scientific reports

OPEN

Accurate prediction of colorectal cancer diagnosis using machine learning based on immunohistochemistry pathological images

Bobin Ning[1], Jimei Chi[2,3], Qingyu Meng[1] & Baoqing Jia[1,4]✉

Colorectal cancer (CRC) ranks as the third most prevalent tumor and the second leading cause of mortality. Early and accurate diagnosis holds significant importance in enhancing patient treatment and prognosis. Machine learning technology and bioinformatics have provided novel approaches for cancer diagnosis. This study aims to develop a CRC diagnostic model based on immunohistochemical staining image features using machine learning methods. Initially, CRC disease-specific genes were identified through bioinformatics analysis, SVM-RFE and Random Forest algorithm utilizing RNA-seq data from both GEO and TCGA databases. Subsequently, verification of these genes was performed using proteomics data from CPTAC and HPA database, resulting in identification of target proteins (AKR1B10, CA2, DHRS9, and ZG16) for further investigation. SVM and CNN were then employed to analyze and integrate the characteristics of immunohistochemical images to construct a reliable CRC diagnostic model. During the training and validation process of this model, cross-validation along with external validation methods were implemented to ensure accuracy and reliability. The results demonstrate that the established diagnostic model exhibits excellent performance in distinguishing between CRC and normal controls (accuracy rate: 0.999), thereby presenting potential prospects for clinical application. These findings are expected to provide innovative perspectives as well as methodologies for personalized diagnosis of CRC while offering more precise references for promising treatment.

Keywords Colorectal cancer, Diagnosis, Machine learning, Immunohistochemistry

Colorectal cancer (CRC) is a prevalent malignant tumor, and its incidence is progressively increasing due to the changing unhealthy lifestyle habits, dietary patterns, and aging[1]. It primarily affects individuals over 50 years old[2]. Despite advancements in early screening and treatment methods that have improved patient survival rates, CRC remains one of the leading causes of death worldwide[3]. Many patients are diagnosed at an advanced stage, which significantly increases the risk of recurrence and metastasis[4]. Therefore, timely diagnosis plays a crucial role for CRC patients as it enables early detection of the disease's presence, determination of its type, grading, and spread. This provides patients with an opportunity for early treatment interventions that can enhance treatment efficacy and improve survival rates. Timely and precisely diagnosis can help prevent tumor progression to advanced stages by reducing the complexity and difficulty associated with treatment while alleviating both physical and psychological burdens on patients[5]. Furthermore, accurate diagnosis aids doctors in developing personalized treatment plans encompassing surgery, chemotherapy, radiotherapy among other modalities to minimize patient discomfort while enhancing their overall quality of life[6].

Currently, commonly employed techniques for the diagnosis of CRC encompass colonoscopy, fecal occult blood examination, blood marker detection, and imaging examinations such as CT scanning[4]. Nevertheless, these methods possess inherent limitations[7]. For instance, colonoscopy necessitates patient cooperation and is

[1]Department of General Surgery, Chinese PLA General Hospital, Beijing, People's Republic of China. [2]Key Laboratory of Green Printing, Institute of ChemistryBeijing Engineering Research Center of Nanomaterials for Green Printing TechnologyNational Laboratory for Molecular Sciences (BNLMS), Chinese Academy of Sciences (ICCAS), Beijing 100190, People's Republic of China. [3]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China. [4]Boqing Jia, Haidian District, No.28, Fuxing Road, Beijing 100853, China. ✉email: baoqingjia@126.com

not universally applicable due to its costliness; fecal occult blood examination may yield false positive or false negative outcomes; blood marker detection lacks specificity and can be influenced by other factors; imaging examinations fail to accurately differentiate between benign and malignant lesions[8]. Consequently, despite their certain utility in diagnosing CRC, these methods still exhibit constraints. It is imperative to comprehensively consider clinical manifestations, risk factors, and medical resources when selecting appropriate diagnostic approaches. In clinical practice, a comprehensive evaluation combined with clinical symptoms, colonoscopy findings, and histopathological examination results typically serves as the gold standard for diagnosing CRC[8].

The field of clinical diagnosis assisted by artificial intelligence is rapidly advancing[9]. Its primary advantage lies in leveraging machine learning and deep learning technologies to analyze vast amounts of medical data, including images, genomes, and clinical records, thereby aiding doctors in making quicker and more accurate decisions regarding diagnosis and treatment[10]. Artificial intelligence has demonstrated significant potential in tumor screening, disease classification, image recognition, and other domains[11]. It not only enhances the precision and efficiency of diagnoses but also fosters the development of personalized medicine by offering patients more precise and tailored treatment plans[12]. With ongoing technological advancements and increased integration into clinical practice, AI-assisted clinical diagnosis will bring about revolutionary changes in healthcare delivery while elevating the standard of medical care provided to enhance patients' quality of life[13].

In this study, we successfully identified diagnostic markers for CRC through bioinformatics analysis, SVM-RFE, and the Random forest method. These markers were systematically analyzed using data mining (TCGA and GEO database) and multi-omics analysis (genomics and proteomics), resulting in the identification of a group of genes with significant differential expression that play crucial roles in the occurrence and development of CRC. This discovery provides new insights into diagnosis possibilities for CRC, offering strong support for potential applications in its diagnosis and treatment. To further enhance accuracy and sensitivity in clinical practice, we conducted machine learning (SVM and CNN) on immunohistochemical images of target proteins, and constructed a comprehensive diagnostic model. Through analyzing large amounts of immunohistochemical image data, we successfully established a diagnostic model that comprehensively considers different marker expression patterns, providing a more accurate auxiliary tool for diagnosing CRC. This breakthrough is expected to bring important progress to clinical practice.

Results

Differentially expressed genes in CRC

We enrolled three datasets (GSE18105, GSE21510, and GSE33114) from the GEO database, which included both CRC and normal tissues. To enhance the credibility and stability of our findings, we initially merged these three datasets to increase the sample size in our study. Through differential analysis (Fig. 1A and 1B), we identified a total of 374 DEGs. Similarly, by performing the same analysis in the TCGA database, we obtained 2,514 DEGs encoding proteins (Fig. 1C). By intersecting the results from both databases, we ultimately identified 232 genes with consistently altered expression in CRC (Fig. 1D).

Disease-characteristic genes of CRC

In Fig. 2A, the X-axis represents the number of trees, while the Y-axis represents the cross-validation error. The three curves in the figure represent errors for the control group, experimental group, and all samples respectively. The all-sample error is depicted by a black line. Our objective was to identify the point with minimal cross-validation error to determine the corresponding number of trees and genes. In Fig. 2B, gene names are represented on the Y-axis, whereas gene importance scores are shown on the X-axis. A higher score indicates greater gene importance. We selected genes with a score exceeding 0.5 for further investigation, resulting in a total of 30 characteristic genes associated with CRC. To further compare and validate the other machine learning methods for identifying disease-related genes, we employed the SVM-RFE algorithm to screen DEGs, resulting in a total of 282 candidate genes (Fig. 2C). Subsequently, we identified 26 CRC-specific genes by intersecting the sets of the genes obtained from both algorithms, which were then utilized in subsequent research (Fig. 2D).

Revalidation of protein levels

When the aforementioned characteristic genes of CRC were inputted into the CPTAC database, we discovered that 6 proteins exhibited consistency with the results of genomics analysis and fulfilled the criteria for subsequent analysis. The disparities in protein expression between tumor and normal tissues, as well as across various tumor stages, are depicted in Fig. 3A-L. Through validation in the HPA database, we ultimately confirmed that 4 proteins displaying significantly differential expression would be incorporated into the construction of machine learning models via IHC. The immunohistochemical maps illustrating these proteins in both normal tissues and bowel cancer tissues can be observed in Fig. 4, while their quantitative outcomes are presented in Table 1.

CRC machine learning model (SVM and CNN) based on immunohistochemical graph construction

The schematic diagram of the CRC diagnosis based on machine learning of immunohistochemical pathological images is presented in Fig. 5A. The framework for dataset division, model training, and prediction evaluation is illustrated in Fig. 5B. Additionally, Fig. 5C displays the ROC curve and binary confusion matrix obtained using the immunohistochemical staining images from the HPA database. AKR1B10, CA2, DHRS9, and ZG16 were utilized as four biomarkers respectively. Among them, DHRS9 exhibited the highest accuracy (0.710) in diagnosing CRC compared to others such as ZG16 which had lower accuracy (0.550) and specificity (0.531), as shown in Table 2. In order to enhance the sensitivity and specificity, we utilized the diagnostic markers and performed joint diagnosis using the SVM algorithm. Based on the ROC curve and binary confusion matrix

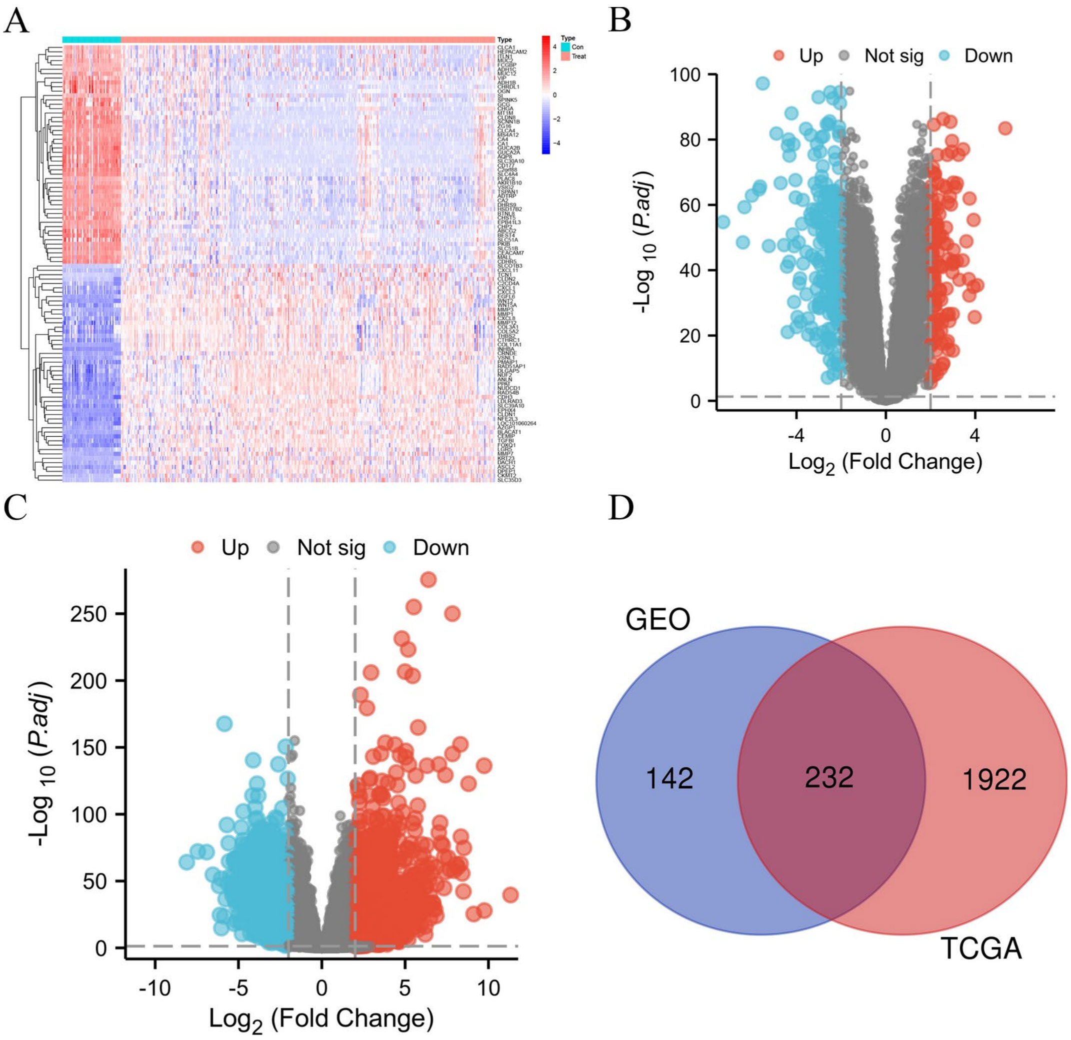

Fig. 1. Differential analysis of genes encoding proteins in CRC. **A**. Heatmap of the most significantly differentially expressed 100 genes in GEO merged CRC cohort; **B**. Volcano plot of differentially expressed genes in GEO merged CRC cohort (|logFC|≥2, p value≤0.05); **C**. Volcano plot of differentially expressed genes encoding proteins in TCGA CRC cohort (|logFC|≥2, p value≤0.05); **D**. Veen plot of DEGs encoding proteins of CRC patients in GEO and TCGA databases.

depicted in Fig. 5D, it is remarkably observed that the model accurately classified all 400 normal colon tissues, with only 5 out of 400 colon cancer tissues being misdiagnosed as normal tissues. This resulted in an exceptional accuracy and specificity of 0.999 (Table 2, Table S1).

In Figure S1, we further compared the results of CRC diagnosis models based on the same set of IHC images using a CNN model. We were surprised to observe that, in contrast to the SVM model's results, while utilizing a single marker, the accuracy rate for diagnosing each marker as a target for diagnosis approached 100% (Figure S1 A ~ D, F ~ I). Among the four markers, DHRS9 exhibited the highest accuracy rate (100%), whereas ZG16 demonstrated the lowest accuracy rate (95%). However, upon considering all four markers collectively, it is evident that the diagnostic rate of the CNN is inferior to the combined diagnostic rate of the SVM (Figure S1E, J, Table S2). At this juncture, the proposed model introduces a novel approach for diagnosing CRC, showcasing the potential of machine learning in analyzing IHC images while providing robust support for clinical diagnosis and practice.

Discussion

The highlight of this study lies in the acquisition of a set of characteristic genes associated with CRC through multi-omics research and machine learning methods. By leveraging the features extracted from IHC images, we have developed a highly robust diagnostic model using machine learning techniques, which significantly contributes to the diagnosis of CRC and advances precision medicine. Multi-omics analysis refers to the comprehensive utilization of diverse biological data types including genomics, transcriptomics, proteomics, metabolomics, etc., aiming to gain a holistic understanding of the complexity and diversity within biological systems[13]. Its advantage lies in its ability to simultaneously consider multiple levels of biological information

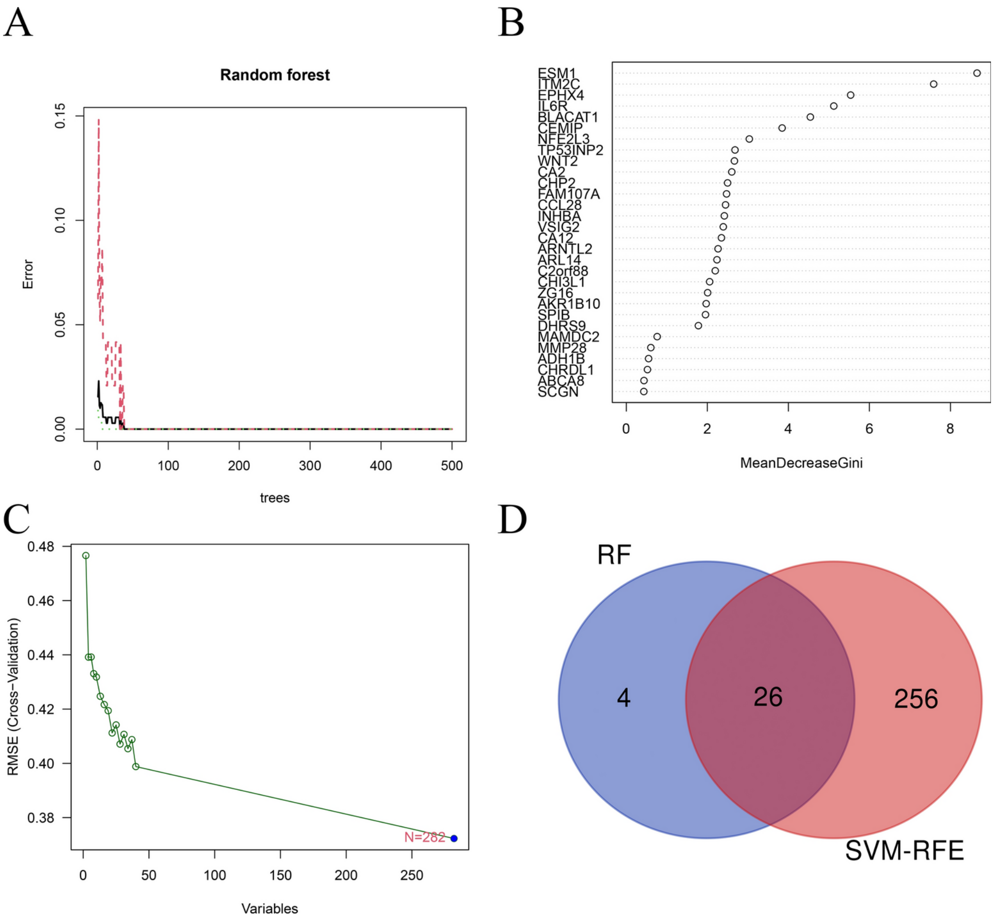

Fig. 2. Identification of characteristic genes of CRC by machine learning. **A**. The construction of Random Forest algorithm based on DEGs; **B**. Identification of CRC characteristic genes based on significance scores; **C**. Identification of CRC characteristic genes based on SVM-RFE; **D**. The veen diagram of Random Forest and SVM-RFE screened genes.

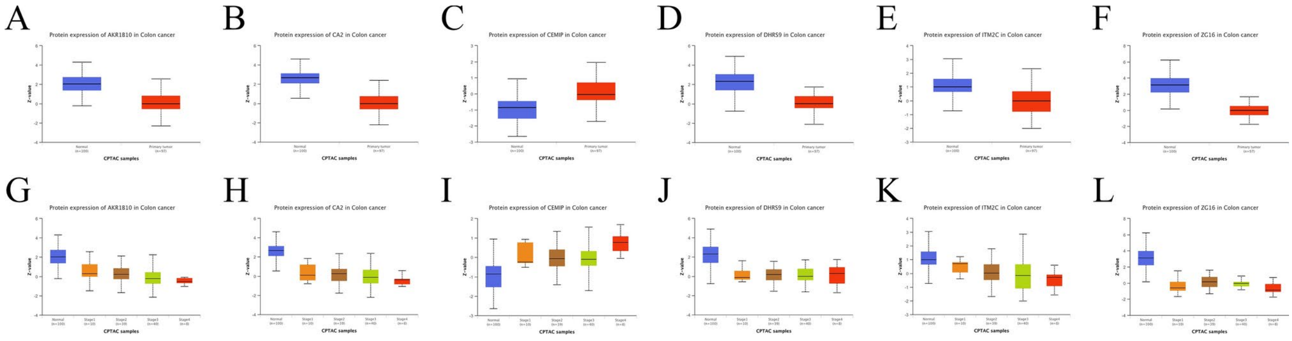

Fig. 3. Validation of differential expression of disease-associated genes at protein level by CPTAC database. Expression differences between normal tissues and tumor tissues, **A**. AKR1B10, **B**. CA2, **C**. CEMIP, **D**. DHRS9, **E**. ITM2C, and **F**. ZG16; expression differences between normal tissues and each stage of tumor, **A**. AKR1B10, **B**. CA2, **C**. CEMIP, **D**. DHRS9, **E**. ITM2C, and **F**. ZG16.

ranging from genes, RNA molecules, proteins to metabolites; thus enabling a more comprehensive and profound comprehension of tumor diseases[14]. In this study, potential characteristic markers for CRC were identified through comprehensive analysis and cross-validation across different datasets. These markers may be implicated in tumor initiation, progression as well as treatment response; thereby facilitating improved diagnosis and treatment outcomes for this disease.

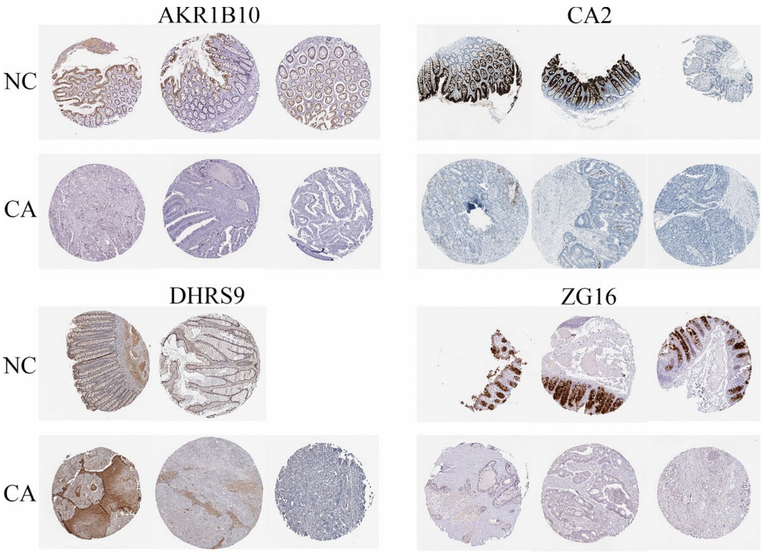

Fig. 4. Characteristic immunohistochemical images of AKR1B10, CA2, DHRS9, and ZG16 in HPA database.

		AKR1B10	CA2	DHRS9	ZG16
Normal					
	Endocrine cells		High		
	Endothelial cells	Not detected	Not detected	Medium	Not detected
	Enterocytes		High		
	Enterocytes—Microvilli		High		
	Fibroblasts		Not detected		
	Glandular cells	Medium		Medium	High
	Goblet cells		High		
	Mucosal lymphoid cells		Not detected		
	Peripheral nerve/ganglion			Medium	Not detected
Cancer					
	High				
	Medium			2	1
	Low			7	2
	Not detected	10	12	3	8

Table 1. Expression of disease-associated proteins in normal and CRC tissues from HPA data.

After conducting comprehensive validation of differential analysis (Fig. 1A–D), employing SVM-RFE and Random Forest algorithm (Fig. 2A–D), assessing protein expression levels (Fig. 3A–L), and performing IHC analysis (Fig. 4), we ultimately identified four specific markers for CRC, namely AKR1B10, CA2, DHRS9, and ZG16. These aforementioned markers exhibited high expression in colon tissues but were either lowly expressed or not expressed at all in CRC. It is noteworthy that there is a paucity of literature on the four biomarkers. Under physiological conditions, AKR1B10 and DHRS9 belong to the aldosterone reductase family and ketone alcohol conversion SDR family respectively[15,16]; they are involved in intracellular reactions associated with aldosterone reduction and ketone alcohol conversion processes which regulate the balance of intracellular metabolites. Previous research has demonstrated that apart from participating in the progression of CRC through their own metabolic reactions, AKR1B10 can also promote its advancement by influencing autophagy responses as well as inflammatory reactions[17–19]. CA2 is a widely expressed carbonic anhydrase isoenzyme in the human body, playing crucial roles in regulating acid–base balance, transporting carbon dioxide, maintaining calcium ion homeostasis, protecting the digestive tract and facilitating nervous system function[20]. Although the molecular biological mechanism of CA2 promoting CRC progression remains unknown, several studies have confirmed that its expression can significantly inhibit the cancer cell growth[21]; furthermore, its downregulation represents an early event in CRC development[22]. ZG16 is a protein secreted by the pancreas involved in various biological processes such as pancreatic and digestive regulation, immune modulation and intestinal microbial regulation[23]. Meng H et al.'s research has demonstrated that ZG16 mainly contributes to CRC progression through tumor immune microenvironment[24].

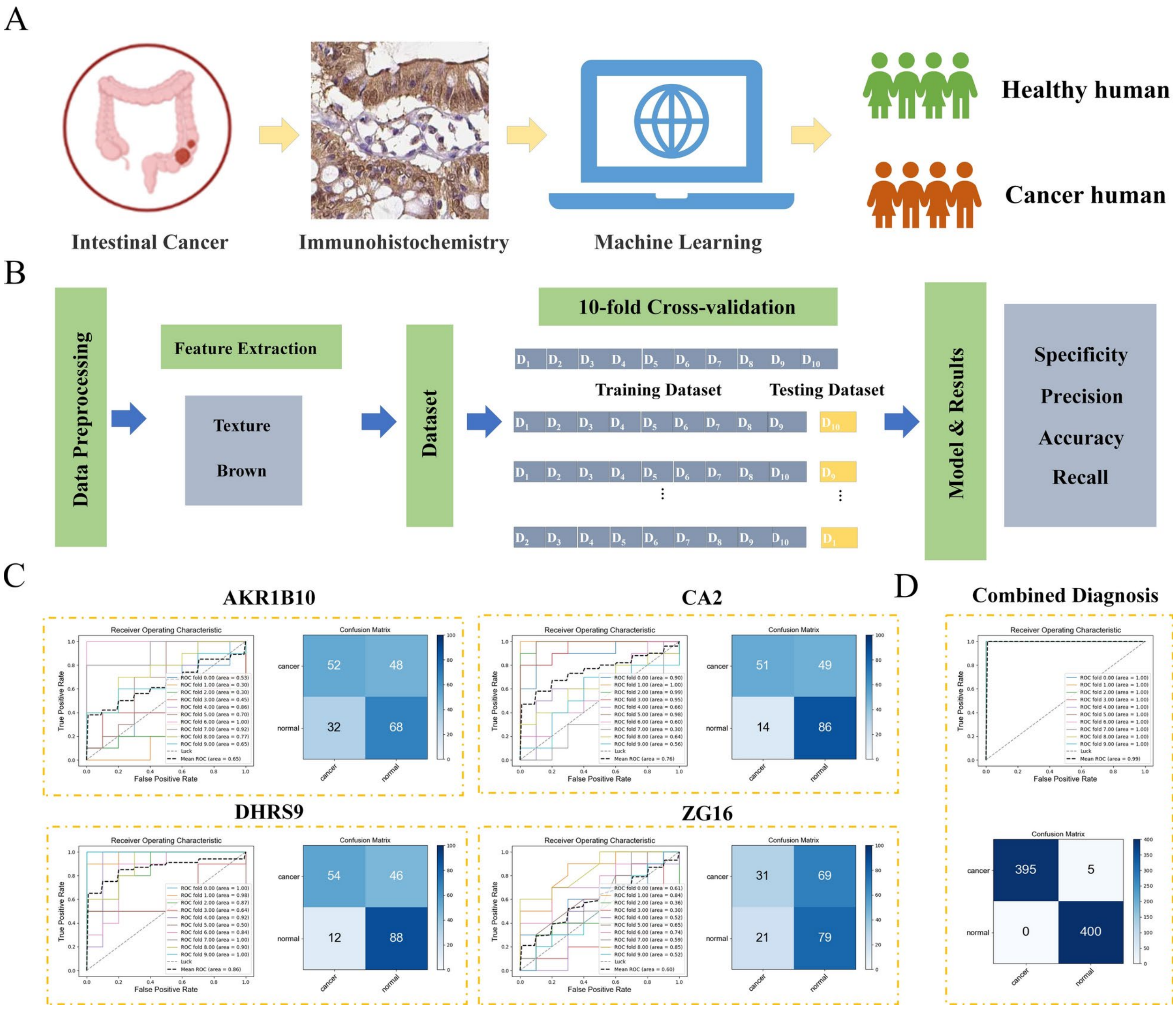

Fig. 5. Pathological diagnosis of CRC by SVM. (**A**) Schematic diagram of machine learning-based immunohistochemical pathological images diagnosis of intestinal cancer. (**B**) Framework for partitioning data sets, training models and assessing predictions. (**C**) ROC curves and binary confusion matrices of AKR1B10, CA2, DHRS9 and ZG16 biomarkers based on immunohistochemical staining images taken by the HPA database, respectively. (**D**) ROC curves and binary confusion matrices of AKR1B10, CA2, DHRS9 and ZG16 combined diagnosis based on the SVM algorithm.

	# of images		AUROC	Accuracy	Precision	Recall	Specificity
	–	+					
AKR1B10	100	100	0.648 (0.644–0.650)	0.600 (0.598–0.602)	0.520 (0.517–0.522)	0.645 (0.642–0.647)	0.532 (0.529–0.534)
CA2	100	100	0.758 (0.755–0.760)	0.685 (0.682–0.687)	0.510 (0.506–0.514)	0.632 (0.628–0.636)	0.665 (0.663–0.667)
DHRS9	100	100	0.865 (0.863–0.866)	0.710 (0.708–0.711)	0.540 (0.537–0.542)	0.899 (0.897–0.900)	0.676 (0.674–0.677)
ZG16	100	100	0.598 (0.596–0.599)	0.550 (0.548–0.551)	0.310 (0.308–0.311)	0.639 (0.636–0.641)	0.531 (0.530–0.532)
All	400	400	0.999 (0.999–0.999)	0.999 (0.999–0.999)	0.999 (0.999–0.999)	0.999 (0.999–0.999)	0.999 (0.999–0.999)

Table 2. Binary diagnostic performance for CRC based on IHC staining images. (95%CI)

In current clinical practice, IHC combined with the clinical manifestations of patients, imaging, serum markers, and colonoscopy serves as the gold standard for diagnosing CRC[25,26]. Its advantages include high specificity and the ability to provide comprehensive information on tissue structure and protein expression, facilitating a more thorough understanding of the pathological characteristics[27]. However, it is important to address and resolve technical complexities, specimen quality issues, and repeatability concerns associated with IHC[28]. Additionally, due to its semi-quantitative nature and susceptibility to subjective judgment by operators, there may be inherent subjectivity and uncertainty in immunohistochemical results[29]. Consequently, variations between different laboratories or operators can compromise diagnostic repeatability and consistency.

Based on machine learning, SVM or CNN , artificial intelligence can process large-scale medical images and biological data, enabling doctors to extract potential lesions and abnormal areas from complex image data[30]. This provides a more comprehensive and objective diagnostic basis while reducing the interference of human factors[31]. Therefore, the application of AI-assistance has brought new opportunities and challenges for tumor diagnosis, which can improve medical standards and patient survival rates. The process diagram of the machine learning-based diagnosis model is illustrated in Fig. 5A. It can be observed from the flow chart presented in Fig. 5B that our classifier's performance was evaluated using support vector machine (SVM) and a tenfold cross-validation approach. Furthermore, to demonstrate the superiority of joint judgment over individual biomarkers' judgments, we incorporated features extracted from other samples within the same class (cancer or normal) into the current image's feature set during joint judgment coding. The code primarily iterates through image files within a designated folder, reads each image, extracts both brown and Gabor filter features, and ultimately concatenates them to form the final feature vector. In each cycle of cross-validation, 9 subsets were selected as the training set while the remaining 1 subset was used as the test set. The performance of the classifier in cancer detection task was evaluated by calculating the confusion matrix, which included metrics such as accuracy, precision, and recall. This process was repeated for each biomarker to individually evaluate their classification performance. For the combined biomarker dataset, features were extracted from a single biomarker and then connected and input into 10 cross-validation processes to assess the impact of combining multiple biomarkers. Finally, the classification results obtained from a single biomarker were compared with those obtained from combined biomarkers to validate their superior joint judgment capability. The AUC area under ROC curve and diagnostic accuracy, specificity, recall, and predictive rate for each of the 4 individual biomarkers were obtained from Fig. 5C and Table S1. According to Fig. 5D and comprehensive evaluation presented in Table 2, it is evident that combining all four biomarkers yields significantly better diagnostic outcomes compared to independent diagnosis alone. Moreover, when considering accuracy, specificity, recall rate, predictive rate,and AUC area together,the combined diagnosis using all four biomarkers can achieve nearly perfect results (close to100%), which holds great significance for clinical diagnosis.

Given that the utilized dataset comprises image modalities, we proceeded to reconfigure the diagnostic model by employing another deep learning architecture such as a convolutional neural network (CNN) to compare its classification accuracy, sensitivity, and specificity with those of the SVM model. CNN offers the advantage of automatic feature extraction and processing of image data, enabling it to capture intricate patterns within the dataset. Our research findings also validate CNN's superiority in feature extraction and classification performance. When assessing the individual diagnostic efficacy of the four markers separately (refer to Table S1, S2), CNN demonstrates nearly superior diagnostic rates compared to SVM across all indicators. However, upon comprehensive evaluation of joint diagnostics involving all four indicators, we observed that while both SVM and CNN achieved outstanding diagnostic rates approaching 100%, SVM exhibited slightly higher effectiveness than CNN. This could potentially be attributed to the relatively small size of our dataset; SVM has demonstrated stable performance in handling small sample high-dimensional feature data due to its robust theoretical foundation and practical results.

The limitations of this study are as follows: Firstly, the data utilized in this study solely originate from the public data platform. To enhance the reliability and validity, future studies should consider incorporating clinical samples. Secondly, given the nature of machine learning and data analysis techniques, it is imperative to expand the sample size in order to obtain more precise and stable outcomes. Lastly, although this study is based on IHC findings without simplifying diagnostic efficiency or steps, there is potential for improvement by collaborating with colleagues of chemical materials to update detection procedures and refine the procedure.

In summary, this study utilized multiple databases and conducted multi-omics analysis to identify four novel and promising protein markers (AKR1B10, CA2, DHRS9, and ZG16) for colorectal cancer. These markers were screened using bioinformatics and machine learning techniques (Random Forest). Furthermore, a CRC diagnosis model was constructed based on IHC employing another machine learning method (Support Vector Machine), achieving remarkable accuracy and specificity of 0.999. This research provides significant opportunities and support for enhancing the comprehensiveness and precision of CRC.

Materials and methods

Data acquisition and difference analysis

We selected three independent datasets (GSE18105, GSE21510, and GSE33114) from the GEO database and merged them using the SVA package in R language to construct a GEO CRC cohort. Given that this study will be based on proteomics analysis and will build a machine learning model using immunohistochemical images, we extracted RNA-seq data of Colon Adenocarcinoma (COAD) from the TCGA database (https://portal.gdc.cancer.gov/). And the protein-coding genes were isolated using the human genome map as a reference to construct a TCGA cohort. Subsequently, we utilized the T-test to analyze differences between the two CRC cohorts by "limma" package. Genes with p-values ≤ 0.05 and $|\text{logFC}| \geq 2$ were identified as differentially expressed genes (DEGs) in CRC. Finally, we obtained the intersection of DEGs exhibiting stable differential expression in both databases for further studies.

Refinement of disease-associated gene screening

We utilized the random forest algorithm to filter DEGs and identify disease-related characteristic genes with specific parameter settings (seed = 123,456 and tree = 500). By optimizing cross-validation error, we constructed an accurate random forest model.

To further identify DEGs associated with CRC, we also employed the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) method. The SVM classifier with a linear kernel was trained on the preprocessed dataset. The choice of a linear kernel is standard for SVM-RFE, as it allows the calculation of feature ranking based on the weights assigned to each feature by the model. The RFE process was initiated by iteratively removing the least significant features (genes) from the model. The final set of selected genes, ranked by their importance, was considered as the most informative features for distinguishing between colorectal cancer and control samples.

The genes identified by both methods can be regarded as characteristic DEGs that hold great potential for disease diagnosis.

The validation of proteomics

We incorporated disease-associated genes of CRC into both the CPTAC and HPA databases for protein-level analysis. In the CPTAC database, we ensured significant differences in target proteins between tumor and normal tissues, while also examining differences among different tumor stages, particularly between normal tissues and Stage 1 (early colorectal cancer patients). Following the retrieval of all images depicting the target protein from the HPA dataset (https://www.proteinatlas.org/) for both colon tissue and colon cancer tissue, we will utilize them to construct our subsequent diagnostic model.

Developing a machine learning model based on CRC Immunohistochemistry (IHC)

The Python programming language and multiple libraries are utilized for implementation. Initially, we import the following libraries: numpy for efficient processing of arrays and matrices, os for seamless manipulation of files and folders, cv2 for advanced image processing capabilities, sklearn for machine learning-related operations, and matplotlib for generating informative graphs. This combination of libraries offers robust functionality and flexibility, enabling us to effectively process data, perform intricate image processing tasks, implement sophisticated machine learning models, and visualize results. Consequently, we successfully develop an analysis and diagnosis model specifically tailored to identify characteristic genes associated with colorectal cancer. The code iterates through the image files within a designated folder. It reads each image file while extracting both brownian features as well as Gabor filter features. Brown channel features are assigned based on the brown color in the RGB color space. By creating a mask that identifies pixels falling within this brown range (set to white—255), all other pixels are set to black (0). Subsequently, this mask is adjusted to a specified size before being normalized. Gabor filter features are obtained by applying the Gabor filter technique to grayscale images. Finally, these extracted brownian features along with Gabor filter features are concatenated together forming the final feature vector. Additionally,the corresponding labels of each image are stored in an array denoted as Y. During each cross-validation cycle,nine subsets from our dataset serve as training sets while one remaining subset functions as a test set. The classifier is then trained using the training set, and predictions are made on the test set. The performance of the classifier in cancer recognition task is evaluated by calculating the confusion matrix and various performance metrics, including accuracy, precision, recall, etc. This process enables us to assess the classifier's generalization ability across different datasets and better evaluate its practical performance.

Given the significant disparities in the overall features of the immunohistochemical images of the four protein markers (AKR1B10, CA2, DHRS9, and ZG16), we merged the image features gathered at each occasion with those from the previous one for comprehensive assessment. Thus, for the four images, we executed three feature fusions. This procedure can guarantee that the features of the four images are thoroughly exploited and the optimal comprehensive diagnostic efficacy is attained.

Due to the efficacy in image-based classification tasks, we further utilized a Convolutional Neural Network (CNN) model to construct a diagnostic model. The dataset was split into training, and validation sets with a ratio of 7:3. The model consisted of several convolutional layers with ReLU activation functions, followed by max-pooling layers to reduce the spatial dimensions. A series of fully connected layers were added after the convolutional base, culminating in a softmax layer for classification. The model was trained using the Adam optimizer with an initial learning rate of 0.001. The categorical cross-entropy loss function was used to measure the performance. And the model's performance was evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score. ROC-AUC curves were also plotted to assess the diagnostic ability of the CNN model.

Data Availability

The datasets generated during and/or analysed during the current study are available in the GEO dataset (https://www.ncbi.nlm.nih.gov/geo/) and the TCGA dataset (https://portal.gdc.cancer.gov/). The immunohistochemis try images utilized in the study were sourced from the HPA database (https://www.proteinatlas.org/). Acquisitio n of the images was conducted in compliance with the stipulations outlined in the "License & Citation".

Received: 19 March 2024; Accepted: 10 October 2024
Published online: 02 December 2024

References

1. Dekker, E., Tanis, P. J., Vleugels, J., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**(10207), 1467–1480 (2019).

2. Baidoun, F. et al. Colorectal Cancer Epidemiology: Recent Trends and Impact on Outcomes. *Curr Drug Targets.* **22**(9), 998–1009 (2021).
3. Patel, S. G., Karlitz, J. J., Yen, T., Lieu, C. H. & Boland, C. R. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. *Lancet Gastroenterol Hepatol.* **7**(3), 262–274 (2022).
4. Biller, L. H. & Schrag, D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. *JAMA.* **325**(7), 669–685 (2021).
5. Shin, A. E., Giancotti, F. G. & Rustgi, A. K. Metastatic colorectal cancer: mechanisms and emerging therapeutics. *Trends Pharmacol Sci.* **44**(4), 222–236 (2023).
6. Mahmoud, N. N. Colorectal Cancer: Preoperative Evaluation and Staging. *Surg Oncol Clin N Am.* **31**(2), 127–141 (2022).
7. Heinimann, K. Hereditary Colorectal Cancer: Clinics, Diagnostics and Management. *Ther Umsch.* **75**(10), 601–606 (2018).
8. Wu, Z. et al. Colorectal Cancer Screening Methods and Molecular Markers for Early Detection. *Technol Cancer Res Treat.* **19**, 1533033820980426 (2020).
9. Sharma, A., Kumar, R., Yadav, G. & Garg, P. Artificial intelligence in intestinal polyp and colorectal cancer prediction. *Cancer Lett.* **565**, 216238 (2023).
10. Mitsala, A., Tsalikidis, C., Pitiakoudis, M., Simopoulos, C. & Tsaroucha, A. K. Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. *A New Era. Curr Oncol.* **28**(3), 1581–1607 (2021).
11. Foersch, S. et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med.* **29**(2), 430–439 (2023).
12. Rompianesi, G., Pegoraro, F., Ceresa, C. D., Montalti, R. & Troisi, R. I. Artificial intelligence in the diagnosis and management of colorectal cancer liver metastases. *World J Gastroenterol.* **28**(1), 108–122 (2022).
13. Qiu, H., Ding, S., Liu, J., Wang, L. & Wang, X. Applications of Artificial Intelligence in Screening, Diagnosis, Treatment, and Prognosis of Colorectal Cancer. *Curr Oncol.* **29**(3), 1773–1795 (2022).
14. Fernandez-Rozadilla, C. et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat Genet.* **55**(1), 89–99 (2023).
15. Salabei, J. K., Li, X. P., Petrash, J. M., Bhatnagar, A. & Barski, O. A. Functional expression of novel human and murine AKR1B genes. *Chem Biol Interact.* **191**(1–3), 177–184 (2011).
16. Napoli, J. L. Physiological insights into all-trans-retinoic acid biosynthesis. *Biochim Biophys Acta.* **1821**(1), 152–167 (2012).
17. Liu, C. et al. AKR1B10 accelerates the production of proinflammatory cytokines via the NF-κB signaling pathway in colon cancer. *J Mol Histol.* **53**(5), 781–791 (2022).
18. Li W, Liu C, Huang Z, et al. AKR1B10 negatively regulates autophagy through reducing GAPDH upon glucose starvation in colon cancer. J Cell Sci. 2021. 134(8).
19. Shen, Y. et al. Impaired self-renewal and increased colitis and dysplastic lesions in colonic mucosa of AKR1B8-deficient mice. *Clin Cancer Res.* **21**(6), 1466–1476 (2015).
20. Viikilä, P. et al. Carbonic anhydrase enzymes II, VII, IX and XII in colorectal carcinomas. *World J Gastroenterol.* **22**(36), 8168–8177 (2016).
21. Nannini G, De Luca V, D&#x27, et al. A comparative study of carbonic anhydrase activity in lymphocytes from colorectal cancer tissues and adjacent healthy counterparts. J Enzyme Inhib Med Chem. 2022. 37(1): 1651–1655.
22. Eldehna, W. M. et al. Design and synthesis of 6-arylpyridine-tethered sulfonamides as novel selective inhibitors of carbonic anhydrase IX with promising antitumor features toward the human colorectal cancer. *Eur J Med Chem.* **258**, 115538 (2023).
23. Meng, H., Li, W., Boardman, L. A. & Wang, L. Loss of ZG16 is associated with molecular and clinicopathological phenotypes of colorectal cancer. *BMC Cancer.* **18**(1), 433 (2018).
24. Meng, H., Ding, Y., Liu, E., Li, W. & Wang, L. ZG16 regulates PD-L1 expression and promotes local immunity in colon cancer. *Transl Oncol.* **14**(2), 101003 (2021).
25. Kryeziu, K., Bergsland, C. H., Guren, T. K., Sveen, A. & Lothe, R. A. Multiplex immunohistochemistry of metastatic colorectal cancer and ex vivo tumor avatars. *Biochim Biophys Acta Rev Cancer.* **1877**(1), 188682 (2022).
26. Bărbălan, A. et al. Immunohistochemistry predictive markers for primary colorectal cancer tumors: where are we and where are we going. *Rom J Morphol Embryol.* **59**(1), 29–42 (2018).
27. Sukswai, N. & Khoury, J. D. Immunohistochemistry Innovations for Diagnosis and Tissue-Based Biomarker Detection. *Curr Hematol Malig Rep.* **14**(5), 368–375 (2019).
28. Magaki, S., Hojat, S. A., Wei, B., So, A. & Yong, W. H. An Introduction to the Performance of Immunohistochemistry. *Methods Mol Biol.* **1897**, 289–298 (2019).
29. Choi, J. H. & Ro, J. Y. The 2020 WHO Classification of Tumors of Soft Tissue: Selected Changes and New Entities. *Adv Anat Pathol.* **28**(1), 44–58 (2021).
30. Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A. & Zou, J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell.* **186**(8), 1772–1791 (2023).
31. Tran, K. A. et al. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* **13**(1), 152 (2021).

Author contributions

Bobin Ning and Jimei Chi contributed to the design of the study and the acquisition and analysis of the data. Qingyu Meng drafted the work and Baoqing Jia revised it.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76083-9.

Correspondence and requests for materials should be addressed to B.J.

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.