scientific reports

Check for updates

OPEN

A fusion model to predict the survival of colorectal cancer based on histopathological image and gene mutation

Binsheng He[1,2,5], Lixia Wang[3,5], Wenjing Zhou[4], Haiyan Liu[1,2], Yingxuan Wang[3], Kebo Lv[3]✉ & Kunhui He[1,2]✉

Colorectal cancer (CRC) is a prevalent gastrointestinal tumor worldwide with high morbidity and mortality. Predicting the survival of CRC patients not only enhances understanding of their life expectancies but also aids clinicians in making informed decisions regarding suitable adjuvant treatments. Although there are many clinical, genomic, and transcriptomic studies on this hot topic, only a few studies have explored the direction of integrating advanced deep learning algorithms and histopathological images. In addition, it is still unclear if combining histopathological images and molecular data can better predict patients' survival. To fill in this gap, we proposed in this study a novel multimodal deep learning computational framework using Multimodal Compact Bilinear Pooling (MCBP) to predict the 5-year survival of CRC patients from histopathological images, clinical information, and molecular data. We applied our framework to the cancer genome atlas (TCGA) CRC data, consisting of 84 samples with histopathological images, clinical information, mRNA sequencing data, and gene mutation data all available. Under the 5-fold cross-validation, the model using only histopathological images achieved an area under the curve (AUC) of 0.743. Whereas, the model combining image and clinical information and the model combining image and gene mutation information achieved AUCs of 0.771 and 0.773 respectively, better than that of the image solely. Our study demonstrates that histopathological images can reasonably predict the 5-year survival of CRC patients, and that the appropriate integration of these images with clinical or molecular data can further enhance predictive performance.

Keywords Colorectal cancer, Overall survival, Histopathological image, Multimodal deep learning, Multimodal compact bilinear pooling

In 2022, there are around 590,000 and 160,000 new cases of colorectal cancer (CRC) in China and the United States (US) respectively, making CRC the second and fourth most common cancer respectively in the two countries[1]. In addition, CRC is the fifth leading cause of cancer death in China and the second leading cause of cancer death in the US[1], posing a big threat to public health. According to the statistics from the FIGHT colorectal cancer website, the 5-year overall survival (OS) rates in localized cancer (stage I and II, accounting for 39% of all CRC), regional cancer (stage III, 35%), and distant cancer (stage IV, 22%) are 89.9%, 71.3%, and 14.2% respectively[2]. The main causes of CRC mortality are frequent recurrence and metastasis, and the main treatment strategy for CRC is surgical resection concomitant with chemotherapy, radiotherapy, immunotherapy, and so on[3–5]. With the improvement in treatment methods and deeper understanding of CRC tumor biology, many diagnostic and prognostic factors for CRC have been identified[6–8]. However, it is very difficult to predict precisely the survival of CRC patients even for experienced doctors, which is of great value to deciding the concomitant therapy after surgery[9]. Thus, it is critical to identify CRC prognostic factors and develop accurate prediction models for CRC survival.

[1]First Clinical College, Changsha Medical University, Changsha 410219, P. R. China. [2]Hunan Provincial Key Laboratory of the Traditional Chinese Medicine Agricultural Biogenomics, Changsha Medical University, Changsha 410219, P. R. China. [3]School of Mathematical Sciences, Ocean University of China, Qingdao 266000, P. R. China. [4]Qingdao Hiser Hospital Affiliated of Qingdao University (Qingdao Traditional Chinese Medicine Hospital), Qingdao 266033, P. R. China. [5]Binsheng He and Lixia Wang have contributed equally to this work. ✉email: kewave@ouc.edu.cn; iuhe0405@163.com

CRC is a disease that is strongly affected by gender, and the mortality rate in men is significantly higher than that in women[10]. It is known that pathological characteristics of resected specimens are powerful predictors of postoperative prognosis. Clinically, the prognosis is predicted by environmental factors, molecular subtype, histological grade, and tumor stage[11]. Due to the limitations of clinical indicators, Lin et al. investigated the prognostic value of lncRNAs. Key genes were screened out and a prognostic risk score model was established by employing univariate Cox regression, lasso, and multivariate Cox regression analysis. And a graphical nomogram was developed to predict the 1-, 3-, and 5-year survival rates of patients with areas under the curve (AUCs) of 0.656 and 0.624 respectively at one and three years[12]. Qian et al. established a miRNA-based CRC prognosis prediction model, the AUCs of the overall survival and relapse-free survival (RFS) models were 0.657 and 0.651, respectively. In addition, the functional annotation of the miRNA genes used showed the underlying mechanism of CRC in this study[13]. Finally, Yang et al. performed the differential analysis of miRNA expression profiles to screen out survival-related biomarkers and constructed a multivariate Cox regression model. The AUC of the five-year survival prediction model on the validation set was 0.697 [14].

Image-based machine learning, especially deep learning, has recently demonstrated expert-level accuracy in medical image classification and prognostic prediction[6,15–26]. Fu et al. used deep learning to quantify H&E-stained images from 28 cancer types and to classify cancer types, simultaneously distinguishing tumor tissues from normal tissues[27]. A retrospective study by Kather et al. found that CNNs can directly assess the human tumor microenvironment and predict prognosis from histopathological images[28]. Wulczyn et al. developed a deep learning system for predicting disease-specific survival in stage II and III CRC, with model five-year disease-specific survival AUCs of 0.70 and 0.69 on two different validation sets, respectively. The limitation of this study is that both validation sets used are from the same dataset and there is no independent validation set[29]. With the development of sequencing technology, other omics information such as genome and transcriptome can help to obtain information on different dimensions of disease and make up for the lack of single omics information. Cheerla et al. investigated the prognostic predictive power of histopathological images combined with other 105 omics in pancreatic cancer data, which achieved an overall C-index of 0.78. The exploration of genomic data in this study was limited to gene expression data, and integration of more diverse data is a goal for future research[30].

Although numerous studies have investigated CRC survival through molecular data and imaging approaches, an effective fusion model is still lacking. This study presents a novel deep learning framework developed to predict the 5-year OS in 84 colorectal cancer patients. Initially, we utilized histopathological images to assess their predictive capability regarding CRC survival. Furthermore, we integrated pathological images with molecular data to create advanced fusion models, significantly improving prediction accuracy. These models not only enable cancer patients to better understand their life expectancy but also assist clinicians in making informed decisions that guide appropriate treatment strategies. Additionally, differential analyses at the clinical, mRNA, and gene mutation levels provided insights into the biological mechanisms influencing survival among CRC patients to some extent.

Materials and methods

Data collecting

In this study, pathological images, clinical data, whole exome sequencing data, and mRNA data of CRC patients in TCGA (https://portal.gdc.cancer.gov/repository/) were collected. The number of samples is shown in Table 1. Survival data for patients including survival time and survival status were divided into two categories, using 5 years as a threshold. Samples with a survival time of more than 5 years were classified as the Long term (LT) group, while samples with a survival time of less than 5 years were classified as the Short term (ST) group. Deleted samples were eliminated. After taking overlapping samples of various types of data, 84 samples with images, mRNA sequencing data, and gene mutation data were obtained, of which 28 survived for more than 5 years and 56 survived for less than 5 years. All research will be based on this sample.

Preprocessing of the image data

Whole Slide Image (WSI) is a full-field digital pathological section image obtained through a pathological section scanner. Since WSI is large and cannot be used directly in deep learning, WSI needs to be cut into small pieces. A pathologist was invited to annotate the tumor area of interest to us, and the annotated area was cut into 512×512 px pieces. When performing sectioning, first determine the outer rectangle of the tumor labeling area, and then use a 512×512 size sliding window to split it into small slices, with no overlap between the small slices. OpenCV is used to do the slicing and filter out small slices that contain less information (e.g. blank rate over 30%)[31].

Due to the use of different microscopes/scanners, or inconsistencies in tissue preparation and storage, shades may be inconsistent between samples[32]. To further reduce the experimental error, this experiment used the Macenko method of the tiatoolbox framework to normalize the colors according to the template file. The Macenko method converts color values in RGB space to the corresponding optical density values and then normalizes the staining using a color deconvolution technique[33].

Image feature extraction

VGGNet has many applications in image detection, localization, classification, and segmentation[34–37]. The network structure is suitable for classification tasks and localization tasks in the ImageNet dataset, and it generalizes well to other datasets. The outstanding contribution of VGGNet is to demonstrate that very small convolutions can effectively improve performance by increasing the network depth. VGGNet uses multiple convolutional layers with smaller convolution kernels (3×3) instead of one convolutional layer with larger

Clinicopathologic variable	Category	Case
Sample type	H&E	360
	Clinic	462
	Mutation	405
	mRNA	531
Age	< 50	8
	>=50	76
Gender	Female	40
	Male	44
Tumor status	T1	3
	T2	5
	T3	65
	T4	11
Lymph node status	N0	40
	N1	21
	N2	23
Distant metastasis	M0	51
	M1	18
	MX	12
	NA	3
Cancer stage	I	8
	II	29
	III	25
	IV	18
	NA	4
pharmaceutical therapy	NO	36
	YES	24
	NA	24

Table 1. Summary of the general information of patients.

convolution kernels. On the one hand, parameters can be reduced. On the other hand, it is equivalent to more nonlinear mapping, which can increase the Fitting ability of the network[38].

Vgg16 is used as the image feature extractor in this study. Importing a pre-trained model based on the ImageNet dataset and then training our model on this. The pre-trained model based on big data learns the common features in the data better, which usually makes the model achieve better results. And it can increase the training speed, making the model converge faster. When training the model, the data is enhanced by scaling, horizontal flipping, etc., to increase the amount of training data and improve the generalization ability of the model.

Figure 1 shows more details of the Vgg16 model, such as the number of layers, and input and output dimensions. Since this study is a dichotomous classification task, the model is changed to a dichotomous classification model by modifying the classification number based on the original VGG16 network. The trained model is saved in this study, and the network structure and parameters before the fully connected layer are saved and used as the image feature extraction model. We specify the values of the hyperparameters, the learning rate is set to 0.001, the loss function is set to cross-entropy loss function, and the epoch is set to 30.

Molecular feature extraction

Due to the high dimension of molecular data, there may be redundant or even irrelevant features, which will seriously affect the model effect. Feature selection can effectively prevent the disaster of dimensionality, reduce the difficulty of learning tasks, and can enhance the generalization ability of the model, and reduce the problem of overfitting. In molecular data preprocessing, labels are encoded for categorical variables and columns with null values greater than 25% in all features are deleted. Random forests are strong classifiers composed of decision tree classifiers and are often used in classification tasks[39]. On the one hand, random forests are used to model molecular data from different omics to compare the survival prediction ability of different omics. On the other hand, random forests are used for feature selection, and feature importance is measured using the Gini index. Appropriate features are selected by ranking their importance.

Feature fusion

There are many methods for the fusion of the two features, such as element-wise sum, element-wise product, concatenation, etc. The Bilinear CNN Model proposed later is a method that uses the outer product of the two features to fuse. The original Bilinear Pooling has the problems that the dimension of the fused feature is too high and the calculation complexity of the outer product is too high. Some studies have proposed Random Maclaurin (RM) and Tensor Sketch (TS) methods to reduce the dimension of bilinear pooling[40]. The multimodal compact bilinear pooling method adopted in this paper is an improvement based on the TS method to adapt to the fusion of different modal features. And MCBP maps the result of the outer product to a low-dimensional space without explicitly calculating the outer product. In this method, the two modal feature vectors extracted by the neural network are obtained by the Count Sketch mapping function respectively. Then, fused features are obtained by Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT)[41].

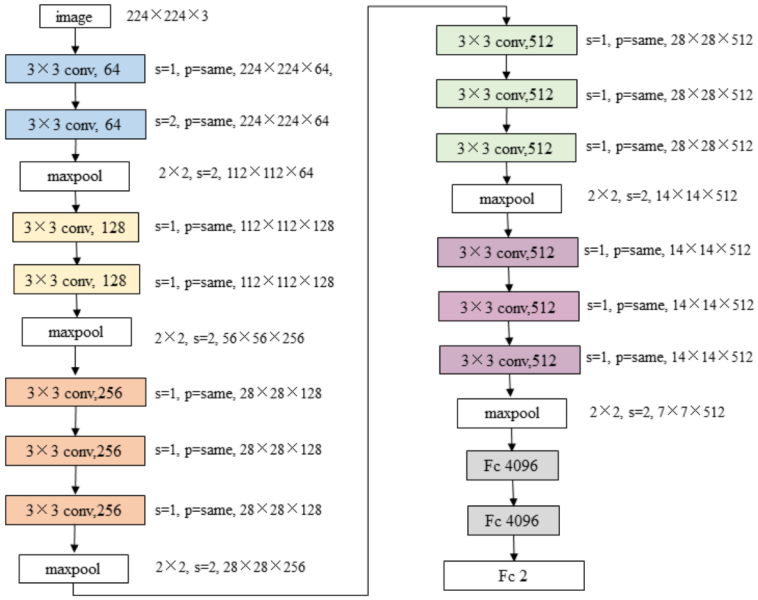

Fig. 1. Structure diagram of Vgg16 network.

Statistical analysis

All statistical analyses were performed based on the R software, and survival-related characteristics were analyzed from three perspectives: clinical, gene expression, and somatic mutation.

Clinical features were coded with label encoding. The log-rank test was used to test whether gender, age, TNM, and tumor stage had significant differences in survival among different groups. For somatic mutations, the R package maftools was used to analyze and visualize somatic mutations including SNVs, SNPs, INS, and DELs based on samples from whole genome sequencing (WES). Fisher's exact test was performed for ST and LT cohorts to detect differentially mutated genes. Co-occurring and exclusive mutations in the top 10 genes with the highest mutation frequencies in both cohorts were analyzed using the CoMEt algorithm[42]. Differentially expressed genes (DEGs) by mRNA sequence in the two groups were analyzed using the R package Deseq2[43]. Genes with an expression fold change greater than 2 and adjusted p value less than 0.05 were up-regulated genes, and genes with an expression fold change of less than −2 and adjusted p value less than 0.05 were down-regulated genes. The screened DEGs were enriched by GO and annotated from three aspects: biological process (BP), molecular function (MF), and cellular component (CC).

Results

The overall framework of the study

The overall framework of the study is shown in Fig. 2. In this study, A framework was established to identify colorectal cancer survival-related biomarkers and predict survival. First, we collected clinical, whole-exome sequencing, and mRNA molecular information of colorectal cancer patients, as well as H&E-stained histopathological images (Fig. 2a). Differential analysis was performed for molecular data to identify biomarkers associated with survival (Fig. 2b). On the other hand, a novel fusion model combining images with molecular data was established to predict the five-year survival of colorectal cancer patients (Fig. 2c).

Clinical correlation with CRC survival

The clinical information of the 84 samples used in the study is shown in Table 1. Most of the patients in the sample are over 50 years old, and the ratio of males and females is close. In TNM staging, the samples are concentrated in the T1, N0, and M0 stages, and in pathological staging, the sample distribution is relatively balanced, and there are relatively a few samples in stage I. To further investigate the relationship between clinical characteristics and survival, a survival analysis was performed, and a Kaplan–Meier plot was drawn as shown in Fig. 3. There are significant differences in the survival of patients with different tumor stages, and the differences between N0 (no regional lymph node metastasis), N1 (1–3 regional lymph node metastasis), and N2 (≥ 4 regional lymph node metastasis) significantly indicate that the surrounding area of the tumor is invaded. The number of lymph nodes correlated with patient survival. Significant differences between M0 (no distant metastases found), M1 (presence of distant metastases), and MX (not evaluable) suggested that distant metastases of tumors were associated with patient survival. In our test, there was no significant difference in survival among patients with different T stages. There was also no significant difference in survival between male and female patients. There were 24 samples

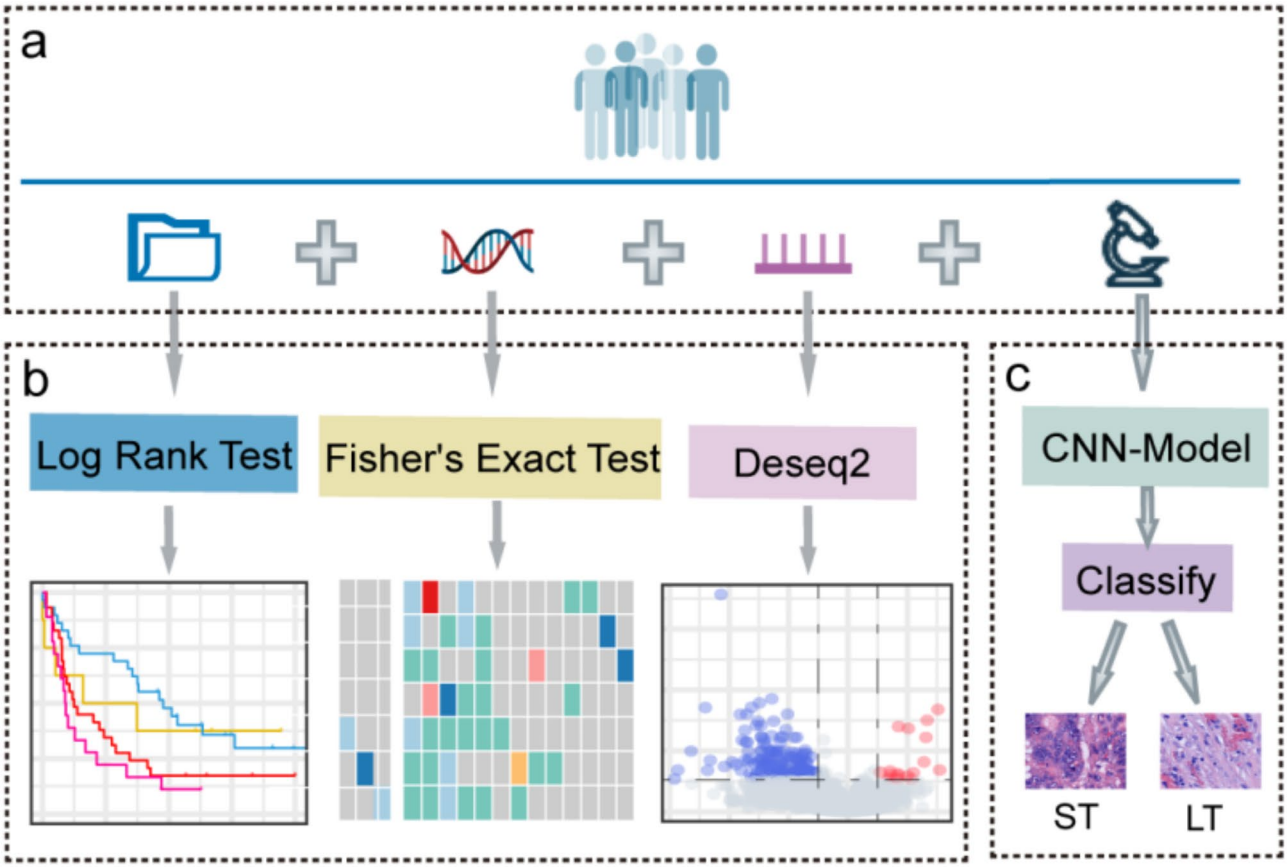

Fig. 2. The overall framework of the study. (**a**) Collection of pathological images and molecular data. (**b**) Differential analysis based on molecular data. (**c**) Building a fusion model to predict survival.

that had not reported information on drug therapy. We only used 60 samples for survival analysis and did not find any significant impact.

Differentially mutated genes between LT and ST groups

All samples were divided into Long-term (LT) and Short-term (ST) groups based on 5-year survival. First, the gene mutations of the two groups of samples were analyzed and visualized separately, and the number of SNPs in all mutation types far exceeded that of DEL and INS (Fig. 4a). Among the six SNV mutations, C > T mutations accounted for the largest proportion (Fig. 4b). In the colorectal cancer patient population, most mutations were missense mutations, and other mutation types were relatively rare (Fig. 4c). To compare the gene mutation status of the samples in the two cohorts, we used Fisher's exact test to find 7 differential genes with different mutation frequencies between the two cohorts. As shown in Fig. 4d, the ST group contained more mutations. To further analyze the mutation status of differential genes in the two cohorts, Fig. 4e shows the detailed mutation status of differential genes in the two cohorts. There are six mutation types in total, and five genes were mutated more than once in the samples. Next, the CoMEt[42] algorithm was used to analyze the co-occurrence and exclusive mutations of the top 10 genes with the highest mutation frequency, as shown in Fig. 4f, there were three unique sets of genes showing mutually exclusive mutations in both cohorts (TP53-PIK3CA, TTN-APC, KRAS-PYR1), suggesting that they may have redundant roles under the same conditions and the selective advantage between them can retain multiple copies of mutations.

Differential genes identified between LT and ST groups

To explore whether changes in high or low expression of certain key genes were associated with survival, we used DESeq2[43] to analyze gene expression differences between the LT and ST groups. With log2 |fold change| ≧ 1 and p-value ≤ 0.05 as the threshold, there were 203 significantly differentially expressed genes in the LT group compared with the ST group. Among them, 19 genes were up-regulated and 184 genes were down-regulated. In Fig. 5a, red dots indicate up-regulated genes, blue dots indicate down-regulated genes and gray dots indicate genes with no significant difference in expression. The expression of the top 10 genes up-regulated and down-regulated in the samples are shown in Fig. 5b, respectively.

The screened differential genes were subjected to Go enrichment analysis to clarify the functions of these differentially expressed genes. The enrichment analysis of the biological process, cellular component, and molecular function was carried out. The biological process obtained by enrichment is shown in Fig. 5c. The

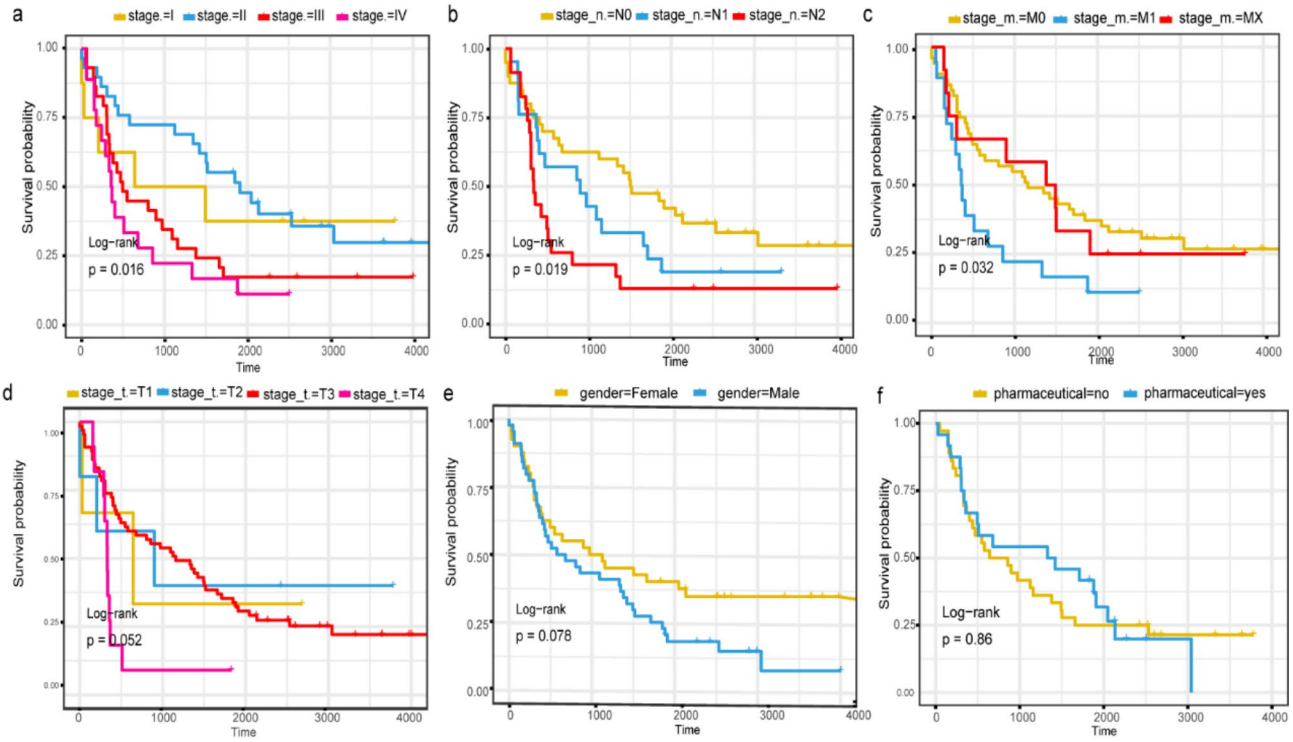

Fig. 3. Kaplan–Meier plot of clinical features. (**a**–**c**) Survival-related clinical features. (**d**–**f**) Clinical features not significantly associated with survival.

results showed that differential genes were enriched in immune-related aspects, such as: immunoglobulin complex; immunoglobulin complex, circulating; and immunoglobulin receptor binding. Figure 5d showed the relationship between genes and enriched GO terms.

A deep learning framework for predicting survival

The complete process of predicting the survival of colorectal cancer based on the novel fusion model based on the Convolutional Neural Network (CNN) is shown in Fig. 6. First, whole sections of pathological tissues of patients with colorectal cancer were collected and preprocessed, such as tumor region annotation, image segmentation, and color normalization. Then a multimodal fusion model was established, in which image features were extracted using a VGG network, and the output dimension of the fully connected layer was modified after the convolutional and pooling layers for fusion with molecular features. Molecular features were selected according to the importance of random forest features, and the selected molecular features and image features were fused into a feature vector using a multimodal compact bilinear fusion method, and then the predicted probability of each tile was output through the BatchNorm layer, and the classification layer. Finally, the prediction results of tiles were aggregated into the prediction results of each patient using the method of Voting Classifier. To measure model performance, the model was evaluated using the average area under the curve (AUC) of 5-fold cross-validation.

Histopathology images can be used to predict survival and outperformed molecular data

The preprocessed images were divided into the training set and validation set according to five-fold cross-validation, and the division ensured that small tiles of the same patient were divided into the same dataset to prevent information leakage. The samples were then trained and validated using the neural network model. The test sample input model can get the predicted probability of each small tile, and then all the small tiles of each sample are aggregated together by taking the average to obtain the predicted probability of each patient. If the predicted probability of the patient was greater than 0.5, it was predicted to be the ST cohort, otherwise, it was predicted to be the LT cohort.

The Vgg16 model was then compared with other existing models under the 5-fold cross-validation. For a fair comparison, we ensured that the partition of the data was the same across all methods. The results were summarized in Table 2. From the table, it can be seen that the vgg16 model has the highest AUC and relatively good performances for other evaluation criteria. Therefore, we chose the vgg16 model for subsequent analysis.

Molecular data were modeled separately and compared with the prediction results of H&E-stained images. Random forests were used to build prediction models for gene mutation and mRNA respectively, and the hyperparameters were selected by grid search. The grouping of 5-fold cross-validation was kept consistent during the comparison. Figure 7a shows the ROC curves predicted by the three omics of the image, mRNA, and gene mutation alone. H&E-stained image has the best prediction effect, with the AUC of 0.743, which is higher

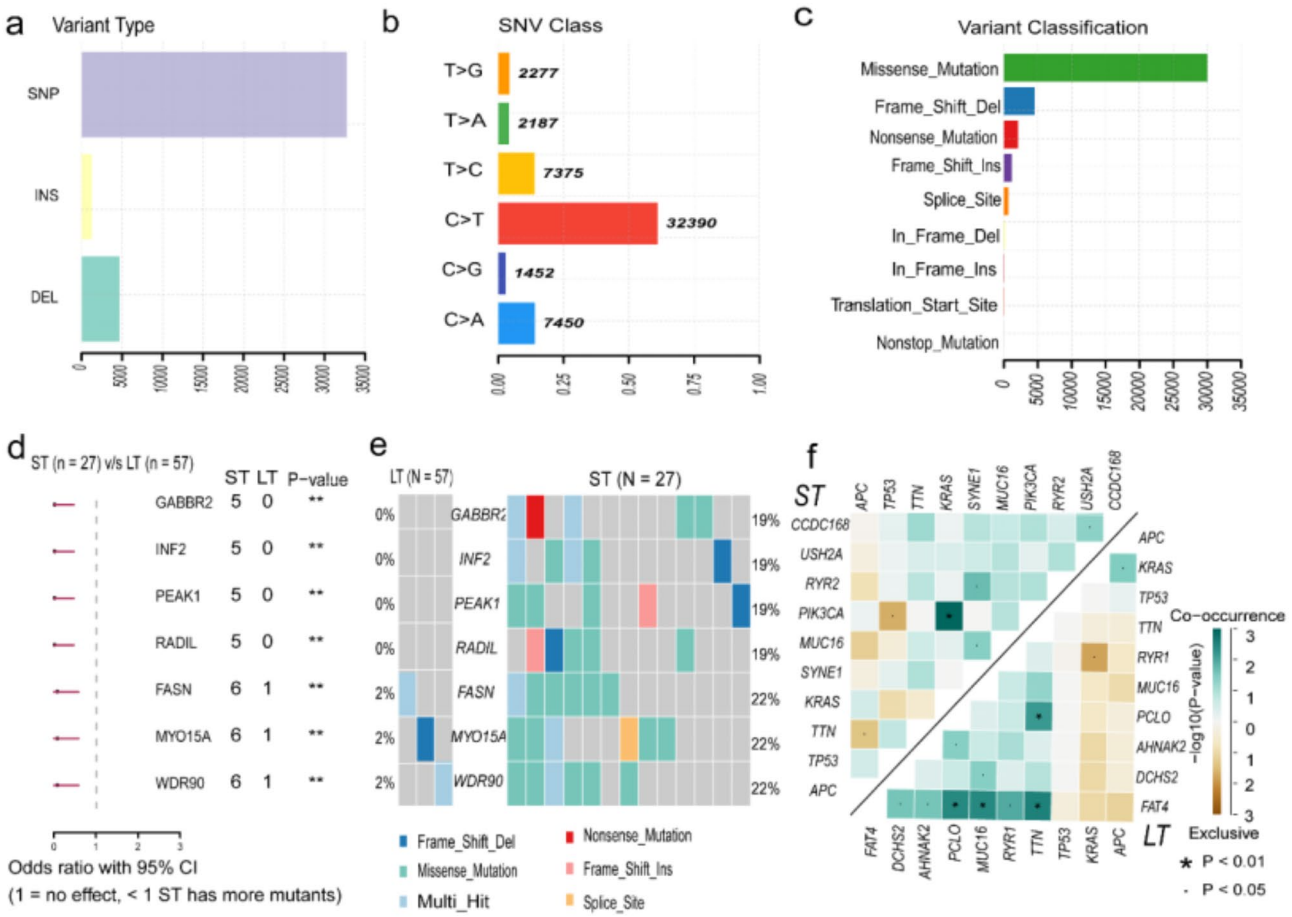

Fig. 4. Gene mutation. (**a**) Classification of mutations resulting from changes in gene structure. (**b**) Six SNV mutations. (**c**) Base substitution mutation. (**d**) Differential genes in the two cohorts. (**e**) Differential gene mutation status. (**f**) Co-occurrence and mutual exclusion of genes with the top 10 mutation frequencies.

than the prediction of mRNA and gene mutation. Figure 7b shows other evaluation metrics, on each of which, the predictive power of images is better than that of molecular data alone.

The introduction of clinical information can improve model performance

Based on the model of images, the effect of introducing molecular features on the performance of the model was explored. The molecular features of feature importance ranking top1, top2, and top3 are fused with image features respectively. Because of the high dimensionality of mRNA and gene mutation features, feature fusions of top50, top100, and top150 were also experimented. The results of 5-fold cross-validation, when only one feature is fused, are shown in Fig. 7c. The introduction of clinical and gene mutation features will improve the model effect to varying degrees. The AUC of the image combined with the gene mutation feature was 0.773, which was improved compared to the prediction result of the pure image (average AUC was 0.743). This showed that images and gene mutations may work together to predict the survival of colorectal cancer patients, and the fusion features improved the prediction performance of the model. In addition, combining images with clinical information can also improve the prediction performance of the model. The average AUC of the model is 0.771, indicating that the introduction of clinical information can also make up for the lack of image information prediction. However, it can be seen from the results that the prediction performance of the image combined with mRNA features was lower than that of the image, which may be because the introduced mRNA features contained less useful information and introduced relatively more noise. Figure 7d showed other evaluation indicators of the model. Overall, the fusion model of the image combined with gene mutation outperformed other models in all indicators.

Discussion

Colorectal cancer is a prevalent malignancy characterized by a high incidence rate and significant cancer-related mortality. Surgical intervention remains the standard treatment for both early and advanced stages of colorectal cancer, with an average postoperative complication rate of 40% [42]. Therefore, it is very important to find effective predictive tools to predict the prognosis of each patient and provide timely and accurate treatment. We explored the biological characteristics related to survival from three levels of clinical, mRNA, and gene mutation, and

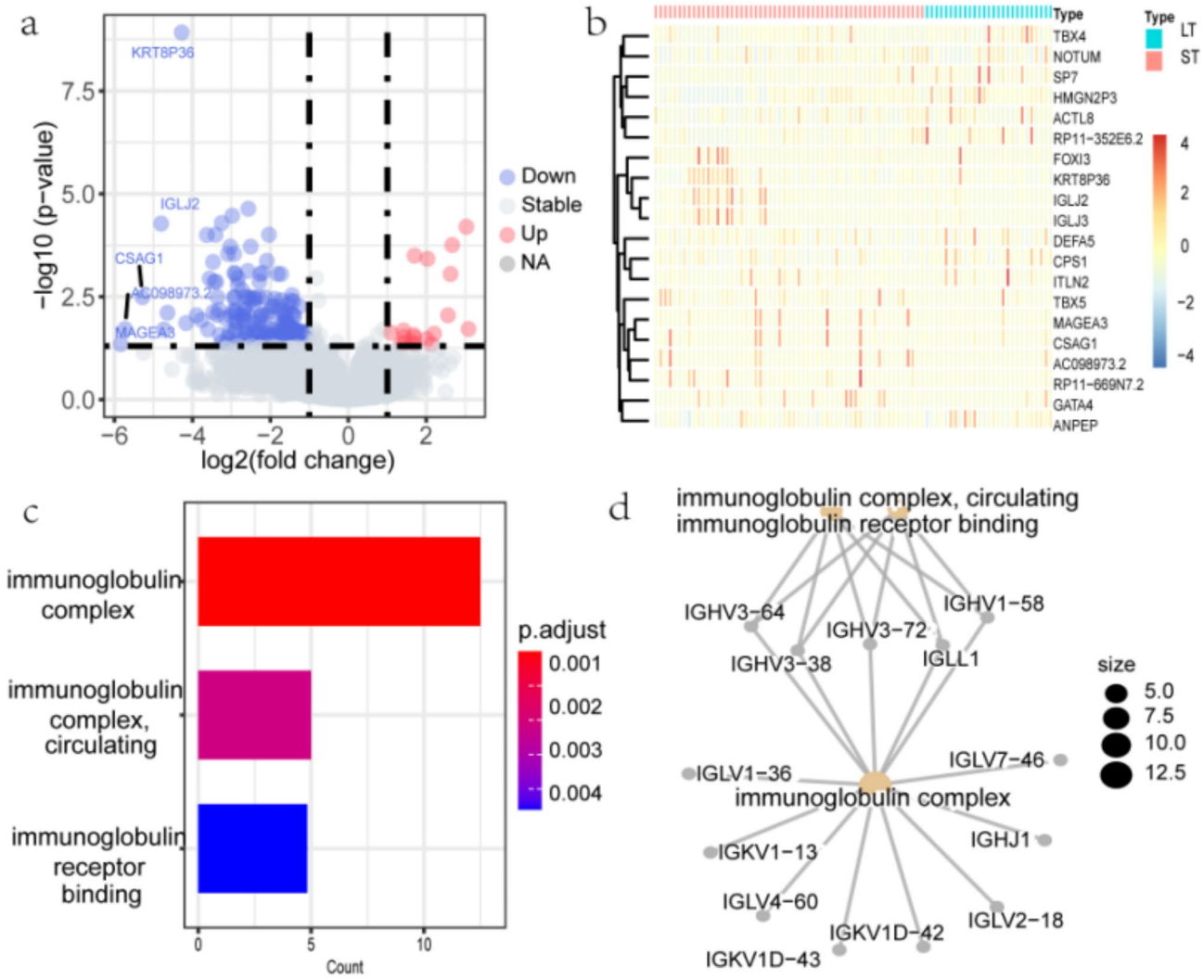

Fig. 5. Differentially expressed genes. (**a**) Volcano plot of differentially expressed genes. (**b**) Heatmap of the top 10 up- and down-regulated genes with the largest fold difference in samples. (**c**) Enriched GO terms for biological process annotations. (**d**) Correspondence between genes and enriched GO terms.

found some biomarkers related to survival. We found that pathological stage, lymph node metastasis, and distant metastasis were significantly associated with overall survival. This has been reported in a previous article, which is consistent with our findings[44]. Some gene mutations are associated with survival. Among the genes identified in the study, FASN, RADIL, INF2, and PEAK1 have been previously reported to be associated with colorectal cancer development and disease progression[45–48]. The differential analysis of mRNA in LT and ST groups found 203 differential genes. The differential genes were enriched in immune-related pathways. This indicates that immune related genes (IRGs) may have a certain impact on the prognosis of colorectal cancer patients. Previous studies have conducted a comprehensive and comprehensive analysis of IRGs involved in CRC, explaining their clinical value, and revealing the potential impact of IRGs in CRC immunotherapy. Studies suggested that some IRGs were significantly associated with CRC progression. We also demonstrated that inexpensive and readily available histopathological images had some predictive power for survival in colorectal cancer patients. We developed a novel fusion model combining H&E-stained images and molecular data, which improved the prediction performance of single omics.

In recent years, advancements in the Internet of Things (IoT) have transformed the diagnosis, treatment, and management of colorectal cancer. The IoT enables remote diagnosis and treatment through real-time telemedicine and consultations. By collecting and transmitting pathological images, radiographic data, and clinical information instantaneously, healthcare providers can monitor the progression of colorectal cancer, facilitating timely interventions that may improve patient outcomes. Specifically, our method provides timely survival predictions within this framework, further enhancing the utility of IoT. Furthermore, the vast amounts of data generated by IoT devices can be utilized to enhance the accuracy of predictive algorithms. As technology continues to advance, the integration of IoT into colorectal cancer management is expected to improve outcomes, increase patient satisfaction, and foster a more efficient healthcare system.

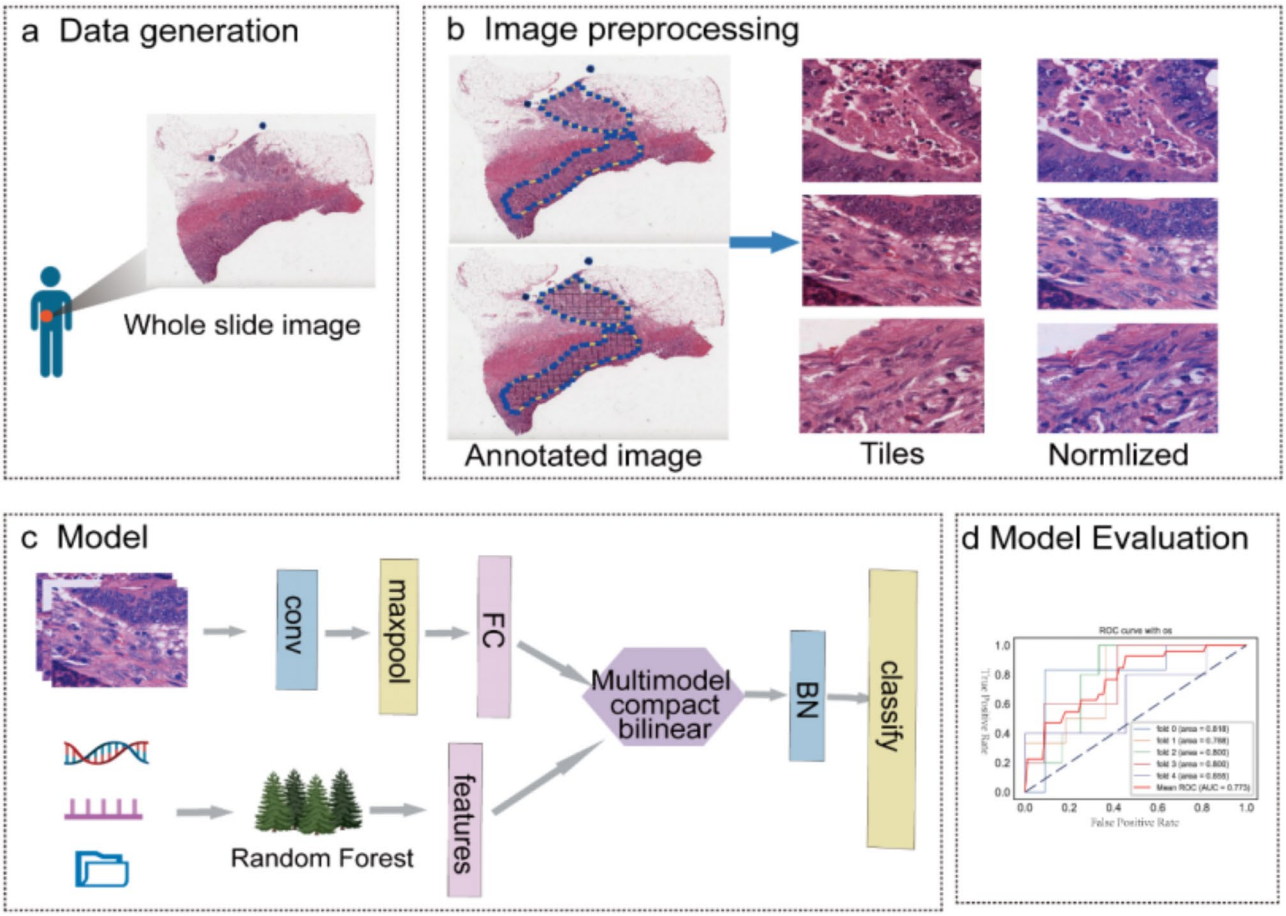

Fig. 6. Computational pipeline for predicting survival. (**a**) Whole slide images are generated from digitized pathological tissue sections. (**b**) Segmentation and color normalization of annotated regions. (**c**) Using MCBP to fuse image and molecular features for classification prediction. (**d**) Evaluating model performance using 5-fold cross-validation.

Model	AUC	ACC	Precision	Recall	f1-Score
AlexNet	0.667	0.595	0.618	0.634	0.589
VGG16	0.743	0.679	0.621	0.607	0.611
ResNet18	0.727	0.663	0.586	0.555	0.548
Inception V3	0.726	0.667	0.583	0.550	0.541

Table 2. Comparison of different neural network models.

Similar to IoT, deep learning has also emerged as a transformative technology in the realm of medical diagnostics, particularly in the detection and classification of CRC. Deep learning operates on neural networks that can automatically extract features from raw histopathology images. This capability is particularly beneficial in the context of CRC, where the complexity and variability of tissue samples can pose challenges for traditional diagnostic methods. For instance, Bousis et al. (2023) highlight that deep learning frameworks can process large datasets of histopathological images, learning to identify subtle patterns that may not be readily visible to human pathologists[49]. This capability can lead to earlier and more accurate diagnoses, potentially improving patient outcomes through timely interventions. Chlorogiannis et al. (2023) further emphasize the readiness for clinical practice implementation, suggesting that the integration of deep learning algorithms could streamline the workflow of pathologists by providing them with second-opinion diagnostics[50]. This not only enhances diagnostic confidence but also allows pathologists to allocate more time to complex cases where human expertise is still crucial.

Several comorbidities can influence surgical outcomes and prognosis in CRC. Colorectal cancer surgery is associated with a range of potential complications. Common complications include infections, anastomotic leaks, hemorrhage, bowel obstructions, thromboembolic events, and delayed recovery. As the medical field increasingly emphasizes personalized care and risk stratification, specific biomarkers are being investigated to

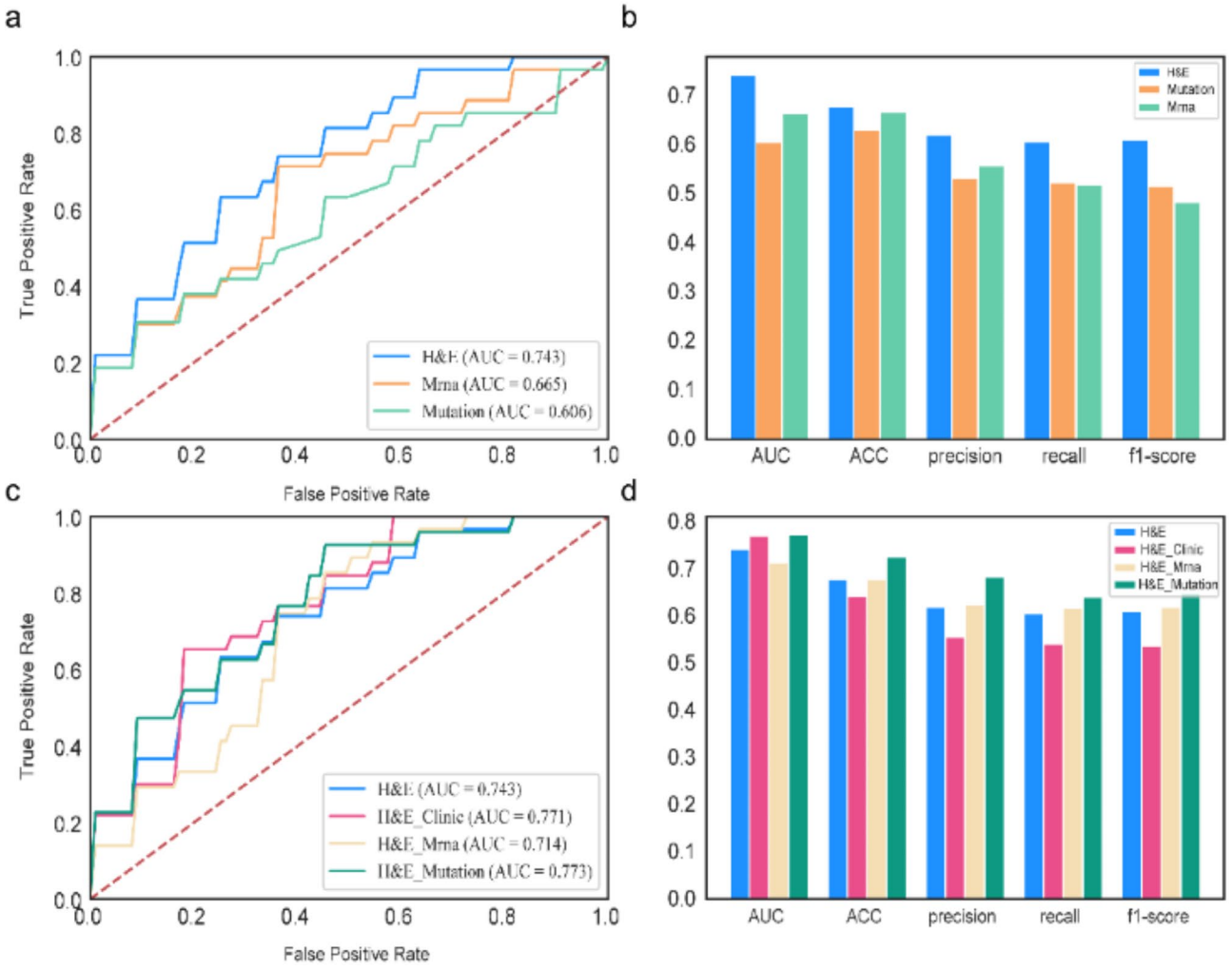

Fig. 7. Performance of Survival Prediction Models. (**a**) Model performance comparison between H&E and molecular data. (**b**) Different evaluation metrics for models for H&E and molecular data. (**c**) The model performance comparison of H&E and H&E combined with molecular features (top1 of clinic, mRNA, and Mutation), respectively. (**d**) Different evaluation metrics for fusion models.

more accurately predict postoperative complications. One such biomarker is Butyrylcholinesterase (BChE). Future research could explore the prediction of such biomarkers using a framework similar to that employed in this study. In addition, obesity is increasingly recognized as a significant factor affecting CRC survival by altering the tumor microenvironment, as evidenced by histopathological images and gene pathways impacted by genetic mutations. Histological examinations revealing lymphovascular invasion are linked to poorer prognosis and may be influenced by the inflammatory state associated with obesity. Obesity often induces a chronic, low-grade inflammatory state characterized by the release of pro-inflammatory cytokines (e.g., TNF-α, IL-6) and adipokines. This inflammatory milieu can promote tumorigenesis, facilitate cancer progression, and render tumors more aggressive. Furthermore, the KRAS mutation is frequently associated with metabolic changes related to obesity. Therefore, understanding the interactions among obesity, genetic mutations, and tumor biology is vital for tailoring targeted therapies, which may ultimately enhance survival rates in obese CRC patients.

In addition to the significance of genetic mutations specific to CRC patients, the inheritance of gene mutations from family members plays a crucial role in predicting the survival and overall prognosis of these patients. The familial transmission of genetic variants can provide valuable insights into a patient's susceptibility to CRC and associated pathogenic mechanisms. For instance, if a patient has a family history of specific gene mutations linked to CRC, it may suggest a hereditary predisposition that could influence not only the development of tumors but also their aggressiveness and response to treatment. Understanding these inherited genetic factors allows for more comprehensive risk assessment and personalized treatment strategies. Furthermore, family studies can aid in identifying potential biomarkers that may be used to monitor disease progression and therapeutic efficacy. By considering both somatic mutations in the tumor and inherited genetic alterations, healthcare providers can enhance their predictive capabilities regarding patient outcomes, ultimately leading to improved survival rates in individuals affected by CRC.

Finally, there are several limitations in our study. First, due to the lack of sufficient follow-up information, there was a large amount of samples with incomplete information, resulting in a limited sample size. Thus, the model might be insufficiently trained and its performance should be further improved, providing more samples. Second, there was a lack of independent validation cohort. As a consequence, it is unclear if the model can be used in samples other than TCGA. Third, due to the lack of sufficient information on radiation therapy in clinical data, we only selected information such as gender, age, and pathological stage of the samples to study their relationship with prognosis. Finally, our fusion model had a certain improvement in performance compared to the single-omics model, but the improvement was limited. There may be two reasons: On one hand, the feature fusion method used needs to be improved and more advanced classification models should be developed or transferred from other biological studies. On the other hand, we only combined H&E-stained images with a single type of molecular features.

In the future, we will first try to increase the sample size of the training cohort, which may be realized by merging a few cohorts. A key issue would be batch effects, which could be relieved by a few batch correction methods. Second, we will collect external validation cohorts by searching the public domain or from hospitals. Third, we will study the effect of treatment methods on prognosis. It might be better to train models under certain treatment methods like radiation therapy. Finally, more advanced multimodal deep learning models and models fusion more types of data will be explored to further improve the prediction accuracy of CRC survival.

Conclusion

In this study, molecular data from multiple levels of ST and LT samples were analyzed to find biomarkers associated with the survival of colorectal cancer patients. On the other hand, a novel fusion model is proposed to fuse images with molecular data to predict the five-year survival of CRC patients. The results of this study can help to strengthen the understanding of the mechanisms related to the survival of patients with CRC, and can provide new ideas for predicting survival.

Data availability

The datasets analyzed during the current study are available in The Cancer Genome Atlas (TCGA) repository, https://portal.gdc.cancer.gov/projects/TCGA-COAD.

Received: 19 December 2024; Accepted: 20 February 2025
Published online: 20 March 2025

References

1. Xia, C. et al. Cancer statistics in China and united States, 2022: profiles, trends, and determinants. *Chin. Med. J.* **135**, 584–590 (2022).
2. CANCER, F. C. (2022). https://fightcolorectalcancer.org/about-colorectal-cancer/general-information/survival-rates-of-colorectal-cancer/.
3. Liu, H. et al. Somatic mutation, and their combinations in inferring tumor Tissue-of-Origin. *Front. Cell. Dev. Biology*. **9**, 619330 (2021). Evaluating DNA Methylation, Gene Expression.
4. He, B. et al. TOOme: A novel computational framework to infer Cancer Tissue-of-Origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* **8**, 394 (2020).
5. Liang, W. et al. KRT17 promotes T-lymphocyte infiltration through the YTHDF2–CXCL10 Axis in colorectal Cancer. *Cancer Immunol. Res.* **11**, 875–894 (2023).
6. Liu, X. et al. Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. *Comput. Biol. Med.* **146**, 105569 (2022).
7. Shi, X., Young, S., Cai, K., Yang, J. & Morahan, G. Cancer susceptibility genes: update and systematic perspectives. *Innov. (Cambridge (Mass))*. **3**, 100277 (2022).
8. Peng, P. et al. Prognostic factors in stage IV colorectal Cancer patients with resection of liver and/or pulmonary metastases: A Population-Based cohort study. *Front. Oncol.* **12**, 850937 (2022).
9. Sjoquist, K. M. et al. Personalizing survival predictions in advanced colorectal cancer: the ARCAD nomogram project. *J. Natl. Cancer Inst.* **110**, 638–648 (2018).
10. White, A. et al. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC cancer*. **18**, 906 (2018).
11. Compton, C. C. Colorectal carcinoma: diagnostic, prognostic, and molecular features. *Mod. Pathology: Official J. United States Can. Acad. Pathol. Inc.* **16**, 376–388 (2003).
12. Cheng, L. et al. Identification and validation of six autophagy-related long non-coding RNAs as prognostic signature in colorectal cancer. *Int. J. Med. Sci.* **18**, 88 (2021).
13. Qian, J. et al. Novel multiple miRNA-based signatures for predicting overall survival and recurrence-free survival of colorectal cancer patients. *Med. Sci. Monitor: Int. Med. J. Experimental*. **25**, 7258 (2019).
14. Yang, G., Zhang, Y. & Yang, J. A five-microRNA signature as prognostic biomarker in colorectal cancer by bioinformatics analysis. *J. Front. Oncol.* **9**, 1207 (2019).
15. Bychkov, D. et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, 3395 (2018).
16. Yang, J. et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* **20**, 333–342 (2022).
17. Ye, Z. et al. Cervical Cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr. Bioinform.* **17**, 164–173 (2022).
18. He, B. et al. A neural network framework for predicting the Tissue-of-Origin of 15 common Cancer types based on RNA-Seq data. *Front. Bioeng. Biotechnol.* **8** (2020).
19. Ma, X. et al. A machine Learning-based diagnosis of thyroid Cancer using thyroid nodules ultrasound images. *Curr. Bioinform.* **15**, 349–358 (2020).
20. He, B. et al. A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Et Biophys. Acta (BBA) - Mol. Basis Disease* **1866**, 165916 (2020).
21. He, B. et al. A new method for CTC images recognition based on machine learning. *Front. Bioeng. Biotechnol.* **8**, 897 (2020).
22. Yang, M. et al. A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* **146**, 105516 (2022).

23. Yao, Y. et al. ICSDA: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data. *Brief. Bioinform.* **23**, bbac448c (2022).
24. Huang, K. et al. Predicting colorectal cancer tumor mutational burden from histopathological images and clinical information using multi-modal deep learning. *Bioinformatics* **38**, btac641 (2022).
25. Xie, X. et al. Evaluating Cancer-Related biomarkers based on pathological images: A systematic review. *Front. Oncol.* **11**, 763527 (2021).
26. Liang, L., Liang, X., Yu, X. & Xiang, W. Bioinformatic analyses and integrated machine learning to predict prognosis and therapeutic response based on E3 Ligase-Related genes in colon cancer. *J. Cancer*. **15**, 5376–5395 (2024).
27. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. cancer*. **1**, 800–810 (2020).
28. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
29. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* **4**, 1–13 (2021).
30. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for Pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).
31. Wu, Z. et al. DeepLRHE: A deep convolutional neural network framework to evaluate the risk of lung Cancer recurrence and metastasis from histopathology images. *Front. Genet.* **11**, 768 (2020).
32. Vahadane, A. et al. IEEE,. in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI) 1012–1015 (2015).
33. Macenko, M. et al. IEEE,. in 2009 IEEE international symposium on biomedical imaging: from nano to macro 1107–1110 (2009).
34. Geng, L., Zhang, S., Tong, J. & Xiao, Z. Lung segmentation method with dilated Convolution based on VGG-16 network. *Comput. Assist. Surg. (Abingdon England)*. **24**, 27–33 (2019).
35. Kadry, S., Taniar, D., Damaševičius, R., Rajinikanth, V. & Lawal, I. A. in 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII) 1–5IEEE, (2021).
36. Tammina, S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *Int. J. Sci. Res. Publications*. **9**, 143–150 (2019).
37. Kaur, T. & Gandhi, T. K. in International Conference on Information Technology (ICIT) 94–98 (IEEE, 2019). (2019).
38. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv* (2014).
39. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
40. Gao, Y., Beijbom, O., Zhang, N. & Darrell, T. in Proceedings of the IEEE conference on computer vision and pattern recognition 317–326 (2016).
41. Fukui, A. et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv* (2016).
42. Leiserson, M. D., Wu, H. T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**, 160 (2015).
43. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
44. Guraya, S. Y. & Pattern Stage, and time of recurrent colorectal Cancer after curative surgery. *Clin. Colorectal Cancer*. **18**, e223–e228 (2019).
45. Huang, L. et al. PEAK1, acting as a tumor promoter in colorectal cancer, is regulated by the EGFR/KRas signaling axis and miR-181d. *Cell Death Dis.* **9**, 271 (2018).
46. Kameyama, H. et al. Genomic characterization of colitis-associated colorectal cancer. *World J. Surg. Oncol.* **16**, 121 (2018).
47. Qian, J. et al. Tanshinone IIA promotes IL2-mediated SW480 colorectal cancer cell apoptosis by triggering INF2-related mitochondrial fission and activating the Mst1-Hippo pathway. *Biomed. Pharmacother.* **108**, 1658–1669 (2018).
48. Wang, H., Xi, Q. & Wu, G. Fatty acid synthase regulates invasion and metastasis of colorectal cancer via Wnt signaling pathway. *Cancer Med.* **5**, 1599–1606 (2016).
49. Bousis, D. et al. The role of deep learning in diagnosing colorectal cancer. *Prz Gastroenterol.* **18**, 266–273 (2023).
50. Chlorogiannis, D. et al. Tissue classification and diagnosis of colorectal cancer histopathology images using deep learning algorithms. Is the time ripe for clinical practice implementation?. Prz *Gastroenterol* 10.5114/pg.2023.130337 (2023).

Author contributions

Conception and design: B.H, L.W, K.L, K.H. Development of methodology: B.H, L.W. Acquisition of data: B.H, L.W. Analysis and interpretation of data: B.H, L.W. All authors contributed to the writing and reviewing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by the Provincial Key R & D Projects of Hunan Provincial Science and Technology Department (No.2022SK2074).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.L. or K.H.

Reprints and permissions information is available at www.nature.com/reprints.

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