BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Developing a multiomics data-based mathematical model to predict colorectal cancer recurrence and metastasis

Bing Li[1], Ming Xiao[1], Rong Zeng[2,3,4] and Le Zhang[1*]

From 17th International Symposium on Bioinformatics Research and Applications
Shenzhen, China 26-28 November 2021 https://alan.cs.gsu.edu/isbra21/?q=node/1

Abstract

Background Colorectal cancer is the fourth most deadly cancer, with a high mortality rate and a high probability of recurrence and metastasis. Since continuous examinations and disease monitoring for patients after surgery are currently difficult to perform, it is necessary for us to develop a predictive model for colorectal cancer metastasis and recurrence to improve the survival rate of patients.

Results Previous studies mostly used only clinical or radiological data, which are not sufficient to explain the in-depth mechanism of colorectal cancer recurrence and metastasis. Therefore, this study proposes such a multiomics data-based predictive model for the recurrence and metastasis of colorectal cancer. LR, SVM, Naïve-bayes and ensemble learning models are used to build this predictive model.

Conclusions The experimental results indicate that our proposed multiomics data-based ensemble learning model effectively predicts the recurrence and metastasis of colorectal cancer.

Keywords Multiomics, Colorectal cancer, Recurrence and metastasis, Data augmentation, Ensemble learning

Background

Colorectal cancer is the fourth most deadly cancer worldwide [1]. Although therapies for colorectal cancer keep improving, the mortality rate remains high. Since cancer metastasis is the most important cause of death of patients with colorectal cancer [1–3], the metastasis status is a very important indicator for the clinical treatment of colorectal cancer.

Surgery is the main clinical treatment used currently, but patients who undergo colorectal cancer resection still have a high probability of developing recurrence and metastasis [4]. Moreover, the postoperative recurrence and metastasis status will continue to affect the disease status and survival time after surgery. Currently,

*Correspondence:
Le Zhang
zhangle06@scu.edu.cn
[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]CAS Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[3]CAS Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China
[4]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

continuous examinations and disease monitoring for patients after surgery are very difficult to conduct. Thus, if we can develop such a mathematical model that predicts postoperative metastasis in patients, we will be able to monitor high-risk patients and provide targeted interventions and precise medical treatments to significantly improve the survival rate of patients.

Previously, most colorectal cancer recurrence and metastasis studies manually select the key features [5–13] from a single omics dataset using various data mining methods, such as correlation coefficient test, chi-square test, t test or Mann-Whitney U test [8, 14–20], resulting in subjectivity and inconsistencies of the selected features. Because our multiomics datasets described in the data source section consist of not only clinical and somatic mutation data but also high-dimensional proteomics (6400 dimensions) and phosphoproteomics (22,000 dimensions) data, our first research question is how to develop such a feature selection and high dimensionality reduction algorithm that processes these high-dimensional multiomics colorectal cancer datasets.

Previous studies usually employed radiological data [14–16], clinical data [5–7, 17] or gene expression data [8, 9] to investigate the recurrence and metastasis of colorectal cancer. However, the occurrence and development of colorectal cancer recurrence and metastasis are so complicated [21] that the use of radiological, clinical or gene expression data alone is not sufficient to comprehensively and deeply explain the mechanism underlying the recurrence and metastasis of colorectal cancer. Recently, Chen Li et al. reported that the analysis of proteomics and phosphoproteomics data from the primary tumour alone successfully identifies metastatic cases [22, 23]. Since the collection of large amounts of multiomics data to optimize the weight of the classifiers of the model used to predict the recurrence and metastasis of colorectal cancer is very expensive and time-consuming, our second research question is how to employ a computational algorithm to perform data augmentation for colorectal cancer predictions.

Also, previous studies have usually employed a data mining algorithm [24–32], such as Cox [6, 8, 11, 12, 33, 34], logistic regression [5, 14, 16, 17], decision tree [17, 35–37] and random forest [15], to model the recurrence and metastasis of colorectal cancer. However, since the predictive accuracy for different omics data is sensitive to the data mining algorithm, the use of a single model does not take advantage of multiomics data to increase the predictive power. Therefore, our third research question is how to build such a predictive model that takes advantage of multiomics data and results in a high predictive accuracy for the recurrence and metastasis of colorectal cancer.

To answer our research questions, this study proposes the following three innovations to determine the recurrence and metastasis of colorectal cancer. First, we integrated multiple statistical tests to select the key features from a multiomics dataset. Second, we employed data augmentation to increase the size of the dataset for model training. Third, we built an ensemble learning model [38, 39] to increase the predictive accuracy.

Next, based on the three innovations listed above, we propose our research plan as described below. First, we integrated Student's t test, Mann-Whitney U test, ANOVA (Analysis of Variance), chi-square test, and Fisher's exact test [40–46] to select the key features from clinical, somatic mutation, proteomics, and phosphoproteomics datasets and then employed PCA (principal component analysis) [47, 48] to perform dimensional reduction. Second, we conducted data augmentation using the SMOTE algorithm to increase the dataset size for model training. Third, we integrated the logistic regression (LR), support vector machine (SVM), and Naive-Bayes algorithms to build an ensemble learning predictive model for the recurrence and metastasis of colorectal cancer.

At last, we selected 3 key features from clinical data, 3 key features from somatic mutations, 89 key features from proteomics and 15 key features from phosphoproteomics. Afterward, we performed dimensional reduction for proteomics and phosphoproteomics features to obtain two principal components. After data augmentation, the sample size increased from 144 to 288, which met the requirement of model training. Finally, we developed a novel multiomics databased ensemble learning model for the prediction of recurrence and metastasis of colorectal cancer that outperformed the classical LR, Naive-Bayes, and SVM models.

Methods

Data source

Our research data were obtained from our previous study [22], which were originally collected from 146 patients with colorectal cancer at Shanghai Hospital, China [22]. Our research data consisted of clinical (clinicopathologic features and prognosis information), somatic mutations (information on somatic single-nucleotide variants (SNVs) and small insertions-deletions (INDELs) identified by WES), proteomics (6,408 quantified protein expression data that were subjected to median normalization by column and log2 transformation) and phosphoproteomics data (22,000 quantified phosphoprotein expression data that were subjected to median normalization by column and log2 transformation). Among the 146 patients, 70 experienced recurrence and metastasis after surgery and were labelled with one; 74 patients were free from recurrence and metastasis and were labelled

with zero; and 2 patients lacked the label. Thus, only samples from 144 patients were used in our study. The informed consent was obtained from all subjects. The experimental protocol was approved by Shanghai Changhai Hospital Ethics Committee (CHEC2017-235, Shanghai, China) [22].

Workflow of the study

Figure 1 describes the workflow of the study. First, we selected the key features from all datasets and then employed PCA to perform dimensional reduction. Next, we conducted data augmentation to increase the sample size for model training. Finally, we integrated the LR, SVM, and Naive-Bayes algorithms to develop an ensemble learning model for colorectal cancer recurrence and metastasis.

Details for feature selection

Fisher's exact test

Construct a contingency table.

	A-positive	A-negative	Total
B-positive	a	b	a+b
B-negative	c	d	c+d
total	a+c	b+d	n

$$p = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{1}$$

Chi-square test

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} \tag{2}$$

Here, n is the number of observations, k is the number of different classes, x_i is the observed value and p_i is the probability of class i.

Student's t test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.1}$$

$$s_p = \sqrt{\frac{(n_1-1)s^2{}_{X_1} + (n_2-1)s^2{}_{X_2}}{n_1+n_2-2}} \tag{3.2}$$

Here, $s^2{}_{X_1}$ and $s^2{}_{X_2}$ are the variances of the two sets and n is the size of the set.

Mann-Whitney U test

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j) \tag{4.1}$$

$$S(X_i, Y_j) = \begin{cases} 1 & Y < X \\ 0.5 & Y = X \\ 0 & Y > X \end{cases} \tag{4.2}$$

ANOVA

$$SS_{total} = SS_{treatment} + SS_{error}$$

$$DF_{total} = DF_{treatment} + DF_{error}$$

$$MS_{treatment} = SS_{treatment}/DF_{treatment}$$

$$MS_{error} = SS_{error}/DF_{error}$$

$$F = \frac{MS_{treatment}}{MS_{error}} = \frac{SS_{treatment}/DF_{treatment}}{SS_{error}/DF_{error}} \tag{5}$$

Here, SS represents the sum of squares, DF represents the degree of freedom and MS is the mean squares.

Results

Feature selection and dimensional reduction

To answer the first research question, we propose a feature selection and dimensional reduction workflow to process the multiomics data as described below.

Feature selection

We proposed a robust feature selection method for multiomics data, and Fig. 2 illustrates two feature selection methods for discrete and continuous data. For discrete data, we used Fisher's exact test (Eq. 1) or the chi-square test (Eq. 2) [22] to determine the correlations between each feature and their label (Fig. 2A). For continuous data, we divided the dataset into two datasets according to the label, and then we integrated Student's t test (Eq. 3) [17, 49], Mann-Whitney U test (Eq. 4) [15] and ANOVA (Eq. 5) [50] to perform feature selection [10, 13, 35, 48, 51, 52, 53, 54] (Fig. 2B). Key equations are listed in Methods.

Table 1 lists the key features for each dataset, and Supplementary Table S1 describes the feature selection procedure.

Dimensional reduction

Since Table 1 shows that the features of proteomics and phosphoproteomics data still had high dimensions, we carried out PCA (Eq. 6) to reduce the dimensions of these two datasets [55].

$$T_L = XW_L \tag{6}$$

In Eq. 6, W_L maps the original data X with p variables to a new space T with p variables that are uncorrelated

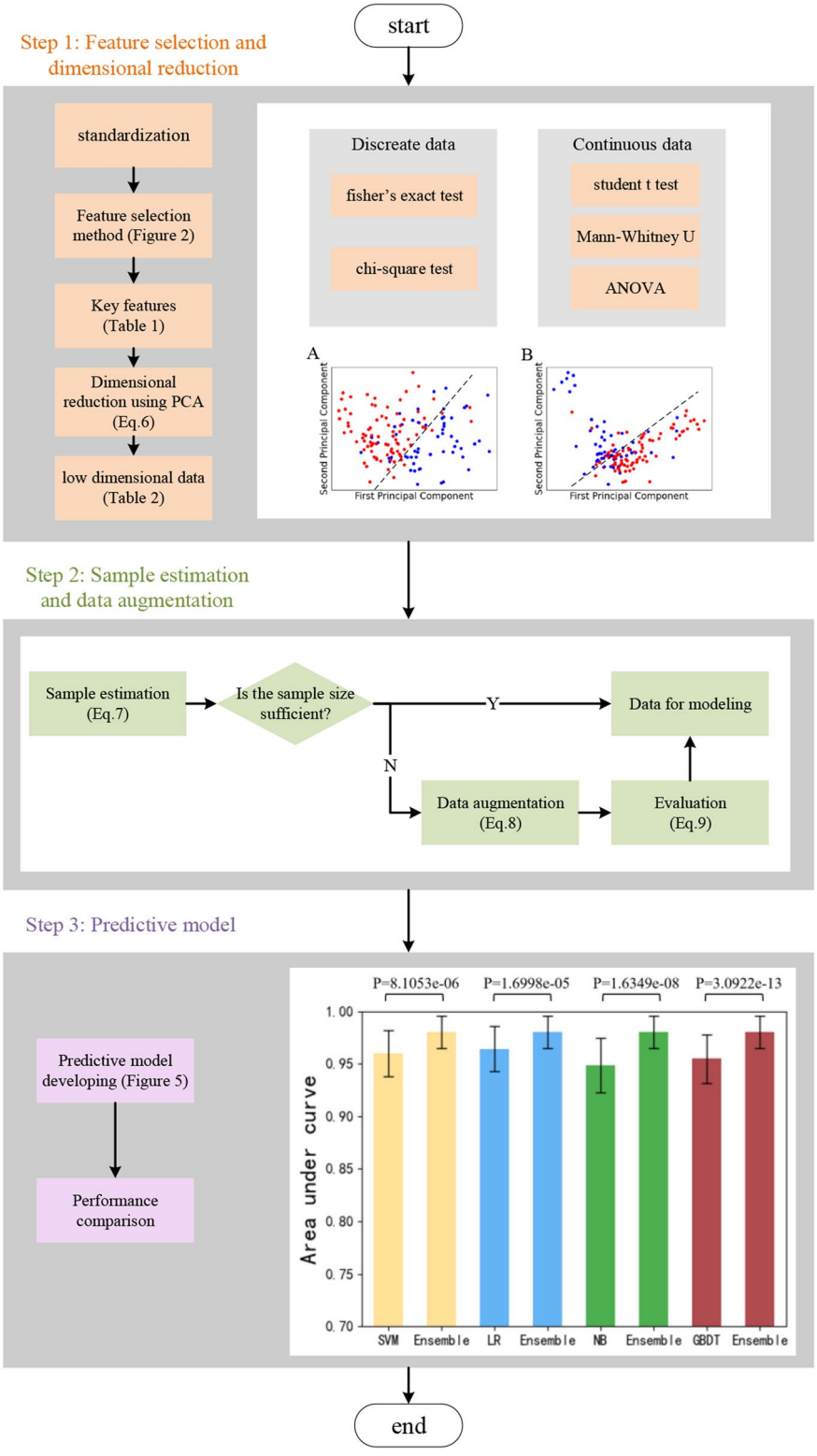

Fig. 1 Workflow of the study. The P values in Step 3 were calculated using the T test [61]

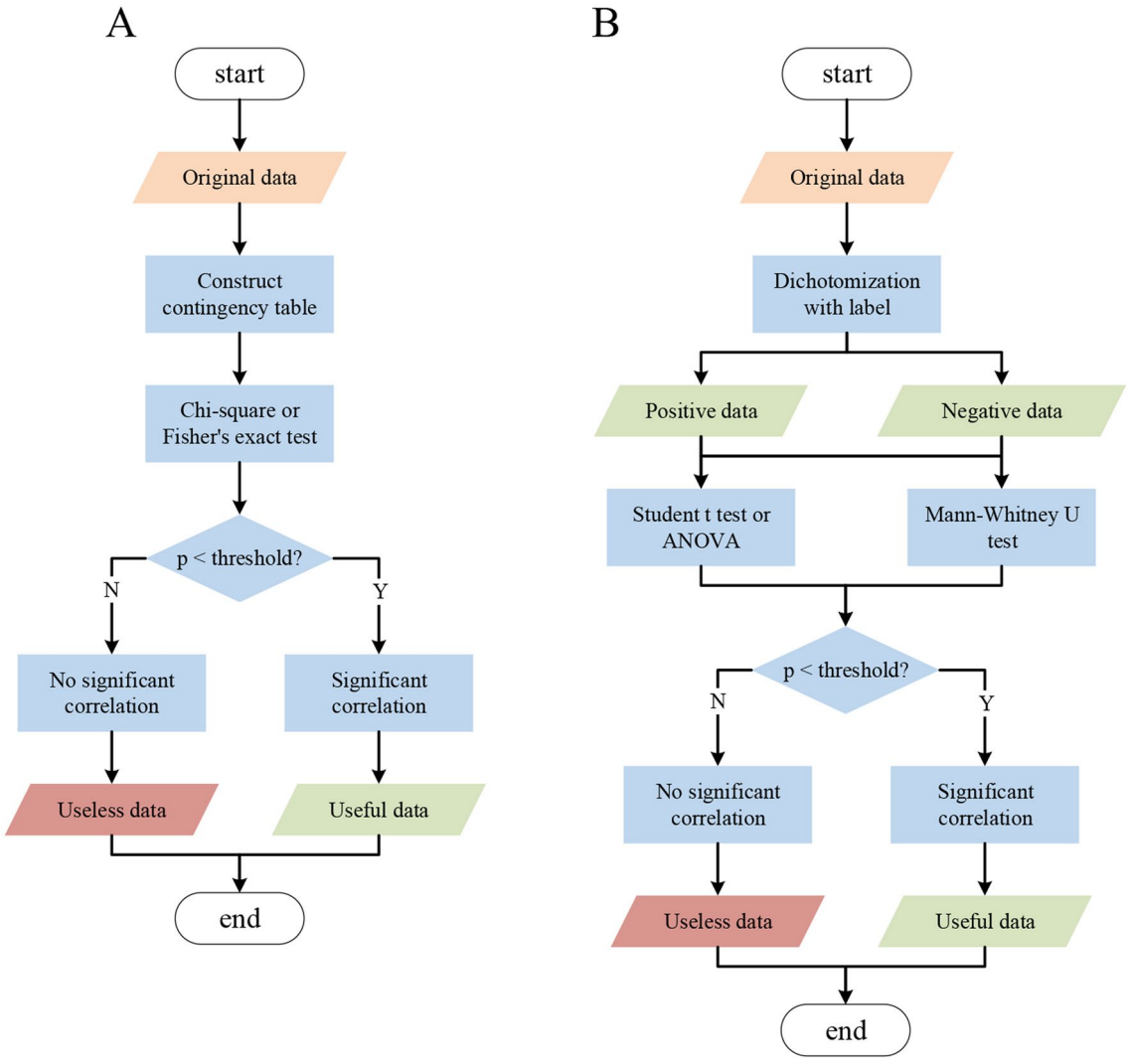

Fig. 2 Feature selection methods for (**A**) discrete and (**B**) continuous datasets

over the dataset, and only the first L principal components are retained after dimensional reduction.

Figure 3 shows the classification results when we chose the first two principal components for dimensional reduction. Since the first two principal components successfully segmented patients with recurrence and metastasis (blue) and patients without recurrence and metastasis (red), we chose the first two principal components to reduce dimensions for proteomics (Fig. 3A) and phosphoproteomics data (Fig. 3B).

After dimensional reduction, the number of features of the clinical data, somatic mutations, proteomics and phosphoproteomics datasets decreased from 110 to 11. Table 2 lists the final features of each dataset, and Supplementary Table S2 describes the dimensional reduction procedure.

Data augmentation

To answer our second question, we used the results of feature selection and dimensional reduction as input (Tables 1 and 2) to estimate if the dataset is sufficient large for model training. If the dataset size was insufficient, we employed data augmentation to increase the dataset size using the method described below.

Sample Estimation

We employed Eq. 7 to compute the optimum sample size (n) for each selected feature with respect to the preset statistical significance [56, 57].

$$n = \frac{\sigma^2\left(Q_1{}^{-1}+Q_2{}^{-1}\right)(\mu_\alpha+\mu_\beta)^2}{\delta^2} \tag{7}$$

Here, σ is the standard deviation; μ_α and μ_β are the critical values of the U-test at the first type of error rate

Table 1 The key features of each dataset

Dataset	Features
Clinical data	Lymph node, Metastasis, Calcium nodus
Somatic mutations	COL6A3, OTOG, KAL1
Proteomics	A0A024R046, A0A024R056, A0A024R0Y5, A0A024R1S8, A0A024R2U7, A0A024R3B5, A0A024R5K1, A0A024R7I3, A0A024R9G4, A0A024RCX8, A0A024RCY1, A0A087WTA8, A0A0A0MRF6, A0A0A0MSM0, A0A0S2Z3J9, A0A140VJC9, A2RUA4, A4D1 × 5, A6NHQ2, A8K2L4, A8K878, B2R4P9, B2R5J1, B2RDF2, B2RDW1, B2ZDQ1, B4DEH0, B7ZB78, C9K0I4, D3DT27, E9PMC9, G1EPM2, O14917, O60493, O75554, P02452, P02461, P02749, P05164, P11908, P14780, P17066, P17600, P18077, P20585, P23378, P24158, P29350, P31947, P35637, P36268, P37802, P40763, P42224, P48061, P49327, P51572, P52209, P52630, P52732, P54707, P63261, P68371, P78346, Q13151, Q13287, Q13884, Q15029, Q2TAM5, Q53SW3, Q546E0, Q6DN03, Q6FIA3, Q7Z434, Q8N8A2, Q8WUM0, Q8WXH0, Q92734, Q96A33, Q96BP3, Q96IS6, Q99497, Q9BTE1, Q9BTT0, Q9NZ08, Q9UI15, Q9UNS2, Q9Y2Z0, Q9Y426
Phosphoproteomics	A0A024R4G1_509_Y, A0A024R4Z6_441_Y, A0A024R9K2_184_S, A0A024RAM4_1817_S, A0A087WVT6_320_S, A0A140VJN8_130_S, B4DPP8_314_T, B4DPP8_320_S, B4DPP8_325_T, P08670_436_T, P62263_137_S, Q6WKZ4_206_S, Q6WKZ4_338_S, Q8IZ21_358_T, Q92625_628_S

Table 2 The results of dimensional reduction

Dataset	Features
Clinical data	Lymph node, Metastasis, Calcium nodus
Somatic mutations	COL6A3, OTOG, KAL1
Proteomics	First two principal components
Phosphoproteomics	First two principal components

Table 3 Optimum sample size N of each dataset

Dataset	Estimated sample size
Clinical data	86
Somatic mutations	231
Proteomics	142
Phosphoproteomics	157

and the second type of error rate; Q_1 and Q_2 are the proportions of each part of the population after the dichotomization procedure; and δ is the difference between the mean of the two datasets.

After performing sample size estimation for every selected feature in Table 1, the optimum sample size n of each dataset is listed in Table 3. Because our original dataset only consisted of 144 labelled data (2.1 Data source section), the dataset was smaller than the optimum sample size n of some datasets (Table 3), indicating that our samples were not sufficient for model training.

SMOTE algorithm

The SMOTE algorithm (Eq. 8) [58] was previously used for oversampling. Here, we employed it for data augmentation. The procedure and key equation are listed below.

Input Dataset

$T = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$, where x_i is the examples and y_i is the labels; number of samples m; number of nearest neighbours k.

Process:

For each data (x_i, y_i) in T:

Find k nearest neighbours with the same label.

$$K = \{ (x_{i1}, y_i), (x_{i2}, y_i), \ldots, (x_{ik}, y_i) \}$$

Choose m neighbours randomly in K.

$$M = \{ (x_{i1}, y_i), (x_{i2}, y_i), \ldots, (x_{im}, y_i) \}$$

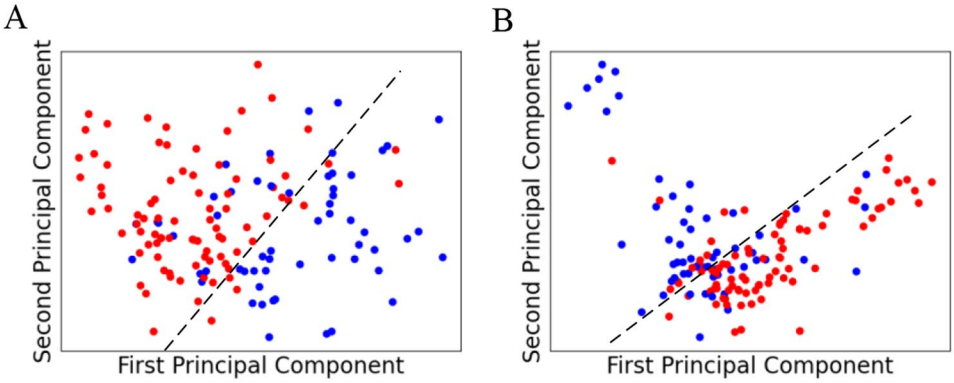

Fig. 3 Illustration of the first two principal components. Here, red points represent patients without recurrence and metastasis, and blue points represent patients with recurrence and metastasis. (**A**) Proteomics data and (**B**) phosphoproteomics data

For each data $(x_{ij},\ y_{ij})$ in M:

$$x_{new} = x_i + rand\,(0,1) * (x_{ij} - x_i) \tag{8}$$

Output Generated new dataset G with label y_i

$$G = \{\,(x_1,\ y_i),\ (x_2,\ y_i),\ \ldots\ ,\ (x_{n*m}, y_i\,)\}$$

We used the SMOTE algorithm to augment the data with pseudo dataset generation by setting $m = 1$ and $k = 5$, as described in detail in Supplementary Table S3. Then, the sample size increased from 144 (original dataset) to 288 (pseudo dataset). Since the size of the pseudo dataset (288) was greater than estimated sample size (231), we consider that it meets the requirement for the sample estimation.

Evaluation of the pseudo dataset quality

We employed the maximum Fisher's discriminant ratio or F1 [59] to validate whether the generated dataset was sufficient for classification and to evaluate the quality of the data augmentation process for the pseudo dataset, as described in a previous study [58]. The F1 value calculated using Eq. 9 shows the degree of overlap. A high F1 value indicates a low degree of overlap in the datasets, which is better for classification [58].

$$f_i = \frac{(\mu_{\,1} - \mu_{\,2})^2}{{\sigma_1}^2 + {\sigma_2}^2} \tag{9.1}$$

$$F_1 = \max\,(f_i) \tag{9.2}$$

We employed Eq. 9.1 and 9.2 to compute f_i for each individual feature i and F1 value, respectively. $\mu_{\,1}$, $\mu_{\,2}$, $\sigma_{\,1}$, and $\sigma_{\,2}$ are the means and standard errors for the two classes, respectively.

As described in a previous study [58], we calculated the F1 value to evaluate the overlap of the two classes. Since the F1 value for the original dataset (Fig. 4A) was less than the F1 value for the SMOTE-generated dataset (Fig. 4B), we consider that the dataset generated by SMOTE has such a lower degree of overlap that is better for classification than the original dataset.

Predictive model

To answer our third question, we developed an ensemble predictive model using three classical classification methods, the performance of which was measured using K-fold cross validation [12, 18, 51, 52, 57]. The development of the ensemble learning model and comparison of the performance between ensemble learning and classical classification are described below.

Ensemble learning model development

Regarding to previous studies [60, 61], we integrate three classical classification methods, LR [62], SVM [63] and Naive-Bayes [64], to develop an ensemble predictive model (Fig. 5) for the recurrence and metastasis of colorectal cancer. The key equations used in this model are listed below.

$$D_t\,(i) = \frac{1}{n} \tag{10}$$

$$\varepsilon_t = \sum_{\substack{i=1 \\ h_t\,(x_i) \neq y_i}}^{n} D_t\,(i) \tag{11}$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{12}$$

$$D_{t+1}\,(i) = \frac{D_t}{sum(D_{t+1})}\begin{cases} e^{-\alpha_t} & h_t\,(x_i) = y_i \\ e^{\alpha_t} & h_t\,(x_i) \neq y_i \end{cases} \tag{13}$$

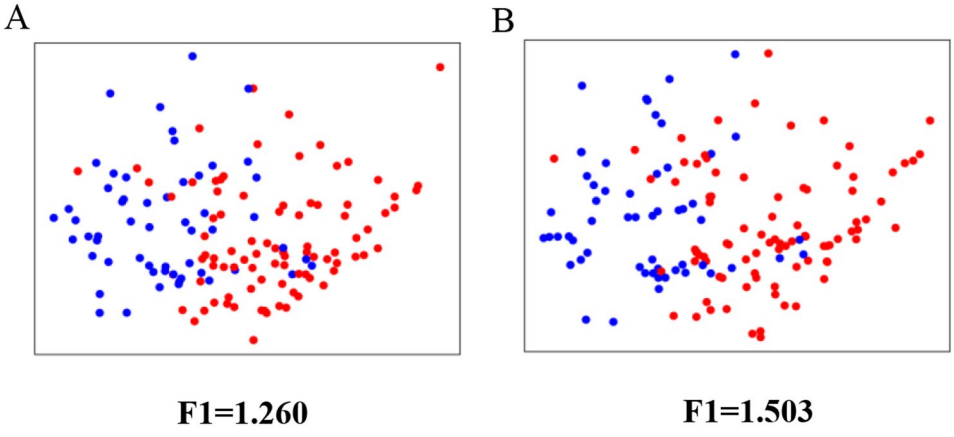

Fig. 4 Illustration of the dataset mapped to two dimensions and the F1 value. Here, red points represent patients without recurrence and metastasis, and blue points represent patients with recurrence and metastasis. (**A**) Original dataset and (**B**) generated dataset

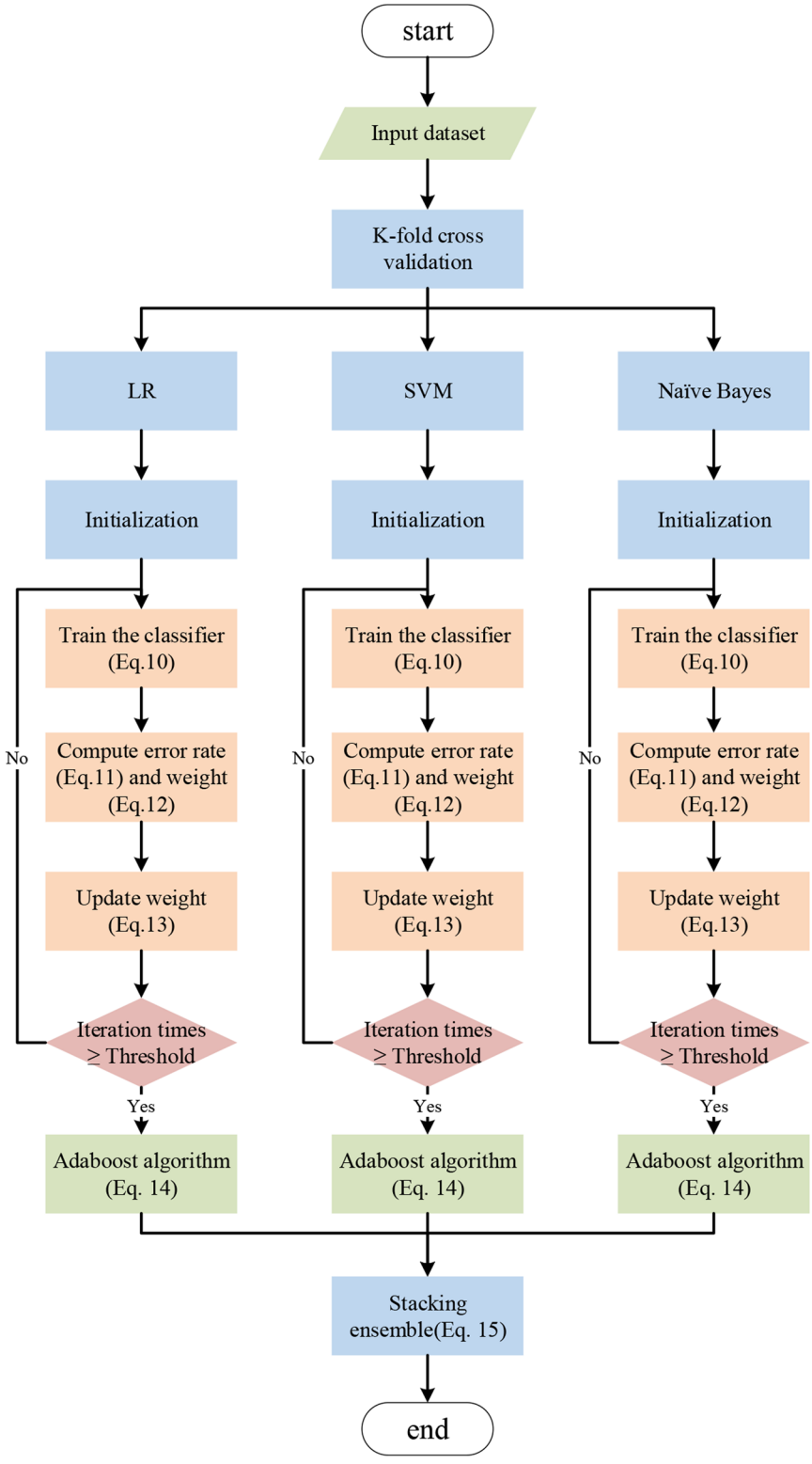

Fig. 5 The workflow of ensemble learning model development

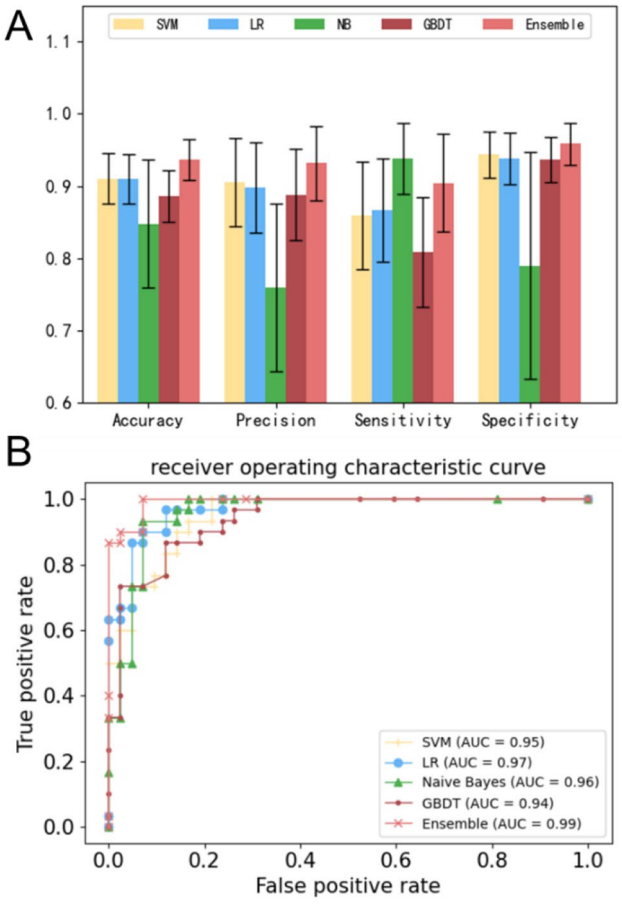

Fig. 6 Model performance. (**A**) Comparison of the classification performance of LR, SVM, Naive-Bayes, and ensemble learning models; (**B**) ROC curves plotted for LR, SVM, Naive-Bayes, and ensemble learning models

$$H_{mT}(x) = \sum_{t=1}^{T} \alpha_t h_t(x_i) \quad (14) \tag{14}$$

$$log\left(\frac{H(x)}{1-H(x)}\right) = c_0 + \sum_{m=1}^{M=3} c_m H_{mT}(x) \tag{15}$$

Here, $D_t(i)$ is the weight distribution, t is the iteration time, i is the index of the sample, and n is the number of samples. ϵ_t and α_t are the error rate and weight of each weak classifier h_t, respectively. For a sample set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x_n represents the samples and $y_n \in \{0, 1\}$ represents the labels; $y_i = 0$ indicates that x_i is not a patient with recurrence and metastasis, and $y_i = 1$ indicates that x_i is a patient with recurrence and metastasis. H_{mT} is the homomorphic integration for each weak classifier h_t; m is the index of the weak classifier, $m = 1, 2, 3$; T is the threshold of the iteration time; $H(x)$ is the ensemble classifier; and c_m is the weight of each weak classifier.

Predictive performance comparison

Figure 6A compares the classification performance for the LR, Naive-Bayes, SVM, GDBT, and ensemble learning models based on four commonly used classification measurements (Supplementary Table S4) [60]. Supplementary Table S5 lists the means and standard deviations for the results presented in Fig. 6A. Supplementary Table S6 lists the P values for LR, NB, SVM, GDBT and ensemble learning models. Figure 6 and Supplementary Tables S5 and S6 show the statistically significantly better classification performance of the ensemble learning model than that of the other four models. Figure 6B shows that if we comprehensively consider both sensitivity and specificity by constructing ROC curves [33], the ROC curve of the ensemble learning model is better than that of LR, Naive-Bayes, SVM and GDBT models.

Discussion

This study aimed to develop a multiomics data-based mathematical model to predict the recurrence and metastasis of colorectal cancer by answering three scientific questions.

To answer the first question, we used multiple data mining methods with the pipelines illustrated in Fig. 2 to explore the key features and employed PCA to reduce the dimensions of those features. Since Table 1 shows not only the selected features with statistically significant differences between positive and negative classes but also manually reviewed evidence indicating that COL6A3 [65] and TNM [66] are related to the development of colorectal cancer, OTOG [67] and KAL1 [68] are related to gastric cancer and oral squamous cell carcinoma, and most of the functions of proteomics and phosphoproteomics features [69] are related to cancer, we consider that these features can be employed as classifiers for our proposed predictive model. Moreover, since Fig. 3 shows that the positive and negative classes were successfully distinguished from each other, we consider that our dimensional reduction is efficient.

To answer the second question, we employed data augmentation to generate the pseudo dataset for model training (Table 3). After data augmentation, we calculated the F1 value [58] to evaluate the quality of the pseudo dataset. As shown in Fig. 4, the pseudo dataset generated by the SMOTE algorithm has a greater F1 value than the original dataset, indicating that the pseudo dataset not only meets the requirement of sample estimation but also ensures the data quality and robustness. Although SMOTE was the most used and effective method for numerical data augmentation, we have also tried other data augmentation methods, such as adding noise to create new data [70], but experiments showed that the data created by this method was not good enough (Shown in Figure S1, F1 value for different methods: Original: 1.260, SMOTE: 1.503, Noise: 1.259). As we explained above, the greater the F1 value, better quality of the generated data.

So, we can see the quality of SMOTE is better than other data augmentation methods.

To answer the third question, we developed an ensemble learning predictive model for the recurrence and metastasis of colorectal cancer. Figure 6 and Supplementary Table S6 show the significantly better performance of the ensemble model than the single classical machine learning model. However, Fig. 6A shows that the sensitivity of the ensemble learning model is not better than that of the Naïve bayes method. A potential explanation is that the ensemble learning model employs accuracy as the objective function to optimize the key weights (Eqs. 12 and 13) for each weak classifier, and thus it does not exhibit the best performance for the other three measurements, especially for sensitivity. On the other hand, Fig. 6B shows that the ROC curves of ensemble learning are better than those of the other three models, implying that the ensemble model still performs better than the single classical machine learning model if we comprehensively consider both sensitivity and specificity.

Conclusion

This study developed a multiomics data-based mathematical model to predict the recurrence and metastasis of colorectal cancer. First, we develop a feature selection and high dimensionality reduction algorithm that processes these high-dimensional multiomics colorectal cancer datasets. Second, we employ a computational algorithm to perform data augmentation for colorectal cancer prediction. Third, we build a predictive model that takes advantage of multiomics data and results in a high predictive accuracy for the recurrence and metastasis of colorectal cancer.

Although we have already achieved substantial progress in predicting colorectal cancer recurrence and metastasis, the unclear connections between proteomics and phosphoproteomics data remain to be solved. Thus, we will integrate more multiomics data and advanced bioinformatics methods into the current predictive model to increase its predictive power in the distant future.

Abbreviations

LR	Logistic Regression
SVM	Support Vector Machine
NB	Naive-Bayes
ANOVA	Analysis of Variance
PCA	Principal Component Analysis
SMOTE	Synthetic Minority Oversampling Technique
SNVs	Single-Nucleotide Variants
INDELs	Insertions-Deletions
WES	Whole-Exome Sequencing
ROC	Receiver Operating Characteristic Curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03012-9.

Supplementary Material 1: Additional file 1– Supplementary Material: Supplementary information for the Supplementary Tables S1-S6, Figure S1 and code availability

Supplementary Material 2: Additional file 2– Supplementary Table S1: The detailed results of feature selection (the p values) are listed in Supplementary Table S1

Supplementary Material 3: Additional file 3– Supplementary Table S2: The results of dimensional reduction are listed in Supplementary Table S2

Supplementary Material 4: Additional file 4– Supplementary Table S3: The results of data augmentation are listed in Supplementary Table S3

About this Supplement

This article has been published as part of BMC Medical Informatics and Decision Making, Volume 25 Supplement 2, 2025: 17th International Symposium on Bioinformatics Research and Applications. The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-25-supplement-2.

Author contributions

LZ and BL conceived the study and developed the model. BL and MX performed the simulations for the model and wrote the manuscript. BL, RZ and LZ performed the analysis for the model. All authors read and approved the final manuscript.

Funding

This work was supported by grants from National Science and Technology Major Project (2021YFF1201200 and 2024ZD0532900), National Natural Science Foundation of China (62372316), and Key Projects of Sichuan Provincial Department of Science and Technology (2024YFHZ0091 and 2025YFHZ0066).

Data availability

The dataset supporting the conclusions of this article is available in the https://ars.els-cdn.com/content/image/1-s2.0-S153561082030413X-mmc2.xlsx.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 8 July 2023 / Accepted: 11 April 2025
Published online: 15 May 2025

References

1. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019;394(10207):1467–80.
2. Turajlic S, Swanton C. Metastasis as an evolutionary process. Science. 2016;352(6282):169.
3. Lambert AW, Pattabiraman DR, Weinberg RA. Emerg Biol Principles Metastasis Cell. 2017;168(4):670–91.
4. Sargent D, Sobrero A, Grothey A, O'Connell MJ, Buyse M, Andre T, Zheng Y, Green E, Labianca R, O'Callaghan C, et al. Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol. 2009;27(6):872–7.
5. Biglarian A, Bakhshi E, Gohari MR, Khodabakhshi R. Artificial neural network for prediction of distant metastasis in colorectal cancer. Asian Pac J Cancer Prev. 2012;13(3):927–30.

6. Miyoshi N, Ohue M, Yasui M, Noura S, Shingai T, Sugimura K, Akita H, Gotoh K, Marubashi S, Takahashi H, et al. Novel prognostic prediction models for patients with stage IV colorectal cancer after concurrent curative resection. ESMO Open. 2016;1(3):e000052.
7. Nagata H, Ishihara S, Oba K, Tanaka T, Hata K, Kawai K, Nozawa H. Development and validation of a prediction model for postoperative peritoneal metastasis after curative resection of Colon cancer. Ann Surg Oncol. 2018;25(5):1366–73.
8. Watanabe T, Kobunai T, Yamamoto Y, Kanazawa T, Konishi T, Tanaka T, Matsuda K, Ishihara S, Nozawa K, Eshima K, et al. Prediction of liver metastasis after colorectal cancer using reverse transcription-polymerase chain reaction analysis of 10 genes. Eur J Cancer. 2010;46(11):2119–26.
9. Yang J, Kim H, Shin K, Nam Y, Heo HJ, Kim GH, Hwang BY, Kim J, Woo S, Choi HS, et al. Molecular insights into the development of hepatic metastases in colorectal cancer: a metastasis prediction study. Eur Rev Med Pharmacol Sci. 2020;24(24):12701–8.
10. Zhang L, Zhang S. Using game theory to investigate the epigenetic control mechanisms of embryo development: comment on: epigenetic game theory: how to compute the epigenetic control of maternal-to-zygotic transition by Qian Wang. Phys Life Rev. 2017;20:140–2.
11. Zhang L, Liu Y, Wang M, Wu Z, Li N, Zhang J, Yang C. EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. J Mol Cell Biol. 2017;9(6):477–88.
12. Xia Y, Yang C, Hu N, Yang Z, He X, Li T, Zhang L. Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model. BMC Genomics. 2017;18(Suppl 1):950.
13. Zhang L, Qiao M, Gao H, Hu B, Tan H, Zhou X, Li CM. Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. Nanoscale. 2016;8(31):14877–87.
14. Lee S, Choe EK, Kim SY, Kim HS, Park KJ, Kim D. Liver imaging features by convolutional neural network to predict the metachronous liver metastasis in stage I-III colorectal cancer patients based on preoperative abdominal CT scan. BMC Bioinformatics. 2020;21(Suppl 13):382.
15. Taghavi M, Trebeschi S, Simoes R, Meek DB, Beckers RCJ, Lambregts DMJ, Verhoef C, Houwers JB, van der Heide UA, Beets-Tan RGH, et al. Machine learning-based analysis of CT radiomics model for prediction of colorectal metachronous liver metastases. Abdom Radiol (NY). 2021;46(1):249–56.
16. Li M, Zhu Y-Z, Zhang Y-C, Yue Y-F, Yu H-P, Song B. Radiomics of rectal cancer for predicting distant metastasis and overall survival. World J Gastroenterol. 2020;26(33):5008–21.
17. Xu Y, Ju L, Tong J, Zhou CM, Yang JJ. Machine learning algorithms for predicting the recurrence of stage IV colorectal Cancer after tumor resection. Sci Rep. 2020;10(1):2519.
18. Zhang L, Liu G, Kong M, Li T, Wu D, Zhou X, Yang C, Xia L, Yang Z, Chen L. Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. Bioinformatics. 2021;37(11):1554–61.
19. Jiang Z, Cheng D, Qin Z, Gao J, Lao Q, Ismoilovich AB, Gayrat U, Elyorbek Y, Habibullo B, Tang D, et al. TV-SAM: increasing Zero-Shot segmentation performance on multimodal medical images using GPT-4 generated descriptive prompts without human annotation. Big Data Min Analytics. 2024;7(4):1199–211.
20. You Y, Tan K, Jiang Z, Zhang L. Developing a Predictive Platform for Salmonella Antimicrobial Resistance Based on a Large Language Model and Quantum Computing. Engineering. 2025.
21. Colorectal cancer. Nat Reviews Disease Primers. 2015;1(1):15066.
22. Li C, Sun YD, Yu GY, Cui JR, Lou Z, Zhang H, Huang Y, Bai CG, Deng LL, Liu P, et al. Integrated omics of metastatic colorectal Cancer. Cancer Cell. 2020;38(5):734–e747739.
23. Reyes A, Marti J, Marfà S, Jiménez W, Reichenbach V, Pelegrina A, Fondevila C, Garcia Valdecasas JC, Fuster J. Prognostic prediction by liver tissue proteomic profiling in patients with colorectal liver metastases. Future Oncol (London England). 2017;13(10):875–82.
24. Ou J, Zhang L, Ru X. Re-examination of statistical relationships between dietary fats and other risk factors, and cardiovascular disease, based on two crucial datasets. Quant Biology. 2024;12(1):117–27.
25. Xiao M, Wei R, Yu J, Gao C, Yang F, Zhang L. CpG island definition and methylation mapping of the T2T-YAO genome. Genomics, Proteomics & Bioinformatics. 2024.
26. Xiao M, Xiao Y, Yu J, Zhang L. PCGIMA: developing the web server for human position-defined CpG Islands methylation analysis. Front Genet. 2024;15:1367731.
27. Zhang L, Song W, Zhu T, Liu Y, Chen W, Cao Y. ConvNeXt-MHC: improving MHC-peptide affinity prediction by structure-derived degenerate coding and the ConvNeXt model. Brief Bioinform. 2024;25(3).
28. Zhang L, Xiong Z, Xiao M. A review of the application of Spatial transcriptomics in neuroscience. Interdiscip Sci. 2024.
29. Gao J, Lao Q, Kang Q, Liu P, Du C, Li K, Zhang L. Boosting your context by dual similarity checkup for In-Context learning medical image segmentation. IEEE Trans Med Imaging. 2024;PP(1):310–9.
30. Huang H, Yang Y, Zhang Q, Yang Y, Xiong Z, Mao S, Song T, Wang Y, Liu Z, Bu H, et al. S100a4 + alveolar macrophages accelerate the progression of precancerous atypical adenomatous hyperplasia by promoting fatty acid metabolism. 2024.
31. You Y, Zhou F, Yue Y, Qiu Y, Wang X, Yu Y, Li B, Li R, Zhang L. The classical iterative HHL-based hemodynamic simulation quantum linear equation algorithm for abdominal aortic aneurysm. Eur Phys J Special Top. 2024.
32. Zhang L, Xiong Z, Xiao M. A review of the application of Spatial transcriptomics in neuroscience. Interdiscip Sci. 2024;16(2):243–60.
33. You Y, Ru X, Lei W, Li T, Xiao M, Zheng H, Chen Y, Zhang L. Developing the novel bioinformatics algorithms to systematically investigate the connections among survival time, key genes and proteins for glioblastoma multiforme. BMC Bioinformatics. 2020;21(Suppl 13):383.
34. Miyoshi N, Ohue M, Noura S, Yasui M, Sugimura K, Tomokuni A, Akita H, Kobayashi S, Takahashi H, Omori T, et al. Prognostic prediction models for colorectal Cancer patients after curative resection. Int Surg. 2016;101(9–10):406–13.
35. Lei Zhang JL, Ming X, Li Yang L, Zhang. Exploring the underlying mechanism of action of a traditional Chinese medicine formula, Youdujing ointment, for cervical cancer treatment. Quant Biology. 2021;0(0):0.
36. Liu G-D, Li Y-C, Zhang W, Zhang L. A brief review of artificial intelligence applications and algorithms for psychiatric disorders. Engineering. 2020;6(4):462–7.
37. Song H, Chen L, Cui Y, Li Q, Wang Q, Fan J, Yang J, Zhang L. Denoising of MR and CT Images Using Cascaded Multi-Supervision Convolutional Neural Networks with Progressive Training. Neurocomputing. 2021.
38. Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33(1):1–39.
39. You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, Deng S, Zhang L. Artificial intelligence in cancer target identification and drug discovery. Signal Transduct Target Ther. 2022;7(1):156.
40. Zhang Q, Zhang H, Zhou K, Zhang L. Developing a physiological Signal-Based, mean threshold and Decision-Level fusion algorithm (PMD) for emotion recognition. Tsinghua Sci Technol. 2023;28(4):673–85.
41. Zhang L, Fan S, Vera J, Lai X. A network medicine approach for identifying diagnostic and prognostic biomarkers and exploring drug repurposing in human cancer. Comput Struct Biotechnol J. 2023;21:34–45.
42. Zhang L, Badai J, Wang G, Ru X, Song W, You Y, He J, Huang S, Feng H, Chen R, et al. Discovering hematoma-stimulated circuits for secondary brain injury after intraventricular hemorrhage by Spatial transcriptome analysis. Front Immunol. 2023;14:1123652.
43. You Y, Zhang L, Tao P, Liu S, Chen L. Spatiotemporal transformer neural network for Time-Series forecasting. Entropy (Basel). 2022;24(11):1651.
44. Xiao M, Ma F, Yu J, Xie J, Zhang Q, Liu P, Yu F, Jiang Y, Zhang L. A computer simulation of SARS-CoV-2 mutation spectra for empirical data characterization and analysis. Biomolecules. 2022;13(1):63.
45. Lai X, Zhou J, Wessely A, Heppt M, Maier A, Berking C, Vera J, Zhang L. A disease network-based deep learning approach for characterizing melanoma. Int J Cancer. 2022;150(6):1029–44.
46. Fan SW, Xiao M, Sun BY, Zhou WZ, Chen QR, Lv WM, Zhang PF, Zhang L. ASTM: developing the web service for anthrax related Spatiotemporal characteristics and meteorology study. Quant Biology. 2022;10(1):67–78.
47. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol. 2019;15(5):e1007069.
48. Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. Brief Bioinform. 2021;22(1):515–25.
49. Gao J, Liu P, Liu G-D, Zhang L. Robust needle localization and enhancement algorithm for ultrasound by deep learning and beam steering methods. J Comput Sci Technol. 2021;36(2):334–46.
50. Kaufmann J, Schering AG. Analysis of variance ANOVA. Wiley Encyclopedia of Clinical Trials; 2007.

51. Xiao M, Liu G, Xie J, Dai Z, Wei Z, Ren Z, Yu J, Zhang L. 2019nCoVAS: developing the web service for epidemic transmission prediction, genome analysis, and psychological stress assessment for 2019-nCoV. IEEE/ACM Trans Comput Biol Bioinf. 2021;18(4):1250–61.
52. Xiao M, Yang X, Yu J, Zhang L. CGIDLA:Developing the web server for CpG Island related density and LAUPs (Lineage-Associated underrepresented Permutations) study. IEEE/ACM Trans Comput Biol Bioinf. 2020;17(6):2148–54.
53. Zhang L, Xiao M, Zhou J, Yu J. Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). Bioinformatics. 2018;34(21):3624–30.
54. Lv J, Deng S, Zhang L. A review of artificial intelligence applications for antimicrobial resistance. Biosaf Health. 2021;3(1):22–31.
55. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.
56. Wu W, Song L, Yang Y, Wang J, Liu H, Zhang L. Exploring the dynamics and interplay of human papillomavirus and cervical tumorigenesis by integrating biological data into a mathematical model. BMC Bioinformatics. 2020;21(Suppl 7):152.
57. Zhang L, Li J, Yin K, Jiang Z, Li T, Hu R, Yu Z, Feng H, Chen Y. Computed tomography angiography-based analysis of high-risk intracerebral haemorrhage patients by employing a mathematical model. BMC Bioinformatics. 2019;20(Suppl 7):193.
58. Fernández Hilario AL, García López S, Herrera Triguero F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. 2018.
59. Tin Kam H, Basu M. Complexity measures of supervised classification problems. IEEE Trans Pattern Anal Mach Intell. 2002;24(3):289–300.
60. Zhang L, Zheng C, Li T, Xing L, Zeng H, Li T, Yang H, Cao J, Chen B, Zhou Z. Building up a robust risk mathematical platform to predict colorectal Cancer. Complexity. 2017;2017:8917258.
61. Lei W, Zeng H, Feng H, Ru X, Li Q, Xiao M, Zheng H, Chen Y, Zhang L. Development of an Early Prediction Model for Subarachnoid Hemorrhage With Genetic and Signaling Pathway Analysis. 2020;11(391).
62. Pearce J, Ferrier S. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model. 2000;133(3):225–45.
63. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9(3):293–300.
64. Salmi N, Rustam Z. Naïve Bayes Classifier Models for Predicting the Colon Cancer. IOP Conference Series: Materials Science and Engineering. 2019;546:052068.
65. Liu W, Li L, Ye H, Tao H, He H. Role of COL6A3 in colorectal cancer. Oncol Rep. 2018;39(6):2527–36.
66. Xu W, He Y, Wang Y, Li X, Young J, Ioannidis JPA, Dunlop MG, Theodoratou E. Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies. BMC Med. 2020;18(1):172.
67. Wu X, Liu M, Zhu H, Wang J, Dai W, Li J, Zhu D, Tang W, Xiao Y, Lin J, et al. Ubiquitin-specific protease 3 promotes cell migration and invasion by interacting with and deubiquitinating SUZ12 in gastric cancer. J Exp Clin Cancer Res. 2019;38(1):277.
68. Liu J, Cao W, Chen W, Xu L, Zhang C. Decreased expression of Kallmann syndrome 1 sequence gene (KAL1) contributes to oral squamous cell carcinoma progression and significantly correlates with poorly differentiated grade. J Oral Pathol Medicine: Official Publication Int Association Oral Pathologists Am Acad Oral Pathol. 2015;44(2):109–14.
69. The UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.
70. Attar AA, Schirle F, Hofmann M. Noise added on interpolation as a simple novel method for imputing missing data from household's electricity consumption. Procedia Comput Sci. 2022;207:2253–62.

Publisher's note