International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Machine Learning-Integrated Analysis of SULF1, CXCL8, and PBLD Expression as Discriminative Biomarkers for Early Detection and Prognosis in Colorectal Cancer

Yang Li[1,2,*], JianFeng Shi[3,4,*], Chao Mei[5], FangYuan Zhou[1], HaoSen Zhao[6], Li Zhang[1]

[1]Department of Laboratory Medicine, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, Liaoning, 121001, People's Republic of China; [2]China Medical University, Shenyang, Liaoning, 110000, People's Republic of China; [3]Department of Cardiology, The Fourth Affiliated Hospital of China Medical University, Shenyang, Liaoning, 110000, People's Republic of China; [4]Department of Cardiology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, Liaoning, 121001, People's Republic of China; [5]Jinzhou Medical University, Jinzhou, Liaoning, 121001, People's Republic of China; [6]The Third Affiliated Hospital of Jinzhou Medical University, Jinzhou, Liaoning, 121001, People's Republic of China

*These authors contributed equally to this work

Correspondence: Li Zhang, Department of Laboratory Medicine, The First Affiliated Hospital of Jinzhou Medical University, No. 2, Section 5, Renmin Street, Guta District, Jinzhou, Liaoning, 121001, People's Republic of China, Email zl252981340@163.com; HaoSen Zhao The Third Affiliated Hospital of Jinzhou Medical University, No. 2, Section 5, Heping Road, Linghe District, Jinzhou, Liaoning, 121001, People's Republic of China, Email zhs_drzhao@outlook.com

Background: Colorectal cancer (CRC) is one of the major cancers that threaten human health. Although the CRC census has been gradually popularized, due to the lack of obvious symptoms in the early stage, it is difficult to detect, and the rapid progression and strong metastasis after onset result in a high incidence of CRC. Therefore, the current research aims to identify more powerful molecular targets and biomarkers for the diagnosis, treatment and clinical research of CRC.

Methods: The limma package was used to analyze datasets GSE4107, GSE110223, and GSE110224 from the Gene Expression Omnibus (GEO) to identify differentially expressed genes (DEGs) in CRC. Functional enrichment analysis of DEGs was performed using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). To further screen for key genes, the DEGs were submitted to the STRING database to construct a protein-protein interaction (PPI) network. Clinical data from The Cancer Genome Atlas (TCGA) database were used to analyze the role of key genes in CRC. Key DEGs were validated using immunohistochemistry, Western blot, and quantitative real-time polymerase chain reaction (RT-qPCR). Survival analysis of key DEGs was performed using the GEPIA database, and survival curves were plotted. The expression levels of DEGs were quantitatively analyzed in samples from 80 CRC patients and 80 healthy controls. Machine learning algorithms were applied to analyze key DEGs and construct a diagnostic model for CRC. A receiver operating characteristic (ROC) curve was plotted to evaluate the performance of the diagnostic model.

Results: A total of 981 (GSE4107), 155 (GSE110223), and 280 (GSE110224) DEGs were identified from the GEO databases, among which 152 DEGs were expressed in at least two datasets. GO and KEGG enrichment analyses revealed that these DEGs were widely involved in biological processes such as the muscle system process and extracellular matrix organization. Downregulated genes were involved in pathways including bile secretion and retinol metabolism. PPI network analysis identified 20 overlapping genes, among which CXCL8 and SULF1 were hub up-regulated genes, while PBLD and 17 others were hub down-regulated genes. mRNA-Seq data and RT-qPCR validation showed that CXCL8 and SULF1 were significantly upregulated in CRC samples, whereas PBLD expression levels were higher in normal tissues compared to CRC tissues. Kaplan-Meier curve analysis indicated that high mRNA expression of SULF1 was significantly associated with poorer overall survival in CRC patients, while high mRNA expression of LRRC19 was associated with better overall survival. In contrast, the mRNA expression of CXCL8 and PBLD showed no significant association with overall survival. Gene expression of SULF1 was significantly correlated with disease-free survival, whereas the gene expression of LRRC19, CXCL8, and PBLD showed no significant correlation with disease-free survival. Immunohistochemical analysis further validated the expression levels of SULF1, CXCL8, and PBLD. The machine learning model demonstrated high efficacy in assisting

Received: 14 July 2025
Accepted: 22 October 2025
Published: 4 December 2025

CRC diagnosis, with an AUC value exceeding 0.8, and the most effective model achieved an AUC value greater than 0.9. Decision curve and calibration curve analyses further confirmed its significant clinical net benefit and good consistency.

Conclusion: These four identified DEGs (SULF1, CXCL8, LRRC19, and PBLD) may contribute to the treatment of CRC as a new therapeutic target and provide valuable biomarkers for cancer metastasis research.The four identified DEGs were combined with machine learning to construct a CRC diagnostic model with high clinical application value.

Keywords: colorectal cancer, biomarker, sulfatase-1, bioinformatic analysis, R language, machine learning

Introduction

Colorectal cancer (CRC) takes hundreds of thousands of lives every year, making it the world's third most lethal cancer.[1] Common symptoms of CRC include diarrhea, constipation, bloody stools, abdominal pain, unexplained weight loss, fatigue, and iron-deficiency anemia due to chronic blood loss. In recent years, the incidence of CRC has shown a marked trend toward younger onset: the proportion of cases among individuals under 55 years increased from 11% in 1995 to 20% in 2019.[2] Although diagnostic and therapeutic techniques for CRC have been progressively improving—including comprehensive strategies such as surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy[3–5]—the five-year overall survival rate of CRC patients remains low.[6] Primary prevention and early detection are crucial strategies for reducing the global burden of CRC and improving patient outcomes.[7,8] Studies indicate that later stages of CRC at diagnosis correlate with poorer survival, while early screening and diagnosis can achieve a five-year overall survival rate of up to 90%.[9,10] Although colonoscopy is regarded as the gold standard for CRC screening, its highly invasive nature leads to low patient compliance and demands substantial medical resources, greatly limiting its widespread application.[11] In contrast, serum biomarkers such as carcinoembryonic antigen and carbohydrate antigen 19–9 offer more convenient testing; however, their sensitivity and specificity are suboptimal, with positivity rates below 30% in early-stage CRC, thus failing to meet clinical screening needs.[12,13] Therefore, from the perspective of molecular biology and laboratory medicine, it is particularly important to find specific biomarkers for CRC.[14–16]

Due to the development of computational science and computer-based technology, bioinformatics that combines biology and informatics has also made considerable progress. Various bioinformatics technologies, such as gene sequencing, proteomics, and metabolomics, have also emerged. Currently, as an important research tool, bioinformatics is gradually being applied to basic mechanistic research by researchers, for example, by helping researchers find differentially expressed genes (DEGs) and the underlying pathways.[17–20] In this context, bioinformatics analytical methods demonstrate great potential and value in identifying tumor-associated genes, elucidating oncogenic mechanisms, optimizing treatment strategies, and improving prognostic assessments. This has enabled in-depth exploration of molecular mechanisms underlying early diagnosis, disease progression, metastasis, and prognostic biomarkers of CRC at the transcriptome level.[21] Moreover, machine learning approaches are increasingly being applied to the diagnosis and efficacy evaluation of various diseases, including CRC.[22,23] Machine learning is a complex interdisciplinary technology involving fields such as statistics and algorithmic complexity theory. It simulates human learning processes through computers, extracting new insights from existing knowledge, continuously improving performance, and enhancing system efficiency and accuracy.[24] Common machine learning methods include support vector machines (SVMs), logistic regression, k-nearest neighbors (KNN), decision trees (DTs), random forests, Least Absolute Shrinkage and Selection Operator regression, and convolutional neural networks. In recent years, with advances in bioinformatics and the establishment and refinement of major biological databases, ML algorithms have fully leveraged their advantages in handling large-scale data, offering new ideas and methods for disease diagnosis and establishing methodological foundations for precision medicine in oncology.[25,26]

This study identified DEGs associated with CRC from cancer-related databases and further screened characteristic diagnostic and prognostic markers related to patient outcomes. Using machine learning approaches, nine ML algorithms were constructed, providing theoretical support for early diagnosis and prognostic assessment of CRC and offering new insights for precision medicine research in CRC.

Materials and Methods

Acquisition of Microarray Data

Three expression profiling datasets (GSE4107,[27] GSE110223,[28] and GSE110224[28]) were obtained from the GEO.[29] Probes were transformed into homologous gene symbols through platform annotation information. The GSE4107 dataset contained 10 normal tissue samples and 12 CRC tissue samples. The GSE110223 dataset contained 13 normal tissue samples and 13 CRC tissue samples. The GSE110224 dataset contained 17 normal tissue samples and 17 CRC tissue samples. Principal component analysis (PCA) is a commonly used method for sample clustering, and is often used for gene expression, diversity analysis, resequencing, and other sample clustering based on various variable information. In the current study, PCA was carried out based on these data to verify and reveal the distribution of the samples.

Kaplan–Meier Survival Analysis

Survival analysis of shared up- or downregulated genes was predicted online using the GEPIA database (http://gepia.cancer-pku.cn/).[30,31] Four genes in the GEPIA platform were searched, and overall survival analysis and disease-free survival were performed based on gene expression. The Log rank test (also known as the Mantel-Cox test) was used for hypothesis testing. The Cox proportional hazard ratio and 95% confidence interval information are also included in the survival chart.

Data Homogenization, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) Functional Annotations of DEGs

Data homogenization was performed using R 4.0.2 software (https://www.r-project.org). The screening of DEGs was performed using the limma package (http://bioconductor.org/packages/release/bioc/html/limma.html) according to the thresholds of |Log FC (fold change)|> 1 and false discovery rate (FDR)< 0.05. The KEGG (http://www.kegg.jp/) and GO (http://www.geneontology.org/) functional annotations of DEGs were performed by DAVID Bioinformatics Resources 6.8 and R 4.0.2 software.

Analysis of the Protein-Protein Interaction (PPI) Network and Hub Genes

To identify the potential hub targets of CRC and the interactions between them, the selected targets were submitted to the STRING network platform (https://string-db.org/) to construct the PPI network.[28] The organism was set to Homo sapiens and the confidence threshold was set to medium (0.400). The top 10 hub genes in the PPI network were identified and ranked by the MCC method using the Cytohubba plug-in.[32]

Human Colorectal Tissue

Formalin fixation and paraffin-embedding samples of 15 paired CRC and adjacent normal tissues were obtained from patients who underwent surgery between October 2019 and December 2020 at the First Affiliated Hospital of Jinzhou Medical University. For immunohistochemistry analysis, tissues were fixed in 4% formalin, embedded in paraffin, and sectioned at a thickness of 4 μm. Informed consent was obtained from all patients before the operation, and all experimental procedures were approved by the the Ethics Committee of the First Affiliated Hospital of Jinzhou Medical University approved all experimental procedures (No.202127).

Immunohistochemistry Analysis

The expression of several major shared up- and downregulated DEGs was analyzed using human CRC and adjacent tissue sections by immunohistochemistry. Tissues were fixed in formalin and embedded in paraffin. The tissue slides were then incubated with anti-CXCL8 (1:50, #MCA497, AbD Serotec, Kidlington, UK), anti-SULF1 (1:100, #BM3925, Boster, Wuhan, China), anti-PBLD1 (1:100, #DF13592, Affinity Biosciences, Jiangsu, China) or anti-LRRC19 (1:100, #200554-T08, Sino Biological, Beijing, China) overnight. After that, the slides were incubated with secondary antibodies for 1 h. Images were obtained by microscopy (Olympus, BX64, Japan), and the percentage of positively stained area was quantified using ImageJ 14.6 (US National Institute of Health, Bethesda, MD).

Table 1 Sequences of Primers

Gene name	Forward 5′-3′	Reverse 5′-3′
CXCL8	TTGGCAGCCTTCCTGATTTCT	TTCTCAGCCCTCTTCAAAAACT
SULF1	GGCTTGATCGGCAACTAGGA	GTTCCTCATCTGCCCTGACC
LRRC19	CACTGGCCTATCGGCTGTAAT	TGACCTTGCTCCATCCATACC
PBLD	TCTGCACACGCTGTTCTCA	CAGCAGCGTCACAGCATAAC
GAPDH	CAAATTCCATGGCACCGTCA	GATGGCATGGACTGTGGTCA

Western Blot

Proteins were extracted from the adjacent and cancer tissues with cell radioimmunoprecipitation assay lysis buffer (Merck Millipore, 92590). Total proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and blotted to polyvinylidene fluoride membranes (Merck Millipore, ISEQ00010) via a wet electrophoretic transfer method. Membranes were blocked with 10% skim milk powder for 1 h at 37 °C with gentle shaking in a water bath shaker, followed by incubation with primary antibodies (CXCL8, 1:1000, SULF1, 1:2000, LRRC19, 1:1000, and PBLD, 1:1000) at 4 °C overnight. After washing with PBS, membranes were incubated with horseradish peroxidase-conjugated secondary antibody for 1.5 h at room temperature with shaking. Protein expression levels were detected using an enhanced chemiluminescence kit (Beyotime, P0018), imaged on a Bio-Rad ChemiDoc TMMP system, and analyzed by ImageJ software.

Quantitative Real-Time Polymerase Chain Reaction (RT-qPCR)

Total RNA from adjacent and cancer tissues was isolated using an RNA Extraction Kit (Accurate Biology, AG21017). Samples were subjected to RT-qPCR using the Evo M-MLV RT Premix for qPCR (Accurate Biology, AG11706) and SYBR Premix Ex TaqTM (Accurate Biology, AG11702), with amplification performed on a Rotor-Gene Q instrument (Qiagen, Germany). The relative expression of the target genes was calculated by the ΔΔCt method.[33] Each sample was analyzed in triplicate. The sequences of primers are shown in Table 1.

Machine Learning

This section presents a machine learning analysis conducted using R, including data preprocessing, feature selection, model training, and evaluation. Multiple machine learning algorithms were employed to construct diagnostic models for colon cancer, including DT, single-layer feedforward neural network, KNN, single-layer neural network, SVM, LightGBM, glmnet, logistic regression, and random forest. The optimal model was selected through cross-validation. A total of 160 patients were randomly divided into a training set comprising 120 patients and a test set comprising 40 patients. Nine machine learning algorithms were applied to build diagnostic models based on the training set. The The Area Under the Curve (AUC) values, calibration curves, and optimal decision thresholds for each model were derived from the training set. Subsequently, the test set data (n = 60) were used to evaluate the performance of the trained models, yielding corresponding AUC values, calibration curves, and optimal decision thresholds.

Statistical Analysis

SPSS 25.0 software and R 3.8.1 software were used for statistical analysis. Using Student's *t* test and one-way analysis of variance, the differences between different groups were statistically compared. Survival curves were compared with the Log rank test to generate Kaplan–Meier survival charts. For all analyses, $P < 0.05$ was considered statistically significant.

Results

Validation of the Datasets and DEGs Identified Between Control and CRC Tissues

The datasets GSE4107 (23,386 genes), GSE110223 (26,389 genes), and GSE110224 (31,970 genes) were obtained from the GEO database. Pearson's test and PCA were used to verify the data in the three datasets. Batch normalization and merging to reduce variability to increase signals and reduce the proportion of false-positive results (Figure 1A). Based on

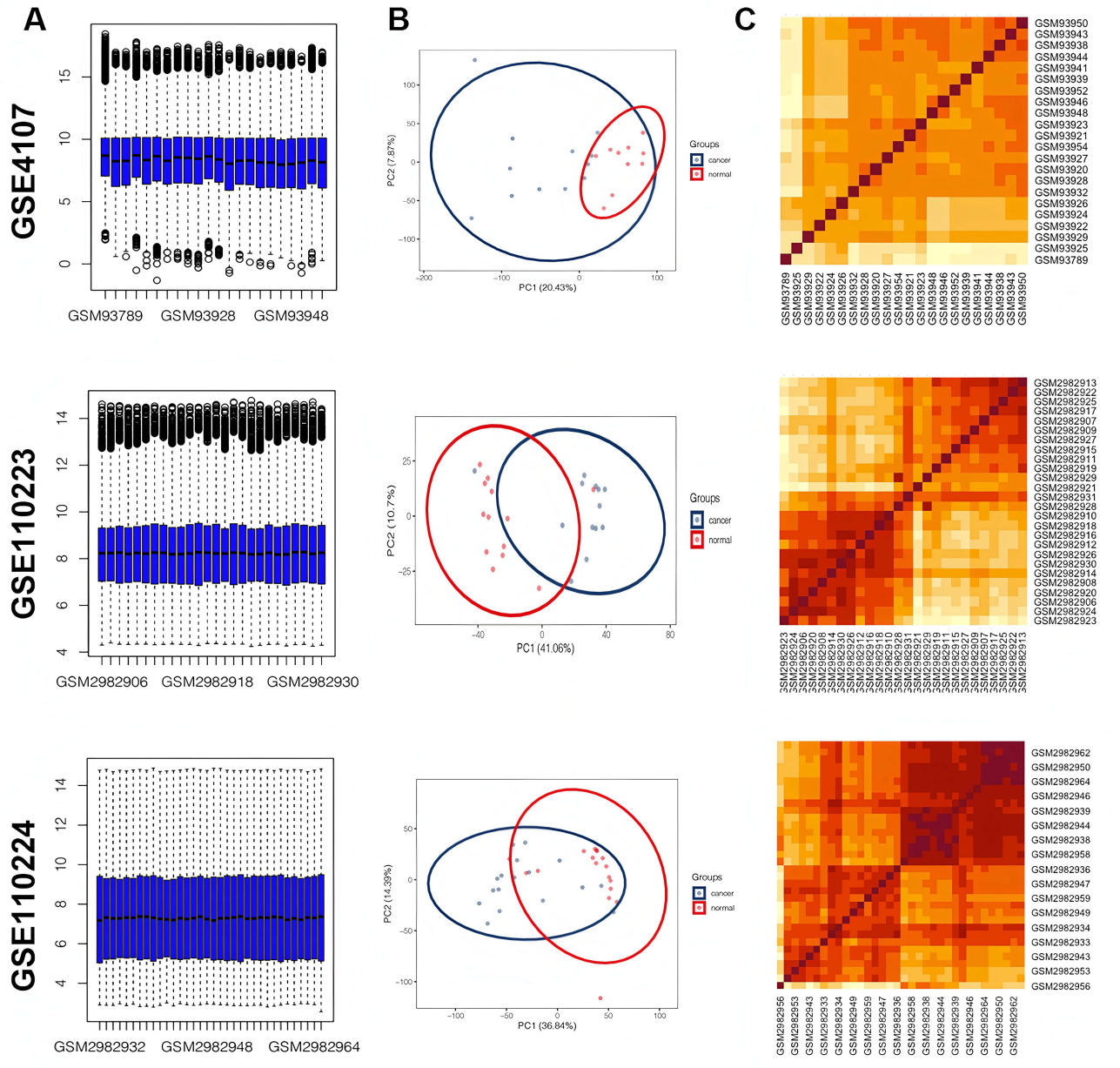

Figure 1 Normalization of raw data in the GSE4107, GSE110223 and GSE110224 datasets,and identification of shared DEGs. (**A**) Normalization in the Gene Expression Omnibus database. Blue represents data after normalization. (**B**) PCA of samples from the GSE4107, GSE110223 and GSE110224 datasets. In the figure, principal component 1 (PC1) and principal component 2 (PC2) are used as the x-axis and y-axis, respectively, to draw the scatter diagram, where each point represents a sample. In such a PCA diagram, the farther the two samples are from each other, the greater the difference is between the two samples in gene expression patterns. (**C**) Pearson's correlation analysis of samples from the three datasets. The color reflects the intensity of the correlation.

PCA, the intragroup data repeatability for GSE4107, GSE110223 and GSE110224 was acceptable (Figure 1B). Additionally, strong correlations were observed among samples within the control group and within the CRC group across all three datasets (Figure 1C).

DEGs were screened by the limma package according to the thresholds of |Log FC|> 1 and FDR< 0.05. According to the hierarchical clustering results, CRC samples and normal samples could be effectively distinguished by hub genes (Figure 2A–C). There were 809 upregulated genes and 172 downregulated genes in 981 DEGs in the GSE4107 dataset. Additionally, the 155 DEGs in the GSE110223 dataset included 44 upregulated genes and 111 downregulated genes. Moreover, the 280 DEGs in the GSE110224 dataset included 102 upregulated genes and 178 downregulated genes. The volcano map results show the DEGs identified from each dataset (Figure 2D–F). In total, 886 differential genes were

https://doi.org/10.2147/IJGM.S553709

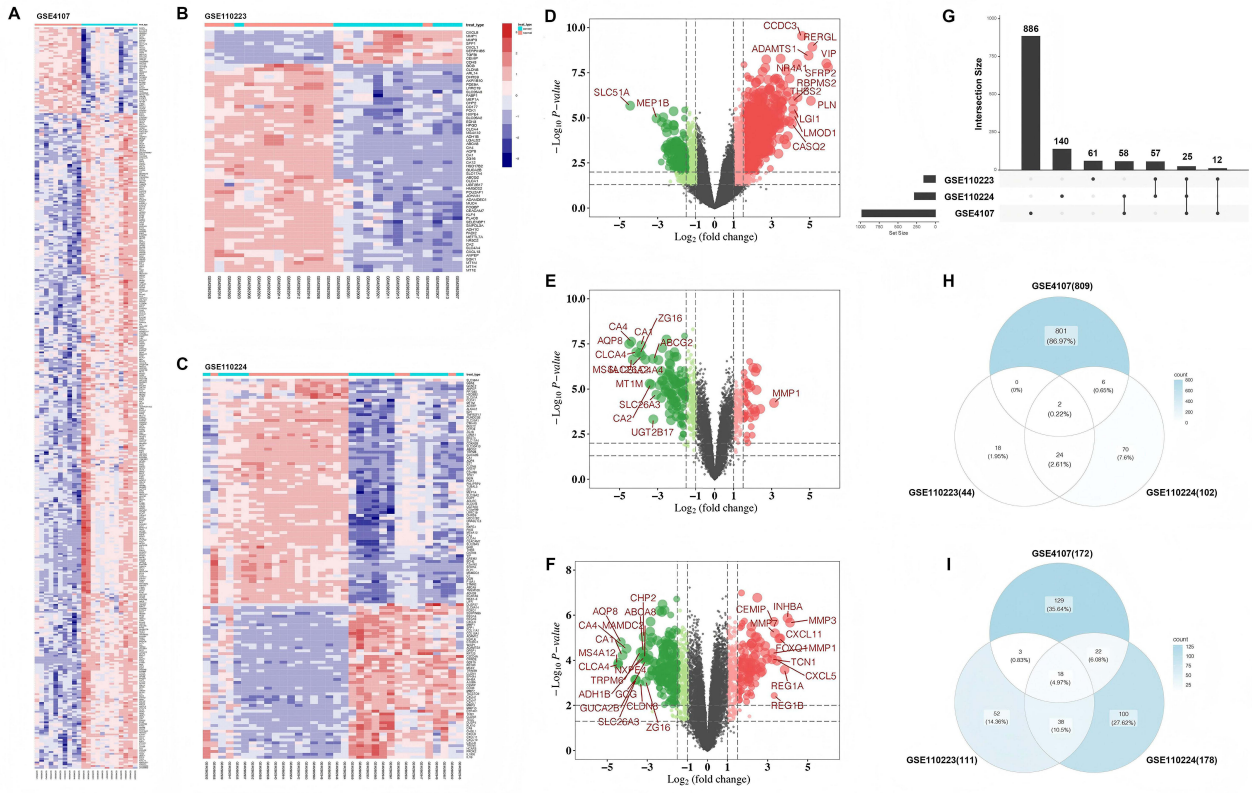

Figure 2 Visualization of differentially expressed genes in the GSE4107, GSE110223, and GSE110224 datasets. (**A–C**) Cluster heat map illustrating the differentially expressed mRNAs in paired human CRC tissues and normal tissues in the three GSE datasets. The upregulated genes are marked in pink, and the downregulated genes are marked in blue. (**D–F**) The volcano plot illustrates the differences between CRC tissues and normal tissues after analysis of the GSE4107, GSE110223 and GSE110224 datasets with GEO2R(The overlapping genes in Figure 2E are MS4A12,SLC26A4). (**G**) Intersection of differential genes in three datasets. (**H** and **I**) Venn diagram demonstrating the intersections of genes between GEO data.

expressed only in the GSE4107 dataset, 140 and 61 differential genes were expressed only in GSE110224 and GSE110223, respectively, and 152 differential genes were expressed in at least two GSE datasets (Figure 2G). In the visual results of the Venn diagram, circles denote the three GSE datasets, and their area of overlap represents the core. These results were standardized. Two Venn diagrams suggested that 2 shared upregulated genes were contained within the 3 datasets (Figure 2H), and 18 shared downregulated genes were contained in the downregulated genes (Figure 2I).

Functional Annotation of DEGs by GO and KEGG Analyses

GO terms and KEGG pathway analysis were performed to investigate the potential functions of the DEGs. As shown in Figure 3A and B, GO analysis of GSE4107 from the DAVID database shows that the upregulated transcripts are mainly involved in biological processes (BPs), such as muscle system processes and muscle contraction. Variations in cell components (CCs) were markedly enriched in the collagen-containing extracellular matrix and contractile fiber. Molecular functions (MFs) are mainly concentrated in the combination of structural components of the extracellular matrix and glycosaminoglycans (Figure 3C). The results of GSE110223 in BPs included extracellular matrix (Figure 3D) and structure organization, while in GSE110224, the response to lipopolysaccharide was the most common (Figure 3G). The variations in CCs of GSE110223 and GSE110224 were markedly enriched in the basal part of cells and collagen-containing extracellular components respectively (Figures 3E and 2H). The variations in MFs of GSE110223 and GSE110224 were mainly focused on cytokine activity, receptor ligand activity and signaling receptor activator activity (Figure 3F and I). GO analysis results from DAVID for GSE4107, GSE110223, and GSE110224 showed that the downregulated transcripts were involved in BPs, CCs and MFs (Figure 4A–I).

KEGG pathway enrichment analysis of the common DEGs was performed. With a $P < 0.05$ cutoff, the most enriched biological pathways associated with CRC were identified. The KEGG enhancement analysis results of GSE4107 revealed

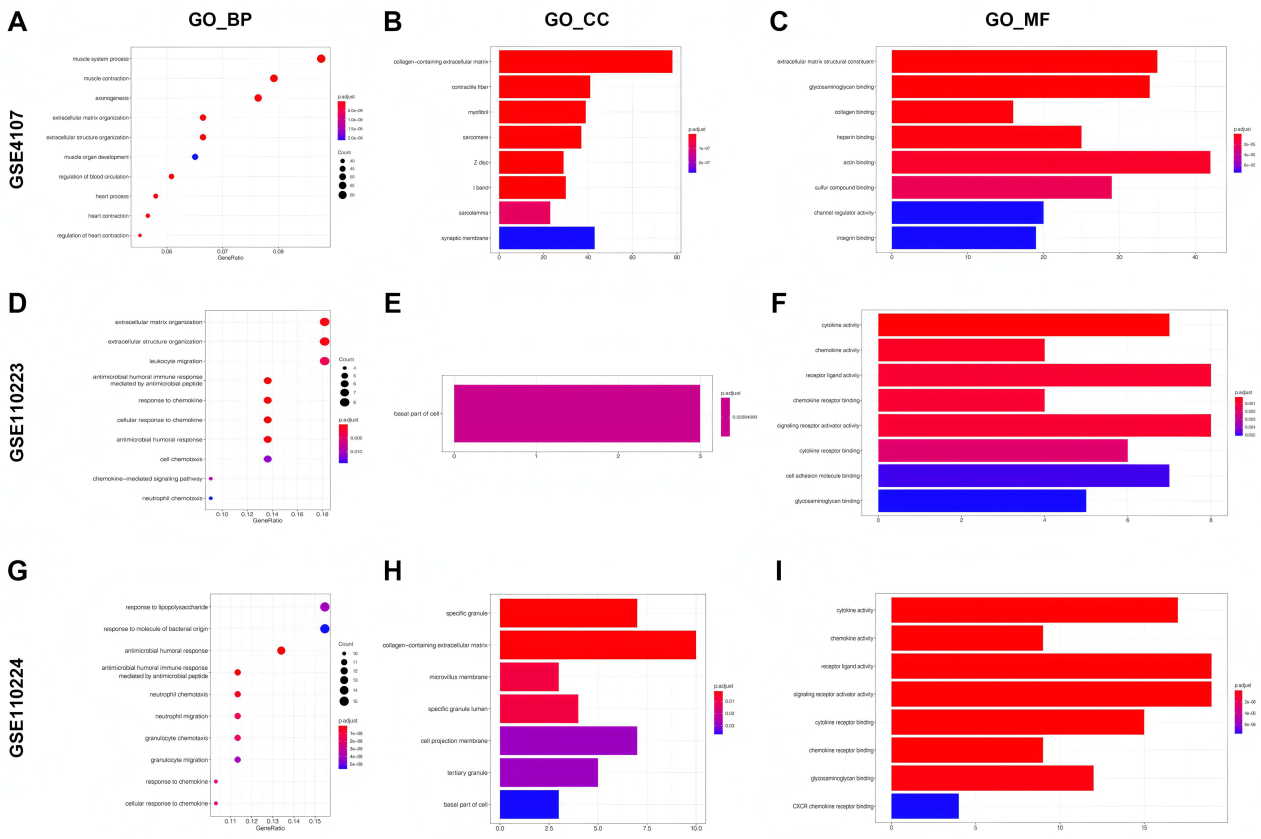

Figure 3 GO annotation of upregulated mRNAs with the top 10 enrichment scores. (**A–C**) Detailed information relating to changes in the BP, CC, and MF of different upregulated genes in CRC and control tissues through GO enrichment analyses in the GSE4107 dataset. (**D–F**) Functional analysis of differentially upregulated genes in the GSE110223 dataset. (**G–I**) Functional analysis of differentially upregulated genes in the GSE110224 dataset.

that all upregulated genes were primarily enriched in vascular smooth muscle contraction, cell adhesion molecules and the calcium signaling pathway. The KEGG pathway analysis of GSE110223 and 110224 showed that the upregulated genes were dominated by the IL-17 signaling pathway, rheumatoid arthritis, the NF-κB signaling pathway, and cytokine-cytokine receptor interactions. In GSE4107, downregulated genes were predominantly enriched in bile secretion, retinol metabolism, and chemical carcinogenesis. In GSE110223 and GSE110224, downregulated genes were enriched in bile secretion, pancreatic secretion, nitrogen metabolism, proximal tubule bicarbonate reclamation, and steroid hormone biosynthesis (Figure 5A–F).

PPI Network Analysis of DEGs

The PPI network of DEGs was constructed by Cytoscape software based on the results of the STRING database. The hub genes from the DEGs with a high degree were identified. We found that 20 genes overlapped among the GSE4107, GSE110223 and GSE110224 datasets. The two upregulated hub genes (Figure 6A) were 3-hydroxy-3-methylglutaryl-CoA synthase 8 (CXCL8) and sulfatase 1 (SULF1), and the eighteen downregulated hub genes (Figure 6B) were 3-hydroxy-3-methylglutaryl-CoA synthase 2 (HMGCS2), ATP binding cassette subfamily G member 2 (Junior blood group) (ABCG2), CD177 molecule (CD177), alanyl aminopeptidase, membrane (ANPEP), alcohol dehydrogenase 1C (class I), gamma polypeptide (ADH1C), aldo-keto reductase family 1 member B10 (AKR1B10), aquaporin 8 (AQP8), calcineurin like EF-hand protein 2 (CHP2), carbonic anhydrase 2 (CA2), chloride channel accessory 1 (CLCA1), dehydrogenase/reductase 9 (DHRS9), leucine rich repeat containing 19 (RC19), peptidyl arginine deiminase 2 (PADI2), phenazine biosynthesis like protein domain containing (PBLD), selenium binding protein 1 (SELENBP1), solute carrier family 4 member 4 (SLC4A4) and zymogen granule protein 16 (ZG16). These results suggest that these dysregulated genes may play pivotal roles in CRC. Detailed information on the DEGs is shown in Table 2.

https://doi.org/10.2147/IJGM.S553709

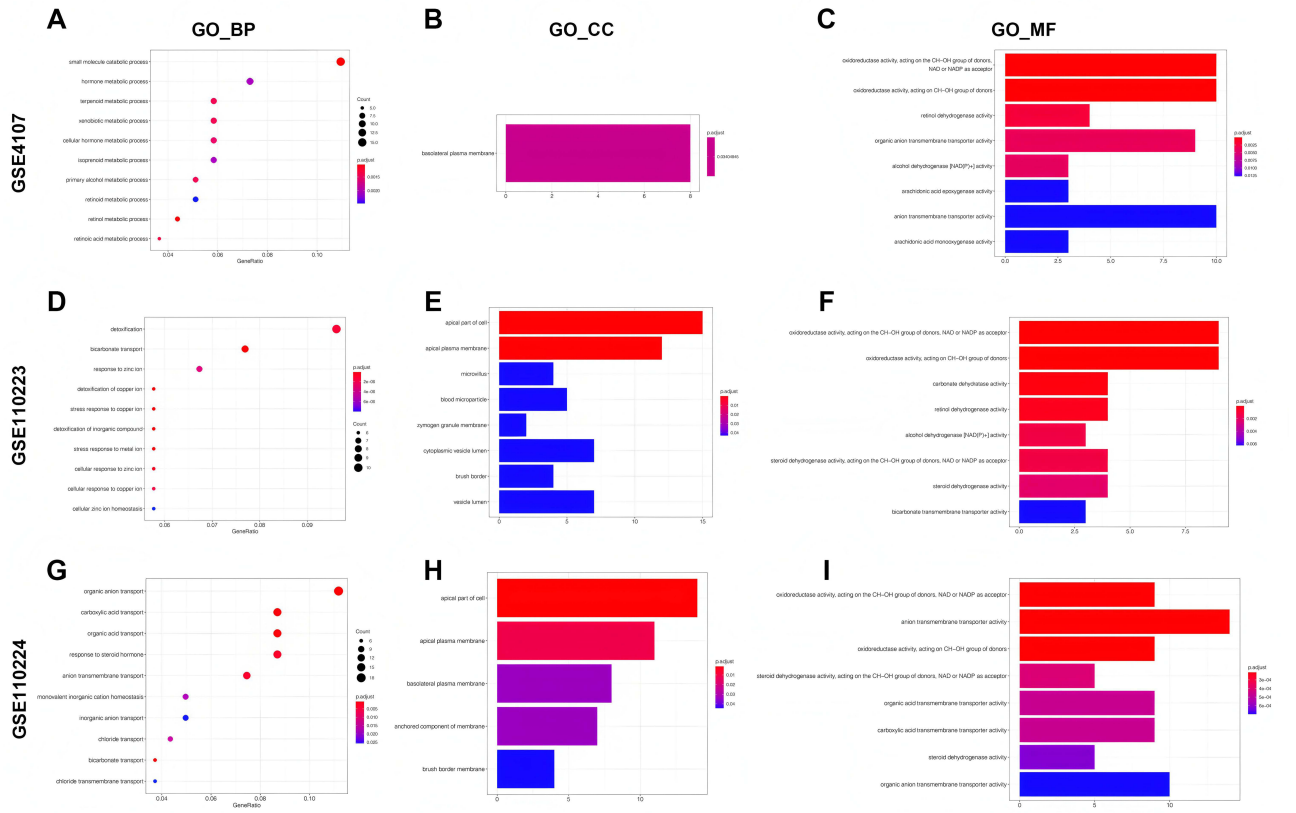

Figure 4 GO annotation of downregulated mRNAs with the top 10 enrichment scores. (**A**–**C**) Detailed information relating to changes in the BP, CC, and MF of different downregulated genes in CRC and control tissues through GO enrichment analyses in the GSE4107 dataset. (**D**–**F**) Functional analysis of differentially downregulated genes in the GSE110223 dataset. (**G**–**I**) Functional analysis of differentially downregulated genes in the GSE110224 dataset.

CXCL8, SULF1, LRRC19 and PBLD Gene Expression

To validate hub gene expression changes in CRC, expression levels were compared between normal and cancer samples using mRNA-Seq data from GSE4107, GSE110223, and GSE110224. CXCL8 and SULF1 were significantly upregulated in CRC samples (Figure 7A and B). Additionally, we found that LRRC19 and PBLD genes were expressed at significantly higher levels in normal tissue than in CRC tissue (Figure 7C and D). Consistent with the mRNA-Seq results, the expression levels of CXCL8, SULF1, LRRC19, and PBLD in cancer/adjacent tissues, which were obtained from, were further demonstrated by RT-qPCR (Figure 7E–H). The Clinicopathologic data from patients were shown in Table 3.

The Association Between the SULF1, CXCL8, LRRC19 and PBLD Gene and Overall Survival and Disease-Free Survival

To further explore whether these hub genes play an essential role in CRC, the clinical data of patients were obtained from the GEPIA database (http://gepia.cancer-pku.cn), and the effects of these 4 genes (SULF1, CXCL8, LRRC19 and PBLD) on the overall survival and disease-free survival of these genes were analyzed by Kaplan-Meier curves. As shown in Figure 5, among these genes, the high mRNA expression of SULF1 was significantly associated with poor overall survival in CRC patients, while high mRNA expression of LRRC19 was associated with better overall survival in CRC patients, with log-rank (Mantel–Cox) P values of 0.017 for SULF1 and 0.044 for LRRC19 (Figure 8A and B). However, mRNA expression of CXCL8 and PBLD showed no significant association with overall survival ($P > 0.05$, Figure 8C and D). Regarding disease-free survival, only SULF1 expression showed a significant correlation, with high SULF1 expression

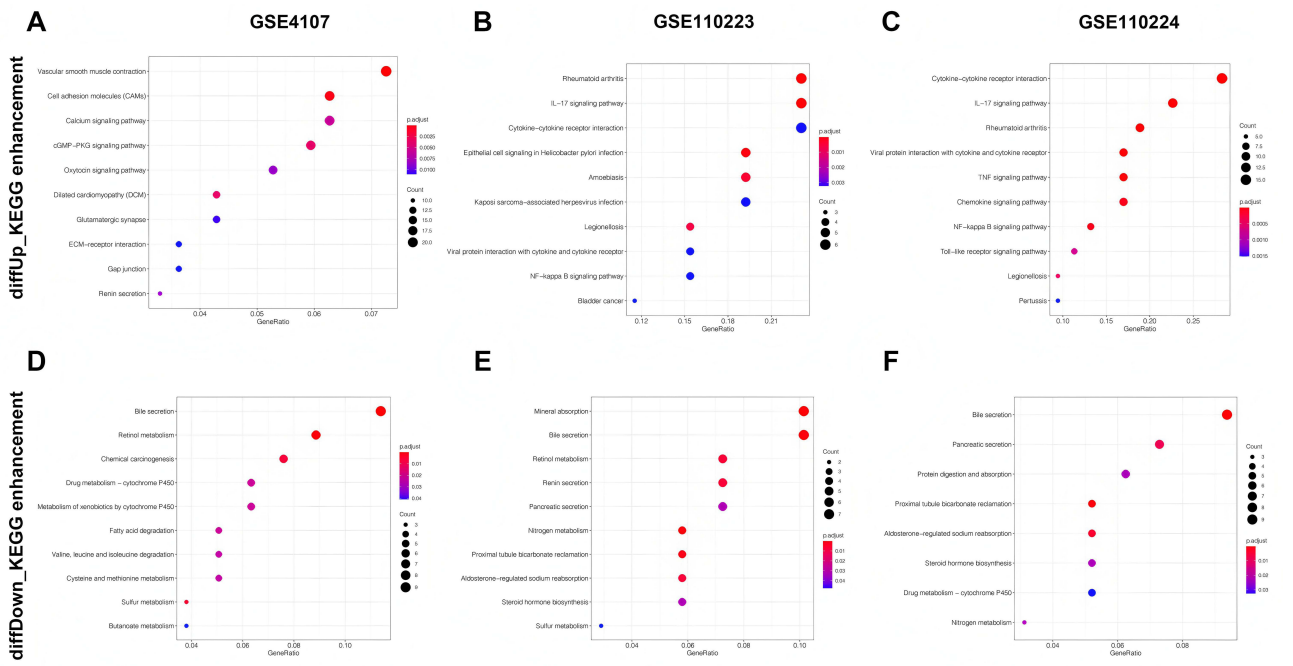

Figure 5 KEGG pathway enrichment analysis of mRNAs with the top 10 enriched scores. (**A**–**C**) KEGG pathway enrichment results of differentially upregulated genes in GSE4107, GSE110223 and GSE110224. (**D**–**F**) KEGG pathway enrichment results of differentially downregulated genes in GSE4107, GSE110223 and GSE110224 that are involved in the regulatory network.

associated with poor prognosis. However, the gene expression of CXCL8, LRRC19 and PBLD had no significant correlation with disease-free survival (Figure 9A–D).

CXCL8, SULF1, LRRC19 and PBLD Gene Expression Verification in Human Tissue Samples

To further verify the expression levels of SULF1, CXCL8, PBLD, and LRRC19, immunohistochemistry analysis in human CRC/adjacent tissue samples was performed. As shown in Figure 10A and B, consistent with the predicted results, the expression levels of SULF1 and CXCL8 were upregulated in cancer tissue. SULF1 was mainly located in colon acinar cells, while CXCL8 was mainly located in the extra-acinar matrix. Moreover, the expression level of PBLD was significantly decreased in cancer tissues compared with adjacent tissues, and this result was consistent with the predicted results (Figure 10C). However, inconsistent with the predicted results, the expression level of LRRC19 in cancer and adjacent tissues did not change significantly (Figure 10D). The expression levels of these proteins were also demonstrated by Western blot (Figure 11C and D). Based on the above results, SULF1, CXCL8, and PBLD demonstrate greater potential as biomarkers for the prevention and diagnosis of CRC.

A Case Study of Clinical Data and Machine Learning

An independent samples *t*-test was conducted on CXCL8, SULF1, and LRRC19 levels in 160 patients using SPSS. The results demonstrated that CXCL8 and SULF1 levels were elevated in the tumour group relative to the normal group, whereas PBLD levels were lower in the tumour group. The model constructed using machine learning exhibited high efficacy in assisting the diagnosis of colon cancer patients, and it had a superior effect on clinical decision-making. The independent samples *t*-test revealed significant differential expression between the normal and tumor groups ($P < 0.05$) (Figure 12A). Based on these findings, we utilized these three biomarkers to develop a machine learning model for the clinical diagnosis of colon cancer, aiming to assess the efficacy and clinical relevance of the constructed model. In the machine learning analysis, the AUC values for each model were computed. The results indicated that most models

https://doi.org/10.2147/IJGM.S553709

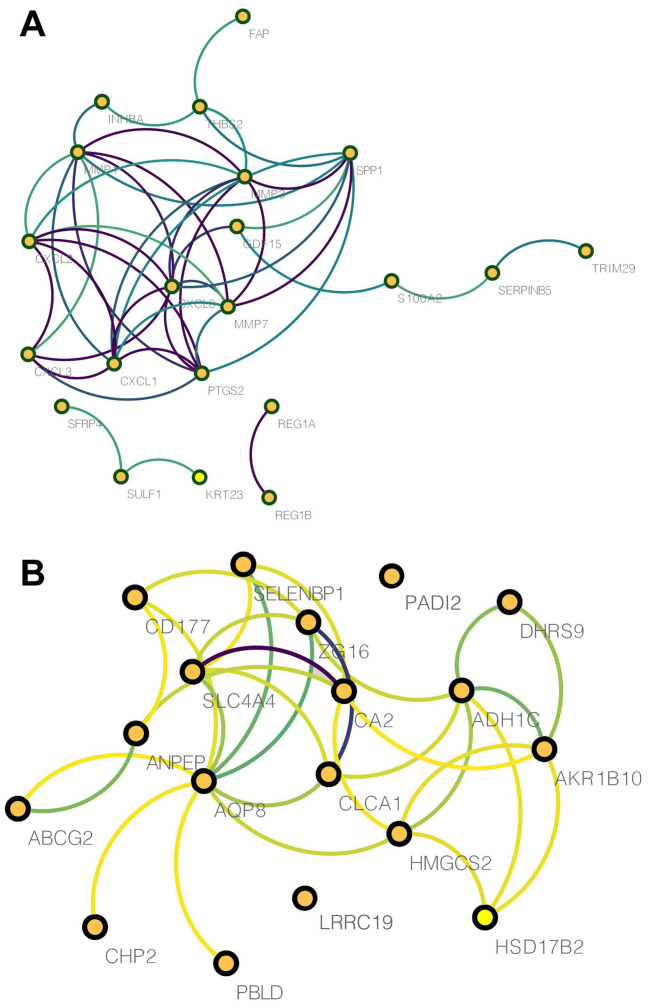

Figure 6 PPI network of DEGs. (**A**) Gene interactions of upregulated hub genes were constructed by Cytoscape software. (**B**) Downregulated hub gene interactions were constructed by Cytoscape software.

demonstrated satisfactory performance, with AUC values exceeding 0.8 (Figure 12B). Notably, the most effective model achieved an AUC value greater than 0.9 (Figure 12C). To evaluate the calibration of the model, we examined the calibration curve, where a deviation from the diagonal line indicates potential prediction bias. In the training set

Table 2 Detailed Information of the up-Regulated DEGs

ID	Gene Name	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT	KEGG_PATHWAY
CXCL8	C-X-C motif chemokine ligand 8 (CXCL8)	GO:0001525~angiogenesis,GO:0002237~response to molecule of bacterial origin, GO:0006928~movement of cell or subcellular component,GO:0006935~chemotaxis,GO:0006954~inflammatory response,GO:0006955~immune response,GO:0007050~cell cycle arrest,GO:0007165~signal transduction,GO:0007186~G-protein coupled receptor signaling pathway,GO:0008285~negative regulation of cell proliferation,GO:00	GO:0005576~extracellular region, GO:0005615~extracellular space,GO:0005622~intracellular,	GO:0005153~interleukin-8 receptor binding,GO:0005515~protein binding,GO:0008009~chemokine activity,	hsa04060:Cytokine-cytokine receptor interaction, hsa04062:Chemokine signaling pathway,hsa04064: NF-kappa B signaling pathway,hsa04620: Toll-like receptor signaling pathway, hsa04621: NOD-like receptor signaling pathway,hsa04622: RIG-I-like receptor signaling pathway, hsa04932: Non-alcoholic fatty liver disease (NAFLD),hsa05120: Epithelial cell signaling in Helicobacter pylori infection,hsa05131: Shigellosis, hsa0513

(*Continued*)

Table 2 (Continued).

ID	Gene Name	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT	KEGG_PATHWAY
SULF1	sulfatase 1(SULF1)	GO:0001822~kidney development,GO:0001937~negative regulation of endothelial cell proliferation,GO:0002063~chondrocyte development, GO:0 003094~glomerular filtration, GO:00069 15~apoptotic process, GO:000815 2~metabolic process, GO:001057 5~positive regulation of vascular endothelial growth factor production,GO:0014846~esophagus smooth muscle contraction,GO:001 6525~negative regulation of	GO:0005615~extrac ellular space, GO:0005783~ endoplasmic reticulum,GO:00057 94~Golgi apparatus,GO:0005 795~Golgi stack,GO:0005886~ plasma membrane,GO:0009 986~cell surface,GO:0045121~membrane raft,	GO:0003824~cataly tic activity, GO:0004065~arylsulfatase activity,GO:0005509~calcium ion binding,GO:0008449~N-acetylglucosamine- 6-sulfatase activity,GO:0008484~sulfuric ester hydrolase activity,	
HMGCS2	3-hydroxy-3-methylglutaryl-CoA synthase 2 (HMGCS2)		GO:0006695~choles terol biosynthetic process,GO:0008299~isoprenoid biosynthetic process, GO:004695 1~ketone body biosynthetic process,	GO:0005739~mitoc hondrion, GO:00057 43~mitochondrial inner membrane,GO:0005759~mitochondrial matrix,	hsa00072:Synthesis and degradation of ketone bodies,hsa00280:Va line, leucine and isoleucine degradation, hsa006 50:Butanoate metabolism, hsa009 00:Terpenoid backbone biosynthesis,hsa01 100:Metabolic pathways,hsa01130: Biosynthesis of antibiotics,
ABCG2	ATP binding cassette subfamily G member 2 (Junior blood group)(ABCG2)		GO:0006810~transp ort, GO:0006855~dr ug transmembrane transport,GO:00068 79~cellular iron ion homeostasis,GO:0015886~heme transport,GO:0042493~response to drug, GO:0042908~x enobiotic transport,GO:00464 15~urate metabolic process,	GO:0005634~nucle us, GO:0005886~pla sma membrane, GO:0016 021~integral component of membrane,GO:0031966~mitochondrial membrane,	hsa02010:ABC transporters,hsa049 76:Bile secretion,
CD177	CD177 molecule(CD177)		GO:0007596~blood coagulation, GO:005 0900~leukocyte migration,	GO:0005886~plasm a membrane, GO:0031 225~anchored component of membrane,GO:0070062~extracellular exosome,	
ANPEP	alanyl aminopeptidase, membrane(ANPEP)	Amino acid transport and metabolism,	GO:0001525~angio genesis, GO:000650 8~proteolysis,GO:0030154~cell differentiation,GO:0043171~peptide catabolic process, GO:004671 8~viral entry into host cell,	GO:0005615~extracellular space, GO:0005765~ lysosomal membrane,GO:0005793~endoplasmic reticulum-Golgi intermediate compartment,GO:0005829~cytosol, GO:0005887~integral component of plasma membrane,GO:0009897~external side of plasma membrane,GO:0016 021~integral component of membrane,GO:0070062~extracellular exosome,	hsa00480:Glutathio ne metabolism, hsa011 00:Metabolic pathways, hsa04614: Renin-angiotensin system,hsa04640:H ematopoietic cell lineage,
ADH1C	alcohol dehydrogenase 1C (class I), gamma polypeptide (ADH1C)	Energy production and conversion,	GO:0006069~ethan ol oxidation,	GO:0005829~cytos ol,	hsa00010:Glycolysis / Gluconeogenesis,h sa00071:Fatty acid degradation,hsa003 50:Tyrosine metabolism,hsa008 30:Retinol metabolism,hsa009 80:Metabolism of xenobiotics by cytochrome P450, hsa00982:Drug metabolism - cytochrome P450, hsa01100:Metabolic pathways,hsa05204: Chemical carcinogenesis,

(*Continued*)

Table 2 (Continued).

ID	Gene Name	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT	KEGG_PATHWAY
AKR1B10	aldo-keto reductase family 1 member B10 (AKR1B10)		GO:0001523~retino id metabolic process,GO:000608 1~cellular aldehyde metabolic process, GO:000758 6~digestion,GO:00 08202~steroid metabolic process, GO:001648 8~farnesol catabolic process,GO:004459 7~daunorubicin metabolic process, GO:004459 8~doxorubicin metabolic process,GO:005511 4~oxidation- reduction process,	GO:0005764~lysoso me, GO:0005829~cy tosol, GO:0070062~ extracellular exosome,	hsa00040:Pentose and glucuronate interconversions,hs a00051: Fructose and mannose metabolism, hsa000 52:Galactose metabolism, hsa005 61:Glycerolipid metabolism, hsa011 00:Metabolic pathways,
AQP8	aquaporin 8(AQP8)		GO:0006810~transp ort, GO:0006833~wa ter transport, GO:00099 92~cellular water homeostasis,GO:00 15793~glycerol transport,GO:00342 20~ion transmembrane transport, GO:00713 20~cellular response to cAMP,	GO:0005886~plasm a membrane, GO:0005 887~integral component of plasma membrane,GO:0016 021~integral component of membrane,GO:0045 177~apical part of cell,	hsa04976:Bile secretion,
CHP2	calcineurin like EF-hand protein 2 (CHP2)	Signal transduction mechanisms / Cytoskeleton / Cell division and chromosome partitioning / General function prediction only,	GO:0008284~positi ve regulation of cell proliferation,GO:00 10922~positive regulation of phosphatase activity,GO:0015031 ~protein transport,GO:00423 07~positive regulation of protein import into nucleus,GO:004594 4~positive regulation of transcription from RNA polymerase II promoter,GO:00708 86~positive regulation of calcineurin-NFAT signaling cascade,GO:007127 7~cellular response to calcium ion,	GO:0005634~nucle us, GO:0005737~cyt oplasm, GO:0005886 ~plasma membrane,	
CA2	carbonic anhydrase 2(CA2)		GO:0001822~kidne y development,GO:0 002009~morphoge nesis of an epithelium,GO:0006 730~one-carbon metabolic process, GO:000926 8~response to pH, GO:0010043~res ponse to zinc ion,GO:0015670~ca rbon dioxide transport,GO:00157 01~bicarbonate transport, GO:00322 30~positive regulation of synaptic transmission, GABAergic, GO:0032 849~positive regulation of cellular pH reduction,GO:00381	GO:0005615~extracellular space, GO:0005737~ cytoplasm,GO:0005 829~cytosol,GO:00 05886~plasma membrane,GO:0005 902~microvillus, GO:0016323~basolater al plasma membrane,GO:0030 424~axon, GO:0043 209~myelin sheath, GO:0045177 ~apical part of cell,GO:0070062~ex tracellular exosome,	hsa00910:Nitrogen metabolism, hsa049 64:Proximal tubule bicarbonate reclamation,hsa049 66: Collecting duct acid secretion, hsa04971: Gastric acid secretion, hsa04972: Pancreatic secretion, hsa04976: Bile secretion,
CLCA1	chloride channel accessory 1 (CLCA1)		GO:0006508~prote olysis, GO:0006810~ transport,GO:00068 16~calcium ion transport, GO:00342 20~ion transmembrane transport,GO:00714 56~cellular response to hypoxia,GO:190247 6~chloride transmembrane transport,	GO:0005615~extracellular space, GO:0005886~ plasma membrane, GO:0005 887~integral component of plasma membrane,GO:0005 902~microvillus, GO:0042589~zymogen granule membrane,	hsa04924:Renin secretion,hsa04972: Pancreatic secretion,

(*Continued*)

Table 2 (Continued).

ID	Gene Name	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT	KEGG_PATHWAY
DHRS9	dehydrogenase/ reductase 9 (DHRS9)	General function prediction only,	GO:0002138~retino ic acid biosynthetic process,GO:0008209~androgen metabolic process, GO:0030855~epithelial cell differentiation,GO:0042448~progester one metabolic process,GO:0042572~retinol metabolic process,GO:0042904~9-cis-retinoic acid biosynthetic process,GO:0055114~oxidation-reduction process,	GO:0005789~endo plasmic reticulum membrane,GO:0030176~integral component of endoplasmic reticulum membrane, GO:0031090~organelle membrane,	hsa00830:Retinol metabolism, hsa01100:Metabolic pathways,
LRRC19	leucine rich repeat containing 19 (LRRC19)		GO:0006469~negati ve regulation of protein kinase activity, GO:0019221~cytokine- mediated signaling pathway,GO:0046426~negative regulation of JAK- STAT cascade,	GO:0005737~cytopl asm, GO:0016021~integral component of membrane,	
PADI2	peptidyl arginine deiminase 2 (PADI2)		GO:0006325~chromatin organization,GO:0010848~regulation of chromatin disassembly,GO:0018101~protein citrullination,GO:0021762~substantia nigra development,GO:0030520~intracellula r estrogen receptor signaling pathway, GO:0036413~histone H3-R26 citrullination,GO:0048096~chromatin- mediated maintenance of transcription,GO:0070100~negative regulation of chemokine- mediated signaling pathway,GO:190162	GO:0005737~cytopl asm, GO:0005829~c ytosol,GO:0035327~transcriptionally active chromatin, GO:0070062~extracellular exosome,	
PBLD	phenazine biosynthesis like protein domain containing(PBLD)		GO:0009058~biosynthetic process,GO:0010633~negative regulation of epithelial cell migration,GO:0010719~negative regulation of epithelial to mesenchymal transition, GO:0030277~maintenance of gastrointestinal epithelium,GO:0030512~negative regulation of transforming growth factor beta receptor signaling pathway, GO:0050680~negative regulation of epithelial cell proliferation,GO:0060392~negative	GO:0005737~cytopl asm, GO:0070062~extracellular exosome,	
SELENBP1	selenium binding protein 1 (SELENBP1)		GO:0015031~protei n transport,	GO:0005615~extracellular space, GO:0005730~ nucleolus,GO:0005829~cytosol,GO:0016020~membrane,GO:0070062~extracellular exosome,	
SLC4A4	solute carrier family 4 member 4 (SLC4A4)		GO:0006810~transp ort, GO:0006814~sodium ion transport,GO:0015698~inorganic anion transport,GO:0015701~bicarbonate transport, GO:0035725~sodium ion transmembrane transport, GO:0051453~regulation of intracellular pH,GO:0098656~an ion transmembrane transport,	GO:0005886~plasm a membrane, GO:0005887~integral component of plasma membrane,GO:0016021~integral component of membrane,GO:0016323~basolateral plasma membrane, GO:0070062~extracellular exosome,	hsa04964:Proximal tubule bicarbonate reclamation,hsa04972: Pancreatic secretion,hsa04976: Bile secretion,

(*Continued*)

Table 2 (Continued).

ID	Gene Name	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT	KEGG_PATHWAY
ZG16	zymogen granule protein 16(ZG16)		GO:0015031~protei n transport,	GO:0005578~protei naceous extracellular matrix,GO:0005796 ~Golgi lumen,GO:0031012 ~extracellular matrix,GO:0042589 ~zymogen granule membrane, GO:0060 205~cytoplasmic membrane- bounded vesicle lumen,	

(Figure 12D), the calibration curve closely aligned with the 45° diagonal but showed a slight elevation above it. This observation suggests a degree of underestimation, indicating that the model may predict probabilities lower than the actual observed probabilities of events. Although this discrepancy is minimal, it may reflect a slight miscalibration of the model during training, potentially attributed to minor overfitting. In the testing set (Figure 12E), we further assessed the clinical utility of the model using Decision Curve Analysis (DCA) (Figure 12F and G), evaluating its net clinical benefit across different threshold probabilities for both the training and validation datasets.

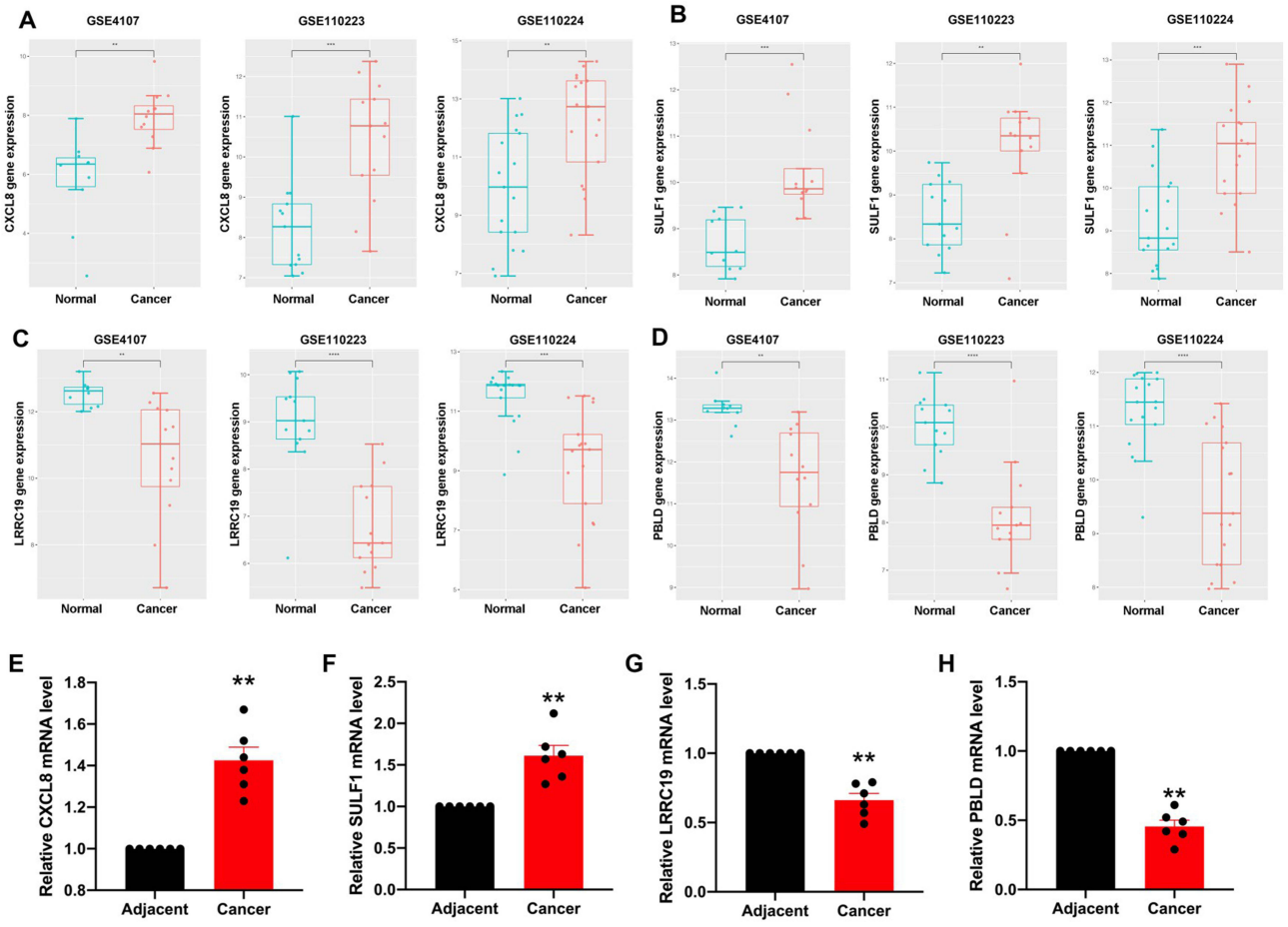

Figure 7 mRNA expression levels of CXCL8, SULF1, LRRC19 and PBLD in CRC across TCGA. (**A**) The expression of CXCL8 in normal and CRC tissues was analyzed in GSE4107, GSE110223 and GSE110224 by using the TCGA database. (**B**) The expression of SULF1 gene expression by using the TCGA database. (**C**) The expression of LRRC19 gene expression by using the TCGA database. (**D**) The expression of PBLD gene expression by using the TCGA database. (**E**–**H**) The RT-qPCR results of CXCL8, SULF1, LRRC19, and PBLD (n=15). Between-group comparisons: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; ** $P < 0.0001$.

Table 3 Clinicopathologic Data from Patients

Access ID	Age	Gender	Tumor Size (cm)	Tumor Location	Histologic Grade	Lymphatic Metastasis	Venous Invasion
4671602	61	Male	7*6	Colon	T4a	Present	Present
4673620	50	Male	4.5*4	Rectum	T3	Absent	Absent
4671722	66	Male	4.5*3.5	Rectum	T3	Absent	Absent
4672724	59	Male	5*2.8	Colon	T4a	Present	Present
4673143	63	Male	4*3	Colon	T3	Absent	Absent
4673502	61	Female	4*4	Colon	T3	Absent	Present
4672637	46	Female	3*3	Colon	T3	Present	Absent
4831533	56	Male	6*4	Colon	T3	Present	Absent
4832474	61	Male	4*3.5	Colon	T3	Present	Absent
4831603	82	Female	4.5*4.3	Colon	T3	Absent	Absent
4831626	71	Female	4.5*4	Colon	T3	Present	Absent
4834073	67	Female	3.9*3	Colon	T3	Absent	Absent
4832133	65	Female	4.5*4.5	Rectum	T3	Absent	Absent
4833061	77	Male	5.3*4.7	Rectum	T2	Absent	Absent
4830659	64	Male	3*2.5	Rectum	T2	Absent	Absent

Discussion

In the present study, two shared upregulated genes and two shared downregulated genes, which are probably associated with CRC progression, were identified from three GSE datasets based on the analysis results. Furthermore, SULF1 was screened as a potential biomarker for CRC based on survival analysis prediction, expression prediction, and immunohistochemical verification results.

Although the mechanism of SULF1 in CRC remains unclear, a large number of studies have confirmed its role in cardiovascular disease and liver cancer. Jin-Ping Lai et al found that the antitumor effect of a combination of doxorubicin and apicidin is increased in SULF1-expressing cells and mice.[34] Additionally, SULF1 facilitates histone H4 acetylation by regulating the activities of HDAC and histone acetyltransferase, further enhancing the induction of apoptosis by the HDAC inhibitors apixidine and scriptaid, as well as tumor growth, migration and angiogenesis inhibitory effects.[35] In addition, Renumathy Dhanasekaran et al demonstrated significant activation of the transforming growth factor-β (TGF-β)/SMAD transcriptional pathway by SULF1 both in vitro and in vivo. Overexpression of SULF1 promotes TGF-β-induced gene expression and epithelial-mesenchymal transition and enhances cell migration/invasiveness.[36] Beyond cancer, Jiangbin Wu et al demonstrated that SULF1 also plays an important role during cardiac fibrosis. SULF1 interference attenuates the activation of myocardial fibroblasts and the deposition of collagen in the primary cardiac fibroblast culture system, while the overexpression of SULF1 facilitates the activation of TGF-β-mediated fibroblasts.[37]

Based on the results of these studies and the results of the current research on biometrics, SULF1 may also play an essential regulatory role in CRC and the specific mechanism is worthy of further exploration.[38]

CXCL8, also known as interleukin-8 (IL-8), plays an important role in the inflammatory response, which negatively regulates cell adhesion molecule production[39] and positively regulate angiogenesis.[40] To date, many studies have shown that CXCL8 can be used as a biomarker.[41–43] In the present study, both bioinformatic and IHC analyses demonstrated that CXCL8 was highly expressed in cancer tissues and was associated with poor prognosis. Notably, unlike SULF1, which was mainly expressed in glandular cells, CXCL8 was mainly distributed in fibroepithelial tissues. Consistent with this, Cheng et al demonstrated that CCL20 and CXCL8 synergistically promote disease progression and poor survival outcomes in CRC patients through synergistic induction of epithelial-mesenchymal transition.[44] Moreover, it has been indicated that high expression of CXCL8 in CRC tissues significantly reduces CRC anoikis, a form of apoptosis that occurs when anchorage-dependent cells either show loss of adhesion or inappropriate adhesion and promotes CRC proliferation and metastasis.[45] Nevertheless, previous studies only proved the correlation between CXCL8 and CRC, but how CXCL8 regulates the progress of CRC and its molecular mechanism are still unclear.

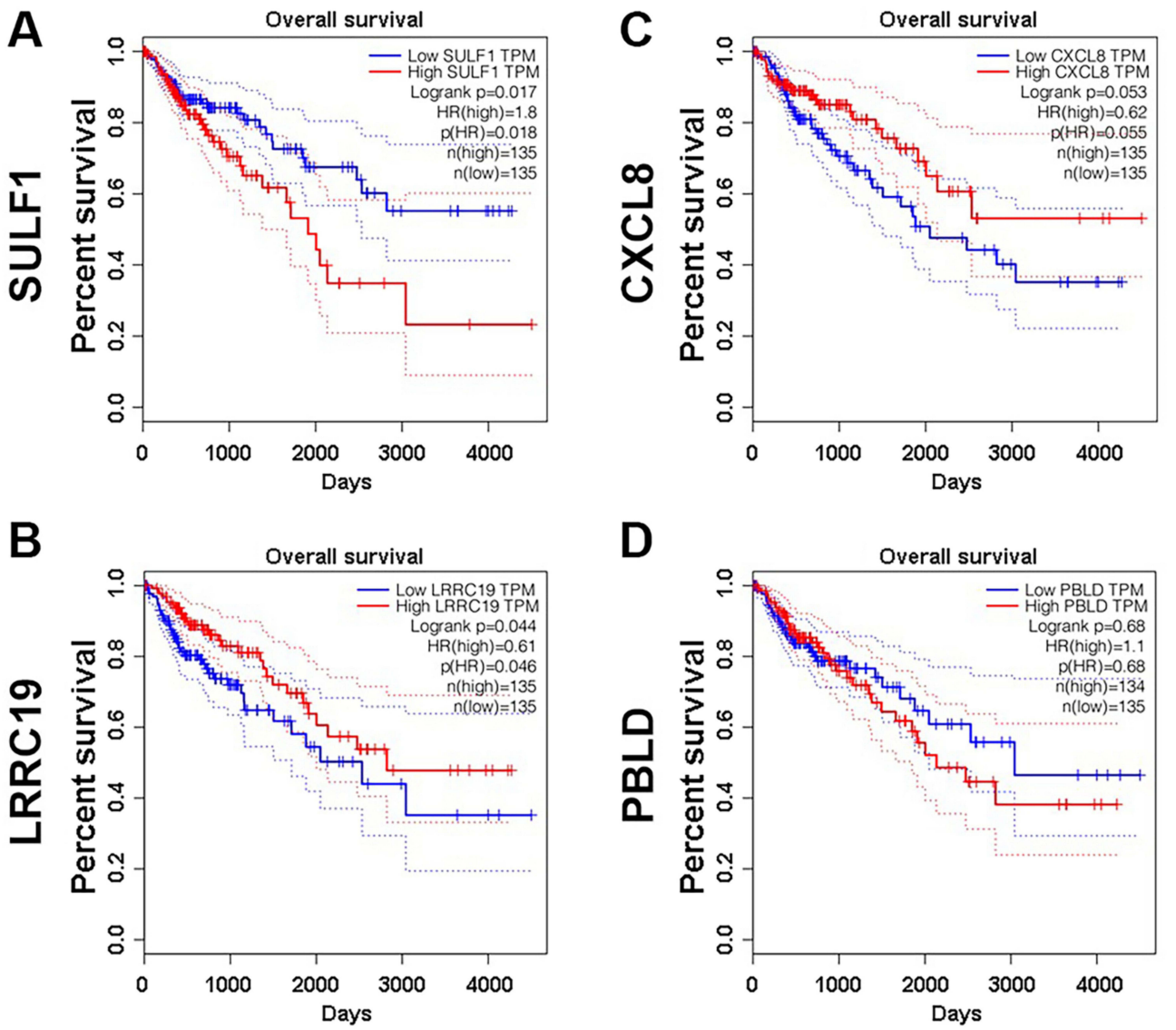

Figure 8 The impact of gene expression on the survival of CRC patients. (**A**) The overall survival rate of CRC patients was estimated by the Kaplan-Meier method according to low or high SULF1 expression. The overall survival rate of CRC patients in the high SULF1 group was significantly lower than that of patients in the low SULF1 group. (**B**) The overall survival rate of CRC patients was estimated by the Kaplan-Meier method. The overall survival rate of patients in the high LRRC19 group was significantly higher than that of patients in the low LRRC19 group. (**C**) The overall survival rate of CRC patients in the high CXCL8 group and low CXCL8 group was not significantly different. (**D**) The overall survival rate of CRC patients in the high PBLD group and low PBLD group was not significantly different.

Another noteworthy gene is PBLD (also known as MAWBP). In the present study, bioinformatics analysis and IHC results showed that its expression in CRC was significantly downregulated compared with that in adjacent tissues. There are currently few studies on the regulatory role of PBLD in tumors. Youyong Lu et al demonstrated that MAWBP could be a new GC-related protein even though its physiological roles remain unexplored, inhibiting EMT and suppressing EMT-aided malignant cell progression.[46,47] Yiran Liang et al indicated that circKDM4C inhibits breast cancer cell progression and doxorubicin resistance by protecting PBLD from miR-548p-mediated degradation.[48] Therefore, the effect of PBLD on the proliferation and migration of CRC and its mechanism in CRC urgently need to be studied.

Studies have shown that PBLD has isomerase activity and can inhibit the migration and proliferation of endothelial cells, as well as suppress the epithelial cell-mesenchymal transition. In the process of tumor growth, it is often accompanied by the excessive proliferation and migration of tumor cells, which results in the low expression of PBLD in CRC. SULF1 can participate in the process of cell apoptosis, proliferation and migration. It is generally

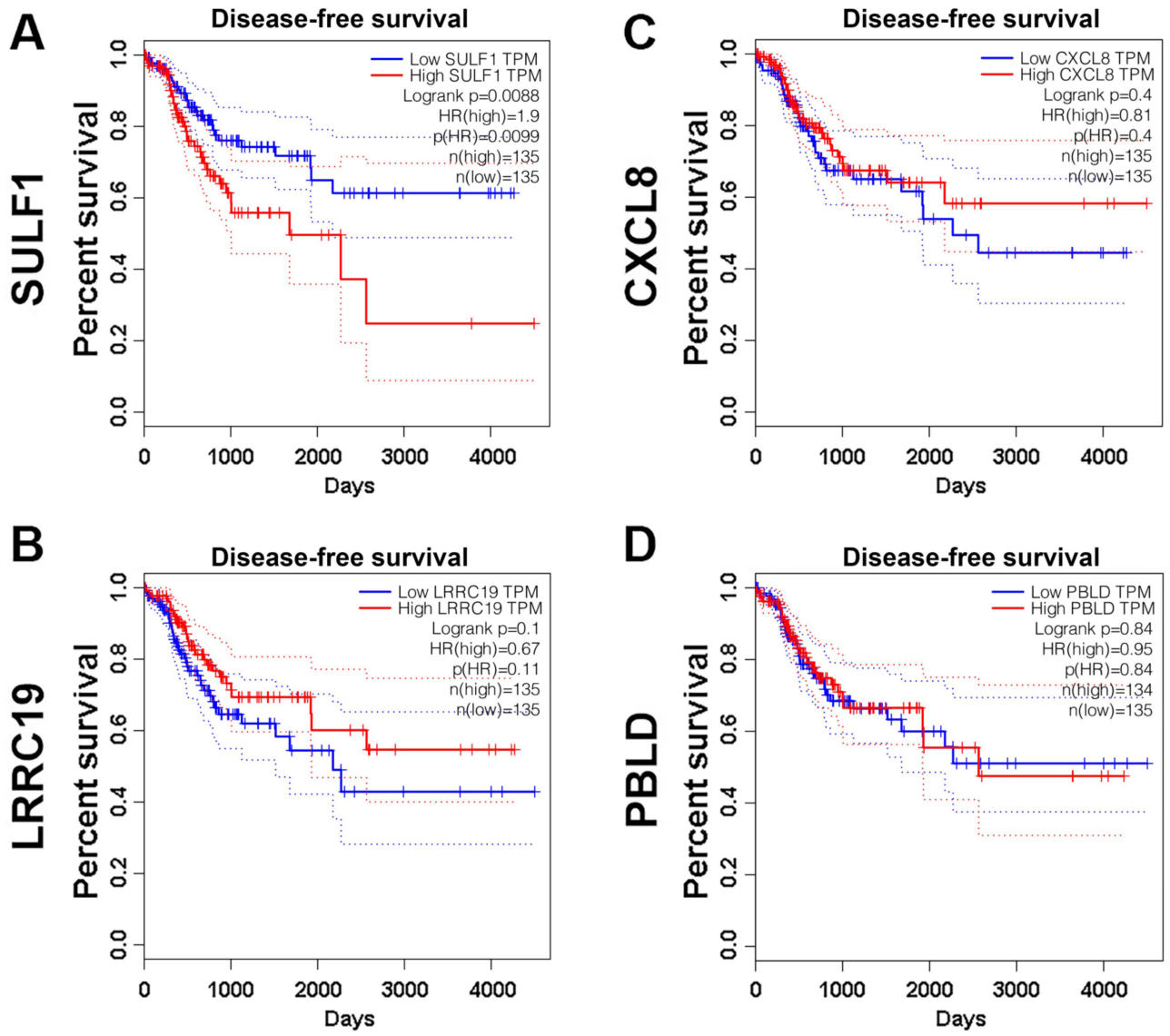

Figure 9 The impact of gene expression on the disease-free survival of CRC patients. (**A**) The disease-free survival rate of CRC patients was estimated by the Kaplan-Meier method according to low or high SULF1 expression. The disease-free survival rate of CRC patients in the high SULF1 group was significantly lower than that of patients in the low SULF1 group. (**B**) The disease-free survival rate of CRC patients in the high LRRC19 group and low LRRC19 group was not significantly different. (**C**) The disease-free survival rate of CRC patients in the high CXCL8 group and low CXCL8 group was not significantly different. (**D**) The disease-free survival rate of CRC patients in the high PBLD group and low PBLD group was not significantly different.

considered a tumor suppressor. However, in studies on liver cancer, it is found that SULF1 is often highly expressed and is associated with poor prognosis. The present study found similar results, so the reasons for the high expression of SULF1 in CRC still need to be further explored. As an inflammatory chemokine, CXCL8 can participate in the positive regulation of blood vessel formation, cell cycle and cell response to a variety of exogenous stimuli, and it can also enter cells to participate in the process of signal transduction. CRC is a kind of intestinal tumor, and the abnormally active intestinal inflammatory response is also closely related to CRC development. Therefore, this may be the reason for the high expression of CXCL8 in CRC.

An independent samples *t*-test was conducted on the expression levels of CXCL8, SULF1, and LRRC19 derived from 160 patients. The analysis revealed significant differences in the expression of these genes between the NORMAL and TUMOR groups ($P < 0.05$). In light of these findings, we sought to develop a machine learning-based diagnostic model

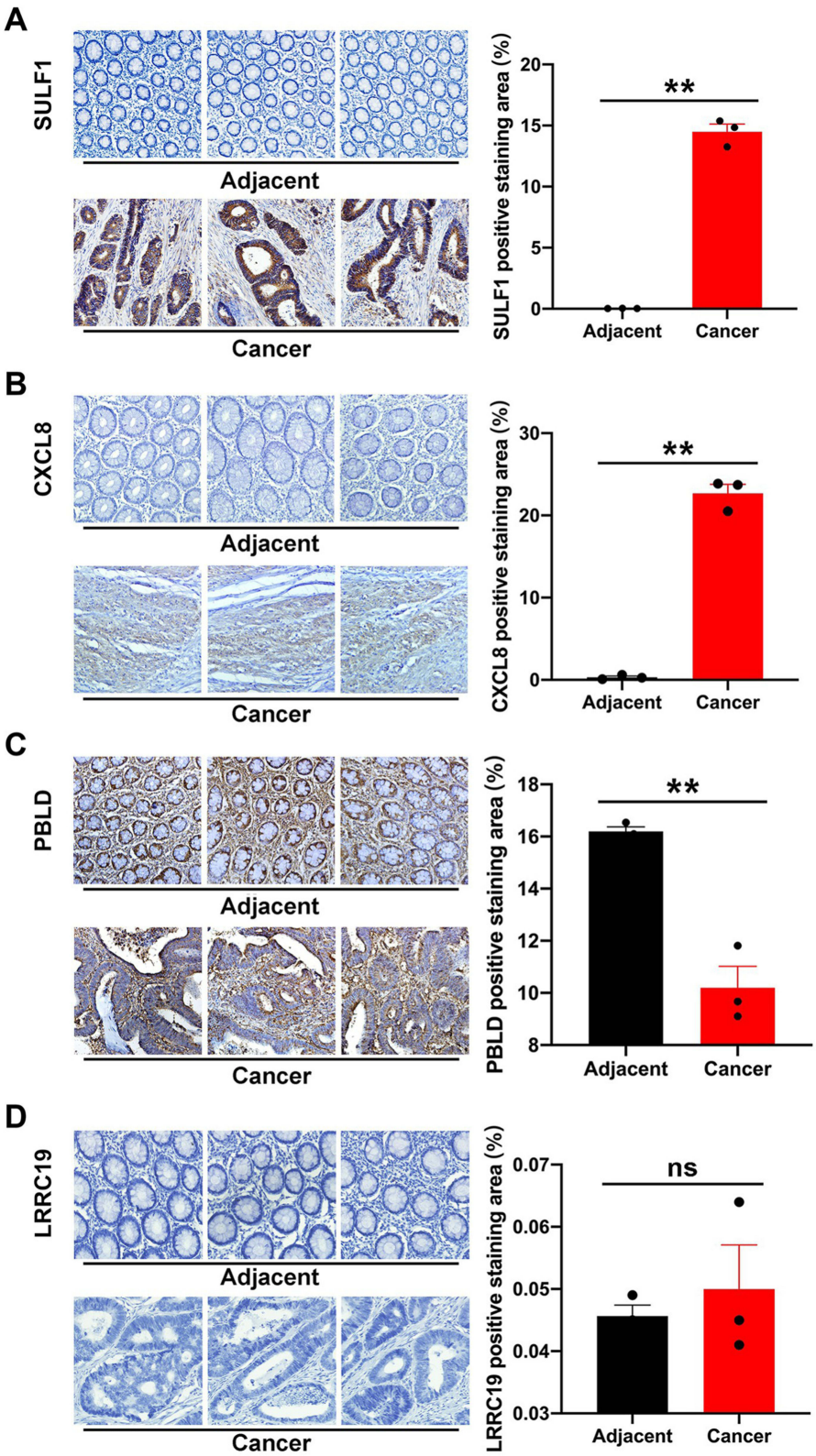

Figure 10 The protein expression levels of SULF1, CXCL8, PBLD and LRRC19 in CRC by immunohistochemistry analysis. (**A**) The protein level of SULF1 was detected in adjacent tissues and CRC tissues. (**B**) The protein level of CXCL8 was detected in adjacent tissues and CRC tissues. (**C**) The protein level of PBLD was detected in adjacent tissues and CRC tissues. (**D**) The protein level of LRRC19 was detected in adjacent tissues and CRC tissues. In A-D, n=15.Between-group comparisons: ns, $P > 0.05$;**,$P < 0.01$.

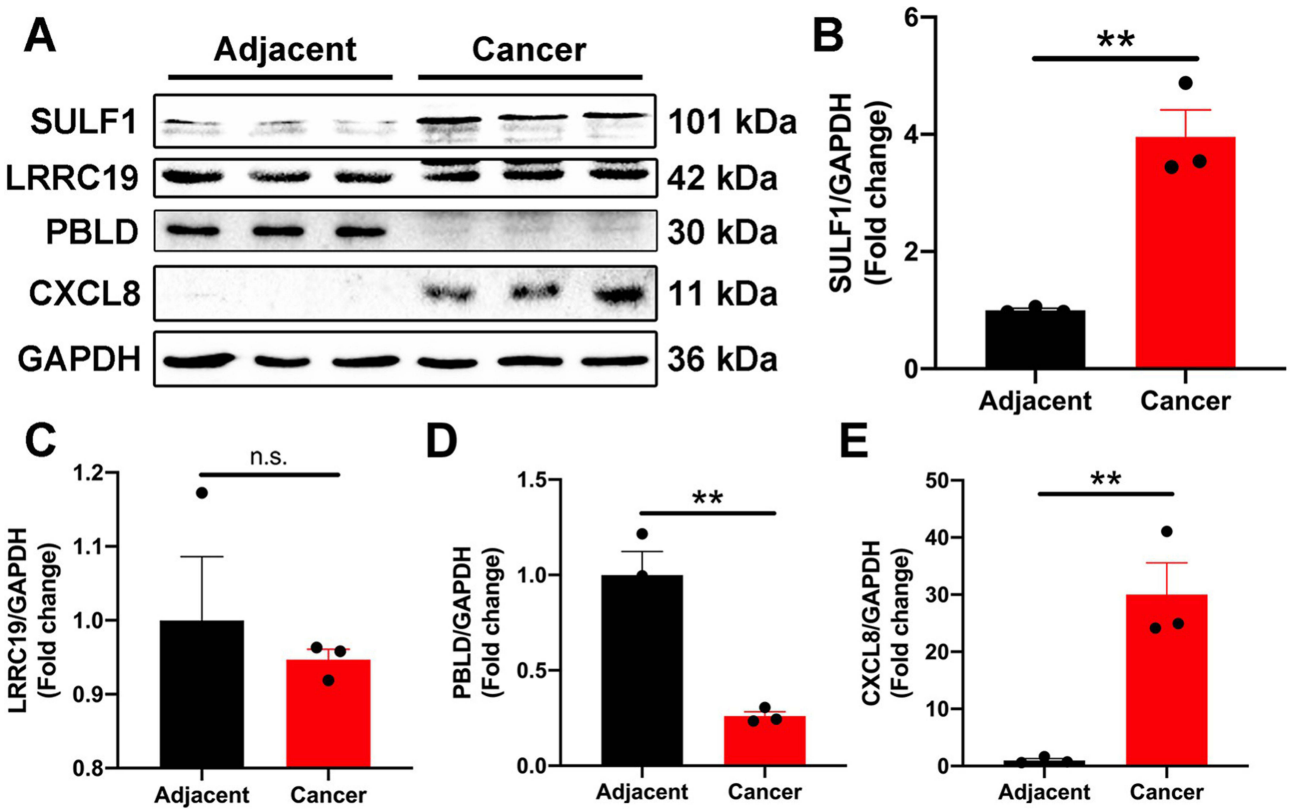

Figure 11 The protein expression levels of SULF1, CXCL8, PBLD and LRRC19 in CRC by Western blot. (**A**) The blots were cut prior to hybridisation with antibodies to differentiate between the protein bands of interest.The protein expression levels of SULF1, LRRC19, PBLD, and CXCL8 were measured in adjacent tissues and CRC tissues (n = 12). (**B**–**E**) Statistical analysis of these proteins. Protein expression was normalized to that of GAPDH.

for colon cancer, utilising the aforementioned three biomarkers, with a view to evaluating its predictive efficacy and clinical relevance.

In the machine learning analysis, the performance of multiple models was assessed by calculating their area under the receiver operating characteristic curve (AUC). The results indicated that the majority of models exhibited robust discriminatory power, with AUC values exceeding 0.8. Of particular note was the best-performing model, which demonstrated exceptional accuracy, achieving an AUC value surpassing 0.9, thereby highlighting its potential utility in clinical applications.The calibration curve is employed to evaluate the calibration of the predictive model, that is to say, the degree of correspondence between the predicted probabilities posited by the model and the actual observed probabilities. In an ideal scenario, the calibration curve would be positioned on a 45°diagonal, indicating that the predicted probabilities generated by the model are in exact alignment with the actual probabilities of occurrence. A divergence from the 45° diagonal line in the calibration curve is indicative of the presence of prediction bias in the model. In the training set, the calibration curve is in close alignment with the diagonal, exhibiting a slight elevation above it, indicative of a mild underestimation of predicted probabilities. This indicates that the model's predictions are somewhat lower than the actual observed probabilities. While this miscalibration is not severe, it may indicate minor overfitting during the training process. In contrast, the calibration curve for the testing set deviates more markedly from the diagonal, exhibiting a general tilt toward a 40° slope and displaying segmentation. In the probability range of 0–0.45, the curve is positioned above the diagonal, indicating an underestimation; whereas in the range of 0.45–1, it is situated below the diagonal, signifying an overestimation. This segmented bias may be attributed to discrepancies in the data distribution characteristics between the validation and training sets, particularly in high-risk samples. These discrepancies contribute to a more pronounced model prediction error at higher probabilities, thereby revealing a degree of bias under these conditions. A comparative analysis demonstrates superior calibration in the training set, while the testing set

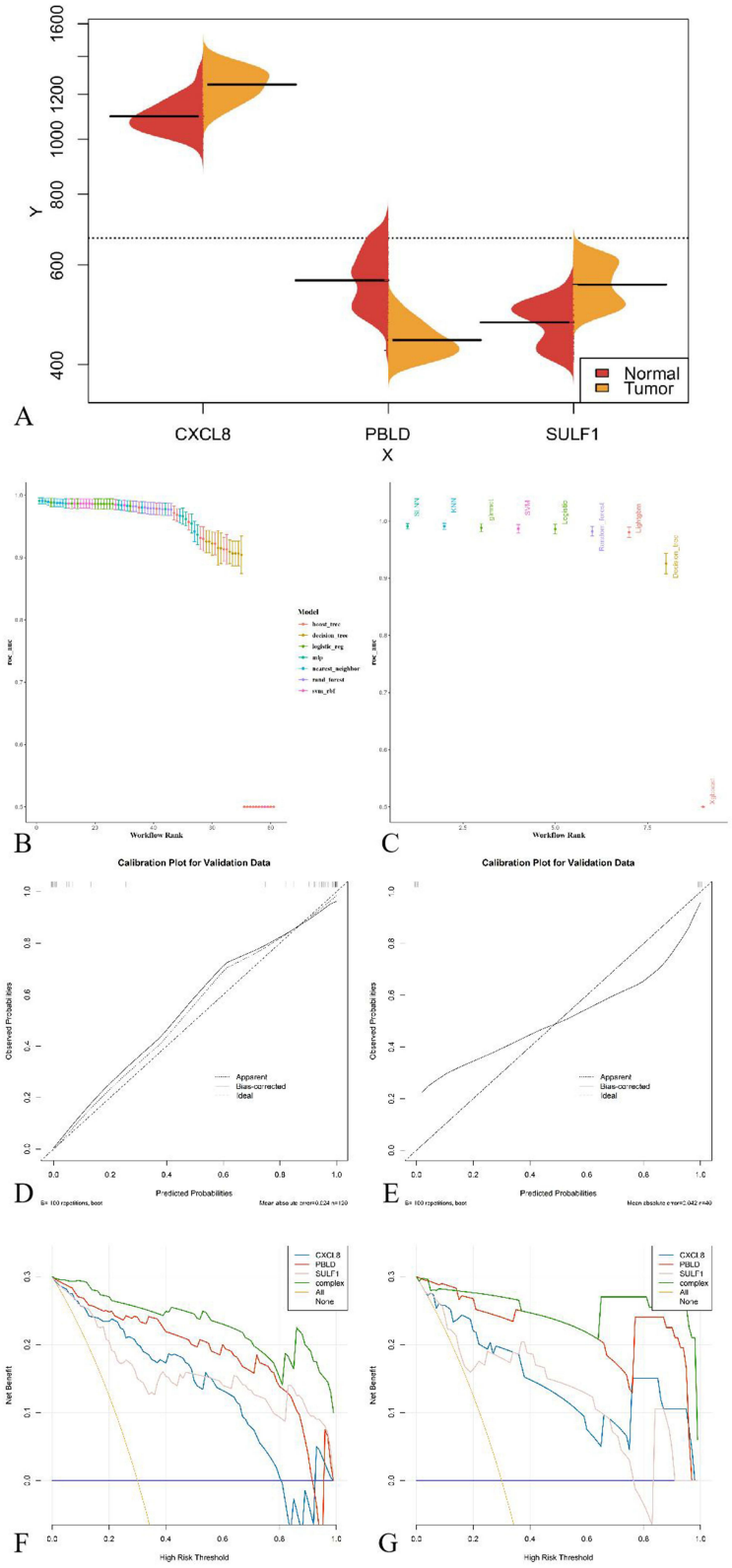

Figure 12 (**A**) Clinical data analysis. (**B**) AUC for all models. (**C**) AUC of the best model. (**D**) Calibration plot for training data. (**E**) Calibration plot for test data. (**F**) DCA for training data. (**G**) DCA for test data.

displays diminished calibration performance, thus providing further evidence of overfitting during training. Furthermore, the observed distributional differences, particularly for high-risk samples, are likely to exacerbate the calibration challenges in the testing set. In order to improve the model's generalisability and robustness, future studies should consider incorporating additional predictive features or leveraging ensemble learning methods, such as bagging or boosting. These strategies could enhance the model's calibration performance and resilience, particularly under conditions of varied data distributions.

DCA was performed in order to evaluate the clinical utility of the predictive model on both the training and validation sets. This method assesses the net clinical benefit across a range of threshold probabilities, providing insights into the model's effectiveness in guiding clinical decision-making. By quantifying the trade-off between true-positive and false-positive predictions at varying thresholds, the DCA helps determine the model's utility in practical applications and its potential to improve patient outcomes in diverse clinical scenarios. The DCA curves illustrate the combined predictive model, the treat-all strategy, and the treat-none strategy. In both the training and validation sets, the integrated prediction model based on CXCL8, PBLD, and SULF1 genes is superior to the treat-all and treat-none strategies, indicating that the model has high clinical utility within a specific threshold range. In particular, in the validation set, the model demonstrated significantly higher net gains than the treat-all and treat-none strategies, as well as the respective prediction strategies for CXCL8, PBLD, and SULF1.In particular, within the validation set, the integrated model demonstrates a net gain that is significantly higher than that of the single gene model and other decision strategies when the threshold probability is within the range of 0.18–1. It is noteworthy that there are discrepancies in the performance of the model between the training and validation sets. In the training set, the model demonstrates satisfactory generalisation ability and robustness within the threshold probability interval of 0.18–0.63. Nevertheless, when the threshold probability exceeds 0.63, the validation set curve is higher than that of the training set, which may be attributed to a slight overfitting issue of the model. Furthermore, the relatively limited sample size of the validation set may also result in calibration errors and fluctuations in model performance. Future studies should consider increasing the sample size further enhance the model's generalisation performance and stability. In conclusion, the comprehensive prediction model constructed based on CXCL8, PBLD and SULF1 genes demonstrated good clinical efficacy and showed high clinical net gain in the validation set, indicating the potential application of the model in CRC prediction.

However, this study has several limitations. First, although multiple datasets were obtained from the GEO database, the sample sources may be limited. Variations in region, ethnicity, and lifestyle habits across different datasets may affect the generalizability of the findings. In addition, the sample sizes of the datasets are relatively limited and may not fully represent the large and heterogeneous population of CRC patients. Second, although LRRC19 expression levels in tumor and adjacent tissues were inconsistent with predictions, no further mechanistic validation was performed. This may overlook the potential role of LRRC19 in CRC. Furthermore, although the machine learning model demonstrated high performance, the calibration curve showed slight elevation in the training set, suggesting possible underestimation and mild overfitting, which could affect the model's predictive accuracy. Although decision curve analysis was used to evaluate clinical utility in the test set, the assessment of net clinical benefit across different threshold probabilities may not be comprehensive enough to fully reflect real-world clinical scenarios. In summary, SULF1, CXCL8, and PBLD, as potentially powerful predictive targets for the progression of CRC, have broad application prospects in CRC diagnosis and drug development. In future studies, we will conduct an in-depth investigation of the role of these genes in the proliferation and metastasis of CRC and clarify their specific molecular mechanism. Regarding PBLD, we will also further clarify its impact on the progression of CRC and its molecular mechanism.The machine learning model, constructed based on SULF1, CXCL8, and PBLD, has been demonstrated to be an efficacious tool for the diagnosis of CRC in clinical decision-making contexts. This provides a new direction for the early and accurate diagnosis of CRC.

Conclusion

Through integrated analysis of multiple gene expression datasets, this study successfully identified and validated genes such as SULF1, CXCL8, and PBLD as discriminative biomarkers for early detection and prognosis of CRC. The machine learning model constructed based on these biomarkers demonstrated high efficacy in clinical diagnosis of CRC

and showed favorable utility in clinical decision-making. As biomarkers for CRC, SULF1, CXCL8, and PBLD hold significant potential for use in the prevention and diagnosis of CRC.

Data Sharing Statement

Data is available from the corresponding author Li Zhang on request.

Ethics Statement

The present study was approved by the Ethics Committee of The First Affiliated Hospital of Jinzhou Medical University (No.202127). All procedures were performed in accordance with the ethical standards of the Institutional Review Board and The Declaration of Helsinki, and its later amendments or comparable ethical standards.

Consent for Publication

Written informed consent for participants was not required for this study in accordance with national legislation and institutional requirements.

Funding

Science and Technology Research Joint Project of Liaoning Provincial Department (No.2023-MSLH-055). Scientific Research of The First Affiliated Hospital of Jinzhou Medical University (No.KYTD-2022006). The Undergraduate Innovation and Entrepreneurship Training Program Project of Liaoning Province (No.S202310160031).

Disclosure

The authors have no conflicts of interest to declare in this work.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(1):7–34. doi:10.3322/caac.21551
2. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(3):233–254. doi:10.3322/caac.21772
3. Van Cutsem E, Cervantes A, Adam R, et al. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann Oncol*. 2016;27(8):1386–1422. doi:10.1093/annonc/mdw235
4. Hossain MS, Karuniawati H, Jairoun AA, et al. Colorectal cancer: a review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers*. 2022;14(7):1732. doi:10.3390/cancers14071732
5. Hernandez Dominguez O, Yilmaz S, Steele SR. Stage IV colorectal cancer management and treatment. *J Clin Med*. 2023;12(5):2072. doi:10.3390/jcm12052072
6. Carlsen L, Huntington KE, El-Deiry WS. Immunotherapy for colorectal cancer: mechanisms and predictive biomarkers. *Cancers*. 2022;14(4):1028. doi:10.3390/cancers14041028
7. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(1):17–48. doi:10.3322/caac.21763
8. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
9. Chatila R, Mansour J, Mugharbil A, et al. Epidemiology and survival of colorectal cancer in lebanon: a sub-national retrospective analysis. *Cancer Control*. 2021;28:10732748211041221. doi:10.1177/10732748211041221
10. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin*. 2020;70(3):145–164. doi:10.3322/caac.21601
11. Du G, Lv H, Liang Y, et al. Population-based colorectal cancer risk prediction using a SHAP-enhanced LightGBM model. *Front Oncol*. 2025;15:1575844. doi:10.3389/fonc.2025.1575844
12. Wang R, Wang Q, Li P. Significance of carcinoembryonic antigen detection in the early diagnosis of colorectal cancer: a systematic review and meta-analysis. *World J Gastrointest Surg*. 2023;15(12):2907–2918. doi:10.4240/wjgs.v15.i12.2907
13. Lee JO, Kim M, Lee JH, et al. Carbohydrate antigen 19-9 plus carcinoembryonic antigen for prognosis in colorectal cancer: an observational study. *Colorectal Dis*. 2023;25(2):272–281. doi:10.1111/codi.16372
14. Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. *Nat Rev Clin Oncol*. 2018;15(3):183–192. doi:10.1038/nrclinonc.2017.186
15. Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med*. 2014;20(6):682–688. doi:10.1038/nm.3559
16. Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*. 2020;395(10221):350–360. doi:10.1016/S0140-6736(19)32998-8
17. Gubin MM, Zhang X, Schuster H, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*. 2014;515(7528):577–581. doi:10.1038/nature13988

18. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat Rev Immunol*. 2018;18(3):168–182. doi:10.1038/nri.2017.131
19. Shergalis A, Bankhead A, Luesakul U, Muangsin N, Neamati N. Current challenges and opportunities in treating glioblastoma. *Pharmacol Rev*. 2018;70(3):412–445. doi:10.1124/pr.117.014944
20. Ståhl PL, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016;353(6294):78–82. doi:10.1126/science.aaf2403
21. Wang J, Wu A, Yang B, Zhu X, Teng Y, Ai Z. Profiling and bioinformatics analyses reveal differential circular RNA expression in ovarian cancer. *Gene*. 2020;724:144150. doi:10.1016/j.gene.2019.144150
22. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*. 2023;186(8):1772–1791. doi:10.1016/j.cell.2023.01.035
23. Liao K, Yang Q, Xu Y, et al. Identification of signature of tumor-infiltrating CD8 T lymphocytes in prognosis and immunotherapy of colon cancer by machine learning. *Clin Immunol*. 2023;257:109811. doi:10.1016/j.clim.2023.109811
24. Lee H, Lee EJ, Ham S, et al. Machine learning approach to identify stroke within 4.5 hours. *Stroke*. 2020;51(3):860–866. doi:10.1161/STROKEAHA.119.027611
25. Wang X, Xu Y, Sun R, Wang S, Wei X. Multi-omics based consensus subtypes, development of prognostic signature, and identification of INHBB as a potential therapeutic target in colorectal cancer. *Funct Integr Genomics*. 2025;25(1):198. doi:10.1007/s10142-025-01691-1
26. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res*. 2007;13(4):1107–1114. doi:10.1158/1078-0432.CCR-06-1633
27. Vlachavas EI, Pilalis E, Papadodima O, et al. Radiogenomic analysis of f-18-fluorodeoxyglucose positron emission tomography and gene expression data elucidates the epidemiological complexity of colorectal cancer landscape. *Comput Struct Biotechnol J*. 2019;17:177–185. doi:10.1016/j.csbj.2019.01.007
28. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607–d13. doi:10.1093/nar/gky1131
29. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(D1):D991–5. doi:10.1093/nar/gks1193
30. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;45(W1):W98–w102. doi:10.1093/nar/gkx247
31. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res*. 2019;47(W1):W556–w60. doi:10.1093/nar/gkz430
32. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol*. 2019;20(1):185. doi:10.1186/s13059-019-1758-4
33. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*. 2001;25(4):402–408. doi:10.1006/meth.2001.1262
34. Lai JP, Sandhu DS, Moser CD, et al. Additive effect of apicidin and doxorubicin in sulfatase 1 expressing hepatocellular carcinoma in vitro and in vivo. *J Hepatol*. 2009;50(6):1112–1121. doi:10.1016/j.jhep.2008.12.031
35. Lai JP, Yu C, Moser CD, et al. SULF1 inhibits tumor growth and potentiates the effects of histone deacetylase inhibitors in hepatocellular carcinoma. *Gastroenterology*. 2006;130(7):2130–2144. doi:10.1053/j.gastro.2006.02.056
36. Dhanasekaran R, Nakamura I, Hu C, et al. Activation of the transforming growth factor-β/SMAD transcriptional pathway underlies a novel tumor-promoting role of sulfatase 1 in hepatocellular carcinoma. *Hepatology*. 2015;61(4):1269–1283. doi:10.1002/hep.27658
37. Wu J, Subbaiah KCV, Xie LH, et al. Glutamyl-prolyl-tRNA synthetase regulates proline-rich pro-fibrotic protein synthesis during cardiac fibrosis. *Circ Res*. 2020;127(6):827–846. doi:10.1161/CIRCRESAHA.119.315999
38. Babel I, Barderas R, Diaz-Uriarte R, et al. Identification of MST1/STK4 and SULF1 proteins as autoantibody targets for the diagnosis of colorectal cancer by using phage microarrays. *Mol Cell Proteomics*. 2011;10(3):M110.001784. doi:10.1074/mcp.M110.001784
39. Tang CP, Zhou HJ, Qin J, Luo Y, Zhang T. MicroRNA-520c-3p negatively regulates EMT by targeting IL-8 to suppress the invasion and migration of breast cancer. *Oncol Rep*. 2017;38(5):3144–3152. doi:10.3892/or.2017.5968
40. Matsuo Y, Ochi N, Sawai H, et al. CXCL8/IL-8 and CXCL12/SDF-1alpha co-operatively promote invasiveness and angiogenesis in pancreatic cancer. *Int J Cancer*. 2009;124(4):853–861. doi:10.1002/ijc.24040
41. Pączek S, Łukaszewicz-Zając M, Gryko M, Mroczko P, Kulczyńska-Przybik A, Mroczko B. CXCL-8 in preoperative colorectal cancer patients: significance for diagnosis and cancer progression. *Int J Mol Sci*. 2020;22(1):21. doi:10.3390/ijms22010021
42. Oladipo O, Conlon S, O'Grady A, et al. The expression and prognostic impact of CXC-chemokines in stage II and III colorectal cancer epithelial and stromal tissue. *Br J Cancer*. 2011;104(3):480–487. doi:10.1038/sj.bjc.6606055
43. Li J, Liu Q, Huang X, et al. Transcriptional profiling reveals the regulatory role of CXCL8 in promoting colorectal cancer. *Front Genet*. 2019;10:1360. doi:10.3389/fgene.2019.01360
44. Cheng XS, Li YF, Tan J, et al. CCL20 and CXCL8 synergize to promote progression and poor survival outcome in patients with colorectal cancer by collaborative induction of the epithelial-mesenchymal transition. *Cancer Lett*. 2014;348(1–2):77–87. doi:10.1016/j.canlet.2014.03.008
45. Xiao YC, Yang ZB, Cheng XS, et al. CXCL8, overexpressed in colorectal cancer, enhances the resistance of colorectal cancer cells to anoikis. *Cancer Lett*. 2015;361(1):22–32. doi:10.1016/j.canlet.2015.02.021
46. Zhang J, Kang B, Tan X, et al. Comparative analysis of the protein profiles from primary gastric tumors and their adjacent regions: MAWBP could be a new protein candidate involved in gastric cancer. *J Proteome Res*. 2007;6(11):4423–4432. doi:10.1021/pr0703425
47. Li DM, Zhang J, Li WM, et al. MAWBP and MAWD inhibit proliferation and invasion in gastric cancer. *World J Gastroenterol*. 2013;19(18):2781–2792. doi:10.3748/wjg.v19.i18.2781
48. Liang Y, Song X, Li Y, et al. circKDM4C suppresses tumor progression and attenuates doxorubicin resistance by regulating miR-548p/PBLD axis in breast cancer. *Oncogene*. 2019;38(42):6850–6866. doi:10.1038/s41388-019-0926-z