Complex & Intelligent Systems (2022) 8:4619–4639
https://doi.org/10.1007/s40747-022-00791-w

ORIGINAL ARTICLE

A two-stage stacked-based heterogeneous ensemble learning for cancer survival prediction

Fangzhou Yan[1] · Yi Feng[2]

Received: 22 March 2022 / Accepted: 18 May 2022 / Published online: 5 July 2022

Abstract

Cancer survival prediction is one of the three major tasks of cancer prognosis. To improve the accuracy of cancer survival prediction, in this paper, we propose a priori knowledge- and stability-based feature selection (PKSFS) method and develop a novel two-stage heterogeneous stacked ensemble learning model (BQAXR) to predict the survival status of cancer patients. Specifically, PKSFS first obtains the optimal feature subsets from the high-dimensional cancer datasets to guide the subsequent model construction. Then, BQAXR seeks to generate five high-quality heterogeneous learners, among which the shortcomings of the learners are overcome by using improved methods, and integrate them in two stages through the stacked generalization strategy based on optimal feature subsets. To verify the merits of PKSFS and BQAXR, this paper collected the real survival datasets of gastric cancer and skin cancer from the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute, and conducted extensive numerical experiments from different perspectives based on these two datasets. The accuracy and AUC of the proposed method are 0.8209 and 0.8203 in the gastric cancer dataset, and 0.8336 and 0.8214 in the skin cancer dataset. The results show that PKSFS has marked advantages over popular feature selection methods in processing high-dimensional datasets. By taking full advantage of heterogeneous high-quality learners, BQAXR is not only superior to mainstream machine learning methods, but also outperforms improved machine learning methods, which indicates can effectively improve the accuracy of cancer survival prediction and provide a reference for doctors to make medical decisions.

Keywords Stacked generalization strategy · Cancer survival prediction · Feature selection · Heterogeneous ensemble learning

. اختيـــار مـــيزة . التنبـــــؤ الســـرطان بقـــاء . التعميـــم اســـــتراتيجية تراكـــب الرئيسيه الكلمات
المتكامـــــل التعلـــم مـن متجانســـة غـير

Introduction

Cancer is a disease with high morbidity and mortality. According to GLOBOCAN 2019 released by the International Agency for Research on Cancer (IARC), there were 18.1 million new cancer cases worldwide in 2018, including 9.6 million cancer deaths [1]. One of the three main tasks of medical prognosis, and survival prediction concerns applying efficient algorithms and techniques to predict the survival status of cancer patients according to the historical dataset of patients with the same type of cancer. Accurate survival prediction can effectively assist doctors in formulating treatment plans and decisions, thereby improving the prognosis effect and reducing decision regret [2].

Traditionally, hospitals use statistical methods to describe and analyze the dataset since the amount of data is small and the data are not too complicated. However, in the big data era, the number and complexity of data grow exponentially. It is difficult for statistical methods to accurately analyze and effectively mine massive amounts of internal data [3]. With the rapid development of machine learning and data mining, researchers have applied various machine learning methods, such as random forest (RF), support vector machine (SVM), decision tree (DT), neural network (NN), etc., to the medical field, which have been shown to be capable of efficiently improving the accuracy of prediction.

✉ Yi Feng
fyfef11@163.com

[1] College of Electrical Engineering, Sichuan University, Chengdu 610064, China

[2] Business School, Sichuan University, Chengdu 610064, China

In recent years, researchers have favored ensemble learning methods because their performance and generalization capability are superior to mainstream machine learning methods in many fields [4–7]. However, few studies have explored the cancer survival prediction problem employing stacked ensemble methods, which limits the further improvement in the accuracy of survival prediction. To develop a cancer survival prediction method with better performance, we proposed an a priori knowledge- and stability-based feature selection (PKSFS) method and developed a novel two-stage heterogeneous stacked ensemble learning model, denoted by BQAXR, and we summarize our contributions as follows:

- For high-dimensional cancer datasets, making good use of the features based on prior knowledge and keeping the stability of BQAXR, PKSFS is proposed to reduce the computational complexity of BQAXR and improve the accuracy of cancer survival prediction. As demonstrated in our numerical studies, compared with the widely used feature selection methods, PKSFS can better guide the subsequent model construction.
- We develop a two-stage heterogeneous stacked ensemble learning model, namely BQAXR, to predict the survival status of gastric cancer and skin cancer patients. In BQAXR, we attempt to improve the deficiencies of the learners and integrate them in two stages through the stacked generalization strategy to further improve the accuracy of cancer survival prediction. Specifically, BQAXR improves the shortcomings of four heterogeneous base learners, and employs a stacked generalization strategy to integrate through advanced meta learner, multi-layer perception based on the rectified Adam optimizer RAdam. To the best of our knowledge, this is the first ensemble learning model for gastric cancer and skin cancer survival prediction, and the experimental results demonstrate the superiority of BQAXR compared with popular machine learning methods.
- Most studies on cancer survival prediction focus on breast cancer [8], colorectal cancer [9], etc. Gastric cancer, as one of the top three cancer diseases in death cases, is ignored. Furthermore, some rare types of cancer, such as skin cancer, are also ignored. Thus, from the perspective of common and uncommon cancer diseases, real cancer datasets including gastric cancer and skin cancer are collected to support this study, and the superiority of the proposed method for cancer survival prediction would be verified on the two different cancer datasets.

The remainder of this paper is as follows: the next section briefly reviews the related literature. In the following section we introduce the research datasets and the PKSFS method, followed by presenting BQAXR in detail. The next section presents and discusses the results of numerical studies conducted to compare BQAXR against some mainstream and improved machine learning methods. In the following section, we discuss the results of this study, and some works related to cancer prediction. Finally, in the last section, we conclude this paper and suggest topics for future research.

Literature review

Survival prediction methods

In the early years, many researchers compared the Cox proportional hazard model with machine learning and deep learning methods for survival prediction problems. Matsuo et al. [10] used the Cox model and the deep learning neural network to predict the survival of cervical cancer patients, and showed that the performance of the neural network is superior to the Cox proportional hazard model. Zhu et al. [11] used an artificial neural network (ANN) and the Cox regression risk model to analyze the prognostic factors of gastric cancer patients, and found that ANN is a more powerful tool in determining the important factors of prognosis. Consistent with Zhu et al. [11], Walczak and Velanovich [2] also found that ANN is superior to the Cox model.

Since then, more and more researchers have applied machine learning and deep learning methods to carry out survival predictions for cancer patients. Tapak et al. [12] applied six machine learning methods (NB, RF, AdaBoost, SVM, least squares-SVM, and Adabag) to predict the survival status of 550 breast cancer patients, and found that SVM is superior to the other machine learning methods. Delen [13] applied three machine learning techniques (DT, NN, and SVM) and one statistic method (LR) to predict the survival probability of prostate cancer, and showed that SVM performs the best and LR performs the worst. Shukla et al. [14] proposed a breast cancer survival prediction model, which uses the self-organizing map (SOM) and density-based spatial clustering of applications with noise (DBSCAN) to generate patient clusters, and then trains multi-layer perception (MLP) using the generated clusters. Zolbanin et al. [15] employed LR, DT, RF, and ANN to predict the overall survival of breast, genital, prostate, and urethral cancer patients, and showed that RF performs the best. Unlike the above studies, we aim to develop survival predictions for cancer patients based on heterogeneous ensemble learning methods and achieve better performance.

Heterogeneous ensemble learning methods

Ensemble learning can be broadly divided into two categories, namely homogeneous ensemble learning and heterogeneous ensemble learning [16]. In recent years, many studies have employed homogenous ensemble learning methods to achieve better performance in cancer survival prediction [8, 9]. However, previous studies have shown that the diversity among the base learners in the heterogeneous ensemble learning methods is higher than the base learners in the homogeneous ensemble learning methods, which has greater potential to achieve higher accuracy [17]. Moreover, the heterogeneous ensemble learning method can reduce the deviation of each base learner due to its inherent assumption of the heterogeneous forms, so the unseen samples get better generalization [17]. As a result, heterogeneous ensemble learning has been widely studied in various fields. Thongkam et al. [18] devised a heterogeneous ensemble learning method combining Adaboost and RF to predict the survival status of breast cancer patients, and showed that the method outperforms Adaboost, RF, and other combined classifiers. Cho and Won [19] combined four different base classifiers using the majority voting strategy and verified the validity of the model using three benchmark cancer datasets. Although the above studies apply the heterogeneous ensemble learning method to predict the survival status of breast cancer patients, they only combine different weak classifiers through relatively simple ensemble mechanisms, resulting in the performance of the developed model not being so good. Consequently, many researchers have tried a variant of the ensemble learning method, known as the stacked ensemble in the literature, with a view to achieving better performance and higher accuracy.

Wolpert [20] first proposed “stacked generalization”, which is a variant of the ensemble learning method that integrates heterogeneous learners through multi-stage. Since then, many researchers in many different fields have paid great attention to the stacked ensemble learning method. Xiao et al. [21] proposed a two-stage stacked ensemble learning model based on deep learning to predict tumor properties by RNA-seq gene in lung adenocarcinoma, gastric cancer, and breast invasive cancer patients. The base learners used in the first stage included the *k*-nearest neighbors (KNN), SVM, DT, RF, and gradient boosting tree. The second stage used a five-layer neural network as meta learner. Chungsoo et al. [6] proposed a two-stage stacked ensemble learning model to predict the cause of death according to the patient’s last medical checkup, where the first stage included two base learners [lasso logistic regression (LLR) and gradient boosting (GB)] and one meta learner (XGBoost) is used in the second stage. Zhai and Chen [4] proposed a two-stage stacked ensemble learning model to predict the daily average PM 2.5 concentration in Beijing, China. The four base learners used in the first stage were Lasso, Adaboost, XGBoost, and MLP, while the second stage applied SVM for ensemble construction. Anifowose et al. [22] and Ali et al. [5] constructed stacked ensemble learning models using different types of SVMs as the learners to solve the respective studied problems.

From the above studies, we see that the stacked ensemble learning model shows good performance in various fields. Despite this, the above studies do not consider how the overall performance of the stacked ensemble learning model can be improved by improving the performance of the selected learners. Specifically, it can be observed that SVM and NN are commonly selected as learners to carry out stacked integration, but the above studies do not consider how to determine the hyperparameters of SVM and the appropriate optimizer of NN.

Specifically, SVM is a pattern recognition method based on the principle of structural risk minimization, and its performance is closely related to its kernel parameters and penalty factors. Therefore, how to select appropriate hyperparameters is the key to improving the accuracy of the SVM classifier. Existing studies have used evolutionary algorithms, such as the genetic algorithm and particle swarm optimization, to search for the optimal hyperparameters of SVM [23, 24], which can improve the performance of SVM. On the other hand, NN is an algorithm inspired by the biological nervous system based on the multi-layer network structure and its performance is closely related to its optimizer. Many excellent optimizers of NN have proposed that in recent years, such as Adam, RAdam [25], LookAhead, nesterov accelerated gradient (NAG) and stochastic gradient descent (SGD). Inspired by the above studies, we try to apply the quantum particle swarm optimization algorithm to optimize the hyperparameters of SVM, and explore different optimizers (Adam, RAdam, SGD, NAD, LookAhead) to obtain high-quality learners, thus enhancing the performance of the stacked ensemble model.

Table 1 summarizes the differences between most of the current research and our research in terms of issues, ensemble methods, number and quality of learners, and performance of the proposed methods. To the best of our knowledge, little previous work has considered a heterogeneous stacked ensemble learning model for predicting the survival status of cancer patients. Meanwhile, few studies have tackled the drawbacks of learners in ensemble construction, which hinder the ensemble model’s performance, and our study bridges these gaps.

Materials and methods

Survival prediction is an important branch of cancer prognosis, which predicts the vital characteristics of cancer patients

Table 1 Literature overview: studies of ensemble methods

Study	Problem	Ensemble method		Stacked ensemble	Improved learner	Learner number	Ensemble member	Performance
		Heterogeneous	Homogeneous					
Wang et al. [9]	Cancer prognosis		√	×	×	21	DT	→ RF, RSB, GB, ADB, RT
Wang et al. [8]	Breast cancer diagnosis		√	×	×	12	SVM	→ SVM, NB, NN, WVBE
Ali et al. [5]	Disease diagnosis		√	√	×	2	L1-SVM, L2-SVM	→ Adaboost, RF, ET
Thongkam et al. [18]	Breast cancer diagnosis	√		×	×	2	RT, ADB	→ SVM, RT, ADB
Cho and Won [19]	Cancer classification	√		×	×	4	MLP, KNN, SVM, SASOM	→ MLP, KNN, SVM, SASOM
Chungsoo et al. [6]	The case of death	√		√	×	3	LLR, GB, XGBoost	→ LLR, GB
Xiao et al. [21]	Cancer diagnosis	√		√	×	6	KNN, SVM, DT, RF, GBDT, DNN	→ KNN, SVM, DT, RF, GBDT, MV
Adem et al. [26]	Cancer diagnosis	√		√	×	6	KNN, SVM, DT, FFNN, RoF, SC	→ KNN, SVM, DT, FFNN, RoF, SC
Bashir et al. [27]	Cancer prognosis	√		×	×	5	NB, DTG, DTI, SVM, MBL	→ NB, DTG, DTI, SVM, MBL
Velusamy and Ramasamy [28]	Disease diagnosis	√		×	×	3	KNN, RF, SVM	→ RF, KNN, SVM;
This study	Cancer prognosis	√		√	√	5	BKNN, QSVM, AMLP, XGB, RMLP	→ DT, LR, SVM, NB, KNN, BKNN, RF, ADB, XGB, LGB, QSVM, GSVM

“ → ” signifies that the proposed method is better than these models

RF random forest, *RSB* random subspace, *GB* gradient boosting, *ADB* Adaboost, *RT* regression tree, *SVM* support vector machine, *L1-SVM* L1 regularized SVM, *L2-SVM* L2 regularized SVM, *ET* extra tree, *NB* naïve Bayes, *NN* neural network, *MLP* multi-layer perception, *DT* decision tree, *KNN* *k*-nearest neighbors, *WVBE* weighted vote-based ensemble, *SASOM* structure adaptive self-organizing map, *MV* majority voting, *FFNN* feed forward neural network, *RoF* rotation forest, *SC* Softmax classification, *BSVM* bagging SVM, *GBDT* gradient boosting decision tree, *GPC* Gaussian process classifier, *BGPC* bagging Gaussian process classifier, *LS* Losso, *XGB* XGBoost, *LGB* LightGBM, *QSVM* quantum particle swarm optimization-based SVM, *GSVM* genetic algorithm-based SVM, *DTG* decision tree using the Gini index, *DTI* decision tree using information gain, *MBL* memory-based learner within a certain period after diagnosis. Many studies set the survival threshold at five years [8, 9], i.e., if a patient is still living in five years (60 months), the case is "alive"; otherwise, it is "dead". In this study, we propose PKSFS and BQAXR to predict the survival status (alive or dead) of different cancer patients, which aims to better understand how patients' cancer is likely to worsen after treatment in the future. First, we apply for and obtain multiple real cancer datasets, and preprocess them. Subsequently, PKSFS is used to obtain the optimal feature subset from high-dimensional cancer datasets, and BQAXR is trained and tested on the cancer datasets with the optimal feature subsets. Finally, the testing results are used to evaluate the performance of BQAXR in terms of machine learning and statistical indicators. The above contents are described in detail as follows.

Data preparation

The quality of data drastically affects the performance of the machine learning models. Therefore, data preparation is an important stage in machine learning, which commonly occupies 80% of the time in the whole machine learning analysis. In this subsection, we introduce the data used in this study and the data pre-processing process.

Data acquisition

In this study, we obtain the real gastric cancer dataset and skin cancer dataset from the Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute (http://www.seer.cancer.gov). SEER collects cancer incidence data from population-based cancer registries that cover 26% of African Americans, 38% of Hispanic Americans, 44% of American Indians and Alaskan Natives, 50% of Asians, and 67% of Hawaiian/Pacific Islanders [15]. At present, the SEER database has been widely used in various analytical research projects [2, 8, 9, 15]. We select the data files from 1973 to 2016 stored in the cancer incidence database, in which the gastric cancer dataset consists of 112,139 records and 146 variables, and the skin cancer dataset consists of 23,624 records and 123 variables. These variables can be divided into seven categories, namely record identification, patient socio-demographics, description of neoplasm, follow-up information, recoded variables, therapy, and case source, which contain detailed information about gastric cancer and skin cancer cases. Since the number of variables in the SEER cancer dataset is as large as 123, we provide main variable names and descriptions in Table 12 in appendix, and more information can be found on the website.

Data pre-processing

Not all the variables and samples can be used for model training since there are problems such as missing values and imbalances between categories, etc. in the original data. To solve the above problems, we adopt some strategies to deal with the original data, which are as follows: (i) variables with attribute missing values greater than 50% are removed from the original datasets, based on which the variables (such as "LYMPH", "RXSSRLNS", "METS", "L2005" etc.) are removed; (ii) variables with only one attribute value and directly related to the patient's survival status are removed from the original datasets, based on which the variables (such as "SCHEMA", "EXTENSION", "SM", etc.) are removed; (iii) samples without determined diagnostic dates and end of follow-up dates are removed from the original datasets; (iv) samples that cancer is not the cause of death and repeated samples are also excluded from the original datasets; (v) the two cancer datasets are imbalanced data which will be processed as balanced datasets using undersample based on the condensed nearest neighbor method [29]; (vi) Wang et al. [9] noted that the variable "ID" contains additional information about the patient, while the collinearity indeed exists between "ID" and other variables, thus the variable "ID" is removed. Finally, 26 (resp., 30) variables and 1165 (resp., 2328) samples are obtained as the experimental data in the gastric (resp., skin) cancer dataset. For target, we denote by "VS" the label for this experiment, which represents the survival status of each cancer patient, including alive (denoted by 0) and dead (denoted by 1).

Due to the types of many variables (such as "gender", "rank" etc.) are all nominal variables in the original datasets, thus one-hot coding is employed to deal with these variables and develop a high-dimensional sparse matrix, where there are 125 (reps., 114) features in the gastric (reps., skin) cancer dataset. Moreover, the types and ranges of the attribute values of some features are different. If the differences are too large in the original dataset, the training model will give high weights to the attributes with high values, which will make the model generate false perceptions and time-consuming. Therefore, the data need to be dimensionless before training, i.e., converting the attribute values of all the features in the original data into uniform specifications, which can speed up the convergence of the algorithm and make the performance of the model more stable. After that, we apply the z-score standardization method for the two datasets. Finally, following Ahmadi et al. [30], we choose 70% of the pre-processed data as the training dataset and the remaining data as the testing dataset.

Table 2 Feature scores using PKSFS in the gastric cancer dataset

N	Feature name	Score	*N*	Feature name	Score
1	EOD_E	1	13	Race_C	0.32
2	YoB	0.78	14	Seer_rKentucky—2000 +	0.31
3	Race_Korean (1988 +)	0.74	15	SEERH_Distant	0.26
4	AAD	0.68	16	Race_White	0.23
5	SEERH_Localized	0.65	17	SN_2nd	0.2
6	RnCDS_Surgery performed	0.59	18	RnCDS_Not recommended	0.2
7	EOD_N	0.54	19	SS2000_Distant	0.2
8	MSAD_Married	0.52	20	Seer_registry_Greater Georgia—2000+	0.1
9	EOD_S	0.47	21	Race_Filipino	0.07
10	SS2000_Localized	0.42	22	FMPI_Yes	0.07
11	NHIA_South or CAeB	0.41	23	Seer_registry_Los Angeles—1992+	0.06
12	NHIA_NSHL	0.34	24	FMPI_No	0.04

Priori knowledge- and stability-based feature selection

A classification model can benefit from feature selection in the following two ways. First, by transforming the original feature subset from a high-dimensional space to low-dimensional space, the computational complexity of the classification model construction process is significantly reduced. Second, removing the invalid or redundant features in the original feature set reduces their adverse effects on the classification model construction process, such as overfitting and low classification accuracy [31, 32]. Therefore, the goal of feature selection is to transform the dimensional space and remove the redundant features from the original feature set, which will not only make the classification better, but also improve the efficiency of training and testing. High-dimensional input variables may cause the model to be extremely unstable during training, and multi-collinearity during model fitting [33]. Given a data set with m features, $2^m - 1$ feature subsets can be generated. If we select the best feature subset among all the $2^m - 1$ feature subsets, it will take considerable time and manpower when m is large. Therefore, an effective and stable feature selection method should be applied to determine the optimal feature subset and facilitate the construction of subsequent models. In this study, we propose PKSFS, i.e., the priori knowledge- and stability-based feature selection method, which takes full advantage of priori knowledge and stability, to determine the best feature subset, where prior knowledge about the feature subsets with the information gain is greater than 0 can be obtained to guide stability selection, and the stability of the method, which effectively measures how different training subsets affect feature preferences, can be assessed in the form of weight scores.

Specifically, PKSFS first obtains the information gain of each feature, and leaves the features with the information gain greater than 0 from the feature sets, and generates a new feature subset. The bootstrap and LR L1-based regularized method are then applied to the new feature subsets in the following way: (i) randomly select a feature subset from the new feature set, (ii) randomly perform bootstrap in the training samples to obtain the training sub-dataset, and (iii) estimate the performance of each feature subset in the training sub-dataset using the LR L1-based regularized method, which scales the penalty of a random feature subset of coefficients in the evaluation process. This process is repeated a certain number of times, and the repeatedly selected features are retained. In brief, the more frequently a feature is selected, the more important it is, and the higher the probability it is retained, and the final retained feature is stable and less sensitive to regularization decisions. In this way, PKSFS not only can select a high-quality feature subset to guide the construction of BQAXR, but also has high efficiency, since the first step of PKSFS filters some features with low information gain and greatly improves the efficiency of the second step. Tables 2, 3 show the scores of each retained feature obtained from PKSFS in gastric cancer and skin cancer dataset respectively, where for some pre-processed nominal variables, their names are written as “variable name_attribute name”.

A two-stage heterogeneous stacked ensemble learning method

For ensemble construction, diversity and consistency between the base learners are the two key factors, indicating that the base learners should be “good and different”. However, for the homogeneous ensemble learning method, the base learner generally is composed of multiple identical

Table 3 Feature scores using PKSFS in the skin cancer dataset

N	Feature name	Score	*N*	Feature name	Score
1	YoB	0.9	12	SEX_Female	0.2
2	GRADE_PD; GIII	0.75	13	YoD	0.19
3	FMPI_Yes	0.51	14	MSaD_Widowed	0.12
4	FMPI_No	0.51	15	PSL_C44.6-Skin of upper limb	0.1
5	GRADE_ IV	0.47	16	GRADE_Well differentiated; Grade I	0.09
6	AAD	0.45	17	SN_One primary only	0.07
7	SN_2nd of 2 or more primaries	0.27	18	MoD_January	0.07
8	MSaD_Married (including common law)	0.24	19	EXTENT	0.07
9	SEERH_Distant	0.23	20	RN	0.06
10	GRADE_ II	0.23	21	NHIA _Non-Spanish-Hispanic-Latino	0.04
11	SEX_Male	0.21	22	PSL_C44.4-Skin of scalp and neck	0.04

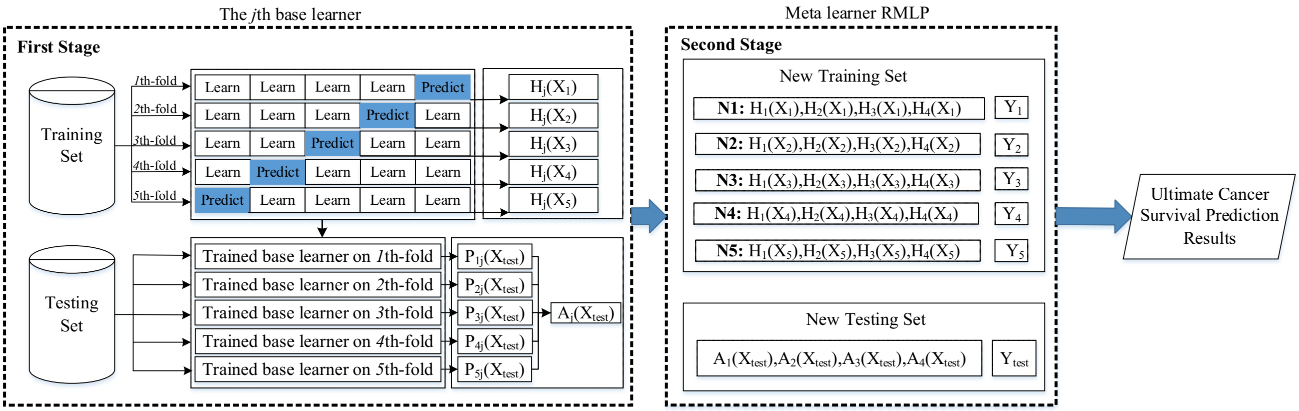

The jth base learner
First Stage
Training Set
1th-fold
2th-fold
3th-fold
4th-fold
5th-fold
Learn
Predict
H_j(X_1)
H_j(X_2)
H_j(X_3)
H_j(X_4)
H_j(X_5)
Testing Set
Trained base learner on 1th-fold
Trained base learner on 2th-fold
Trained base learner on 3th-fold
Trained base learner on 4th-fold
Trained base learner on 5th-fold
P_1j(X_test)
P_2j(X_test)
P_3j(X_test)
P_4j(X_test)
P_5j(X_test)
A_j(X_test)
Meta learner RMLP
Second Stage
New Training Set
N1: H_1(X_1),H_2(X_1),H_3(X_1),H_4(X_1)
N2: H_1(X_2),H_2(X_2),H_3(X_2),H_4(X_2)
N3: H_1(X_3),H_2(X_3),H_3(X_3),H_4(X_3)
N4: H_1(X_4),H_2(X_4),H_3(X_4),H_4(X_4)
N5: H_1(X_5),H_2(X_5),H_3(X_5),H_4(X_5)
Y_1
Y_2
Y_3
Y_4
Y_5
New Testing Set
A_1(X_test),A_2(X_test),A_3(X_test),A_4(X_test)
Y_test
Ultimate Cancer Survival Prediction Results

Fig. 1 The framework of BQAXR

algorithms, which will weaken the diversity of the ensemble learning method. In other words, the same structure of the base learners makes it difficult for them to overcome their weaknesses and drawbacks. For the heterogeneous ensemble learning method, although heterogeneous learners greatly improve the diversity of the model, the weaknesses and shortcomings of the base learners have not been improved, which will hamper the effectiveness of the ensemble models. Therefore, we develop the two-stage heterogeneous stacked ensemble learning method BQAXR, which integrates heterogeneous base learners in two-stage through the stacked generalization strategy. In BQAXR, base learners consist of four different algorithms, which can further improve the diversity between base learners, and the shortcomings of four heterogeneous base learners are improved, thus achieving an improvement in cancer survival prediction accuracy. Figure 1 shows the overall framework of BQAXR, which mainly consists of four heterogeneous base learners in the first stage, i.e., the bagging algorithm based on the k-nearest neighbors algorithm (BKNN), support vector machine based on the quantum particle swarm optimization algorithm (QSVM), multi-layer perception based on the adaptive moment estimation optimizer Adam (AMLP), extreme gradient boosting (XGBoost), and one meta learner in the second stage, i.e., multi-layer perception based on the rectified Adam optimizer RAdam (RMLP).

Specifically, in the first stage, we randomly split the pre-processed training dataset into five subsets D_1, D_2, D_3, D_4, andD_5, where $D_i = (X_i, Y_i)$, $i = 1,\ldots,5$, and X_i is the ith training subset with the best features obtained by PKSFS and Y_i is the label set corresponding to the ith training subset. We denote by $Z = (X_{\text{test}}, Y_{\text{test}})$ the pre-processed testing dataset. After that, the base learners BKNN, QSVM, AMLP, and XGBoost, denoted as jth base learner ($j = 1, 2, 3, 4$), respectively, will, in turn, perform fivefold cross-validation, i.e., each base learner will be trained and predicted five times. As a consequence, a total of 20 (nfold × 4) operations are conducted in the first stage, where nfold denotes the number of folds. Let $H_j(X_i)$ be the prediction result of the jth base learner after the ith-fold cross-validation, where the ith-fold cross-validation means that the training subset D_i is used as the testing set, and the remaining subsets are used as the training set, and $H_j(X_i) \in [0, 1]$. After all the base learners have completed the fivefold cross-validation, $N_i = (H_1(X_i), H_2(X_i), H_3(X_i), H_4(X_i))$ is used to denote all predictions of all base learners in the ith-fold cross-validation, that is the predictions of all base learners on D_i. At the same time, after each base learner performing the ith-fold cross-validation, the trained base learner will make predictions in the pre-processed testing dataset, and the prediction result of the trained jth base learner is denoted by $P_{ij}(X_{\text{test}}) i = 1, 2, 3, 4, 5 j = 1, 2, 3, 4$, where $P_{ij}(X_{\text{test}}) \in [0, 1]$. Therefore, the subset of the prediction results for the jth base learner in each fold is recorded as $[P_{1j}(X_{\text{test}}), P_{2j}(X_{\text{test}}), P_{3j}(X_{\text{test}}), P_{4j}(X_{\text{test}}), P_{5j}(X_{\text{test}})]$, and the average of the five prediction results A_j is used as the final prediction result $(A_j(X_{\text{test}}) = 1/5\Big[P_{1j}(X_{\text{test}}) + P_{2j}(X_{\text{test}}) + P_{3j}(X_{\text{test}}) + P_{4j}(X_{\text{test}}) + P_{5j}(X_{\text{test}})\Big])$.

Upon completion of the first stage, we obtain a new training set $D_{\text{newtrain}} = (X_{\text{newtrain}}, Y_{\text{newtrain}})$, where $X_{\text{newtrain}} = \{N_1, N_2, N_3, N_4, N_5\}$, $Y_{\text{newtrian}} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$; a new testing set $D_{\text{newtest}} = (X_{\text{newtest}}, Y_{\text{newtest}})$, where $X_{\text{newtest}} = \{A_1(X_{\text{test}}), A_2(X_{\text{test}}), A_3(X_{\text{test}}), A_4(X_{\text{test}})\}$, Y_{newtest} is Y_{test}. In the second stage. We train RMLP on the new training set, apply the trained meta learner to predict the survival status of cancer patients on the new testing set, and finally output the prediction results. The theory of the four base learners of the first stage and the meta learner of the second stage are elaborated in detail as follows.

Base learners pool in the first stage

k-nearest neighbors algorithm based on bagging algorithm (BKNN): k-nearest neighbors algorithm (KNN) has been shown to be one of the popular choices in stacked ensemble learning [21]. In the stacked ensemble model, the differences in the performance of the used learners should not be too large, otherwise, the stacked generalization strategy for heterogeneous learners will perform poorly. Through our numerical studies in the two cancer datasets, we find that the performance of the KNN is worse than that of the other three base learners QSVM, AMLP, and XGBoost in the first stage, which will decrease the effect of stacked ensemble construction to some extent. Therefore, we propose an improved k-nearest neighbors algorithm (BKNN) as one of the base learners in the first stage, which can significantly improve the performance of KNN and narrow the performance gap with the other three base learners. In BKNN, we combine the bagging algorithm, which is an ensemble algorithm composed of multiple independently weak classifiers [34], with the k-nearest neighbors. The basic idea of BKNN is to randomly extract sample subsets (bootstrap) from the original sample set as training sample sets. After that, each KNN is trained independently using the different training sample sets, where the Euclidean distance is used to measure the distances between the samples. The results of all KNNs are summarized using weighted voting.

Support vector machine based on quantum particle swarm optimization (QSVM): SVM is one of the most widely used learners in the stacked ensemble model [4, 5, 21, 22]. The classification performance of SVM is mainly determined by two key factors. One is the selection of the kernel function and the other is the determination of the hyperparameters. To solve these two problems, we propose the improved support vector machine QSVM to find the best SVM as a base learner for stacked ensemble construction, since the excellent performance of the base learner can effectively improve the stacked effect.

To search for the hyperparameters C and γ, we use the quantum particle swarm optimization algorithm (QPSO) developed by Sun et al. [35], which is quantum-behaving inspired by particle swarm optimization trajectory analysis and quantum mechanics. It effectively solves the deficiency of particle swarm optimization (PSO) that global convergence is not guaranteed due to the redundant parameters in PSO. In PSO, the state of the particle is determined by the parameters' positions and velocities, while the state of the particle is determined by the wave function $W_{(v,t)}$ in QPSO, where $W_{(v,t)}$ denotes the energy and momentum of particle v at time t. Thus, QPSO can effectively reduce the number of parameters to reduce the sensitivity of the algorithm for parameters. The probability density function $\left|W_{(v,t)}\right|^2$ is applied to obtain the probability distribution function of the particle's position, the form of which depends on the potential field where the particles are located, and particles move according to the following iterative equation:

$$v_{i,j}(t+1) = p_{i,j} + \beta \times \left|\text{mbest}_j(t) - v_{i,j}(t)\right| \times \ln\left(\frac{1}{u_{ij}(t)}\right), \text{ if } z \geq 0.5 \tag{1}$$

$$v_{i,j}(t+1) = p_{i,j} - \beta \times \left|\text{mbest}_j(t) - v_{i,j}(t)\right| \times \ln\left(\frac{1}{u_{ij}(t)}\right), \text{ if } z < 0.5 \tag{2}$$

$$\text{mbest}_j(t) = 1/M \sum_{i=1}^{M} \text{pbest}_{i,j}(t), \tag{3}$$

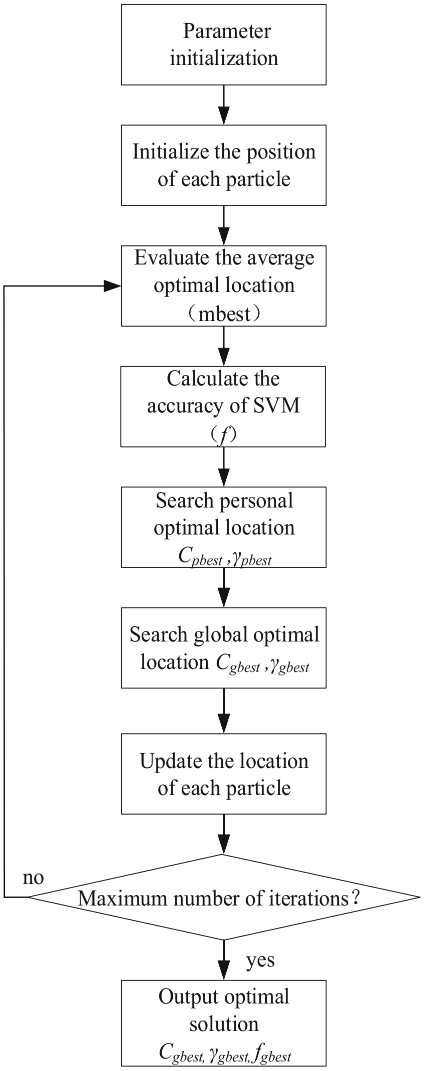

Parameter initialization
Initialize the position of each particle
Evaluate the average optimal location （mbest）
Calculate the accuracy of SVM （f）
Search personal optimal location $C_{pbest}, \gamma_{pbest}$
Search global optimal location $C_{gbest}, \gamma_{gbest}$
Update the location of each particle
no
Maximum number of iterations?
yes
Output optimal solution $C_{gbest}, \gamma_{gbest}, f_{gbest}$

Fig. 2 A flowchart of QSVM

$$p_{i,j}(t) = \delta_{i,j}(t) \times \text{pbest}_{i,j}(t) + \left(1 - \delta_{i,j}(t)\right) \times \text{gbest}_j(t), \tag{4}$$

where mbest is the average optimal location, which represents the mean of all the optimal locations (pbest) of each particle among the population, and gbest denotes the global optimal location and M is the size of the population. The coefficients z, u, and δ are randomly generated numbers using the uniform probability distribution in the range [0,1], respectively. The parameter β is called the contraction–expansion coefficient, which can be used to control the convergence rate of the algorithm and it is the only parameter in QPSO. Figure 2 presents the framework of QSVM. The values of the parameters of QSVM are set as follows: the number of particles = 20, number of iterations = 100, β=0.7, and the ranges of C and γ are both [1, 100] and [1,1000] respectively.

Multi-layer perception based on adaptive moment estimation optimizer (AMLP): MLP based on the adaptive moment estimation optimizer (AMLP) is MLP combined with the adaptive moment estimation method Adam, which can calculate the adaptive learning rate of each parameter. In the structure of MLP, the leftmost layer is the input layer, and the number of neurons input is equal to the number of features. The rightmost layer is the output layer, which is responsible for generating the prediction values. The middle layer is a hidden layer consisting of hidden neurons. The neurons between the layers pass through a non-linear activation function. The common activation functions include Sigmoid, ReLu, Tanh, etc., among which ReLu learns much faster in MLP than the other activation functions [21], hence, we use ReLu as the activation function in AMLP. Since cancer survival prediction is a binary-class prediction problem, the loss function of AMLP is the binary cross entropy. Thus, the goal of AMLP is to use the Adam optimizer to update the parameters W and b in the MLP to minimize the binary cross entropy.

Adam is an integrated algorithm based on momentum and root mean square prop (RMSprop), which consists of the variable $v(v^t_{dW}, v^t_{db})$ given in momentum, and the weighted moving average variable S (S^t_{dW}, S^t_{db}), given in the RMSprop. First, AMLP initializes all the variables v^o_{dW}, v^o_{db}, S^o_{dW}, andS^o_{dW} to be zero. During the tth iteration, all the variables are updated as follows:

$$\begin{cases} v^t_{dW} = \beta_1 v^{t-1}_{dW} + (1-\beta_1)dW \\ v^t_{db} = \beta_1 v^{t-1}_{db} + (1-\beta_1)db \\ S^t_{dW} = \beta_2 v^{t-1}_{dW} + (1-\beta_2)(dW)^2 \\ S^t_{db} = \beta_2 v^{t-1}_{db} + (1-\beta_2)(db)^2, \end{cases} \tag{5}$$

where β_1 and β_2 are two hyperparameters, being the first moment and second moment, respectively, both of which are within [0,1], and (dW, db) and $((dW)^2, (db)^2)$ denote the differential and the exponentially weighted average of squares, respectively. Following Kingma and Ba [36], we set $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

After updating the hyperparameters β_1 and β_2 using momentum and RMSprop, respectively, AMLP calculates the bias corrections for all the variables as follows:

$$\begin{cases} v^{t,\text{corrected}}_{dW} = \frac{v^t_{dW}}{1-\beta^t_1} \\ v^{t,\text{corrected}}_{db} = \frac{v^t_{db}}{1-\beta^t_1}, \end{cases} \tag{6}$$

$$\begin{cases} S_{dW}^{t,\text{corrected}} = \frac{v_{dW}^{t}}{1-\beta_2^t} \\ S_{db}^{t,\text{corrected}} = \frac{v_{db}^{t}}{1-\beta_2^t}, \end{cases} \tag{7}$$

Finally, AMLP uses the bias-corrected variables to update the parameters W and b according to the following rule in the tth iteration:

$$\begin{cases} W_t = W_{t-1} - \alpha \frac{v_{dW}^{t,\text{corrected}}}{\sqrt{S_{dW}^{t,\text{corrected}}+\varepsilon}}, \\ b_t = b_{t-1} - \alpha \frac{v_{db}^{t,\text{corrected}}}{\sqrt{S_{db}^{t,\text{corrected}}+\varepsilon}}, \end{cases} \tag{8}$$

where α represents the learning rate that ranges from 0 to 1, and ε is a value that prevents the denominator from being 0. Following Kingma and Ba [36], we set $\varepsilon = 10^{-8}$.

Based on preliminary computational tests, we construct a three-layer perception model, where the input layer and hidden layer include 24 and 8 neurons for the gastric cancer dataset, and include 22 and 16 neurons for the skin cancer dataset, respectively, and the output layer contains one neuron for both datasets. We set the initial value of the learning rate at 0.0001 and the number of iterations at 1000.

Extreme gradient boost classifier (XGBoost): Developed by Chen and Guestrin [37], XGBoost has been extensively applied in stacked ensemble learning as a base learner [4, 34]. XGBoost belongs to the tree ensemble model, which is a machine learning system based on the improved gradient boosting decision tree (GBDT) [38]. The basic idea of XGBoost is to develop a new decision tree in a gradient direction of the residuals to minimize the loss function. XGBoost supports both row sampling and column sampling, introduces a second-order Taylor expansion for the loss function, and uses second-order partial derivatives in training, which makes XGBoost converge faster.

The loss function of XGBoost consists of two parts, namely the training loss and the sum of the complexity of each tree, is as follows:

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \widehat{y_i}) + \sum_{h=1}^{H} \Omega(y_h), \tag{9}$$

where y_i denotes the actual value of the ith sample, $\widehat{y_i}$ is the predicted value of the ith sample, $\Omega(y_h)$ represents the regularization term generated by each tree h, and H is the number of trees.

XGBoost contains a large number of parameters. In this study we apply the Bayesian optimization algorithm to determine the values of the four key hyperparameters, i.e., n_estimators, max_depth, learning_rate, and gamma, of XGBoost in the two cancer datasets, where the search ranges for each of the four hyperparameters are [1,500], [1,15], [0,1] and [0,1].

Meta learner in the second stage

MLP based on the rectified Adam optimizer (RMLP) is an improved MLP developed in the second stage, which uses the new training and testing sets generated in the first stage for training and predicting. The developed RMLP combines the ideas of RAdam and MLP. RAdam is a new Adam variant proposed by Liu et al. [25] in 2019, which draws on the advantages of both Adam and SGD. It introduces a term that corrects the variance of the adaptive learning rate and dynamically turns the adaptive learning rate on or off according to the dispersion of the variance. Therefore, RMLP does not require adjusting the hyperparametric learning rate. Table 4 presents the pseudo-code that describes how RAdam updates the parameters in MLP. Based on preliminary computational tests, we construct a five-layer perception model for two cancer datasets, where the input layer includes 4 neurons, the first hidden layer includes 64 neurons, the second hidden layer is the dropout layer, and the third hidden layer includes 12 neurons.

Evaluation indicator

In this study, we use six machine learning classification indicators and three statistical indicators to assess the performance of the machine learning model BQAXR and comparison methods, including accuracy, recall, precision, F1-score, AUC, p-value, Cohen's kappa, and Matthews' correlation coefficient.

Among the six machine learning classification indicators, accuracy, recall, precision, F1-score, and AUC are closely related to the following four states: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These indicators can be calculated according to the four states, and the specific formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{12}$$

$$F_\beta - \text{measure} = \left(1 + \beta^2\right) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{13}$$

$$\mathrm{AUC} = 1 - \frac{1}{m^+m^-}\sum_{x^+\in M^+}\sum_{x^-\in M^-}\left(\left(f\left(x^+\right) < f\left(x^-\right) + \frac{1}{2}\left(f\left(x^+\right) = f\left(x^-\right)\right)\right)\right). \tag{14}$$

For $F_\beta -$ measure, following Mahajan et al. [39], we set the parameter $\beta = 1$ and denote it as the F1-score. For AUC, m^+ and m^- are positive and negative examples, and M^+ and M^- represent a collection of positive and negative examples, respectively.

The evaluation of a model by only machine learning indicators cannot fully reflect the scientificness and objectivity of the model. Therefore, in this study, we use three statistical indicators to assess the superiority of the model in statistics. The specific description of each statistical indicator is as follows:

(1) p-value: p-value is a measure of hypothesis testing results in statistics, which is obtained through conducting the paired t test on the performance of the algorithms. Specifically, there are statistical differences if the p value is less than 0.05 and vice versa.

(2) Cohen's kappa: A commonly used agreement measure, Cohen's kappa denotes the degree of agreement between the actual result and predicted results on the classification problem, which is calculated as follows [31]:

$$k = \frac{(p_0 - p_e)}{(1 - p_e)}, \tag{15}$$

where p_0 is the actual ratio and p_e is the theoretical ratio.

(3) Matthews' correlation coefficient (MCC): MCC is a measure of the quality of the binary-class model in machine learning, which is essentially a correlation coefficient value between -1 and + 1. MCC is calculated

Table 4 The pseudo-code for parameter updating in RMLP

```
The pseudo-code for parameter updating in RMLP
Input: Hyperparameters $\beta_1$ and $\beta_2$, and step size $\alpha_t$, $t = 1,2,\ldots,T$
Output: Parameters $W_t$ and $b_t$
Initialize the exponentially moving 1st moment and 2nd moment
Compute the maximum length of the approximated simple moving average $\rho_\infty \leftarrow \frac{2}{1-\beta_2} - 1$
while $t = 1,2,\ldots,T$ do
    Update the exponential moving 1st moment and 2nd moment according to Eq. (5)
    Compute bias-corrected moving average according to Eq. (6)
    Compute the length of the approximated simple moving average $\rho_t \leftarrow \rho_\infty - 2t\frac{\beta_2^t}{(1-\beta_2^t)}$
    If $\rho_t > 4$, then
        Compute the bias-corrected second moment estimation according to Eq. (7)
        Compute the variance rectification $r_t \leftarrow \sqrt{\frac{(\rho_t-4)(\rho_t-2)\rho_\infty}{(\rho_\infty-4)(\rho_\infty-2)\rho_t}}$
        Update the parameters $W_t \leftarrow W_{t-1} - \alpha_t r_t \frac{v_{dW}^{t,corrected}}{s_{dW}^{t,corrected}}$
        Update the parameters $b_t \leftarrow b_{t-1} - \alpha_t r_t \frac{v_{db}^{t,corrected}}{s_{db}^{t,corrected}}$
    else
        Update the parameters $W_t \leftarrow W_{t-1} - \alpha_t v_{dW}^{t,corrected}$
        Update the parameters $b_t \leftarrow b_{t-1} - \alpha_t v_{db}^{t,corrected}$
end while
Return $W_t$ and $b_t$
```

as follows [40]:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{16}$$

To avoid over-fitting of the model and better evaluate the generalization ability of the model, all the machine learning algorithms considered in the numerical studies apply five-fold cross-validation with the same settings, and the number of iterations is set at 20 to avoid contingency. The evaluation indicators introduced are used to evaluate the model performance. All the training and testing experiments are performed in the Python software and Python third-party libraries (Numpy, Pandas, etc.). The experimental device is an Intel CoreTM i7 processor @ 1.80 GHZ running under the Windows 10 16G operating system.

Numerical results

In this section, we present a series of analyses on numerical studies from different perspectives to assess the quality and performance of PKSFS and BQAXR, including by comparison of the different feature selection methods, comparison of the different ensemble mechanisms and stacked strategies, and comparison between BQAXR and advanced classification methods.

Comparison of different feature selection methods

Feature selection methods have a direct impact on the construction of a model. To verify that PKSFS can better guide the BQAXR's construction, we compare and analyze different feature selection methods including three traditional feature selection methods (filtering, wrapper, and embedded), without feature selection and the hybrid feature selection method proposed by Han et al. [41], respectively. Specifically, we select the information gain feature selection method (IG) from the filtering; the feature selection based on genetic algorithm (GA) from wrapper, and the feature selection method based on random forest (RF) from embedded. The hybrid feature selection method (HFS) proposed by Han et al. [41] is a feature selection method that combines filtering and embedded. In this paper, all the above feature selection methods are applied to construct the BQAXR model and carry out the comparative experimental analysis. The experimental results are listed in Table 5.

From Table 4, the differences in the performance of different feature selection methods in the two cancer datasets are discernible. The following analysis is made based on their performance in machine learning indicators. First, PKSFS has the best performance in all indicators, followed by the WFS, and the GA has the worst performance. However, it should be noted that the performance of WFS is like that of PKSFS, but the number of features for WFS is 125 (resp., 114) while the number of feature selection for PKSFS is 24 (resp., 22) in the gastric (resp., skin) cancer dataset, which indicates that PKSFS can reduce the complexity of the constructed model while guaranteeing the accuracy, thus better reduce the computational cost. The performance of HFS and GA is even worse than that of the WFS, which reveals that they are difficult to select the important features effectively from the high-dimensional datasets, resulting in the loss of too much important feature information. Specifically, PKSFS is 0.70% (resp., 1.35%), 2.09% (resp., 2.99%), and 0.30% (resp., 0.41%) higher in terms of accuracy, recall, and AUC in the gastric (resp., skin) cancer dataset, respectively, than the WFS. Through the above analysis, we conclude that PKSFS is a useful method to eliminate redundant features as well as select important features from high-dimensional datasets.

Comparison of different ensemble mechanisms

Although ensemble learning methods are attractive because of their good performance, the performance of different ensemble mechanisms can vary greatly. Thus, it is necessary to explore their performance operating in different ensemble mechanisms. The proposed method BQAXR is constructed under the stacked ensemble mechanism, and in this subsection, we first compare BQAXR with other four ensemble mechanisms: (i) Soft voting ensemble mechanism (SV): SV refers to totaling the predicted probabilities of the four base learners in each class label for a particular sample, and outputting the class label with a high probability as the predicted class label; (ii) Hard voting ensemble mechanism (HV): HV refers to selecting the class label with the most prediction results as the predicted label class from the class labels predicted by the four base learners for a particular sample, and the predicted class label of the last base learner XGBoost is chosen in case of a tie; (iii) Maximum ensemble mechanism (MaxE): MaxE refers to select the largest predicted value from the four base learners to make the final decision. (iv) Minimum ensemble mechanism (MinE): Like MaxE, MinE refers to select the smallest predicted value from the four base learners to make the final decision. We use the above four ensemble mechanisms to integrate four heterogeneous base learners BKNN, QSVM, AMLP, and XGBoost, and denote these four heterogeneous ensemble models as S, H, A, I. Table 6 presents the experimental results of these four heterogeneous ensemble models and BQAXR, where D_SP, D_HP, D_AP, D_IP denote the differences between BQAXR

Table 5 Experimental results on comparing different feature selection methods

Dataset	Indicator	WF	IG	GA	RF	HFS	PKSFS
Gastric cancer	Accuracy	0.8139	0.8142	0.7690	0.8090	0.7725	**0.8209**
	Recall	0.7891	0.7919	0.7771	0.8143	0.7665	**0.8100**
	Precision	0.8332	0.8217	0.7704	0.8204	0.7856	**0.8352**
	AUC	0.8173	0.8133	0.7712	0.8112	0.7736	**0.8203**
	Number of features	125	27	19	63	**15**	24
Skin cancer	Accuracy	0.8201	0.8233	0.8017	0.7984	0.8111	**0.8336**
	Recall	0.8611	0.8793	0.8336	0.8432	0.8474	**0.8910**
	Precision	0.8460	0.8223	0.8105	0.8112	0.8300	**0.8332**
	AUC	0.8173	0.8127	0.7977	0.8041	0.8053	**0.8214**
	Number of features	114	27	17	57	**16**	22

The bold value is the best performance on this metric

Table 6 Experimental results on comparing different ensemble mechanisms

Dataset	Indicator	S	H	A	I	BQAXR	D_SP	D_HP	D_AP	D_IP
Gastric cancer	Accuracy	0.8039	0.8063	0.7811	0.7910	**0.8209**	0.0170	0.0146	0.0398	0.0299
	Recall	0.8110	0.7919	**0.8672**	0.7252	0.8100	– 0.0010	0.0181	– 0.0572	0.0848
	Precision	0.8087	0.8325	0.7477	**0.8466**	0.8352	0.0265	0.0027	0.0875	– 0.0114
	F1-score	0.8099	0.8117	0.8030	0.7812	**0.8224**	0.0125	0.0107	0.0194	0.0412
	AUC	0.8040	0.8120	0.7754	0.7930	**0.8203**	0.0163	0.0083	0.0449	0.0273
Skin cancer	Accuracy	0.8232	0.8124	0.8030	0.8214	**0.8336**	0.0104	0.0212	0.0306	0.0122
	Recall	0.8664	0.8659	**0.8910**	0.8173	**0.8910**	0.0246	0.0251	0.0000	0.0737
	Precision	0.8291	0.8217	0.7932	**0.8733**	0.8332	0.0041	0.0115	0.0400	– 0.0401
	F1-score	0.8473	0.8432	0.8393	0.8444	**0.8611**	0.0138	0.0179	0.0219	0.0168
	AUC	0.8152	0.8093	0.7874	**0.8214**	**0.8214**	0.0089	0.0148	0.0367	0.0028

The bold value is the best performance on this metric

S SV, *H HV*, *A MaxE*, *I MinE*, *D_SP* the differences between BQAXR and SV, *D_HP* the differences between BQAXR and HV, *D_AP* the differences between BQAXR and MaxE, *D_IP* the differences between BQAXR and MinE

and S, H, A, I, respectively. In two cancer datasets, the overall performance of BQAXR in five indicators is better than the other four heterogeneous ensemble models.

Comparison of different stacked strategies

In this study, we expect to further improve the performance of the stacked ensemble model by obtaining high-quality learners, thus the shortcomings of the learners are improved in BQAXR. Therefore, in addition to different integrated learning methods for comparative analysis, it is very necessary to verify that improved learners can be more beneficial for cancer survival prediction.

In this subsection, we compare and analyze the stacked ensemble models built by the improved learners and the stacked ensemble models built by the unimproved learners respectively, where the unimproved version consists of KNN, SVM, MLP and XGBoost in the first stage, and the improved version consists of BKNN, SVM, AMLP, and XGBoost. In the second stage, SVM, RMLP and LR are selected as meta learners to construct six stacked ensemble models with different structures, respectively (see Table 7 for details), where SVM and LR are chosen as the meta learner for comparison because they are widely used in existing studies related to stacked ensemble learning [26, 42]. Furthermore, Fig. 3 shows the performance of BKNN, QSVM, AMLP and XGBoost as base learners in the first stage and LR, SVM and RMLP as meta learners in the second stage, respectively, in terms of the five machine learning indicators.

The first stage consists of base learners, and the second stage consists of one meta learner.

Based on the experimental results listed in Table 6 and shown in Fig. 3, the following observations can be drawn:

Table 7 Experimental results on comparing different stacked strategies

Dataset	First stage	Second stage	Accuracy	Recall	Precision	F1-score	AUC
Gastric cancer	KNN + SVM + MLP + XGB (unimproved)	LR	0.7897	0.8000	0.7933	0.7967	0.7893
	BKNN + QSVM + AMLP + XGB (improved)		0.8130	0.8080	0.8240	0.8160	0.8130
	KNN + SVM + MLP + XGB (unimproved)	SVM	0.7868	0.7889	0.7955	0.7922	0.7868
	BKNN + QSVM + AMLP + XGB (improved)		0.8060	0.7940	0.8240	0.8090	0.8070
	KNN + SVM + MLP + XGB (unimproved)	RMLP	0.7954	0.7972	0.8039	0.8005	0.7954
	BKNN + QSVM + AMLP + XGB (improved)		**0.8209**	**0.8100**	**0.8352**	**0.8224**	**0.8203**
Skin cancer	KNN + SVM + MLP + XGB (unimproved)	LR	0.7971	0.8812	0.7911	0.8337	0.7818
	BKNN + QSVM + AMLP + XGB (improved)		0.8200	0.8810	0.8200	0.8490	0.8090
	KNN + SVM + MLP + XGB (unimproved)	SVM	0.8020	0.8713	0.8037	0.8361	0.7904
	BKNN + QSVM + AMLP + XGB (improved)		0.8140	0.8710	0.8190	0.8440	0.8040
	KNN + SVM + MLP + XGB (unimproved)	RMLP	0.8070	0.8713	0.8008	0.8341	0.7870
	BKNN + QSVM + AMLP + XGB (improved)		**0.8336**	**0.8910**	**0.8332**	**0.8611**	**0.8214**

The bold value is the best performance on this metric

i. In the two cancer datasets, no matter which meta learner is selected in the second stage, the improved base learners in the first stage have better performance than the unimproved base learners.

ii. In both cancer datasets, regardless of whether the improved base learners are selected in the first stage, when the RMLP is selected as the meta learner in the second stage, the effect of the stacked ensemble model has the best performance, followed by SVM and LR with the worst performance.

iii. In the gastric (resp., skin) cancer dataset, BQAXR is 2.55% (resp., 2.66%), 1.28% (resp., 1.97%) and 2.49% (resp., 3.44%) higher in accuracy, recall and AUC, respectively, than KSMXR (first stage: KNN + SVM + MLP + XGBoost; second stage: RMLP), the best performing of all comparison models.

In the second stage, RMLP is proposed as a meta learner to train and test the new dataset generated in the first stage. Optimizer is one of the important factors affecting the quality of multi-layer perception. To verify that the multi-layer perceptron based on RAdam can produce a better stacking effect, we choose BKNN, QSVM, AMLP and XGBoost as the base learners in the first stage, and explore the performance generated by MLP with five different optimizers as the meta learners in the second stage. The five optimizers include stochastic gradient descent (SGD), Nesterov momentum gradient optimizer (Nesterov accelerated gradient descent, NAG), Adam, RAdam and Ranger (https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer). Figure 4 shows the changes (200 epochs) in the accuracy of the five different optimizers in the two cancer datasets from left to right, respectively. In the gastric cancer dataset (Fig. 4a), RAdam is better than Adam and SGD, followed by NAG, and the worst is Ranger. In the

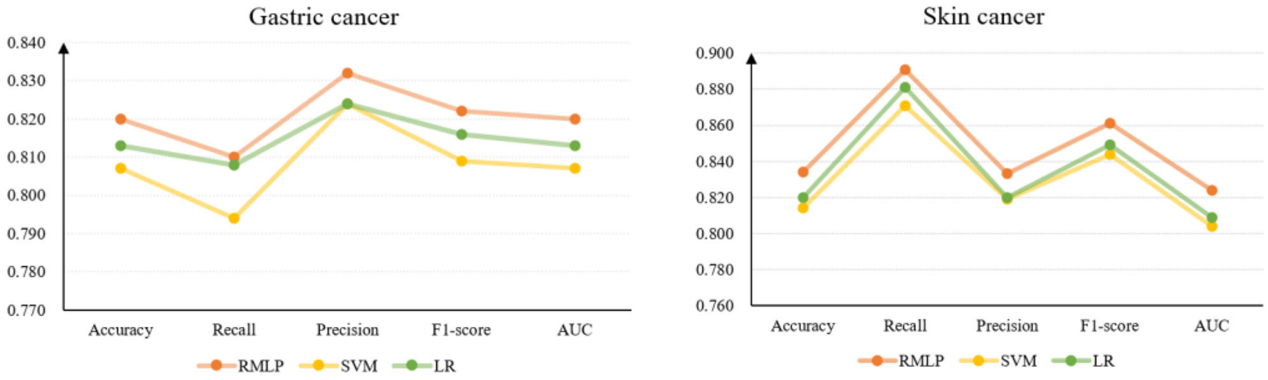
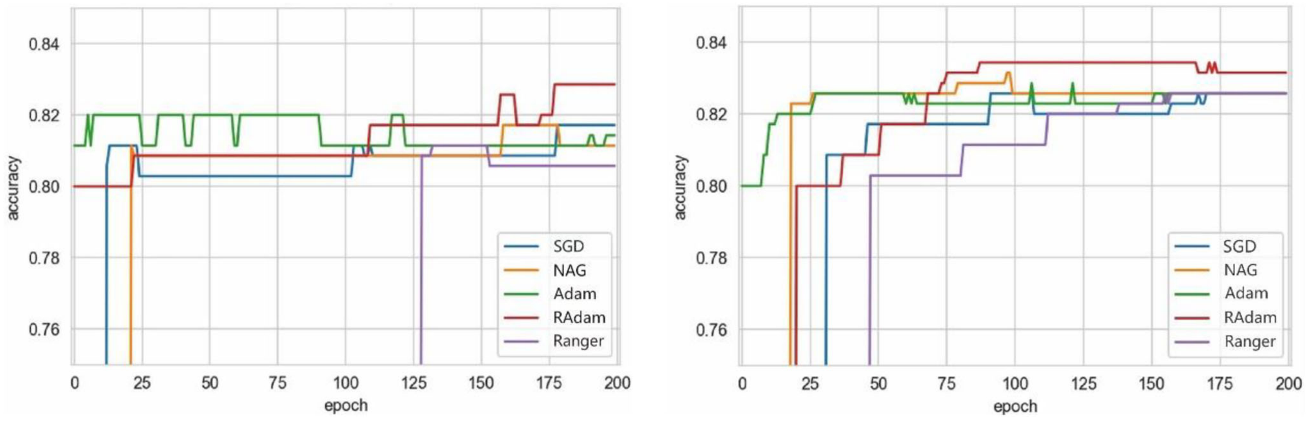


Fig. 3 Effects of different meta learners on the second stage



(a) Gastric cancer

(b) Skin cancer

Fig. 4 Effects of 200 epochs with different optimizers

skin cancer dataset (Fig. 4b), RAdam is significantly better than the other optimizers after 75 epochs, and the accuracy of NAG, SGD, Adam, and Ranger tends to be similar and stable after 175 epochs.

Comparison between BQAXR and advanced classification methods

In this subsection, we compare BQAXR with twelve advanced classification methods, including the single classifiers (DT, LR, SVM, NB, and KNN), the ensemble classifiers (RF, Adaboost, XGBoost, and light gradient boosting machine (LightGBM), and the improved classifiers (GSVM, QSVM, and BKNN) in terms of the machine learning indicators and statistical indicators (Cohen's kappa and MCC). After that, for two cancer datasets, we perform the paired t test with the classifiers (KNN, BKNN, SVM, QSVM, MLP, XGBoost, and BQAXR) in terms of accuracy, recall, precision, and AUC, respectively.

Table 8 summarizes the experimental results on the machine learning indicators in the two cancer datasets. From Table 8, it can be observed that BQAXR has the best performance and strongest generalization ability, followed by the improved classifiers, and the single classifiers perform the worst among all classifiers. Specifically, in the gastric (resp., skin) cancer dataset, BQAXR is approximately 7% (resp., 5%) on average higher than all the single classifiers. BQAXR is 1.89% (resp., 2.42%), 0.79% (resp., 3.97%), 2.05% (resp., 0.99%), 1.4% (resp., 2.4%), and 1.72% (resp., 2.03%) higher than XGBoost (resp., Adaboost), in terms of accuracy, recall, precision, F1-score, and AUC, respectively, which performs the best among the ensemble classifiers. BQAXR is also 2.38% (resp., 1.64%) and 4.57% (resp., 3%) higher than QSV in terms of accuracy and recall on the two cancer datasets. Through the above analysis, BQAXR has better performance than the advanced machine learning methods.

To further verify the superiority of BQAXR, Cohen's kappa and MCC are calculated for BQAXR and the advanced machine learning algorithms, including XGBoost, Adaboost, SVM, etc. As for Cohen's kappa, six levels are typically used to represent consistent performance, i.e., poor ($k < 0$), slight ($0 \leq k < 0.2$), fair ($0.2 \leq k < 0.4$), moderate ($0.4 \leq k < 0.6$), substantial ($0.6 \leq k < 0.8$), and almost perfect ($0.8 \leq k < 1$). The experimental results in the two cancer datasets are shown in Table 9. From Table 9, we see that Cohen's kappa values of BQAXR are the highest in the two cancer datasets, 0.620 and 0.670, respectively. According to the above description of Cohen's kappa, BQAXR reaches a substantial level. Moreover, BQAXR performs well in terms of MCC in the two cancer datasets as shown in Table 9.

Table 10 displays the experimental results on the paired t tests. From Table 10, we see that BQAXR has a p value of less than 0.05 in terms of accuracy, recall, and AUC for both cancer datasets, indicating that there are statistical differences between BQAXR and any of the other compared machine learning methods.

Discussion

The above experimental analyses from multiple perspectives demonstrate that PKSFS and BQAXR yield satisfactory results for cancer survival prediction. Specifically, in the two cancer datasets, BQAXR under the feature subset obtained by PKSFS reaches the best effect compared with the single classifiers (DT, LR, NB, SVM, and KNN) and advanced homogeneous ensemble classifiers (Adaboost, XGBoost,

Table 8 Experimental results on comparing different algorithms

Dataset	Type	Model	Accuracy	Recall	Precision	F1-score	AUC
Gastric cancer	Single classifier	DT	0.7123	0.5722	0.7132	0.6350	0.7281
		LR	0.7706	0.7439	0.7968	0.7694	0.7703
		SVM	0.7711	0.7393	0.7917	0.7646	0.7709
		NB	0.7534	0.6780	0.8110	0.7386	0.7550
		KNN	0.7551	0.7440	0.7952	0.7687	0.7556
	Ensemble classifier	RF	0.7827	0.7781	0.7954	0.7867	0.7830
		Adaboost	0.7738	0.7494	0.7893	0.7688	0.7744
		XGBoost	0.8020	0.8021	0.8147	0.8084	0.8031
		LightGBM	0.7867	0.7640	0.7746	0.7693	0.7881
	Improved classifier	QSVM	0.7971	0.7643	0.8276	0.7947	0.7982
		GSVM	0.7892	0.7705	0.8151	0.7922	0.7888
		BKNN	0.7900	0.7941	0.8022	0.7981	0.7979
	Proposed classifier	BQAXR	**0.8209**	**0.8100**	**0.8352**	**0.8224**	**0.8203**
Skin cancer	Single classifier	DT	0.7905	0.8216	0.8184	0.8200	0.7911
		LR	0.8053	0.8166	0.8406	0.8284	0.8040
		SVM	0.8062	0.8463	0.8222	0.8341	0.7976
		NB	0.8000	0.8660	0.8031	0.8334	0.7878
		KNN	0.7804	0.7180	0.8257	0.7681	0.7910
	Ensemble classifier	RF	0.8160	0.8564	0.8124	0.8338	0.7984
		Adaboost	0.8094	0.8513	0.8233	0.8371	0.8011
		XGBoost	0.8026	0.8322	0.8282	0.8302	0.7983
		LightGBM	0.8029	0.8456	0.8182	0.8317	0.7989
	Improved classifier	QSVM	0.8172	0.8610	0.8290	0.8447	0.8090
		GSVM	0.7987	0.8431	0.8165	0.8296	0.7973
		BKNN	0.8108	0.8661	0.8181	0.8414	0.8014
	Proposed classifier	BQAXR	**0.8336**	**0.8910**	**0.8332**	**0.8611**	**0.8214**

The bold value is the best performance on this metric

LightGBM, and RF), and is better than three improved classifiers (GSVM, QSVM, and BKNN) and other heterogeneous stacked ensemble models. For the gastric (resp., skin) cancer dataset, the best accuracy, recall, precision, F1-score, and AUC of BQAXR are 82.1% (resp., 83.4%), 81.0% (resp., 89.1%), 83.5% (resp., 83.4%), 82.2% (resp., 86.1%), and 82.0% (resp., 82.1%), respectively.

At present, most of the studies related to cancer survival prediction focus on breast cancer, prostate cancer, and cervical cancer, etc., and gastric cancer and skin cancer are ignored. Therefore, comparing the model proposed in this paper with the previous studies is a huge challenge. To verify the effectiveness of the proposed method, based on previous cancer survival prediction and disease diagnosis studies, the performance of the proposed method is compared with nine ensemble learning methods proposed in cancer survival prediction and disease diagnosis. Specifically, three heterogeneous ensemble methods were proposed by Bashir et al. [27], Velusamy and Ramasamy [28] and Thongkam et al. [18]; three homogeneous ensemble methods were proposed by Wang et al. [8], Zein et al. [43] and Gu et al. [44]; three stacked ensemble methods proposed by Chungsoo et al. [6], Ali et al. [5] and Xiao et al. [21].

Table 11 shows the accuracy and AUC of the above nine ensemble methods and the proposed method in the two cancer datasets. Specifically, the proposed method performed best on the two cancer datasets, followed by the stacked ensemble model proposed by Chungsoo et al. [6], and the heterogeneous ensemble model proposed by Bashir et al. [27] performed the worst. In the gastric (resp., skin) cancer dataset, the proposed method is 2.25% (reps., 1.11%) and 2.14% (resp., 0.74%) higher in terms of accuracy and AUC than the method proposed by Chungsoo et al. [6].

Table 9 Experimental results of MCC and kappa with different algorithms

Classifier	Gastric cancer		Skin cancer	
	MCC	Cohen's kappa	MCC	Cohen's kappa
DT	0.4795	0.4770	0.6074	0.6074
KNN	0.5560	0.4990	0.5789	0.5713
RF	0.5906	0.5906	0.5883	0.5880
LR	0.5589	0.5588	0.6168	0.6140
Adaboost	0.5847	0.5846	0.6029	0.6060
XGBoost	0.5908	0.5907	0.5818	0.5810
SVM	0.5851	0.5850	0.5929	0.5883
NB	0.5144	0.5071	0.5767	0.5638
BKNN	0.5873	0.5894	0.5996	0.6045
QSVM	0.5962	0.5943	0.6287	0.6268
GSVM	0.5884	0.5901	0.624	0.6213
LightGBM	0.6051	0.6050	0.6237	0.6233
BQAXR	**0.6200**	**0.6220**	**0.6701**	**0.6721**

The bold value is the best performance on this metric

Table 10 Experimental results of the paired *t* test on comparing different algorithms

Dataset		Gastric cancer						Skin cancer					
		BKNN	SVM	QSVM	AMLP	XGBoost	BQAXR	BKNN	SVM	QSVM	AMLP	XGBoost	BQAXR
KNN	Accuracy	0.034	0	0	0	0	0	0	0	0	0	0.013	0
	Recall	0	0	0	0	0	0	0.023	0.010	0	0.144	0	0
	Precision	0	0.021	0	0	0	0	0	0	0	0	0	0.004
	AUC	0	0	0	0	0	0	0	0.006	0	0	0.441	0
BKNN	Accuracy	–	0.121	0	0.001	0.001	0	–	0	0	0	0	0
	Recall	–	0	0.016	0.001	0.001	0	–	0	0.001	0.017	0	0
	Precision	–	0.213	0.001	0	0	0	–	0.375	0.001	0	0.316	0.001
	AUC	–	0	0	0.001	0	0	–	0.131	0.028	0	0	0
SVM	Accuracy	–	–	0.004	0.014	0	0	–	–	0.010	0	0	0
	Recall	–	–	0.012	0	0	0.003	–	–	0.023	0.323	0	0.001
	Precision	–	–	0	0	0	0	–	–	0.011	0.023	0.4	0
	AUC	–	–	0	0.422	0	0	–	–	0	0	0.004	0
QSVM	Accuracy	–	–	–	0.012	0.032	0	–	–	–	0.002	0.017	0
	Recall	–	–	–	0	0.012	0	–	–	–	0	0.023	0
	Precision	–	–	–	0.004	0.033	0.01	–	–	–	0.038	0.032	0.001
	AUC	–	–	–	0.743	0.042	0.026	–	–	–	0.376	0.042	0.001
AMLP	Accuracy	–	–	–	–	0.02	0	–	–	–	–	0.007	0
	Recall	–	–	–	–	0.001	0	–	–	–	–	0	0
	Precision	–	–	–	–	0.012	0.001	–	–	–	–	0	0.034
	AUC	–	–	–	–	0.042	0.03	–	–	–	–	0.032	0.015
XGBoost	Accuracy	–	–	–	–	–	0.001	–	–	–	–	–	0
	Recall	–	–	–	–	–	0	–	–	–	–	–	0
	Precision	–	–	–	–	–	0	–	–	–	–	–	0
	AUC	–	–	–	–	–	0.002	–	–	–	–	–	0.007

Table 11 Experimental results on comparing different ensemble learning models

Type	References	The structure of the proposed model	Gastric cancer		Skin cancer	
			Accuracy	AUC	Accuracy	AUC
Heterogeneous ensemble	[27]	NB + DTG + DTI + SVM + MBL	0.7560	0.7514	0.7753	0.7832
	[28]	KNN + RF + SVM	0.7848	0.7819	0.7905	0.7804
	[18]	Adaboost + RF	0.7802	0.7803	0.7709	0.7603
Homogeneous ensemble	[8]	Ensemble SVM	0.7881	0.7879	0.8013	0.8000
	[43]	LightGBM	0.787	0.788	0.8032	0.7963
	[44]	XGBoost	0.8020	0.8031	0.8025	0.8002
Stacked ensemble	[6]	First: LLR + GB; Second: XGBoost	0.7984	0.7989	0.8225	0.8140
	[5]	Stacking SVM	0.7883	0.7869	0.8143	0.8018
	[21]	First stage: KNN + DT + RF + GBDT + SVM; Second: MLP	0.7885	0.7901	0.7900	0.7971
	This paper	BQAXR	**0.8209**	**0.8203**	**0.8336**	**0.8214**

The bold value is the best performance on this metric

Conclusions

Due to the complexity and high incidence of cancer, survival prediction has been one of the three important tasks of cancer prognosis. To address the problem of low accuracy in cancer survival prediction, we first propose an a priori knowledge- and stability-based feature selection method PKSFS to obtain the optimal feature subset from the high-dimensional cancer dataset, which effectively reduces the computational complexity and helps improve the accuracy of survival prediction. Subsequently, the two-stage heterogeneous stacked ensemble learning model BQAXR is proposed for cancer survival prediction, which integrates five heterogeneous high-quality learners in two stages through the stacked generalization strategy, where the weaknesses and shortcomings of the learners are overcome by some improved methods in BQAXR. To verify the accuracy and reliability of the model for cancer survival prediction, extensive experiments are conducted on the real gastric cancer dataset and skin cancer dataset. The experimental results show that PKSFS is beneficial to the construction of BQAXR, and the proposed method can significantly improve the efficiency and accuracy of cancer survival prediction compared with many state-of-the-art machine learning methods. The output information of BQAXR can be used as an important reference for practical cancer prognosis, which can not only effectively assist doctors in the development of treatment plans, but also improve patient satisfaction, and reduce hospital operating costs and patient treatment costs.

Future research may consider the following extensions of our research. First, it is necessary to further validate the validity and usefulness of our proposed method in more real cancer datasets. Second, it is worth incorporating more valuable features into the model, such as the dietary habits of patients with gastric cancer and the sunshine times of patients with skin cancer, to further improve accuracy in cancer survival prediction. Finally, it would be of interest to consider more information into the cancer survival prediction model, such as detailed DNA and RNA-seq information, which can further improve the survival rates of cancer patients by systematically tracking their conditions.

Appendix

See Table 12.

Table 12 List of features and their descriptions in the initial dataset

Number	Variable	Description	Number	Variable	Description
1	ID	Patient ID	35	CSSSF9	CS site-specific factor 9
2	Seer_registry	SEER registry	36	CSSSF10	CS site-specific factor 10
3	MSAD	Marital status at diagnosis	37	CSSSF11	CS site-specific factor 11
4	Race	Race/ethnicity	38	CSSSF12	CS site-specific factor 12
5	NHIA	NHIA Derived Hisp Origin	39	CSSSF13	CS site-specific factor 13
6	Sex	Sex	40	CSSSF15	CS site-specific factor 15
7	AAD	Age at diagnosis	41	CSSSF16	CS site-specific factor 16
8	YoB	Year of birth	42	CSSSF25	CS site-specific factor 25
9	SN	Sequence number	43	AJCCT	Derived AJCC T 7th ed
10	YoD	Year of diagnosis	44	AJCCN	Derived AJCC N 7th ed
11	MoD	Month of diagnosis	45	AJCCM	Derived AJCC M 7th ed
12	PSL	Primary Site-labeled	46	SS1977	Derived SS1977
13	Laterality	Laterality	47	AJCC	Derived AJCC Stage Group 7th ed
14	HT_ICD	Histologic Type ICD-O-3	48	SS2000	Derived SS2000
15	HC_ICD	Behavior code ICD-O-3	49	RXSSPS	RX Summ–Surg Prim Site
16	Grade	Grade	50	RXSRLE	RX Summ–Reg LN Examined
17	DC	Diagnostic Confirmation	51	RXSSRLS	RX Summ–Reg LN Sur
18	ToRS	Type of reporting source	52	RXXOR	RX Summ–Oth Reg/Dis
19	EOD_E	EOD 10-extent	53	RnCDS	Reason no cancer-directed surgery
20	EOD_N	EOD 10-nodes	54	RN	Record number
21	EOD_S	EOD 10-size	55	L2005	Louisiana 2005
22	CSTS	CS tumor size	56	SEERH	SEER historic stage A
23	CSE	CS extension	57	FMPI	First malignant primary indicator
24	CSS	CS Schema	58	PBIR	Primary by international rules
25	CSLN	CS lymph nodes	59	SS2000	Summary stage 2000
26	CSMAD	CS mets at dx	60	SoRD	Surgery of othreg/dis sites
27	CSSSF1	CS site-specific factor 1	61	TN_SITU	Number of in situ/malignant tumors
28	CSSSF2	CS site-specific factor 2	62	TN_BB	Number of in benign/borderline tumors
29	CSSSF3	CS site-specific factor 3	63	TM1	Tumor marker 1
30	CSSSF4	CS site-specific factor 4	64	TM2	Tumor marker 2
31	CSSSF5	CS site-specific factor 5	65	TM3	Tumor marker 3
32	CSSSF6	CS site-specific factor 6	66	SM	Survival months
33	CSSSF7	CS site-specific factor 7	67	VS	Vital status
34	CSSSF8	CS site-specific factor 8	–	–	–

References

1. Ferlay J, Colombet M, Isabelle S (2019) Estimating the global cancer incidence and mortality in 2019: GLOBOCAN sources and methods. Int J Cancer 144:1941–1953
2. Walczak S, Velanovich V (2018) Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks. Decis Support Syst 106:110–118
3. Andreu-Perez J, Poon CCY, Merrifield RD et al (2015) Big data for health. IEEE J Biomed Health Inform 19:1193–1208. https://doi.org/10.1109/JBHI.2015.2450362
4. Zhai B, Chen J (2018) Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing. China Sci Total Environ 635:644–658. https://doi.org/10.1016/j.scitotenv.2018.04.040
5. Ali L, Niamat A, Khan JA et al (2019) An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7:54007–54014. https://doi.org/10.1109/ACCESS.2019.2909969
6. Chungsoo K, You SC, Reps JM et al (2020) Machine-learning model to predict the cause of death using a stacking ensemble

method for observational data. J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocaa277

7. Mingyu Z, Sutong W, Yanzhang W, Dujuan W (2022) An interpretable prediction method for university student academic crisis warning. Complex Intell Syst 8:323–336. https://doi.org/10.1007/s40747-021-00383-0
8. Wang H, Zheng B, Yoon SW, Ko HS (2018) A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res 267:687–699
9. Wang Y, Wang D, Ye X et al (2019) A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. Inf Sci 474:106–124. https://doi.org/10.1016/j.ins.2018.09.046
10. Matsuo K, Purushotham S, Jiang B et al (2019) Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. Am J Obstet Gynecol 220:381.e1-381.e14. https://doi.org/10.1016/j.ajog.2018.12.030
11. Zhu L, Luo W, Su M et al (2013) Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients. Biomed Rep 1:757–760. https://doi.org/10.3892/br.2013.140
12. Tapak L, Shirmohammadi-Khorram N, Amini P et al (2019) Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Health 7:293–299. https://doi.org/10.1016/j.cegh.2018.10.003
13. Delen D (2009) Analysis of cancer data: a data mining approach. Expert Syst 26:100–112. https://doi.org/10.1111/j.1468-0394.2008.00480.x
14. Shukla N, Hagenbuchner M, Win KT, Yang J (2018) Breast cancer data analysis for survivability studies and prediction. Comput Methods Programs Biomed 155:199–208. https://doi.org/10.1016/j.cmpb.2017.12.011
15. Zolbanin HM, Delen D, Hassan Zadeh A (2015) Predicting overall survivability in comorbidity of cancers: a data mining approach. Decis Support Syst 74:150–161. https://doi.org/10.1016/j.dss.2015.04.003
16. Nguyen TT, Nguyen MP, Pham XC, Liew AW-C (2018) Heterogeneous classifier ensemble with fuzzy rule-based meta learner. Inf Sci 422:144–160. https://doi.org/10.1016/j.ins.2017.09.009
17. Chen L, Wang C, Song S (2022) Software defect prediction based on nested-stacking and heterogeneous feature selection. Complex Intell Syst. https://doi.org/10.1007/s40747-022-00676-y
18. Thongkam J, Xu G, Zhang Y (2008) AdaBoost algorithm with random forests for predicting breast cancer survivability. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, Hong Kong, pp 3062–3069
19. Cho S-B, Won H-H (2003) Machine learning in DNA microarray analysis for cancer classification. Australian Computer Society, pp 189–198
20. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
21. Xiao Y, Wu J, Lin Z, Zhao X (2018) A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed 153:1–9. https://doi.org/10.1016/j.cmpb.2017.09.005
22. Anifowose F, Labadin J, Abdulraheem A (2015) Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. Appl Soft Comput 26:483–496. https://doi.org/10.1016/j.asoc.2014.10.017
23. Cui S, Wang D, Wang Y et al (2018) An improved support vector machine-based diabetic readmission prediction. Comput Methods Programs Biomed 166:123–135. https://doi.org/10.1016/j.cmpb.2018.10.012
24. Tao Z, Huiling L, Wenwen W, Xia Y (2019) GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Appl Soft Comput 75:323–332. https://doi.org/10.1016/j.asoc.2018.11.001
25. Liu L, Jiang H, He P et al (2019) On the variance of the adaptive learning rate and beyond. http://arxiv.org/abs/1908.03265
26. Adem K, Kiliçarslan S, Cömert O (2019) Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. Expert Syst Appl 115:557–564. https://doi.org/10.1016/j.eswa.2018.08.050
27. Bashir S, Qamar U, Khan FH (2015) Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Qual Quant 49:2061–2076. https://doi.org/10.1007/s11135-014-0090-z
28. Velusamy D, Ramasamy K (2021) Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. Comput Methods Programs Biomed 198:105770. https://doi.org/10.1016/j.cmpb.2020.105770
29. Hart PE (1968) The condensed nearest neighbor rule. IEEE Trans Inf Theory 14:515–516. https://doi.org/10.1109/TIT.1968.1054155
30. Ahmadi E, Garcia-Arce A, Masel DT (2019) A metaheuristic-based stacking model for predicting the risk of patient no-show and late cancellation for the neurology appointments. IIE Trans 9:272–291. https://doi.org/10.1080/24725579.2019.1649764
31. Peng L, Zhang H, Zhang H, Yang B (2017) A fast feature weighting algorithm of data gravitation classification. Inf Sci 375:54–78. https://doi.org/10.1016/j.ins.2016.09.044
32. Drotár P, Gazda M, Vokorokos L (2019) Ensemble feature selection using election methods and ranker clustering. Inf Sci 480:365–380. https://doi.org/10.1016/j.ins.2018.12.033
33. Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl Based Syst 140:103–119. https://doi.org/10.1016/j.knosys.2017.10.028
34. Xia Y, Liu C, Li Y, Liu N (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241. https://doi.org/10.1016/j.eswa.2017.02.017
35. Jun S, Wenbo X, Bin F (2005) Adaptive parameter control for quantum-behaved particle swarm optimization on individual level. In: 2005 IEEE international conference on systems, man and cybernetics. IEEE, Waikoloa, pp 3049–3054
36. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. http://arxiv.org/abs/1412.6980 [Cs]
37. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proc 22nd ACM SIGKDD int conf knowl discov data min—KDD 16, pp 785–794. https://doi.org/10.1145/2939672.2939785
38. Araz OM, Olson D, Ramirez-Nafarrate A (2019) Predictive analytics for hospital admissions from the emergency department using triage information. Int J Prod Econ 208:199–207. https://doi.org/10.1016/j.ijpe.2018.11.024
39. Mahajan PD, Maurya A, Megahed A et al (2020) Optimizing predictive precision in imbalanced datasets for actionable revenue change prediction. Eur J Oper Res 285:1095–1113. https://doi.org/10.1016/j.ejor.2020.02.036
40. Hsu W-Y (2018) A decision-making mechanism for assessing risk factor significance in cardiovascular diseases. Decis Support Syst 115:64–77. https://doi.org/10.1016/j.dss.2018.09.004
41. Han L, Luo S, Yu J et al (2015) Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J Biomed Health Inform 19:728–734. https://doi.org/10.1109/JBHI.2014.2325615

42. Kaya A (2018) Cascaded classifiers and stacking methods for classification of pulmonary nodule characteristics. Comput Methods Programs Biomed 166:77–89. https://doi.org/10.1016/j.cmpb.2018.10.009

43. Zein JD, Wu C, Attaway AH et al (2021) Novel machine learning can predict acute asthma exacerbation. Chest 159(5):1747–1757

44. Gu D, Su K, Zhao H (2020) A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif Intell Med 107:101858. https://doi.org/10.1016/j.artmed.2020.101858

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.