of lung cancer patients [13].

Quite a few multi-gene panels were clinically or commercially available for the diagnosis or prognosis estimations of breast cancers [14]. A panel of 50 differentially-expressed genes (PAM50) was constructed to define four heterogeneous subtypes of breast cancers [15]. PAM50 was widely used to estimate the recurrence risk and treatment prognosis of hormonal therapy and chemotherapy [15]. The OncoType DX genomic test was used in the clinical practice

The associate editor coordinating the review of this manuscript and approving it for publication was Bin Liu.

to guide the treatment decisions for invasive breast cancer patients [16], [17]. The robust clinical prediction results also suggested that the OncoType DX has the potential to be involved in the clinical routine practice for breast cancer patients [18]. EndoPredict is a multi-gene panel to estimate the distance recurrence (DR) risk and the prognosis of adjuvant chemotherapy for the female early-stage breast cancer patients with estrogen receptor positive (ERp) and human epidermal growth factor receptor 2 (HER2) negative statuses [19], [20]. The Breast Cancer Index is an RT-PCR-based assay on the FFPE tissues to predict the distance recurrence risk for the ER-positive breast cancer patients [14].

The OMIC technologies generate a much larger number of features than that of the samples in a biomedical modeling study. This could cause the model overfitting problem, and the feature dimension has to be decreased. A feature selection algorithm may be utilized to detect the phenotype-associated features by optimizing the specific optimization goal. There were two main groups of feature selection algorithms, filters and wrappers [21], [22]. A filter ranked the features with a specific metrics, e.g. Pvalue for T-test (Ttest) [23]. While a wrapper screened for a feature subset using a heuristic rule and returned the feature subset with the best optimization goal [24], [25]. A wrapper usually performed slower than a filter, but achieved a much better prediction accuracy.

The main contribution of this study was to investigate the survival problem from a new perspective. We tried to answer the question of how long a specific patient would live after the diagnosis, instead of the surviving percentage on a time point of a cohort after the diagnosis in the conventional survival analysis. A series of proof-of-principle experiments were carried out to demonstrate that the novel problem setting was solvable by the regular machine learning approaches.

II. MATERIALS AND METHODS

A. SUMMARY AND PREPROCESSING OF THE DATASET

The methylomic data was retrieved from the breast cancer project of The Cancer Genome Atlas (TCGA). There were 928 samples with methylomes in the project TCGA-BRCA. Each sample was probed for 485,577 methylation features using the Illumina 450k BeadChip [26, 27]. The clinical data of each sample was extracted from the TCGA consortium data portal [28].

Firstly, this study investigated the survival time of the breast cancer patients. So the normal samples and control group were removed from the dataset. Then in order to remove those reference or stably methylated features, the top-ranked 100,000 features with the largest variances were kept for further analysis. Thirdly, one patient may have multiple samples in the project TCGA-BRCA, and one sample from each patient was randomly chosen.

After these preprocessing, only 221 samples with the clinical survival time data were obtained and each sample has 100,000 methylated features. There were 71 samples who died within five years, and were denoted as the negative samples. The other 150 patients were denoted as the positive samples. 103 out of the 221 samples have known deceasing dates but 32 patients lived longer than five years.

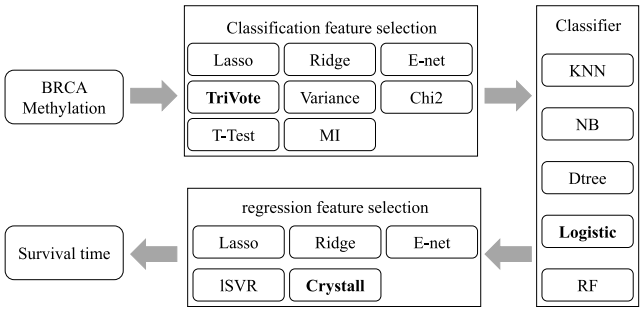


FIGURE 1. Experimental design of this study. The 221 methylomes were the input data to the experimental procedure.

B. PERFORMANCE EVALUATION METRICS

This study formulated the survival time prediction as a binary classification problem and a regression problem. The binary classification problem was to discriminate the breast cancer patients who lived longer than five years (positive samples) from those who didn't (negative samples). The classification performance was evaluated by the metrics accuracy (Acc), sensitivity (Sn) and specificity (Sp). Sn and Sp were defined as the percentages of correctly predicted positive and negative samples, respectively. And the overall accuracy Acc was the percentage of all the correctly predicted samples. The metrics balanced accuracy bAcc was defined as (Sn+Sp)/2, considering the dataset was not strictly balanced, as similar in [29]–[32].

The regression problem was to predict how long a negative sample survived after the diagnosis. The regression performance was evaluated by the mean absolute error (MAE), as similar in [33], [34]. MAE was defined as $\left(\sum_{i=1}^{n} |y(i) - y'(i)|\right)/n$, where y(i) and y'(i) were the real and predicted survival times, and n was the number of samples.

All the evaluations were carried out by the 10-fold cross validation strategy. A larger classification Acc or bAcc suggested a better model. And a regression model with a smaller MAE was better than that with a large MAE. Considering the "large p small n" paradigm in the Omic studies [35]–[37], the principle of parsimony (Ocam's razor) preferred a simpler model [38]. So a prediction model was better than that with a similar performance metrics and a larger number of features.

C. EXPERIMENTAL PROCEDURE

The experimental procedure in this study was illustrated in Figure 1. Firstly, we evaluated various classification-based feature selection algorithms on the binary classification problem of whether a patient lived longer than five years or not. The status of five-year survival was widely investigated in the conventional survival analysis studies [39], [40], and a patient was generally considered as being "cancer free" if this patient lived for five years and longer [41]–[44].

Each feature selection algorithm was applied separately on the dataset and the subset of chosen features was used to build the classification model using one of the five classifiers with the 10-fold cross validation strategy. These classification models were evaluated for their classification performances and usually the best model was delivered as the final model.

Then the survival time of a patient being predicted to die within five years was estimated using a regression model. All the five regression algorithms were applied on the regression dataset and the built regression models were evaluated for their regression performances. The best regression model was delivered as the final regression model.

D. BINARY CLASSIFICATION OF FIVE-YEAR SURVIVAL

Feature selection was necessary to reduce the data dimensions of five-year-survival prediction. Even after the preprocessing step, a methylome still had 100,000 features for each sample.

Five binary classifiers were utilized to build the binary prediction model of whether a patient lived longer than five years or not. The classifier k-nearest neighbor (KNN) assigned the query sample to the majority class label of the k nearest neighbors [45]–[47]. Naïve Bayes (NB) assumed the inter-feature independence and calculated the probability of each class label that the query sample belonged to [47]–[49]. Decision tree was a simple supervised learned and made the decisions based on the decisions made on the internal nodes [50]–[52]. Logistic regressor (LR) was a regression-based binary classifier [53]–[55]. And the classifier random forest (RF) assembled the decisions of multiple decision trees and provided the final integrated prediction [56], [57]. We compared the performances of these five classifiers on a subset of features. The classification performance was calculated using the 10-fold cross validation strategy.

Eight feature selection algorithms were utilized. The regression-based feature weighting algorithms L1 penalization (Lasso) [58], [59], L2 penalization (Ridge) [60] and Elastic Net (E-net) [61], TriVote [62] were evaluated. Lasso [63], Ridge [60] and E-net [64] were widely used to select features by regularization. This study removed the features with small absolute values of the model coefficients, by assuming that they carried weak contributions to the class labels. TriVote was a three-step feature selection algorithm. Firstly, TriVote used a linear SVM model to select a feature subset. Then, an SVM-RFE model was trained to select a subset of the remaining features. Finally, TriVote used the linear SVM classifier again to evaluate a subset of feature from the second step.

This study also evaluated some feature ranking metrics, including T-test [65], mutual information (MI) [66], chi-squared test (Chi2) [67], Variance [68]. The statistical significance Pvalues of T-test [69] and Chi2 [70] were used to measure the importance of each feature. A smaller p-value represented a better feature. We ranked the features by Pvalues and selected the top-ranked k features as the final feature subset. The two algorithms MI and Variance ranked the features by the calculated values of MI and Variance. The top-ranked k features with the best classification performance were delivered as the final feature subset. We compared the classification performances of the eight feature selection algorithms using a user-specified classifier and choose the best feature selection algorithm to reduce the dimension of the features.

E. REGRESSION OF THE SURVIVAL TIME

Four existing regression-based feature selection algorithms were utilized to find the methylomic features associated with the survival time. The L1- and L2-penalized regressors Lasso [71] and Ridge [72] were used. The elastic net (E-net) [73], [74] and linear support vector regressor (lSVR) [75], [76] were also widely used to select Omic features. These algorithms evaluated the phenotype associations by the feature importance scores, and a larger feature importance score was assumed as a better phenotype association.

We ranked the features by the absolute values of these features' coefficients in the above-mentioned regression algorithms, and selected the top-ranked k features with the best regression performance. After the step of feature selection, the corresponding regressor was used to predict how long the patient would live for. For example, if we used Lasso to select features, we would also use Lasso to predict the patient's survival time.

F. COMPARISON WITH THE EXISTING GENE PANELS FOR BREAST CANCERS

Breast cancer is one of the most investigated cancer types and various gene panels have been scientifically released or even commercially available for the purposes of diagnosis and prognosis.

PAM50 is a breast cancer profiling assay to estimate how probable a breast cancer may metastasize based on the expression levels of 50 genes [77]. This gene is also widely used to subtype breast cancers [78], [79].

EndoPredict offers a risk score of recurring as distant metastasis for early-stage breast cancer patients with Estrogen-Receptor (ER) positive and HER2 negative [80]. Activities of 12 genes are profiled in breast cancer cells to calculate the risk score and the threshold 3.3287 is optimized to discriminate the patients with higher than 10% risks [81].

The RT-PCR-based assay Breast Cancer Index evaluates the activities of seven genes to calculate the risk that the non-negative, hormone-receptor positive breast cancers may recur 5 to 10 years later [82]. This assay may be applied to the FFPE (Formalin-Fixed Paraffin-Embedded) tissue samples and its risk score is based on the ratio between two genes HOXB13:IL17BR and the five-gene (BUB1B, CENPA, NEK2, RACGAP1, RRM2) molecular grade index [83].

Oncotype DX is a widely-used prognostic assay for the ER-positive breast cancers [84], [85]. This assay utilizes the expression patterns of 16 cancer-associated genes and 5 house-keeping genes to evaluate the possibilities of a breast cancer to grow and to respond to the chemotherapeutic treatments. This assay demonstrates

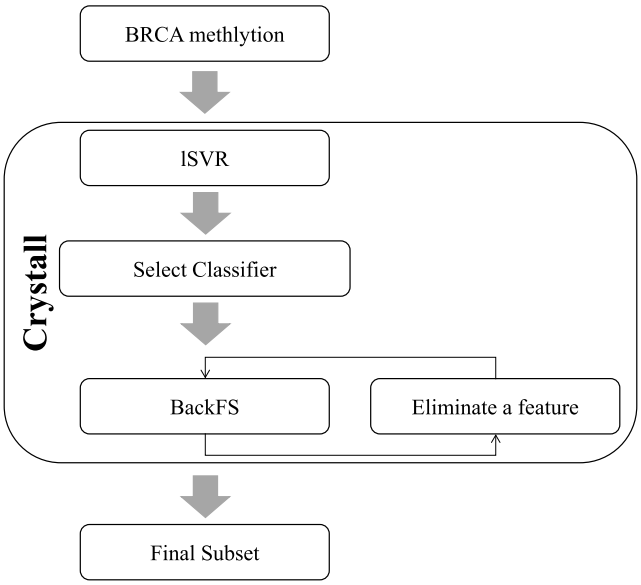


FIGURE 2. Illustration of the proposed algorithm crystall. Firstly, a regressor lSVR is used to filter the features with little associations with the regression label. Then a "select classifier" strategy is used to refine the feature subset. The last step uses the backFS framework to find the features with the smallest mean absolute error (MAE).

a sex disparity and may need refining for male patients [86].

The biomarker genes in the above gene panels were used to build the survival time regression models of a breast cancer patient, for the comparison with the proposed method Crystall.

III. CRYSTALL, A FEATURE SELECTION ALGORITHM TO ESTIMATE THE CANCER SURVIVAL TIME

Survival time prediction is an important question for a patient with a lethal disease. This is different from the conventional survival analysis which tries to calculate the percentage of alive patients within a cohort on a time point [39], [40]. Nie *et al.* utilized the 3-dimensional convolution neural network (CNN) to extract features for an SVM model and predicted with an accuracy 89.9% whether a patient had a long or short overall survival time [87]. Lsik *et al.* employed a random walk-based algorithm to predict whether a patient had a long- or short-term survival by integrating transcriptome, proteome and protein-protein interaction data [88]. The proposed method achieved the accuracies between 66% and 78% for three cancer types. Chato *et al.* proposed a wavelet transform-based denoising method to improve the prediction of short/mid/long-term survival for the brain tumor patients using the MRI images [89].

This study investigated the survival time prediction problem, that most existing studies didn't provide quantitative solutions. A regression model was formulated to estimate how long a patient lived within five years, as described in the above section. The number of methylomic features was much larger than that of the samples, and the feature dimension could be reduced using feature selection algorithms. This study proposed a three-step feature selection algorithm Crystall to optimize this regression model. The mean absolute error (MAE) was used as the optimization goal function.

Firstly, a large number of features were removed by their coefficients in the trained linear support vector regression (lSVR) model. Each methylome consisted of nearly half a million methylation features. It's highly time-consuming to train such a model using all the features, and the model may also be easily overfitted. The absolute value of a feature's coefficient in a lSVR model reflected the contribution of this feature to the model [90]. So only those features with the largest absolute values of their lSVR coefficients were kept for further analysis. This step efficiently removed a large number of features with minor contributions to the regression problem.

The second step of Crystall evaluated the remaining features using a linear regression model and removed the features with small model coefficients. This step assumed that features important to the overall regression problem should have large coefficients in this model, too. The functional module SelectFromModel() in the Python package scikit-learn version 0.21.3 was used to implement the "Select Classifier" strategy by evaluating the features' correlation with the survival time.

The last step of Crystall further refined the subset of selected features using the BackFS strategy [91]. The first two steps of Crystall eliminated features with small absolute values of the model coefficients, instead of the regression performance metric MAE. BackFS carried out a brute-force screening of features that may potentially increase the regression performance. In summary, a feature was iteratively removed, if its removal generated the smallest mean absolute error (MAE).

The proposed feature selection framework Crystall was theoretically time-efficient than the conventional single-step algorithms, while still achieved satisfying regression performances. This was based on the assumption that the majority of the features important to the regression problem should be selected by multiple feature selection algorithms. And this assumption was evaluated in the experiments in the following sections.

IV. RESULTS AND DISCUSION

A. MAIN AIM OF THIS STUDY

This study aimed to predict the life duration of a breast cancer patient. A cancer patient was considered as "cancer free", if she or he lived longer than 5 years. So this study firstly predicted a binary classification problem of whether a breast cancer patient lived at least 5 years or not. The life duration of a breast cancer patient who died within 5 years was then formulated as a regression problem. Because the number of methylomic features was much higher than that of samples in a clinical cohort, feature selection algorithm had to be used to find a subset of features with the best predictive performances. Different feature selection algorithms were

originally designed for different types of data, and they may be used with cautions about the pre-assumptions.

The optimization goals of this study were to find two subsets of features with the best accuracy and minimum mean absolute error for the binary classification problem and the regression problem, respectively.

The proposed feature selection algorithm Crystall tried to recommend a subset of features with the minimum mean absolute error for the regression problem.

B. BINARY CLASSIFICATION OF 5-YEAR SURVIVAL

This study focused on the feature selection algorithm Crystall for the survival time regression problem of breast cancer patients who died within five years. Firstly, we utilized a recently published feature selection algorithm TriVote [62] to investigate the binary classification problem of whether a patient survived 5 years or not. The previous study demonstrated that TriVote achieved very good and stable classification performances with its chosen OMIC features on both transcriptomic and methylomic datasets [62]. The Python package TriVote calculated the 10-fold cross validation classification performances of four representative binary classifiers, i.e., Nearest Neighbor (NN) [92], Support Vector Machine (SVM) [93], [94], Naive Bayes (NBayes) [49], [95], and Decision Tree (DTree) [96].

Both prediction accuracy and feature number are important for an OMIC-based prediction panel [37, 98]. The classifier Logistic Regression achieved the prediction accuracy Acc=1.0000 with the minimum number 40 of features, as shown in Figure 3. And both Sn and Sp reached 1.0000. Figure 3 demonstrated that some classifiers may achieve 1.0000 for Sn or Sp separately. The metrics bAcc=(Sn+Sp)/2 was a good performance metrics for a dataset with imbalanced inter-class samples [31], [98]. Due to the imbalance in our dataset, the metrics geometric mean (G-mean) and area under the ROC curve (AUC) were also used to measure the model performance. So the Logistic Regression model with the 40 features was chosen for the binary classification model for prediction whether a breast cancer patient may survive 5 years or not.

C. COMPARISON WITH THE OTHER FEATURE SELECTION ALGORITHMS

A performance comparison was carried out between TriVote and the seven other feature selection algorithms, as shown in Figure 4. The group of three regression-based feature selection algorithms Lasso, Ridge and Elastic Net (E-net) were utilized as wrappers of the Python functional module SelectFromModel() with the default parameters. This study also evaluated the second group of four filter feature selection algorithms, i.e., T-test (Ttest), Variance-based ranking (Var), Chi-squared Test (Chi2) and Mutual Information (MI). The features were selected by the Incremental Feature Selection strategy (IFS) for a filter [37]. The classification performance of a given feature subset is calculated by the 10-fold cross validation strategy.

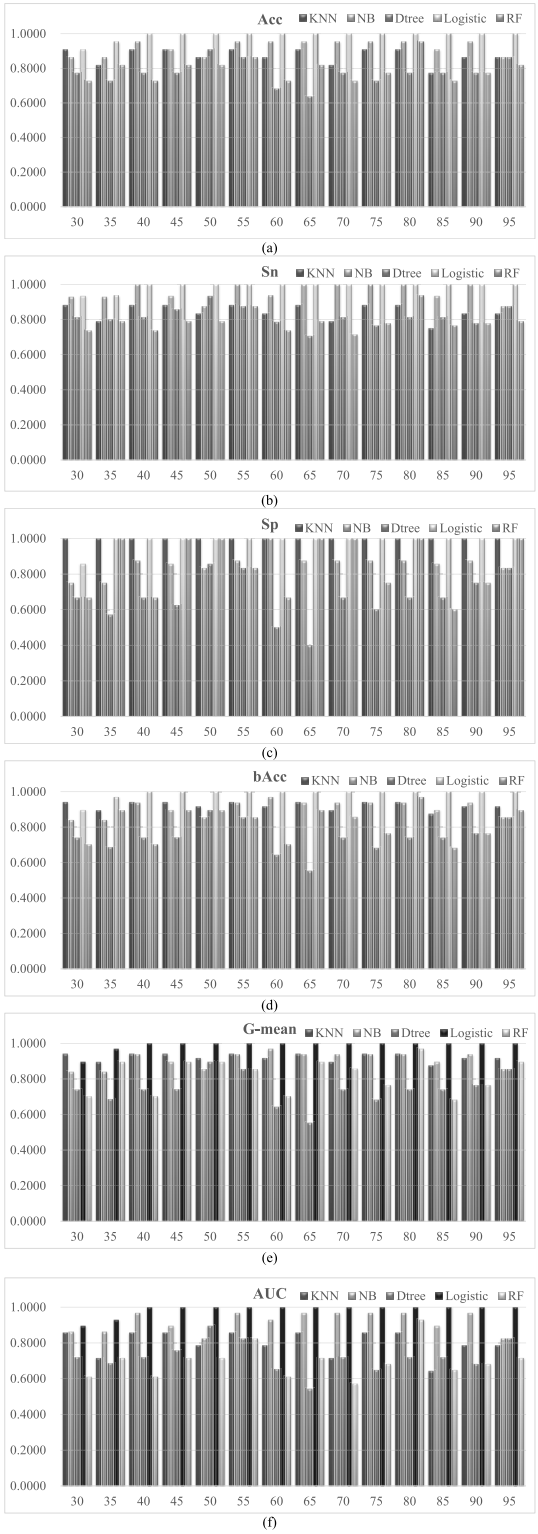


FIGURE 3. Binary prediction performances of the 5-year survival. The horizontal axis gave the number of features selected by the feature selection algorithm TriVote. The vertical axis gave the values of the binary prediction metrics (a) Acc, (b) Sn, (c) Sp and (d) bAcc (e) G-mean (f) AUC for the five binary classifiers provided in the Python package TriVote.

Filter feature selection algorithms didn't perform well on this regression problem, as shown in Figure 4. Ttest, Var and

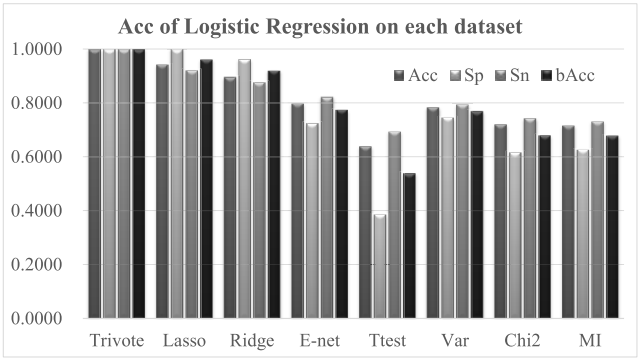


FIGURE 4. Comparison between TriVote and the 7 other feature selection algorithms. The horizontal axis gave the algorithms used to select features. And the horizontal axis gave the feature selection algorithms and the vertical axis gave the value of Logistic regression prediction accuracy (Acc) of the feature subset selected by the algorithm.

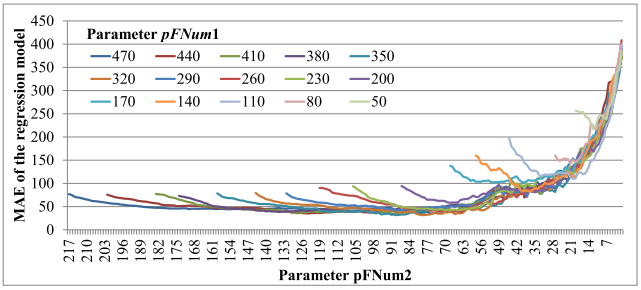


FIGURE 5. Parameter optimization of Crystall. The horizontal axis is the parameter pFNum2, which is the number of remaining features in each iteration of feature removal in the third step of Crystall. The vertical axis is the optimization goal MAE of Crystall. Each line represents a different value choice of the parameter pFNum1, which is the number of top-ranked features selected in the first step of Crystall.

MI may be applied on the numerical variables, and Chi2 was usually applied on nominal or category variables. Although the methylomic features were numerical, Chi2 seemed to have outperformed Ttest on selecting the numerical features with an improvement 0.0814 in Acc. The literature also showed that Chi2 achieved similar performances in evaluating the numerical features as Ttest in both OMIC and imaging data [99]–[102].

Figure 4 illustrated that the group of three regression-based feature selection algorithms generally performed better than the filter-based feature selection algorithms. At least an improvement of 0.0136 was achieved in Acc by the three regression-based feature selection algorithms, i.e., Lasso, Ridge, and E-net. And no existing feature selection algorithms outperformed the algorithm TriVote on the binary classification problem of the dataset TCGA-BRCA [103]–[105].

D. OPTIMIZATION OF THE PARAMETERS OF CRYSTALL

The two parameters pFNum1 and pFNum2 of Crystall are evaluated for the best choices of their values, as shown in Figure 5. The parameters pFNum1 and pFNum2 are the numbers of the top-ranked features in the first step and the number of remaining features after the iterations of feature removal in the third step of Crystall.

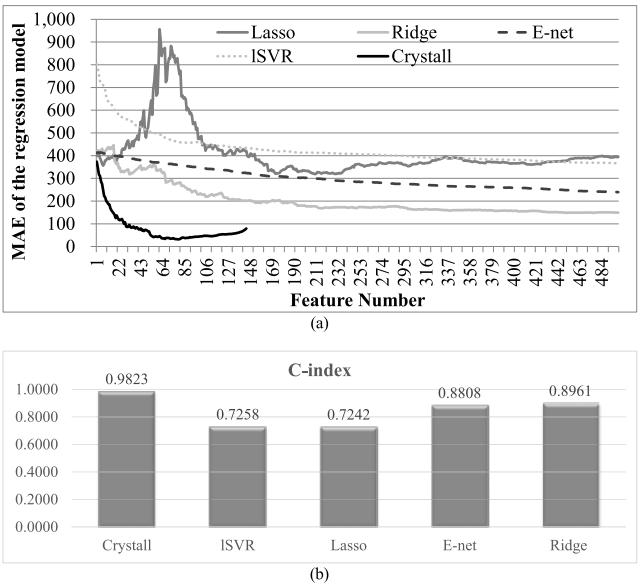


FIGURE 6. Regression performance comparison of Crystall and the other four regression-based feature selection algorithms. (a) Comparison of the regression performance metric MAE of these five feature selection algorithms. The horizontal axis is the number of features. The vertical axis gives the performance metric MAE. (b) Comparison of the regression performance metric C-index of the five feature selection algorithms. The horizontal axis lists the algorithms, and vertical axis is the value of C-index.

The overall trend for all the values of the parameter pFNum1 was not linearly correlated with the number of remaining features (pFNum2). A good model tends to have a small MAE. The data suggested that pFNum1$\leq$200 achieved a much worser MAE than the larger values of pFNum1. And the overall best MAE=31.62 was achieved by pFNum1 =320 in the first step of Crystall. In this case, The second step of Crystall recommended 144 features, among which 79 features were finally selected by Crystall.

E. REGRESSION PERFORMANCE COMPARISON OF FEATURE SELECTION ALGORITHMS

Two regression performance metrics are utilized to evaluate the proposed algorithm Crystall and the other four existing feature selection algorithms. The performance metric MAE is defined in the above section. And the metric Harrell's C-index is the generalized version of the area under the ROC curve (AUC) [106], and has been widely used to evaluate the regression models [107], [108].

The regression performance of Crystall is compared with the four existing regression algorithms in the incremental feature selection (IFS) framework [109]–[111], as shown in Figure 6. Lasso and Ridge regressors are the L1- and L2-regularizations [112], respectively. Elastic net (E-net) is the weighted combination of both L1 and L2 regressions [113]. The linear support vector regressor (lSVR) is another popular regressor for the biomedical prediction problems [114], [115]. The features selected by each algorithm are used to train a regression model by the same regression algorithm.

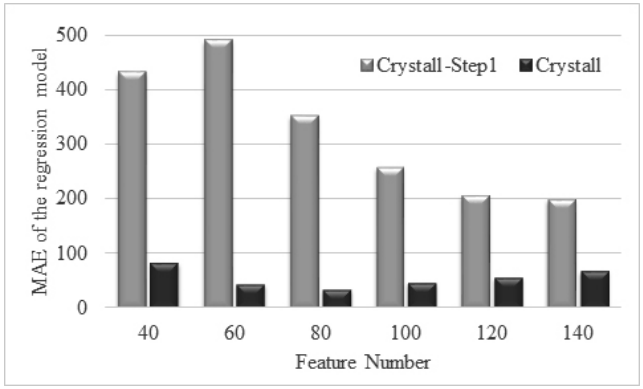


FIGURE 7. The regression performances of Crystall and its first step. The performance metric MAE is used as the vertical axis, and the horizontal axis lists the crystall and its first step (crystall-step1).

Crystall selected 79 features for the final lSVR model and achieved MAE=31.62 and C-index=0.9823, as shown in Figure 6. All the four existing algorithms achieved MAE>140, which are much worse than Crystall, as shown in Figure 6 (a). Figure 6 (b) illustrates that Crystall achieved C-index=0.9823, outperforming the other four algorithms by at least 0.0862 in C-index. So Crystall performs better than these four popular regression-based feature selection algorithms.

F. COMPARE CRYSTALL WITH ITS FIRST STEP

Crystall consists of three consecutive steps of feature selections, and this section evaluates how the first step of Crystall (denoted as Crystall-Step1) performs, as shown in Figure 7. The regression performance metric MAE is used to compare the features selected by Crystall-Step1 and Crystall. The previous studies demonstrated that the module BackFS performs very well on various feature selection problems, but its time complexity is very high [91]. So a wrapper is integrated as the second step of Crystall. Figure 7 illustrates that Crystall-Step1 achieves much larger values of MAE when being compared on the same number of features with Crystall. Crystall achieves the smallest MAE=32.6700 for 80 features, while Crystall-Step1 achieves the smallest MAE=198.7219 for 140 features.

G. EVALUATION OF THE EXISTING GENE PANELS

Four popular gene panels for breast cancers are evaluated for their prediction capabilities of the patients' survival time, as shown in Figure 8. The methylomic features corresponding to the genes in each gene panel are collected and the BackFS strategy is applied to refine the features of each gene panel.

As discussed in the above sections, Crystall delivered a regression model with 79 methylomic features and MAE=31.62. Figure 8 (a) illustrates a similar pattern for all the four gene panels. When the features are removed one-by-one by their weights in the trained lSVR, the regression metric MAE gets a slight improvement (decreasing) and then is increased significantly. This supports the observation of Crystall in Figure 5. Figure 8 (b) shows that the four existing

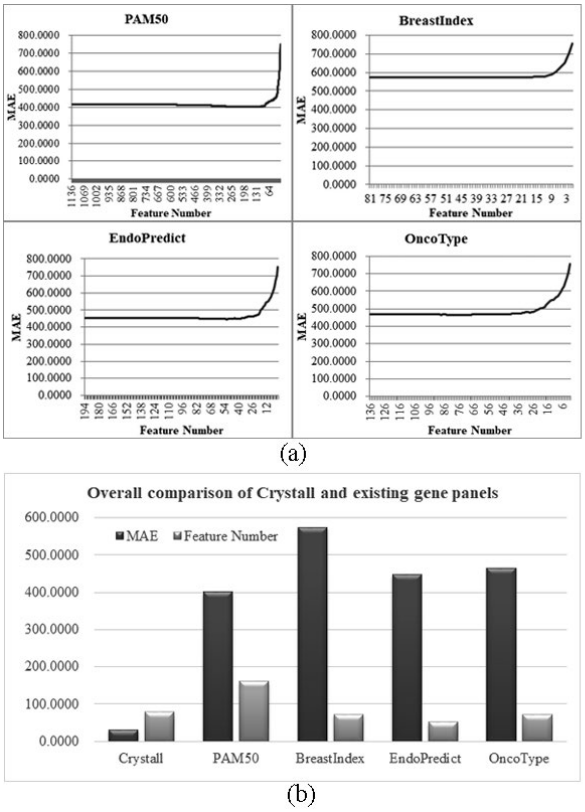


FIGURE 8. Prognostic evaluation of the Crystall model and the four existing gene panels. (a) The BackFS curves of the four existing gene panels. The horizontal axis is the number of features selected by BackFS on that gene panel. The vertical axis is the regression performance metric MAE. (b) Comparison of the best model optimized from crystall and the four existing gene panels. The horizontal axis lists the algorithms. The vertical axis is the value of the two metrics MAE and number of features (Feature Number).

gene panels didn't perform very well on estimating the survival time of each breast cancer patient, even after removing the redundant features.

H. BIOMARKERS FOR PREDICTING WHETHER A PATIENT SURVIVE 5 YEARS

This study uses 40 methylomic features to predict whether a breast cancer patient survivals 5 years after the diagnosis. These 40 features are evaluated for their individual differential expressions using the popular statistical method Ttest, as shown in Table 1 and Figure 9. These 40 features are selected by the algorithm TriVote among the 10,000 features filtered by variance. Some features are even ranked as 48,799 (cg17099656, Pvalue=3.53e-1) and 98,508 (cg26647197, Pvalue=9.79e-1) by Ttest.

So we carried out a recursive feature eliminating strategy on these 40 features, as shown in Figure 9. Figure 9 (a) illustrates that the feature with the largest Pvalue is eliminated in each iteration. We may see that the removal of even the feature cg26647197 (Ttest rank=98,508, Pvalue=9.79e-1) reduces the prediction model's accuracy to 0.9276. If each iteration removes the feature with the smallest Pvalue, the first removal

TABLE 1. Summary of the 40 biomarkers for predicting the 5-year survival of a patient. Columns "Feature" and "Gene" are the methylomic feature name and the corresponding gene symbol of each feature. Columns "Rank" and "Pvalue" are ranks and statistical Pvalue by Ttest.

Feature	Gene	Rank	Pvalue	Feature	Gene	Rank	Pvalue
cg26647197	PCDHGA4	98508	9.79E-01	cg09933929	DPP6	6265	1.44E-02
cg17099656	CERK	48797	3.53E-01	cg00733288	CARS2	5833	1.29E-02
cg15590007	ALOX5	36211	2.25E-01	cg12016809	C21orf56	5157	1.04E-02
cg07812289	MGRN1	33482	1.99E-01	cg19297232	SMPD3	4215	7.55E-03
cg26097573	PRNT	26120	1.35E-01	cg24852779	ALG13	3032	4.64E-03
cg17143179	PRDM1	23077	1.11E-01	cg08787332	SDK1	2707	3.93E-03
cg01423393	KIAA1875	21546	1.00E-01	cg06850283	FGF10	2359	3.17E-03
cg13791254	FOXE1	21002	9.60E-02	cg14391586	SOX18	1372	1.40E-03
cg14791502	SDK1	20175	9.02E-02	cg05845178	IL17RE	1053	9.19E-04
cg01636599	FGF22	18622	7.93E-02	cg09331011	GNAL	978	8.13E-04
rs264581	-	16642	6.66E-02	cg24141198	NRAS	964	7.98E-04
cg04064054	PIGQ	13570	4.85E-02	cg02047547	ITPRIPL2	666	4.55E-04
cg05391892	PLEKHM3	13550	4.85E-02	cg17016000	RIN2	333	1.67E-04
cg15948785	PTPRN2	11074	3.52E-02	cg04677410	MAML2	324	1.65E-04
cg11294241	RIMBP2	10766	3.37E-02	cg00776960	IGDCC4	310	1.58E-04
cg24937727	RGL3	8518	2.34E-02	cg06912252	C9orf125	236	1.04E-04
cg15295732	LOC100128977	8511	2.34E-02	cg24030449	FGF20	193	7.48E-05
cg21746532	SNRPN	7700	2.00E-02	cg12574296	NCCRP1	20	2.23E-06
cg06574335	LYPD6B	6871	1.66E-02	cg25117092	MED12L	4	2.06E-07
cg20238128	MIR548H4	6302	1.45E-02	cg06318796	ARID5B	2	2.74E-08

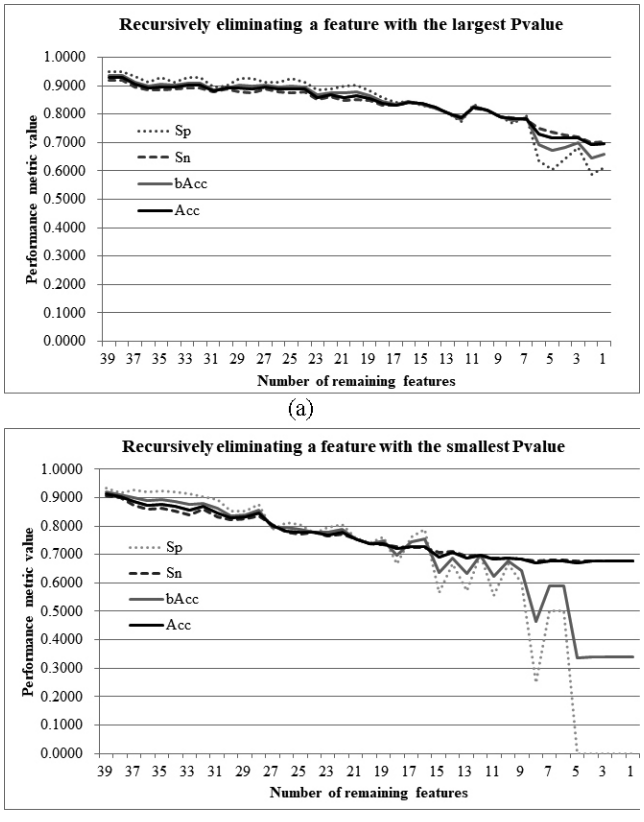


FIGURE 9. Performance evaluation by removing features from the subset of biomarkers for predicting the 5-year survival of breast cancer patients. (a) Recursively removing a feature with the largest Pvalue. (b) Recursively removing a feature with the smallest Pvalue. The horizontal axis is the number of remaining features. And the vertical axis is the value of the four classification performance metrics, i.e., Sp, Sn, bAcc and Acc.

reduces the model's Acc to 0.9140, as shown in Figure 9 (b). So every one of the 40 features has its essential contribution to the prediction model of whether a breast cancer patient may survive five years after the diagnosis.

TABLE 2. The 79 methylomic biomarkers and their annotations for the regression model of the breast cancer patients' survival time. The methylomic biomarkers are in the column "Feature". Columns "Chr" and "Position" are where each methylation residue locates. Column "Gene" gives the gene symbol covering this methylomic feature.

Feature	Chr	Position	Gene	Weight	Feature	Chr	Position	Gene	Weight
cg26228266	1	35351340	DLGAP3	-119.6904	cg15701794	10	77156421		-91.2512
cg26030770	1	36912487	OSCP1	-85.1677	cg10862468	10	135342218	CYP2E1	78.5831
cg06121193	1	90282411		-129.2678	cg11694519	11	111783563	CRYAB	64.9162
cg22817352	1	204653629	LRRN2	-115.0575	cg10156499	11	112161291		-125.5192
cg02985240	1	235468148	ARID4B	49.7529	cg07629355	11	133825519	IGSF9B	133.9032
cg14156405	1	241520286	RGS7	162.2215	cg03621974	12	7650334	CD163	82.1135
cg25819275	1	248684766	OR2G6	-61.0403	cg20402783	12	54381013	HOXC10	144.7154
cg27425262	2	113953981	PSD4	-112.1335	cg04740679	13	114016140	GRTP1	47.0245
cg13832669	2	119609429		85.2176	cg02210934	14	1948874		138.2427
cg12863169	2	119769323		80.6743	cg14479910	14	36975263	SFTA3	-32.1937
cg11467141	2	132348705		-153.0857	cg00846021	14	36977901	SFTA3	-168.4044
cg08079908	2	176997277		112.3647	cg21495612	14	96342250		-132.0937
cg17664269	3	3080621	CNTN4	-74.1399	cg24034005	14	97059192		78.6282
cg19974283	3	149510376	C3orf16	46.5195	cg26875073	15	25200490	SNRPN	-139.5795
cg06573459	3	153840654	SGEF	81.5758	cg01837657	15	90039822	RHCG	-110.5620
cg18096251	5	2205553		22.7030	cg00061551	16	5121151	ALG1	147.3037
cg26333652	5	2750758	IRX2	51.8462	cg08271366	16	82816457	CDH13	8.2453
cg16304215	5	76928810	OTP	-62.2688	cg21405799	16	84002843	NECAB2	-175.9651
cg08806496	5	92918943	NR2F1	-65.6768	cg17344932	17	38183730	MED24	-92.0297
cg08395122	5	140800983	PCDHGA4	-70.0637	cg26893861	17	41843967	DUSP3	102.1793
cg05646373	6	26044460	HIST1H2BB	-154.9283	cg18454685	17	48639239	CACNA1G	-186.9488
cg05404698	6	28956247		109.1777	cg19466818	17	56409534	MIR142	-90.1682
cg05383619	6	32553920	HLA	120.3421	cg26700919	18	13375474	C18orf1	-110.9664
cg27260772	6	50791202	TFAP2B	-121.4315	cg14985989	19	496428	MADCAM1	36.7137
cg19980771	6	110798022	SLC22A16	82.3068	cg18517266	19	5827811	NRTN	-31.5527
cg13972706	6	144671686	UTRN	37.5955	cg12019614	19	11353996	DOCK6	-135.3977
cg03322161	6	161188322		134.0808	cg07379574	19	18724173	TMEM59L	121.1527
cg27339550	7	6654880	ZNF853	-115.6800	cg01464835	19	30016147	VSTM2B	123.1618
cg01411921	7	25893858		163.4330	cg23489630	19	44645078	ZNF234	98.9173
cg19236675	7	76624761	PMS2L11	-78.3238	cg03217253	19	53758055	ZNF677	-5.9389
cg07026433	7	129781152		65.6828	cg23685712	20	52790063	CYP24A1	30.8698
cg10570241	7	158785291		-76.3645	cg02143877	20	52790141	CYP24A1	172.2749
cg01540522	8	1948874	KBTBD11	136.1052	cg19551589	20	62579792	UCKL1	-142.0310
cg23279503	8	35649016	UNC5D	114.2117	cg12667511	21	31744537	KRTAP13-2	153.2367
cg03911745	8	49486641		115.0507	rs845016	NA	0	NA	138.7995
cg27247689	8	73445394		113.5659	rs9839873	NA	0		73.9940
cg08508337	8	144660607	NAPRT1	-133.2965	cg25832925	X	25021346		52.3666
cg00816387	10	1975591		51.1634	cg06302025	X	102000758	BHLHB9	92.7961
cg00456343	10	49348564		-86.5080	cg14931215	Y	21867702	KDM5D	79.3636
cg02249577	10	52434778		-48.3049					

More than half (21) of the 40 biomarker genes are known to be associated with breast cancers. The methylomic biomarker cg24141198 locates in the 3' UTR (untranslated region) of the protein-coding gene N-Ras, which is annotated be an oncogene in many cancer types, including melanoma and thyroid cancer, etc [116]. N-Ras is also observed to be prognostically associated with breast cancers [118, 119]. Another biomarker cg06850283 is within the promoter proximal region TSS1500 of the protein-coding gene Fibroblast Growth Factor 10 (FGF10). The antisense RNA molecule 1 of FGF10 (abbreviated as FGF10-AS1) is a long non-coding RNA, which is associated with the prognosis of triple-negative breast cancer (TNBC) patients [119].

I. BIOMARKERS FOR REGRESSING THE SURVIVAL TIME OF A BREAST CANCER PATIENT

The regression model in this study is the linearly weighted sum of the above 79 features, as shown in Table 2. A feature with a larger absolute value of its weight contributes more than a feature with a smaller value. The top three biomarkers are cg18454685 (Calcium Voltage-Gated Channel Subunit Alpha1 G, abbreviated as CACNA1G), cg21405799 (N-Terminal EF-Hand Calcium Binding Protein 2, abbreviated as NECAB2) and cg02143877 (Cytochrome P450 Family 24 Subfamily A Member 1, abbreviated as CYP24A1), according to the GeneCards annotations [120].

The two biomarkers CACNA1G and CYP24A1 are involved in the prognosis of breast cancers according to the

literature [121]. The protein CACNA1G and the other family members of the voltage- gated calcium channels (VGCCs) tend to be lowly expressed in many types of cancers, including lung and breast cancers [122]. And CACNA1G is associated with the prognosis of bladder cancer and colorectal cancer [124, 125]. The protein CYP24A1 is involved in the breast cancer cell proliferation through interacting with the cellular apoptosis susceptibility protein (CAS) [125]. The repressed expression of CYP24A1 is involved in the prognosis of breast cancers [126].

The secondly-ranked biomarker NECAB2 is not associated with either breast cancer and prognosis in the literature database PubMed [121]. NECAB2 modulates the expressions and functionalities of various cell surface receptors [128, 129]. Based on the biomarker CACNA1G discussed in the above, the hypothesis may be worth of experimental confirmation that NECAB2 impacts the prognosis of breast cancers through interacting with the system of the voltage-gated calcium channels [122].

V. DISCUSSION

This study tries to answer the question of how long a breast cancer patient may live after the diagnosis. Most of the existing studies focus on the statistical associations of molecular biomarkers with the 5-year survival rates of breast cancer patients. The above personalized question remains largely unresolved. This study formulates the question into two machine learning problems. The first problem is whether a patient will live longer than 5 years or not, and the second problem is the length of a patient's remaining life for those who will die within 5 years.

The best binary classification model achieves Acc=1.0000 using 40 methylomic features for the first problem. The best regression model with 79 methylomic features achieves the regression performance MAE=31.62 days for the question how long a patient will live.

Due to the limitations in the cohort size and ethics composition, the proposed models may need further tuning with more independent validation samples and a more balanced ethics composition. Due to the limited availability of other datasets with the similar number samples and methylome profiling technology, we didn't find an independent validation dataset for the proposed prediction models. The generality of this study will be validated with more available datasets.

After the proposed models were further validated by the independent datasets in the future studies, a breast cancer patient may be diagnosed using a methylome profile about whether this patient could live for at least 5 years or shorter than 5 years, and how long this patient may live if the diagnosis is shorter than 5 years. Those patients predicted to live shorter than 1 year may need more frequent follow-up Computed Tomography (CT) imaging examinations.

CONFLICT OF INTEREST

The authors declare that they do not have conflicts of interests in this study.

REFERENCES

[1] R.-M. Feng, Y.-N. Zong, S.-M. Cao, and R.-H. Xu, "Current cancer situation in China: Good or bad news from the 2018 global cancer statistics?" *Cancer Commun.*, vol. 39, no. 1, p. 22, Apr. 2019, doi: 10.1186/s40880-019-0368-6.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA, Cancer J. Clin.*, vol. 68, no. 1, pp. 7–30, Jan. 2018, doi: 10.3322/caac.21442.

[3] W. Chen, R. Zheng, P. Baade, and S. Zhang, "Cancer statistics in China, 2015," *CA, Cancer J. Clin.*, vol. 66, no. 2, pp. 32–115, Mar-Apr. 2016, doi: 10.3322/caac.21338.

[4] C. Rodriguez, A. Aranda, P. Vicioso, B. Hernando, L. Parreño, J. Ryder, H. Fredebohm, and A. Queipo-Ortuño, "Detection of TP53 and PIK3CA mutations in circulating tumor DNA using next-generation sequencing in the screening process for early breast cancer diagnosis," *J. Clin. Med.*, vol. 8, no. 8, p. 1183, Aug. 2019, doi: 10.3390/jcm8081183.

[5] W. Liang, Y. Zhao, W. Huang, Y. Gao, W. Xu, J. Tao, M. Yang, L. Li, W. Ping, H. Shen, X. Fu, Z. Chen, P. W. Laird, X. Cai, J.-B. Fan, and J. He, "Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA)," *Theranostics*, vol. 9, no. 7, pp. 2056–2070, 2019, doi: 10.7150/thno.28119.

[6] X. Li, M. Yang, Q. Zhang, Y. Fan, and T. Zhu, "Whole exome sequencing in the accurate diagnosis of bilateral breast cancer: A case study," *J. Breast Cancer*, vol. 22, no. 1, pp. 131–140, Mar. 2019, doi: 10.4048/jbc.2019.22.e10.

[7] B. Liu, X. Gao, and H. Zhang, "Bioseq-analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Res.*, vol. 47, no. 20, p. e127, Nov. 2019.

[8] S. Srivastava, A. Srivastava, S. Tiwari, and A. K. Mishra, "Life quality index assessment in breast cancer patients," *Indian J. Surg. Oncol.*, vol. 10, no. 3, pp. 476–482, Sep. 2019, doi: 10.1007/s13193-019-00923-8.

[9] M. Yang, C. Liu, and X. Yu, "Skeletal-related adverse events during bone metastasis of breast cancer: Current status," *Discov Med*, vol. 27, no. 149, pp. 211–220, May 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31361984.

[10] L. Chang, L. S. Weiner, S. J. Hartman, S. Horvath, D. Jeste, P. S. Mischel, and D. M. Kado, "Breast cancer treatment and its effects on aging," *J. Geriatric Oncol.*, vol. 10, no. 2, pp. 346–355, Mar. 2019, doi: 10.1016/j.jgo.2018.07.010.

[11] T. Kolben, "Impact of guideline-based use of uPA/PAI-1 on patient outcome in intermediate-risk early breast cancer," *Breast Cancer Res. Treat*, vol. 155, no. 1, pp. 15–109, Jan. 2016, doi: 10.1007/s10549-015-3653-3.

[12] H. G. Kaplan, J. A. Malmgren, and M. Atwood, "T1N0 triple negative breast cancer: Risk of recurrence and adjuvant chemotherapy," *Breast J*, vol. 15, no. 5, pp. 454–460, Sep-Oct. 2009, doi: 10.1111/j.1524-4741.2009.00789.x.

[13] O. P. Geerse, D. Brandenbarg, H. A. M. Kerstjens, A. J. Berendsen, S. F. A. Duijts, H. Burger, G. A. Holtman, J. E. H. M. Hoekstra-Weebers, and T. J. N. Hiltermann, "The distress thermometer as a prognostic tool for one-year survival among patients with lung cancer," *Lung Cancer*, vol. 130, pp. 101–107, Apr. 2019, doi: 10.1016/j.lungcan.2019.02.008.

[14] B. Gyárffy, C. Hatzis, T. Sanft, E. Hofstatter, B. Aktas, and L. Pusztai, "Multigene prognostic tests in breast cancer: Past, present, future," *Breast Cancer Res.*, vol. 17, no. 1, p. 11, Jan. 2015, doi: 10.1186/s13058-015-0514-2.

[15] B. Wallden, J. Storhoff, T. Nielsen, N. Dowidar, C. Schaper, S. Ferree, S. Liu, S. Leung, G. Geiss, J. Snider, T. Vickery, S. R. Davies, E. R. Mardis, M. Gnant, I. Sestak, M. J. Ellis, C. M. Perou, P. S. Bernard, and J. S. Parker, "Development and verification of the PAM50-based prosigna breast cancer gene signature assay," *BMC Med. Genomics*, vol. 8, no. 1, p. 54, Aug. 2015, doi: 10.1186/s12920-015-0129-6.

[16] C.-Y. Lin, K. Mooney, W. Choy, S.-R. Yang, K. Barry-Holson, K. Horst, I. Wapnir, and K. Allison, "Will oncotype DX DCIS testing guide therapy? A single-institution correlation of oncotype DX DCIS results with histopathologic findings and clinical management decisions," *Mod. Pathol.*, vol. 31, no. 4, pp. 562–568, Apr. 2018, doi: 10.1038/modpathol.2017.172.

[17] T. P. McVeigh and M. J. Kerin, "Clinical use of the oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer," *Breast Cancer, Targets Therapy*, vol. 9, pp. 393–400, May 2017, doi: 10.2147/BCTT.S109847.

[18] L. Xin, Y.-H. Liu, T. A. Martin, and W. G. Jiang, "The era of multigene panels comes? The clinical utility of oncotype DX and MammaPrint," *World J. Oncol.*, vol. 8, no. 2, pp. 34–40, 2017, doi: 10.14740/wjon1019w.
[19] R. Buus, I. Sestak, R. Kronenwett, C. Denkert, P. Dubsky, K. Krappmann, M. Scheer, C. Petry, J. Cuzick, and M. Dowsett, "Comparison of endopredict and EPclin with oncotype DX recurrence score for prediction of risk of distant recurrence after endocrine therapy," *J. Nat. Cancer Inst.*, vol. 108, no. 11, Nov. 2016, Art. no. djw149, doi: 10.1093/jnci/djw149.
[20] P. Dubsky, J. Brase, R. Jakesz, and M. Rudas, "The EndoPredict score provides prognostic information on late distant metastases in ER+/HER2- breast cancer patients," *Br. J. Cancer*, vol. 109, no. 12, pp. 2959–2964, Dec. 10 2013, doi: 10.1038/bjc.2013.671.
[21] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, "McTwo: A two-step feature selection algorithm based on maximal information coefficient," *BMC Bioinf.*, vol. 17, no. 1, p. 142, Mar. 2016, doi: 10.1186/s12859-016-0990-0.
[22] M. R. Yousefi, J. Hua, C. Sima, and E. R. Dougherty, "Reporting bias when using real data sets to analyze classification performance," *Bioinformatics*, vol. 26, no. 1, pp. 68–76, Jan. 2010, doi: 10.1093/bioinformatics/btp605.
[23] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, "Exploring the filter bubble: The effect of using recommender systems on content diversity," in *Proc. 23rd Int. Conf. World wide web*, 2014, pp. 677–686.
[24] S. Redkar, S. Mondal, A. Joseph, and K. S. Hareesha, "A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing," *Mol. Informat.*, vol. 39, no. 5, May 2020, Art. no. 1900062, doi: 10.1002/minf.201900062.
[25] J. J. Cai, "ScGEAToolbox: A MATLAB toolbox for single-cell RNA sequencing data analysis," *Bioinformatics*, vol. 4, pp. 1948–1949, Nov. 2019, doi: 10.1093/bioinformatics/btz830.
[26] N. A. Marzouka, J. Nordlund, C. L. Backlin, G. Lonnerholm, A. C. Syvanen, and J. Carlsson Almlof, "CopyNumber450kCancer: Baseline correction for accurate copy number calling from the 450k methylation array," *Bioinformatics*, vol. 32, no. 7, pp. 1080–1082, Apr. 2016, doi: 10.1093/bioinformatics/btv652.
[27] J.-P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood, and K. D. Hansen, "Functional normalization of 450k methylation array data improves replication in large cancer studies," *Genome Biol.*, vol. 15, no. 11, p. 503, Dec. 2014, doi: 10.1186/s13059-014-0503-2.
[28] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk, "International cancer genome consortium data portal—A one-stop shop for cancer genomics data," *Database*, vol. 2011, Sep. 2011, Art. no. bar026, doi: 10.1093/database/bar026.
[29] A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky, "Gene identification in novel eukaryotic genomes by self-training algorithm," *Nucleic Acids Res.*, vol. 33, no. 20, pp. 6494–6506, 2005, doi: 10.1093/nar/gki937.
[30] V. Solovyev and A. Salamov, "The gene-finder computer tools for analysis of human and model organisms genome sequences," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 5, 1997, pp. 294–302. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9322052
[31] Y. Zhang, S. Yang, Y. Liu, Y. Zhang, B. Han, and F. Zhou, "Integration of 24 feature types to accurately detect and predict seizures using scalp EEG signals," *Sensors*, vol. 18, no. 5, p. 1372, Apr. 2018, doi: 10.3390/s18051372.
[32] X. Feng, J. Li, H. Li, H. Chen, F. Li, Q. Liu, Z.-H. You, and F. Zhou, "Age is important for the early-stage detection of breast cancer on both transcriptomic and methylomic biomarkers," *Frontiers Genet.*, vol. 10, p. 212, Mar. 2019, doi: 10.3389/fgene.2019.00212.
[33] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "TyG-er: An ensemble regression forest approach for identification of clinical factors related to insulin resistance condition using electronic health records," *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103358, doi: 10.1016/j.compbiomed.2019.103358.
[34] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, Jun. 15 2016, doi: 10.1093/bioinformatics/btw074.
[35] R. S. Zoh, A. Sarkar, R. J. Carroll, and B. K. Mallick, "A powerful Bayesian test for equality of means in high dimensions," *J. Amer. Stat. Assoc.*, vol. 113, no. 524, pp. 1733–1741, Oct. 2018, doi: 10.1080/01621459.2017.1371024.
[36] B. Li, N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li, "Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods," *Frontiers Genet.*, vol. 9, p. 237, Jul. 2018, doi: 10.3389/fgene.2018.00237.
[37] Y. Ye, R. Zhang, W. Zheng, S. Liu, and F. Zhou, "RIFS: A randomly restarted incremental feature selection algorithm," *Sci. Rep.*, vol. 7, no. 1, Oct. 2017, Art. no. 13013, doi: 10.1038/s41598-017-13259-6.
[38] C. Pichereaux, E. E. Hernández-Domínguez, M. D. S. Santos-Diaz, A. Reyes-Agáero, M. Astello-García, F. Guéraud, A. Negre-Salvayre, O. Schiltz, M. Rossignol, and A. P. Barba de la Rosa, "Comparative shotgun proteomic analysis of wild and domesticated opuntia spp. Species shows a metabolic adaptation through domestication," *J. Proteomics*, vol. 143, pp. 353–364, Jun. 2016.
[39] M. Chiara, P. Mandreoli, M. A. Tangaro, A. M. D'Erchia, S. Sorrentino, C. Forleo, D. S. Horner, F. Zambelli, and G. Pesole, "VINYL: Variant prIoritizatioN bY survivaL analysis," *Bioinformatics*, 2020, Art. no. btaa1067, doi: 10.1093/bioinformatics/btaa1067.
[40] S. Kim, K. Kim, J. Choe, I. Lee, and J. Kang, "Improved survival analysis by learning shared genomic information from pan-cancer data," *Bioinformatics*, vol. 36, no. 1, pp. 389–398, Jul. 2020, doi: 10.1093/bioinformatics/btaa462.
[41] K. Cootjans, J. Vanhaecke, M. Dezillie, J. Barth, H. Pottel, and F. Stockmans, "Joint survival analysis and clinical outcome of total joint arthroplasties with the ARPE implant in the treatment of trapeziometacarpal osteoarthritis with a minimal follow-up of 5 years," *J. Hand Surg.*, vol. 42, no. 8, pp. 630–638, Aug. 2017, doi: 10.1016/j.jhsa.2017.05.007.
[42] D. M. Green, M. A. Zevon, P. A. Reese, G. S. Lowrie, and A. M. Michalek, "Factors that influence the further survival of patients who survive for five years after the diagnosis of cancer in childhood or adolescence," *Med. Pediatric Oncol.*, vol. 22, no. 2, pp. 91–96, 1994, doi: 10.1002/mpo.2950220206.
[43] H. P. Petridis, A. Zekeridou, M. Malliari, D. Tortopidis, and P. Koidis, "Survival of ceramic veneers made of different materials after a minimum follow-up period of five years: A systematic review and meta-analysis," *Eur. J. Esthetic Dentistry, Off. J. Eur. Acad. Esthetic Dentistry*, vol. 7, no. 2, pp. 52–138, 2012.
[44] M. Charlson, T. P. Szatrowski, J. Peterson, and J. Gold, "Validation of a combined comorbidity index," *J. Clin. Epidemiol.*, vol. 47, no. 11, pp. 1245–1251, Nov. 1994, doi: 10.1016/0895-4356(94)90129-5.
[45] H. Kim, L. Kim, and C.-H. Im, "Machine-learning-based detection of craving for gaming using multimodal physiological signals: Validation of test-retest reliability for practical use," *Sensors*, vol. 19, no. 16, p. 3475, Aug. 2019, doi: 10.3390/s19163475.
[46] S. Joof, T. Yilmaz, M. Çayören, B. Önal, and I. Akduman, "Microwave dielectric property based classification of renal calculi: Application of a kNN algorithm," *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103366, doi: 10.1016/j.compbiomed.2019.103366.
[47] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
[48] Attallah, Sharkas, and Gadelkarim, "Fetal brain abnormality classification from MRI images of different gestational age," *Brain Sci.*, vol. 9, no. 9, p. 231, Sep. 2019, doi: 10.3390/brainsci9090231.
[49] C. Wang, W. Pu, D. Zhao, Y. Zhou, T. Lu, S. Chen, Z. He, X. Feng, Y. Wang, C. Li, S. Li, L. Jin, S. Guo, J. Wang, and M. Wang, "Identification of hyper-methylated tumor suppressor genes-based diagnostic panel for esophageal squamous cell carcinoma (ESCC) in a chinese han population," *Frontiers Genet.*, vol. 9, p. 356, Sep. 2018, doi: 10.3389/fgene.2018.00356.
[50] Ibrahim, Parrish, Brown, and McDonald, "Decision tree pattern recognition model for radio frequency interference suppression in NQR experiments," *Sensors*, vol. 19, no. 14, p. 3153, Jul. 2019, doi: 10.3390/s19143153.
[51] H. Cai, X. Pang, D. Dong, Y. Ma, Y. Huang, X. Fan, P. Wu, H. Chen, F. He, Y. Cheng, S. Liu, Y. Yu, M. Hong, J. Xiao, X. Wan, Y. Lv, and J. Zheng, "Molecular decision tree algorithms predict individual recurrence pattern for locally advanced nasopharyngeal carcinoma," *J. Cancer*, vol. 10, no. 15, pp. 3323–3332, 2019, doi: 10.7150/jca.29693.
[52] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
[53] Z. Zhang, X. Shi, X. Xiang, C. Wang, S. Xiao, and X. Su, "Bootstrap confidence intervals for the optimal cutoff point to bisect estimated probabilities from logistic regression," *Stat. Methods Med. Res.*, vol. 4, Art. no. 962280219864998, Jul. 2019, doi: 10.1177/0962280219864998.

[54] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[55] P. F. Verhulst, "Resherches mathematiques sur la loi d'accroissement de la population," *Nouveaux Memoires Academie Roy. Sci.*, vol. 18, no. 1, pp. 1–41, 1845.

[56] F. Yin, X. Shao, L. Zhao, X. Li, J. Zhou, Y. Cheng, X. He, S. Lei, J. Li, and J. Wang, "Predicting prognosis of endometrioid endometrial adenocarcinoma on the basis of gene expression and clinical features using random forest," *Oncol. Lett.*, pp. 1597–1606, Jun. 2019, doi: 10.3892/ol.2019.10504.

[57] H. Tin Kam, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

[58] H. Climente-González, C.-A. Azencott, S. Kaski, and M. Yamada, "Block HSIC lasso: Model-free biomarker detection for ultra-high dimensional data," *Bioinformatics*, vol. 35, no. 14, pp. 427–435, Jul. 2019, doi: 10.1093/bioinformatics/btz333.

[59] J. Cai, W.-G. He, L. Wang, K. Zhou, and T.-X. Wu, "Osteoporosis recognition in rats under low-power lens based on convexity optimization feature fusion," *Sci. Rep.*, vol. 9, no. 1, Jul. 2019, Art. no. 10971, doi: 10.1038/s41598-019-47281-7.

[60] S. Paul and P. Drineas, "Feature selection for ridge regression with provable guarantees," *Neural Comput.s*, vol. 28, no. 4, pp. 716–742, Apr. 2016, doi: 10.1162/NECO_a_00816.

[61] S. Mostafaei, H. Abdollahi, S. Kazempour Dehkordi, I. Shiri, A. Razzaghdoust, S. H. Z. Moghaddam, A. Saadipoor, F. Koosha, S. Cheraghi, and S. R. Mahdavi, "CT imaging markers to improve radiation toxicity prediction in prostate cancer radiotherapy by stacking regression algorithm," *La Radiol. Med.*, vol. 125, no. 1, pp. 87–97, Jan. 2020, doi: 10.1007/s11547-019-01082-0.

[62] C. Xu, J. Liu, W. Yang, Y. Shu, Z. Wei, W. Zheng, X. Feng, and F. Zhou, "An OMIC biomarker detection algorithm TriVote and its application in methylomic biomarker detection," *Epigenomics*, vol. 10, no. 4, pp. 335–347, Apr. 2018.

[63] V. Fonti and E. Belitser, "Feature selection using lasso," Res. Paper Bus. Anal., VU Amsterdam, Amsterdam, The Netherlands, Tech. Rep., 2017, vol. 30, pp. 1–25.

[64] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc., Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[65] J. Tang, Y. Wang, J. Fu, Y. Zhou, Y. Luo, Y. Zhang, B. Li, Q. Yang, W. Xue, Y. Lou, Y. Qiu, and F. Zhu, "A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies," *Briefings Bioinf.*, vol. 21, no. 4, pp. 1378–1390, Jul. 2020, doi: 10.1093/bib/bbz061.

[66] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Mining prognosis index of brain metastases using artificial intelligence," *Cancers*, vol. 11, no. 8, p. 1140, Aug. 2019, doi: 10.3390/cancers11081140.

[67] H.-J. Cho, S. Lee, Y. G. Ji, and D. H. Lee, "Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0207204, doi: 10.1371/journal.pone.0207204.

[68] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinf.*, vol. 18, no. 1, p. 169, Mar. 2017, doi: 10.1186/s12859-017-1578-z.

[69] N. Zhou and L. Wang, "A modified t-test feature selection method and its application on the hapmap genotype data," *Genomics Proteomics Bioinf.*, vol. 5, no. 3, pp. 242–249, Jan. 2007, doi: 10.1016/S1672-0229(08)60011-X.

[70] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," in *Proc. Int. Workshop Data Mining Biomed. Appl.* Berlin, Germany: Springer, Apr 2006, pp. 106–115.

[71] R. Liang, Y. Zhi, G. Zheng, B. Zhang, H. Zhu, and M. Wang, "Analysis of long non-coding RNAs in glioblastoma for prognosis prediction using weighted gene co-expression network analysis, Cox regression, and L_1-LASSO penalization," *Onco Targets*, vol. 12, pp. 157–168, 2019, doi: 10.2147/OTT.S171957.

[72] K. H. Hellton and N. L. Hjort, "Fridge: Focused fine-tuning of ridge regression for personalized predictions," *Stat. Med.*, vol. 37, no. 8, pp. 1290–1303, Apr. 2018, doi: 10.1002/sim.7576.

[73] A. Basu, R. Mitra, H. Liu, S. L. Schreiber, and P. A. Clemons, "RWEN: Response-weighted elastic net for prediction of chemosensitivity of cancer cell lines," *Bioinformatics*, vol. 34, no. 19, pp. 3332–3339, Oct. 2018, doi: 10.1093/bioinformatics/bty199.

[74] Y. Zhang, D. J. Topham, J. Thakar, and X. Qiu, "FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis," *Bioinformatics*, vol. 33, no. 13, pp. 1944–1952, Jul. 2017, doi: 10.1093/bioinformatics/btx104.

[75] C.-H. Zhai, J.-B. Xuan, H.-L. Fan, T.-F. Zhao, and J.-L. Jiang, "The application of SVR model in the improvement of QbD: A case study of the extraction of podophyllotoxin," *Drug Develop. Ind. Pharmacy*, vol. 44, no. 9, pp. 1506–1511, Sep. 2018, doi: 10.1080/03639045.2018.1467924.

[76] Q. Zhang, X. Hu, and B. Zhang, "Comparison of L_1-norm SVR and sparse coding algorithms for linear regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1828–1833, Aug. 2015, doi: 10.1109/TNNLS.2014.2377245.

[77] R. Liu, W. Zhang, Z. Q. Liu, and H. H. Zhou, "Gene modules associated with breast cancer distant metastasis-free survival in the PAM50 molecular subtypes," *Oncotarget*, vol. 7, no. 16, pp. 21686–21698, Apr. 2016, doi: 10.18632/oncotarget.7774.

[78] G. Viale, f. the MINDACT investigators, F. A. de Snoo, L. Slaets, J. Bogaerts, L. van 't Veer, E. J. Rutgers, M. J. Piccart-Gebhart, L. Stork-Sloots, A. Glas, L. Russo, P. Dell'Orto, K. Tryfonidis, S. Litière, and F. Cardoso, "Immunohistochemical versus molecular (BluePrint and MammaPrint) subtyping of breast carcinoma. Outcome results from the EORTC 10041/BIG 3-04 MINDACT trial," *Breast Cancer Res. Treatment*, vol. 167, no. 1, pp. 123–131, Jan. 2018, doi: 10.1007/s10549-017-4509-9.

[79] N. Tobin, J. Harrell, J. Lövrot, S. Brage, and M. Stolt, "Molecular subtype and tumor characteristics of breast cancer metastases as assessed by gene expression significantly influence patient post-relapse survival," *Ann. Oncol.*, vol. 26, no. 1, pp. 8–81, Jan. 2015, doi: 10.1093/annonc/mdu498.

[80] M. Filipits, P. Dubsky, M. Rudas, R. Greil, M. Balic, Z. Bago-Horvath, C. F. Singer, D. Hlauschek, K. Brown, R. Bernhisel, R. Kronenwett, J. M. Lancaster, F. Fitzal, and M. Gnant, "Prediction of distant recurrence using EndoPredict among women with ER+, HER2 node-positive and node-negative breast cancer treated with endocrine therapy only," *Clin. Cancer Res.*, vol. 25, no. 13, pp. 3865–3872, Jul. 2019, doi: 10.1158/1078-0432.CCR-19-0376.

[81] L. Hochheiser, J. Hornberger, M. Turner, and G. H. Lyman, "Multigene assays: Effect on chemotherapy use, toxicity and cost in estrogen receptor-positive early stage breast cancer," *J. Comparative Effectiveness Res.*, vol. 8, no. 5, pp. 289–304, Apr. 2019.

[82] B. Schroeder, Y. Zhang, O. Stál, T. Fornander, A. Brufsky, D. C. Sgroi, and C. A. Schnabel, "Risk stratification with breast cancer index for late distant recurrence in patients with clinically low-risk (T1N0) estrogen receptor-positive breast cancer," *NPJ Breast Cancer*, vol. 3, no. 1, p. 28, Dec. 2017, doi: 10.1038/s41523-017-0037-3.

[83] L. A. Habel, L. C. Sakoda, N. Achacoso, X.-J. Ma, M. G. Erlander, D. C. Sgroi, L. Fehrenbacher, D. Greenberg, and C. P. Quesenberry, "HOXB13:IL17BR and molecular grade index and risk of breast cancer death among patients with lymph node-negative invasive disease," *Breast Cancer Res.*, vol. 15, no. 2, Mar. 2013, doi: 10.1186/bcr3402.

[84] M. A. Khan, L. Henderson, D. Clarke, S. Harries, and L. Jones, "The warwick experience of the oncotype DX breast recurrence Score assay as a predictor of chemotherapy administration," *Breast Care*, vol. 13, no. 5, pp. 369–372, 2018, doi: 10.1159/000489131.

[85] C. Markopoulos, "Overview of the use of Oncotype DX((R)) as an additional treatment decision tool in early breast cancer," *Expert Rev Anticancer Ther*, vol. 13, no. 2, pp. 94–179, Feb. 2013, doi: 10.1586/era.12.174.

[86] F. Wang, S. Reid, W. Zheng, T. Pal, I. Meszoely, I. A. Mayer, C. E. Bailey, B. H. Park, and X.-O. Shu, "Sex disparity observed for oncotype DX breast recurrence score in predicting mortality among patients with early stage ER-positive breast cancer," *Clin. Cancer Res.*, vol. 26, no. 1, pp. 101–109, Jan. 2020, doi: 10.1158/1078-0432.CCR-19-2424.

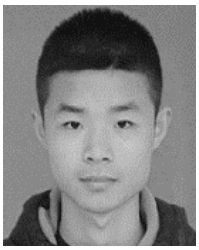
[87] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2016, pp. 212–220.

[88] Z. Isik, M. E. J. C. i. B. Ercan, and Medicine, "Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients," *Comput. Biol. Med.*, vol. 89, pp. 397–404, Oct. 2017.

[89] L. Chato, E. Chow, and S. Latifi, "Wavelet transform to improve accuracy of a prediction model for overall survival time of brain tumor patients based On MRI images," in *Proc. IEEE Int. Conf. Healthcare Inform.*, Jun. 2018, pp. 441–442.
[90] E. E. Bron, M. Smits, W. J. Niessen, and S. Klein, "Feature selection based on the SVM weight vector for classification of dementia," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1617–1626, Sep. 2015, doi: 10.1109/JBHI.2015.2432832.
[91] X. Gao, S. Liu, H. Song, X. Feng, M. Duan, L. Huang, and F. Zhou, "AgeGuess, a methylomic prediction model for human ages," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 80, Mar. 2020, doi: 10.3389/fbioe.2020.00080.
[92] L. A. Zhang, R. S. Parker, D. Swigon, I. Banerjee, S. Bahrami, H. Redl, and G. Clermont, "A One-nearest-neighbor approach to identify the original time of infection using censored baboon sepsis data," *Crit. Care Med.*, vol. 44, no. 6, pp. 432–442, Jun. 2016, doi: 10.1097/CCM.0000000000001623.
[93] T. Chen, C. Zhang, Y. Liu, Y. Zhao, D. Lin, Y. Hu, J. Yu, and G. Li, "A gastric cancer LncRNAs model for MSI and survival prediction based on support vector machine," *BMC Genomics*, vol. 20, no. 1, p. 846, Nov. 2019, doi: 10.1186/s12864-019-6135-x.
[94] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.017.
[95] A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei, "Private naive bayes classification of personal biomedical data: Application in cancer data analysis," *Comput. Biol. Med.*, vol. 105, pp. 144–150, Feb. 2019, doi: 10.1016/j.compbiomed.2018.11.018.
[96] M. Guckenberger, Y. Lievens, A. Bouma, and L. Collette, "Characterisation and classification of oligometastatic disease: A European society for radiotherapy and oncology and European organisation for research and treatment of cancer consensus recommendation," *Lancet Oncol*, vol. 21, no. 1, pp. 18–28, Jan. 2020, doi: 10.1016/S1470-2045(19)30718-1.
[97] H. Hung and S. Huang, "Sufficient dimension reduction via random-partitions for the large-p-small-n problem," *Biometrics*, vol. 75, no. 1, pp. 245–255, Mar. 2019, doi: 10.1111/biom.12926.
[98] C. Lou, J. Zhao, R. Shi, Q. Wang, W. Zhou, Y. Wang, G. Wang, L. Huang, X. Feng, and F. Zhou, "SefOri: Selecting the best-engineered sequence features to predict DNA replication origins," *Bioinformatics*, vol. 36, no. 1, pp. 49–55, Jan. 2020, doi: 10.1093/bioinformatics/btz506.
[99] K. Davagdorj, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, "XGBoost-based framework for smoking-induced noncommunicable disease prediction," *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6513, Sep. 2020, doi: 10.3390/ijerph17186513.
[100] A. K. Shukla and D. Tripathi, "Identification of potential biomarkers on microarray data using distributed gene selection approach," *Math. Biosci.*, vol. 315, Sep. 2019, Art. no. 108230, doi: 10.1016/j.mbs.2019.108230.
[101] Q. Lin, Y. Ji, Y. Chen, H. Sun, D. Yang, A. Chen, T. Chen, and X. M. Zhang, "Radiomics model of contrast-enhanced MRI for early prediction of acute pancreatitis severity," *J. Magn. Reson. Imag.*, vol. 51, no. 2, pp. 397–406, Feb. 2020, doi: 10.1002/jmri.26798.
[102] A. J. Sahoo and Y. Kumar, "Seminal quality prediction using data mining methods," *Technol. Health Care*, vol. 22, no. 4, pp. 531–545, 2014, doi: 10.3233/THC-140816.
[103] X.-Q. Chen, F. Zhang, Q.-C. Su, C. Zeng, F.-H. Xiao, and Y. Peng, "Methylome and transcriptome analyses reveal insights into the epigenetic basis for the good survival of hypomethylated ER-positive breast cancer subtype," *Clin. Epigenetics*, vol. 12, no. 1, p. 16, Jan. 2020, doi: 10.1186/s13148-020-0811-1.
[104] N. Jalaleddine, L. El-Hajjar, H. Dakik, A. Shaito, J. Saliba, R. Safi, K. Zibara, and M. El-Sabban, "Pannexin1 is associated with enhanced epithelial-to-mesenchymal transition in human patient breast cancer tissues and in breast cancer cell lines," *Cancers*, vol. 11, no. 12, p. 1967, Dec. 2019, doi: 10.3390/cancers11121967.
[105] M. Popeda, T. Stokowy, N. Bednarz-Knoll, A. Jurek, M. Niemira, A. Bielska, A. Kretowski, L. Kalinowski, J. Szade, A. Markiewicz, and A. J. Zaczek, "NF-kappa b signaling-related signatures are connected with the mesenchymal phenotype of circulating tumor cells in non-metastatic breast cancer," *Cancers*, vol. 11, no. 12, p. 1961, Dec. 2019, doi: 10.3390/cancers11121961.
[106] R. Van Oirbeek and E. Lesaffre, "An application of Harrell's c-index to PH frailty models," *Stat. Med.*, vol. 29, no. 30, pp. 3160–3171, Dec. 2010.
[107] S. H. Strand, L. Schmidt, S. Weiss, M. Borre, H. Kristensen, A. K. I. Rasmussen, T. F. Daugaard, G. Kristensen, H. V. Stroomberg, M. A. Røder, K. Brasso, P. Mouritzen, and K. D. Sørensen, "Validation of the four-miRNA biomarker panel MiCaP for prediction of long-term prostate cancer outcome," *Sci. Rep.*, vol. 10, no. 1, Jul. 2020, Art. no. 10704, doi: 10.1038/s41598-020-67320-y.
[108] X. Zou, Z. Hu, C. Huang, and J. Chang, "A seven-gene signature with close immune correlation was identified for survival prediction of lung adenocarcinoma," *Med. Sci. Monitor*, vol. 26, May 2020, Art. no. e924269, doi: 10.12659/MSM.924269.
[109] Z.-M. Zhang, J.-X. Tan, F. Wang, F.-Y. Dao, Z.-Y. Zhang, and H. Lin, "Early diagnosis of hepatocellular carcinoma using machine learning method," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 254, Mar. 2020, doi: 10.3389/fbioe.2020.00254.
[110] Z. Lv, J. Zhang, H. Ding, and Q. Zou, "RF-PseU: A random forest predictor for RNA pseudouridine sites," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 134, Feb. 2020, doi: 10.3389/fbioe.2020.00134.
[111] Y. Zhang, C. Chen, M. Duan, S. Liu, L. Huang, and F. Zhou, "BioDog, biomarker detection for improving identification power of breast cancer histologic grade in methylomics," *Epigenomics*, vol. 11, no. 15, pp. 1717–1732, Nov. 2019, doi: 10.2217/epi-2019-0230.
[112] B. Byram, K. Dei, J. Tierney, and D. Dumont, "A model and regularization scheme for ultrasonic beamforming clutter reduction," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 62, no. 11, pp. 1913–1927, Nov. 2015, doi: 10.1109/TUFFC.2015.007004.
[113] Y. Huang, C. Schell, T. B. Huber, N. Hersch, R. Merkel, G. Gompper, and B. Sabass, "Traction force microscopy with optimized regularization and automated Bayesian parameter selection for comparing cells," *Sci. Rep.*, vol. 9, no. 1, Jan. 2019, Art. no. 539, doi: 10.1038/s41598-018-36896-x.
[114] R. Mohanty, A. M. Sinha, A. B. Remsik, K. C. Dodd, B. M. Young, T. Jacobson, M. Mcmillan, J. Thoma, H. Advani, V. A. Nair, T. J. Kang, K. Caldera, D. F. Edwards, J. C. Williams, and V. Prabhakaran, "Early findings on functional connectivity correlates of behavioral outcomes of brain-computer interface stroke rehabilitation using machine learning," *Frontiers Neurosci.*, vol. 12, p. 624, Sep. 2018, doi: 10.3389/fnins.2018.00624.
[115] V. Suomi, J. Järvinen, T. Kiviniemi, A. Ylitalo, and M. Pietilä, "Full feature selection for estimating KAP radiation dose in coronary angiographies and percutaneous coronary interventions," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103725, doi: 10.1016/j.compbiomed.2020.103725.
[116] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers," *Nature Rev. Cancer*, vol. 18, no. 11, pp. 696–705, Nov. 2018, doi: 10.1038/s41568-018-0060-1.
[117] F. Coussy, L. de Koning, and M. Lavigne, "A large collection of integrated genomically characterized patient-derived xenografts highlighting the heterogeneity of triple-negative breast cancer," *Int J Cancer*, vol. 145, no. 7, pp. 1902–1912, Oct 1 2019, doi: 10.1002/ijc.32266.
[118] F. Leo, S. Bartels, L. Mägel, T. Framke, G. Básche, D. Jonigk, M. Christgen, U. Lehmann, and H. Kreipe, "Prognostic factors in the myoepithelial-like spindle cell type of metaplastic breast cancer," *Virchows Archiv.*, vol. 469, no. 2, pp. 191–201, Aug. 2016, doi: 10.1007/s00428-016-1950-9.
[119] C. Fan, L. Ma, and N. Liu, "Comprehensive analysis of novel three-long noncoding RNA signatures as a diagnostic and prognostic biomarkers of human triple-negative breast cancer," *J. Cellular Biochem.*, vol. 120, no. 3, pp. 3185–3196, Mar. 2019, doi: 10.1002/jcb.27584.
[120] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, and D. Lancet, "GeneCards version 3: The human gene integrator," *Database*, vol. 2010, Aug. 2010, Art. no. baq020, doi: 10.1093/database/baq020.
[121] N. Fiorini, K. Canese, R. Bryzgunov, and I. Radetska, "PubMed Labs: An experimental system for improving biomedical literature search," *Database*, vol. 2018, p. 5, Jan. 2018, doi: 10.1093/database/bay094.
[122] N. N. Phan, C.-Y. Wang, C.-F. Chen, Z. Sun, M.-D. Lai, and Y.-C. Lin, "Voltage-gated calcium channels: Novel targets for cancer therapy," *Oncol. Lett.*, vol. 14, no. 2, pp. 2059–2074, Aug. 2017, doi: 10.3892/ol.2017.6457.

[123] R. García-Baquero, P. Puerta, and M. Beltran, "Methylation of a novel panel of tumor suppressor genes in urine moves forward noninvasive diagnosis and prognosis of bladder cancer: A 2-center prospective study," *J. Urol.*, vol. 190, no. 2, pp. 723–730, Aug. 2013, doi: 10.1016/j.juro.2013.01.105.
[124] K. Shima, T. Morikawa, Y. Baba, and K. Nosho, "MGMT promoter methylation, loss of expression and prognosis in 855 colorectal cancers," *Cancer Causes Control*, vol. 22, no. 2, pp. 301–309, Feb. 2011, doi: 10.1007/s10552-010-9698-z.
[125] M. Ye, R. Han, J. Shi, X. Wang, A. Z. Zhao, F. Li, and H. Chen, "Cellular apoptosis susceptibility protein (CAS) suppresses the proliferation of breast cancer cells by upregulated cyp24a1," *Med. Oncol.*, vol. 37, no. 5, p. 43, Apr. 2020, doi: 10.1007/s12032-020-01366-w.
[126] H. Cai, Y. Jiao, Y. Li, Z. Yang, M. He, and Y. Liu, "Low CYP24A1 mRNA expression and its role in prognosis of breast cancer," *Sci. Rep.*, vol. 9, no. 1, Sep. 2019, Art. no. 13714, doi: 10.1038/s41598-019-50214-z.
[127] L. Canela and V. Fernández-Dueñas, "The association of metabotropic glutamate receptor type 5 with the neuronal Ca^{2+}-binding protein 2 modulates receptor function," *J Neurochem*, vol. 111, no. 2, pp. 555–567, Oct. 2009, doi: 10.1111/j.1471-4159.2009.06348.x.
[128] L. Canela, R. Luján, C. Lluís, J. Burgueño, J. Mallol, E. I. Canela, R. Franco, and F. Ciruela, "The neuronal Ca^{2+}-binding protein2 (NECAB2) interacts with the adenosine A_2A receptor and modulates the cell surface expression and function of the receptor," *Mol. Cellular Neurosci.*, vol. 36, no. 1, pp. 1–12, Sep. 2007, doi: 10.1016/j.mcn.2007.05.007.

SHUAI LIU received the bachelor's degree in computer sciences from Jiangnan University, in 2016, and the master's degree from Jilin University, in 2019. His research interest includes the development of feature selection algorithms for the biomedical big data.



HAN LI received the B.E. degree in software engineering from Jilin University, Changchun, China, in 2019. He is currently pursuing the Ph.D. degree with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. His research interests include machine learning and computational biology.

QICHEN ZHENG is currently pursuing the bachelor's degree with Jilin University.

LU YANG received the B.S. degree in computer science and technology from Jilin University, Changchun, China, in 2019. She is currently pursuing the M.S. degree in computer application technology with the Guilin University of Electronic Technology. Her research interests include feature selection algorithms and biomedical big data mining.

MEIYU DUAN received the B.S. degree in computer science and technology from Yanbian University, Yanji, China, in 2018. She is currently pursuing the M.S. degree in computer application technology with Jilin University. Her research interests include feature selection algorithms and biomedical big data mining.

XIN FENG received the Ph.D. degree from the School of Computer Science, Jilin University, in 2019. She is currently a Postdoctoral Fellow with Jilin University. Her main research interest includes big data analysis and mining.

FEI LI was with the Bioknow Health Informatics Laboratory (HILab), from 2016 to 2019. He is currently focusing on NLP and gene network with HILab.

LAN HUANG received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, Changchun, China, in 2003. She is currently working as a Professor with the College of Computer Science and Technology, Jilin University. Her research interests include data mining and business intelligence.

FENGFENG ZHOU (Senior Member, IEEE) received the bachelor's and Ph.D. degrees in computer sciences from the University of Science and Technology of China, in 2000 and 2005, respectively. His laboratory at Jilin University focuses on the development of feature selection algorithms for the biomedical big data.