Computers in Biology and Medicine 145 (2022) 105409

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis

Ying Su [a], Xuecong Tian [a], Rui Gao [c], Wenjia Guo [b], Cheng Chen [a,*], Chen Chen [c,d], Dongfang Jia [a], Hongtao Li [b], Xiaoyi Lv [a,e,**]

[a] College of Software, Xinjiang University, Urumqi, 830046, Xinjiang, China
[b] Affiliated Tumor Hospital of Xinjiang Medical University, Urumqi, 830011, China
[c] College of Information Science and Engineering, Xinjiang University, Urumqi, 830046, China
[d] Cloud Computing Engineering Technology Research Center of Xinjiang, Kelamayi, 834099, China
[e] Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi, 830046, Xinjiang, China

ARTICLE INFO

Keywords:
Machine learning
Colon cancer
Prognosis
WGCNA
Staging
PPI

ABSTRACT

Advanced metastasis of colon cancer makes it more difficult to treat colon cancer. Finding the markers of colon cancer (Colon Cancer) can diagnose the stage of cancer in time and improve the prognosis with timely treatment. This paper uses gene expression profiling data from The Cancer Genome Atlas (TCGA) for the diagnosis of colon cancer and its staging. In this study, we first selected the gene modules with the greatest correlation with cancer by Weighted Gene Co-expression Network Analysis (WGCNA), extracted the characteristic genes for differential expression results using the least absolute shrinkage and selection operator algorithm (Lasso) and performed survival analysis, and then combined the genes in the modules with the Lasso-extracted feature genes were combined to diagnose colon cancer versus healthy controls using RF, SVM and decision trees, and colon cancer staging was diagnosed using differentially expressed genes for each stage. Finally, Protein-Protein Interaction Networks (PPI) networks were done for 289 genes to identify clusters of aggregated proteins for survival analysis. Finally, the RF model had the best results in the diagnosis of colon cancer versus control group fold cross-validation with an average accuracy of 99.81%, F1 value reaching 0.9968, accuracy of 99.88%, and recall of 99.5%, and an average accuracy of 91.5%, F1 value reaching 0.7679, accuracy of 86.94%, and recall in the diagnosis of colon cancer stages I, II, III and IV. The recall rate reached 73.04%, and eight genes associated with colon cancer prognosis were identified for GCNT2, GLDN, SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5 and CPT2.

1. Introduction

Colorectal cancer is the third most common cause of cancer death for both men and women in the United States, and the second most common when men and women are combined [1]. According to the American Society of Clinical Oncology, an estimated 149,500 adults in the United States will be diagnosed with colorectal cancer in 2021. This includes 104,270 new cases of colon cancer and 45,230 new cases of rectal cancer, which accounts for about 70% of colorectal cancers [2]. An important test for CRC screening is the fecal occult blood test (FOBT), which includes a guaiac FOBT and a fecal immunology test [3]. Colonoscopy is considered the gold standard method for CRC screening, with the advantages of high sensitivity, specificity and direct visualization, and it plays an important role in cancer and precancerous lesions (biopsy and removal of polyps) [4]. This specifically includes endorectal ultrasonography (USG), abdominal ultrasonography (USG), computed tomography (CT) and nuclear magnetic resonance (NMR). However, these methods are only effective in the case of severe focal lesions [5].

In recent years, tumor markers have been widely used in the field of cancer diagnosis and treatment. The ideal tumor marker should have strong specificity for tumor screening, diagnosis, efficacy and prognosis assessment, recurrence detection, etc., and can detect microscopic lesions and quantitatively reflect tumor load [6]. Cancer staging is used to describe how far the cancer has spread in the body. It helps determine

* Corresponding author.
** Corresponding author. College of Software, Xinjiang University, Urumqi, 830046, Xinjiang, China.
E-mail addresses: chenchengoptics@gmail.com (C. Chen), xjuwawj01@163.com (X. Lv).

https://doi.org/10.1016/j.compbiomed.2022.105409
Received 8 January 2022; Received in revised form 20 February 2022; Accepted 12 March 2022
Available online 19 March 2022

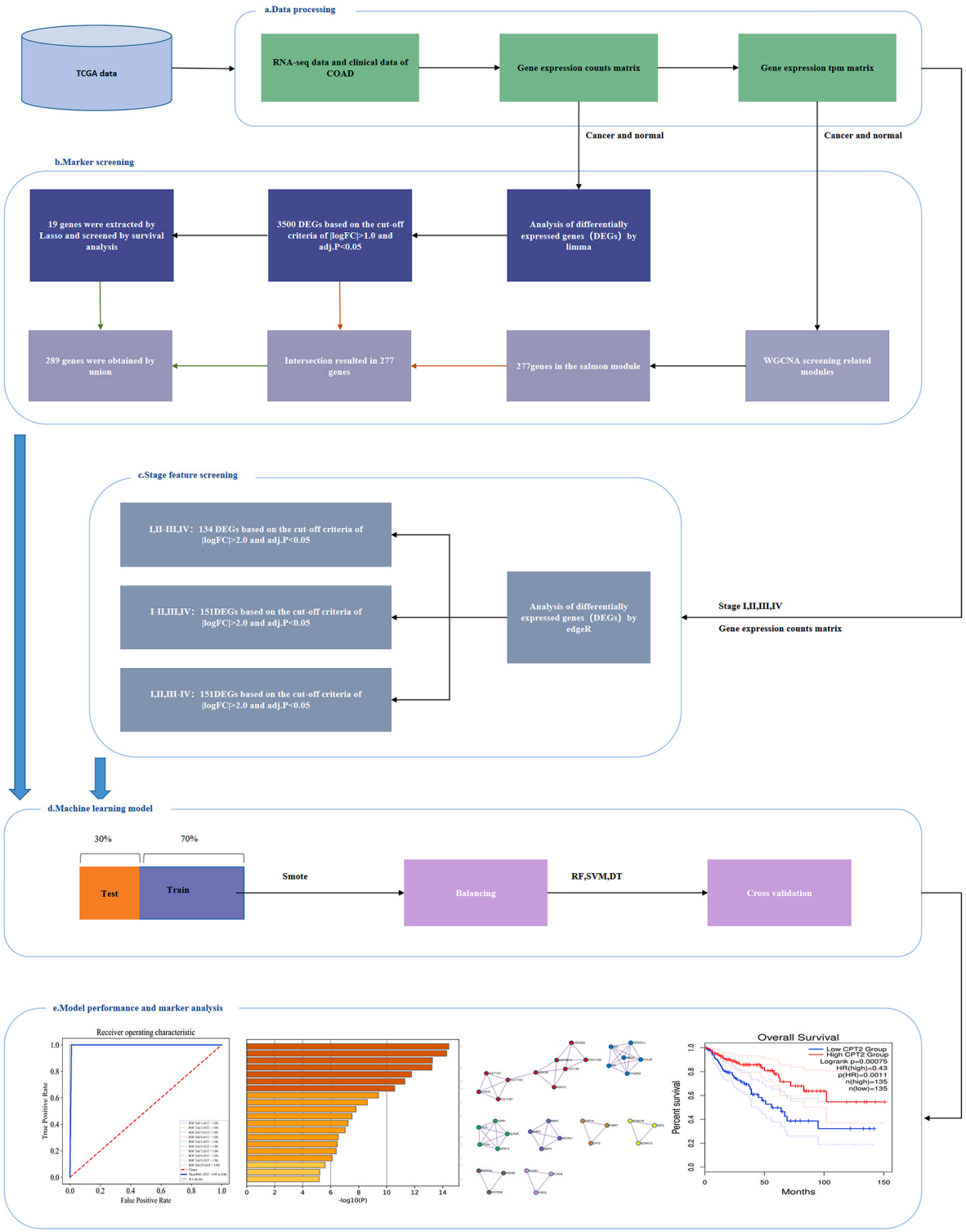

Fig. 1. Workflow diagram.

the severity of the cancer and how best to treat it, and is also used by doctors in survival statistics [7]. According to the American Joint Committee on Cancer Tumor-Lymph Node-Metastasis (TNM), most cancers have five distinct stages - Stage 0, I, II, III and IV [8]. The stage of the cancer will tell us the location and size of the cancer, how much it has grown in nearby tissues, and whether it has spread to nearby lymph nodes or other parts of the body, and will influence the presence of markers of cancer spread [9]. Among patients with the worst grade of colon cancer, if diagnosed at stage I, the 5-year survival rate for those aged 18 and 65 years is 91% and survivable with appropriate treatment, while the 5-year survival rates at stages II, III and IV are about 82%, 66% and 10%, respectively [10]. Colon cancer mutations are the most common and deadly of many cancers, and detecting the disease in its early stages greatly increases a patient's chances of survival [11].

Machine learning can detect hard-to-identify patterns from large, noisy or complex datasets. This capability is particularly well suited for applications in data analysis in the medical field, especially those involving complex proteomic and genomic expression data related

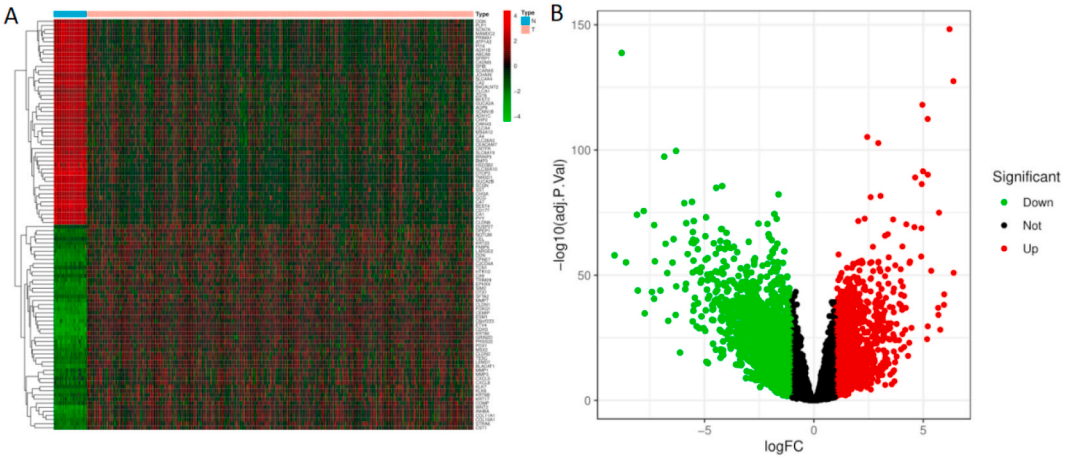

Fig. 2. Graph of differentially expressed gene results. (A) Heat map of differentially expressed genes. Each column of the heat map represents the sample, each row represents the gene, red indicates high expression, green indicates low expression, and the color shade represents the gene expression level. (B) Differential expression volcano map. The rows of the volcano plot represent log | FC | and the columns represent -log10 (adjusted P value).

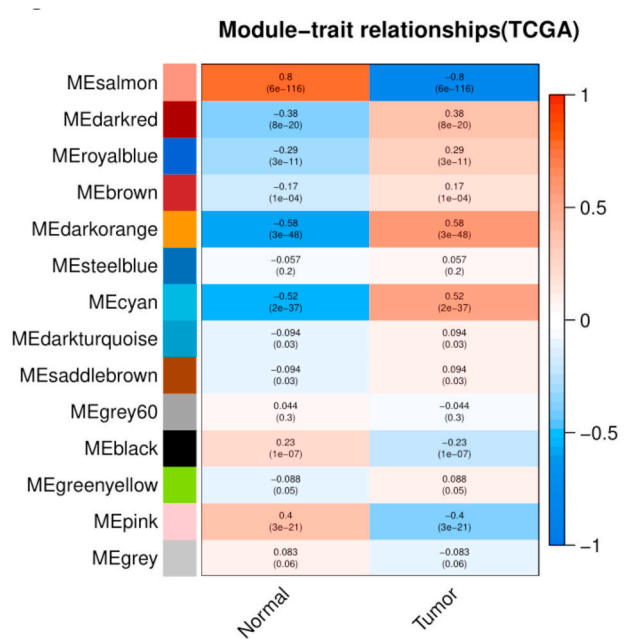

Fig. 3. Colon cancer genes and clinical correlation. Each row represents a gene module and each column represents a clinical feature. Each module contains the correlation coefficient and P-value between the gene module and the clinic.

applications, often used in recent years for cancer diagnosis and detection [12]. Machine learning algorithms have been widely used in the medical field including SVM [13–15], random forests [16] and decision trees [17,18]. Specific applications such as Xie et al. used spectral data to build SVM models for rapid and noninvasive screening of keratitis with an average accuracy of 100% [19] and Chen Fangfang et al. used SVM and DT models for rapid detection of glioma with model accuracy around 90% [20], which further indicates that machine learning has better applicability in medical disease diagnosis. SVM is used as a supervised learning model for classification and regression problems, and can solve both linear and nonlinear problems. A random forest classifier is a set of decision trees from a randomly selected subset of the training set that aggregates votes from different decision trees to determine the final class of the test object. Decision trees are popular machine learning models for classification and regression tasks.

Gene network-based cancer predictive biomarker screening has yielded some good results in the biomedical field, such as the subtype-specific network biomarker approach to identify breast cancer survivorship constructed by Sheikh Jubair et al. which has high predictive performance in identifying breast cancer patients' survivorship [21]. Shiyan Li et al. developed a model to assess the prognosis of cervical cancer patients using the weighted gene co-expression network (WGCNA) combined with the LASSO approach and demonstrated that the model is valid and stable [22]. In this study, we first did differential expression analysis between colon cancer and healthy controls, and used WGCNA to correlate healthy and cancer samples to obtain gene modules associated with cancer, and combined the features extracted by LASSO machine algorithm to diagnose colon cancer. Weighted Gene Co-expression Network Analysis (WGCNA) has been applied to the analysis of various cancers [23,24], such as: bladder cancer [25], breast cancer [26], and lung cancer [27], and can help identify the underlying mechanisms involved in specific biological processes as well as explore candidate biomarkers. The LASSO feature selection technique has been used in many applications in the biological field. Lasso is a well-known feature selection method that considers an L1 type penalty, which adds a constraint on the sum of all absolute values of the feature coefficients to ensure both global optimality and computational efficiency [28]. The study of feature selection methods for microbial and microbiome data at the 2021 IEEE International Conference on Bioinformatics and Biomedicine found that LASSO consistently outperformed other methods in several key classification metrics, most notably AUC, and that the LASSO framework can generate more meaningful feature selection algorithms relative to similar feature selection methods for features [29]. In addition, the Lasso algorithm was well applied to the field of cancer by Neha Shree Maurya et al. who used LASSO and other

Table 1
Number of genes contained in the 14 modules.

Black	Brown	Cyan	Darkorange	Darkred	Darktuquoise	Greenyellow
831	1542	350	288	309	4365	508
Grey	**Grey60**	**Pink**	**Royalblue**	**Saddlebrown**	**Salmon**	**Steelblue**
2482	288	1012	158	79	277	58

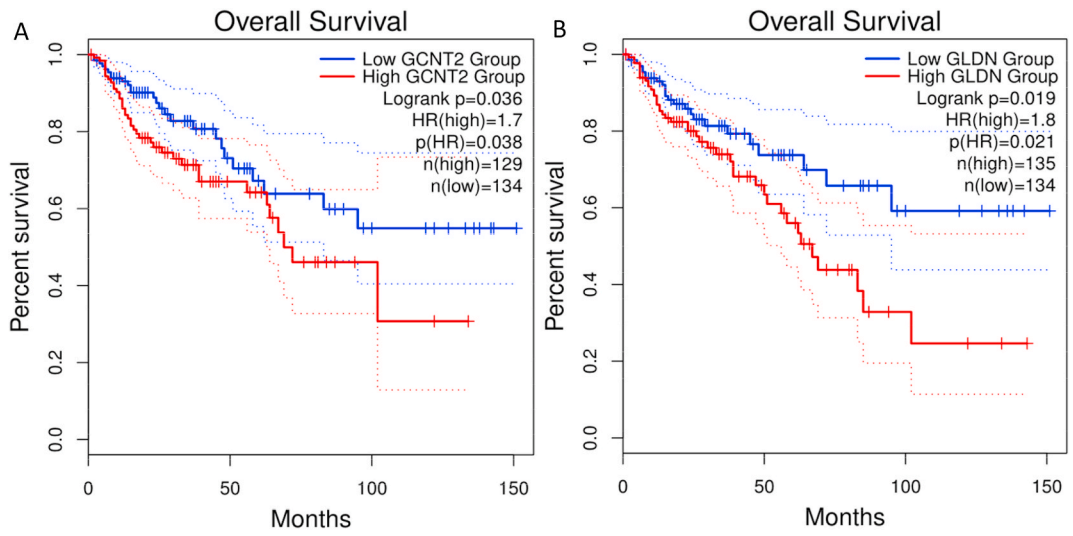

Fig. 4. Relationship between GCNT2 (A) and GLDN (B) genes and overall survival of colon cancer patients. The horizontal axis indicates the survival time and the vertical axis indicates the survival rate.

Table 2
Full names of the 19 genes screened.

Full Name	Abbreviation
C-Type Lectin Domain Family 3 Member B	CLEC3B
Transmembrane And Immunoglobulin Domain Containing 1	TMIGD1
Proteolipid Protein 1	PLP1
Glucosaminyl (N-Acetyl) Transferase 2 (I Blood Group)	GCNT2
MAM Domain Containing 2	MAMDC2
Glucagon Like Peptide 2 Receptor	GLP2R
Lymphatic Vessel Endothelial Hyaluronan Receptor 1	LYVE1
Transmembrane Protein 100	TMEM100
Scavenger Receptor Class A Member 5	SCARA5
Carbonic Anhydrase 1	CA1
Carbonic Anhydrase 2	CA2
CD177 Molecule	CD177
Alcohol Dehydrogenase 1B (Class I), Beta Polypeptide	ADH1B
Gremlin 2, DAN Family BMP Antagonist	GREM2
Membrane Spanning 4-Domains A12	MS4A12
Joining Chain Of Multimeric IgA And IgM	JCHAIN
UDP-GlcNAc:BetaGal Beta-1,3-N-Acetylglucosaminyltransferase 7	B3GNT7
Solute Carrier Family 16 Member 9	SLC16A9
Gliomedin	GLDN

methods to extract signature genes to discover TMEM236, a novel biomarker for the diagnosis of colorectal cancer [30]. Secondly, differential expression analysis was done by dividing the first three stages of colon cancer into two groups according to early stage cancer versus advanced or metastatic cancer [31] and stage I versus the last three stages, respectively. Colon cancer was classified into two groups, I, II and III, IV, according to whether the cancer had spread to nearby lymph nodes or elsewhere, and differential expression analysis was performed for the staging groups. Finally, colon cancer and colon cancer staging were classified using machine learning SVM, random forest and decision tree, and the final feature genes used for classification were screened for prognostic genes after Protein-Protein Interaction Networks (PPI) analysis. Here, the results of the model constructed by our extracted features achieved better results for colon cancer and its early and late diagnosis, and the screened prognostic genes provided more ideas for the treatment of colon cancer.

Table 3
Results of each model for the diagnosis of colon cancer.

Model	Accuracy	Precision	Recall	F1	Specificity	ROC_auc
SVM	98.46%	94.50%	96.92%	95.29%	98.94%	99.67%
RF	99.81%	99.88%	99.50%	99.68%	100%	100%
DT	99.62%	99.06%	98.64%	98.73%	99.79%	98.64%

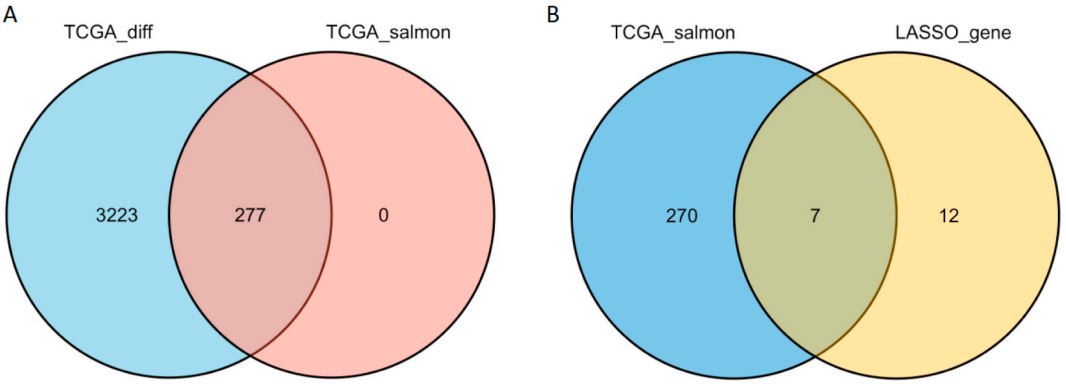

Fig. 5. Venn diagram. (A) Venn diagram of differential genes with Mesalmon gene module, taken as intersection. (B) Venn diagram of Mesalmon gene module with machine learning screening features, taking the merged set.

Table 4
Results of each model for the diagnosis of colon cancer staging.

Stage	Model	Accuracy	Precision	Recall	F1	Specificity	ROC_auc
I-II,III,IV	SVM	63.88%	55.24%	58.40%	53.56%	50.10%	62.60%
	RF	85.25%	75.24%	72.64%	73.46%	52.74%	80.38%
	DT	80.63%	68.02%	73.33%	69.42%	60.89%	73.33%
I,II-III,IV	SVM	65.95%	64.92%	63.53%	63.02%	78.72%	65.41%
	RF	79.71%	79.75%	78.54%	78.15%	87.30%	86.35%
	DT	73.89%	72.73%	72.98%	72.57%	75.96%	72.98%
I,II,III-IV	SVM	71.11%	48.12%	46.88%	46.98%	79.75%	49.37%
	RF	91.52%	86.94%	73.04%	76.79%	98.16%	82.37%
	DT	86.71%	72.04%	76.52%	73.02%	90.78%	76.52%

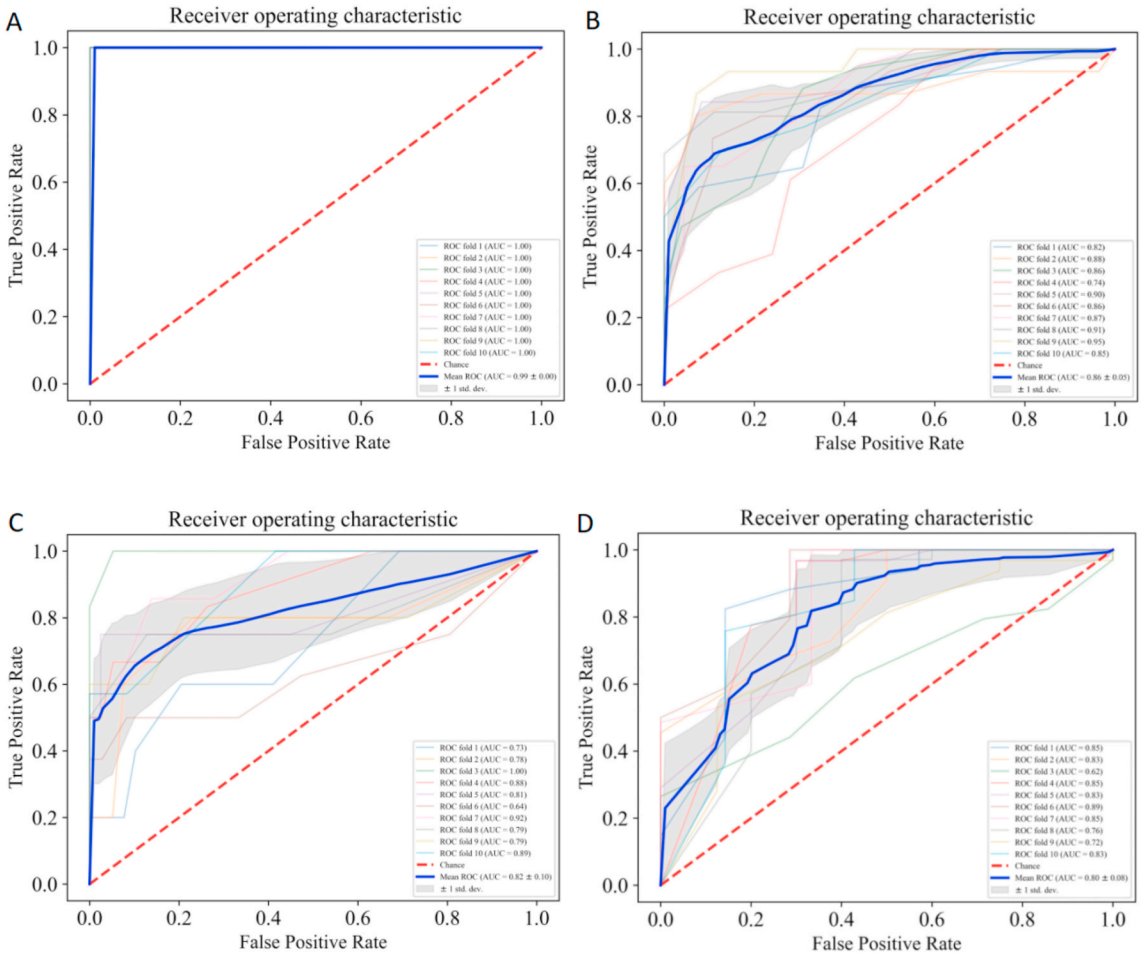

Fig. 6. ROC curves of random forest (RF) after crossover in each diagnosis. (A) ROC curves of RF in the diagnosis of colon cancer. (B) ROC curves of RF in the diagnosis of colon cancer stage I with stage II, III, IV. (C) ROC curves of RF in colon cancer stages I and II versus III and IV. (D) Diagnostic ROC curves of RF in colon cancer stages I, II, III and IV.

2. Materials and methods

2.1. Experimental method

The methods used in this experiment are all illustrated by the workflow diagram 1 below (Fig. 1).

2.2. Materials and data

Colon cancer (COAD) gene expression data were obtained from TCGA (https://portal.gdc.cancer.gov/repository). Transcriptomic gene expression data from the TCGA-COAD project on the TCGA official website were selected and 521 samples were downloaded, including 480 tumor tissue samples and 41 corresponding control tissue samples. In this trial, clinical data of 459 colon cancer patients were downloaded and key survival data and staging information were extracted. The collated COAD series matrices were annotated using the gene annotation file on the Ensemble (http://asia.ensembl.org) website to obtain gene expression matrices. Using the R language edgeR [32] package, the expression matrix was filtered into a matrix of tpm values according to the tpm transformation formula, resulting in an expression matrix of 13, 970 genes with 521 samples.

2.3. Differential expression analysis

Differential expression analysis (DEG) of colon carcinoma tissues versus control tissues was done using the R language limma [33] package with settings $|\log FC|>1.0$ and adj.P.Val<0.05 to screen for

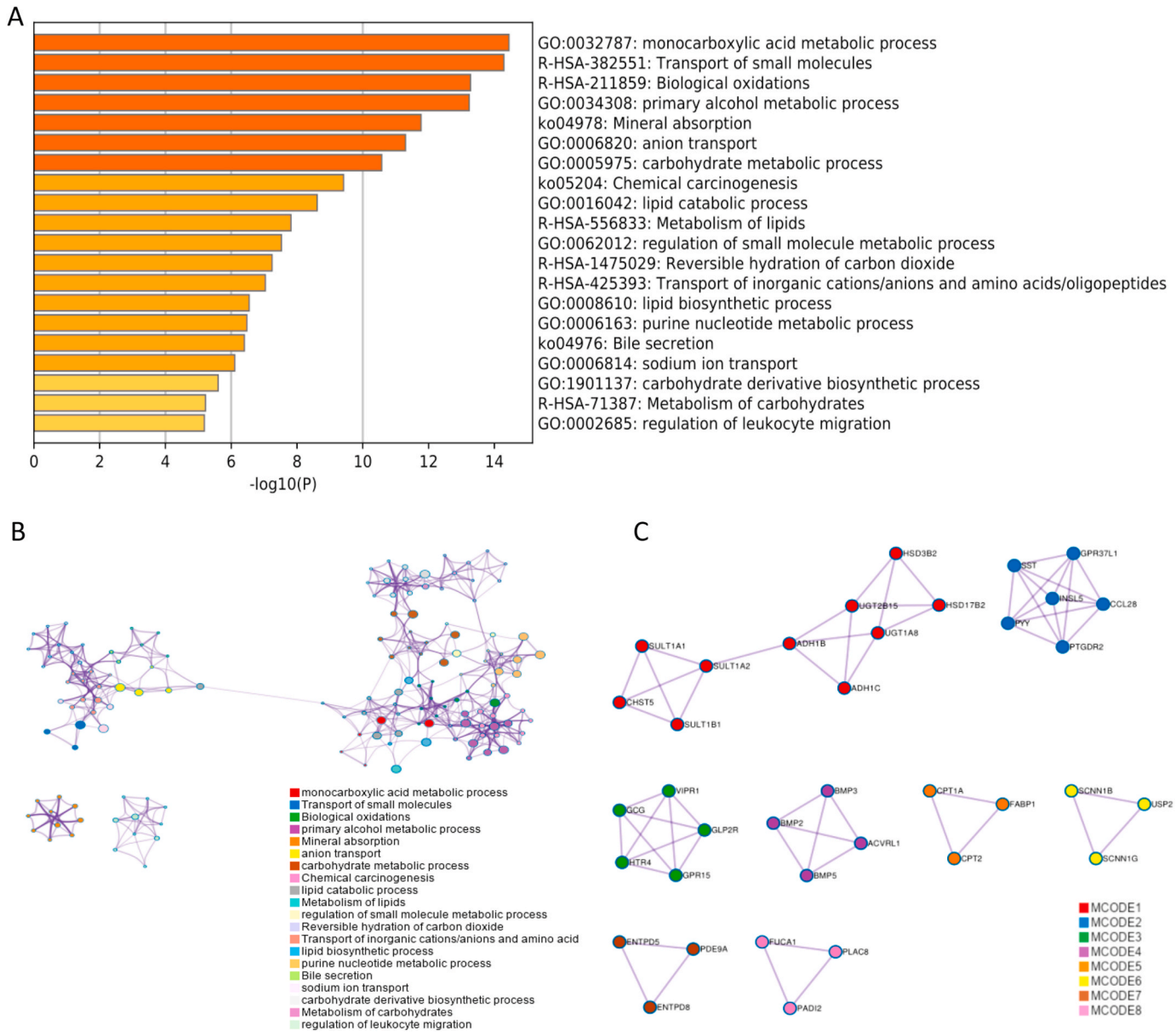

Fig. 7. Gene list analysis. (A) Graph of enrichment analysis of genes. Rows represent enriched pathways and columns represent -log10(P) values. (B) Network diagram of the relationship between enriched biological pathways, different colors represent different biological pathways. (C) Densely connected protein clusters searched by MCODE algorithm, different colors represent different protein clusters.

differentially expressed genes and generate a differential expression matrix. Heat maps of gene expression profiles were plotted by the R package ggplot2 and the differential expression results were visualized as volcano plots. For differential expression between cancer tissues of samples from the first two stages and the second two stages of colon cancer (of which there were 294 samples from I and II and 192 samples from III and IV), from the first three stages and IV (of which there were 434 samples from I, II and III and 64 samples from IV) and from stage I and the second three stages (of which there were 85 samples from I and 368 samples from II, III and IV) For analysis, in order to screen out significant differential genes as much as possible, the experiments were performed using the edgeR package, setting the difference multiplier $|\log FC|>2.0$, adj.P.Val<0.05.

2.4. Weighted gene Co-expression network analysis

Gene expression matrices with tpm values were used as input data for WGCNA, and genes with small fluctuations in all samples were removed for gene co-expression network analysis of the TCGA-COAD dataset using the WGCNA package in R. To explore the genetic modules affecting the association between colon cancer and healthy samples, we performed a correlation analysis using cancer and healthy clinical traits.

2.5. LASSO extraction of feature genes

The expression matrix generated from the differential expression of colon cancer tissues and healthy samples was used as input to set up a random seed to screen for signature genes using the R language glmnet package. The relationship between Lasso-extracted signature genes and prognosis was analyzed by the GEPIA2 (http://gepia2.cancer-pku.cn/#index) online analysis tool.

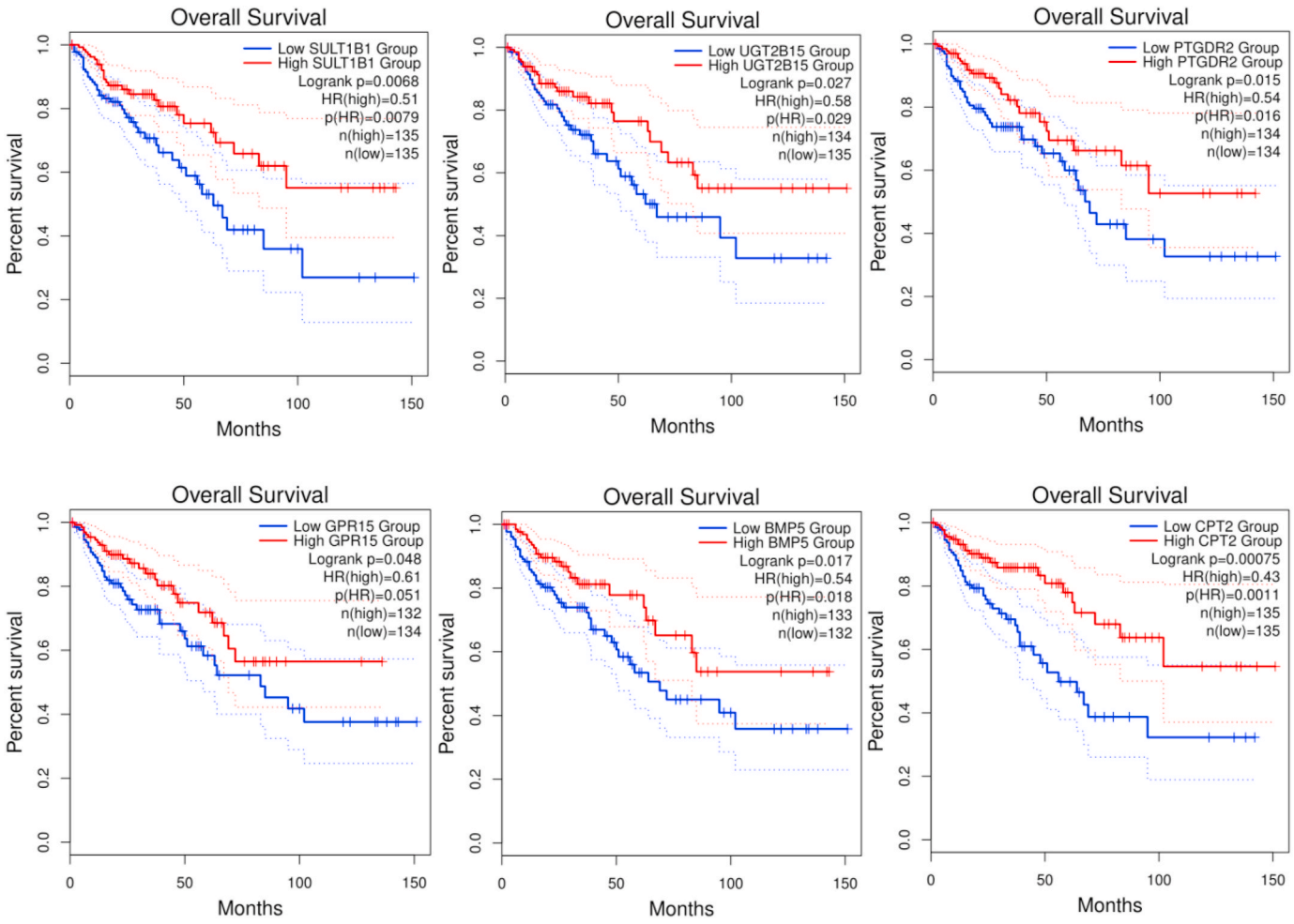

Fig. 8. Association of SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5 and CPT2 genes with overall survival of patients with colon cancer.

2.6. Screening of characteristic genes

The most correlated module Mesalmon analyzed using WGCNA was intersected with the genes from differential expression analysis, and these genes were merged with the feature genes extracted by LASSO in order to combine the advantages of machine learning to extract features and the credibility of raw letter analysis.

2.7. Building the model

For the diagnosis of colon cancer and health, 521 samples from the TCGA-COAD project, including 480 tumor tissue samples and 41 corresponding control tissue samples, were used in this experiment, and 289 genes after feature extraction were selected as classification features for diagnosis. A total of 486 samples were used for the diagnosis of first two and second two stages of colon cancer, including 294 samples in early stage and 192 samples in late stage, and 134 differentially expressed genes were obtained as classification features using differential analysis; 498 samples were used for the diagnosis of first three and stage IV of colon cancer, including 434 samples in first three stages and 64 samples in late stage, and 151 differentially expressed genes were obtained as classification features using differential analysis; 453 samples were used for the diagnosis of stage I and stage III of colon cancer, including 85 samples in early stage and 368 samples in late stage. A total of 453 samples were used for stage I and post-stage III diagnosis of colon cancer, including 85 samples in early stage and 368 samples in post-stage III, and 189 differentially expressed genes were obtained as classification features using differential expression analysis.

This experiment uses Support Vector Machine (SVM) with rbf kernel function, Random Forest (RF) and Decision Tree as classifiers. All models first normalized the data of colon cancer diagnosis and staging first, and then randomly divided into training set and test set, where 70% of the training set and 30% of the test set. To balance the sample inhomogeneity, we did sample balancing for the training set only using smote and tested the trained model with data from the test set. Finally, cross-cross validation was used and the AUC (area enclosed with the coordinate axis under the ROC curve) results averaged over ten times were taken as the final measure.

2.8. Enrichment analysis, PPI analysis and survival analysis

The 289 genes screened were analyzed enriched by the online tool Metascape (https://metascape.org/gp/index.html#/main/step1) and a protein interaction network (PPI) between genes was established, and the PPI analysis was followed by the use of the Molecular Complexation Detection (MCODE) algorithm10 to identify densely connected network components, after which the genes contained in the network components were subjected to survival analysis to find genes in major pathways that affect colon cancer survival. Survival analysis of genes in the obtained protein clusters was performed using GEPIA2 (http://gepia2.cancer-pku.cn/#index).

3. Result

3.1. Identification of differentially expressed genes

The differential expression analysis of colon carcinoma tissues versus control tissues finally yielded 3500 differentially expressed genes, whose heat map (Fig. 2A) and volcano map (Fig. 2B) are shown below. Differential expression analysis between colon cancer I, II and III, IV cancer tissues and healthy samples yielded 134 differentially expressed genes, 151 differentially expressed genes between first three and IV stage samples, and 189 differentially expressed genes between stage I and last three stage samples.

3.2. Key gene modules identified by WGCNA

A total of 14 gene modules were obtained by WGCNA analysis, and the correlation heat map is shown in Fig. 3, and the number of genes in each module is shown in Table 1. Among all these modules Mesalmon, Medarkorange and Mestelblue correlations of $|P|>0.5$, the gene module correlation plot with clinical traits showed that Mesalmon had the highest correlation of 0.8 with health traits, and the module contained 277 genes in total.

3.3. 19 signature genes extracted by LASSO

Lasso algorithm extracted a total of 20 feature genes, and we found that PLAAT2 gene could not be retrieved during survival analysis, and we excluded it, meanwhile we found that only GCNT2 and GLDN were associated with colon cancer prognosis among 19 genes, and the survival results are shown in Fig. 4. The final 19 gene names are shown specifically in Table 2.

3.4. Characteristic genes for classification

The Mesalmon gene module of WGCNA was intersected with the genes taken from differential expression analysis to obtain a total of 277 genes, and the results were consistent with the genes included in Mesalmon (Fig. 5A). The Mesalmon gene module and the feature genes extracted from LASSO were taken and the results were obtained for a total of 289 genes (Fig. 5B).

3.5. Results of different machine learning classification models

The results showed that for the diagnosis of colon cancer all three models showed better results, with the random forest (RF) model performing the best, with more than 99% of all assessment metrics of the model (Table 3). For the diagnosis of colon cancer staging, the three models were the best in the classification of the first three stages of I, II, and III with stage IV. Again, the random forest (RF) model performed the best among the three staging models, with an accuracy of 79.71% in the classification of early and late stages, 91.52% in the classification of the first three stages with stage IV, and 85.25% in the classification of stage I with the latter three stages. For the diagnosis of colon cancer (Table 2) and the diagnosis of colon cancer stage (Table 4) are shown in the following table, respectively. The ROC curve of the RF model is shown in Fig. 6.

3.6. Enrichment analysis, PPI analysis and survival analysis results

Metascape online tool analysis showed that the GO biological processes of these 289 genes involve monocarboxylic acid metabolic processes, primary alcohol metabolic processes, anion transport, carbohydrate metabolic processes, lipid catabolic processes, regulation of small molecule metabolic processes, etc. The reaction group genome involves small molecule transport, biological oxidation, lipid metabolism, reversible hydration of carbon dioxide, etc. The KEGG pathway includes mineral uptake, chemical carcinogenesis, and bile secretion processes. The enrichment is shown in Fig. 7A, and the enrichment pathway network diagram is shown in Fig. 7B. The network of protein interaction components was obtained for a total of 8 protein groups and 37 genes, and the network diagram is shown in Fig. 7C. Survival analysis revealed that only 6 of the 37 genes were associated with the prognosis of colon cancer, namely SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5 and CPT2 genes, and their survival curves are shown in Fig. 8.

4. Discussion

Colon cancer is considered the third leading cause of death in women and the second leading cause of death in men worldwide. Colon cancer is poorly treated after carcinogenesis and metastasis, and early diagnosis of colon cancer is beneficial in improving the survival rate of patients compared to those diagnosed at a later stage, and with certain treatments lower morbidity and better survival can be achieved [34]. Many European countries mainly perform interval colonoscopy or fecal occult blood testing to screen for colon cancer, but colonoscopy relies on the experience of clinicians and fecal occult blood testing is costly both have limitations [35]. To solve these problems and diagnose colon cancer and its staging in a timely manner, we established a machine learning diagnostic model, which was trained and tested after analysis and screening of 289 signature genes obtained from the expression profile data of cancer and control tissues, and the classification accuracy reached 99.81%.Meanwhile, we constructed a staging model for colon cancer using the differential genes between stages, and achieved an accuracy rate of 91.52% in the diagnosis of the first three stages with stage IV staging. This better diagnostic result may be related to the fact that machine learning itself can spontaneously acquire more laws through learning and continuously improve its own performance in the process of learning, for example, Ying Xie et al. used metabolomics and machine learning methods to screen early lung cancer diagnostic biomarkers with diagnostic accuracy as high as 98.9% [36], which fully demonstrates the potential of machine learning in potential of the application of machine learning in cancer classification. The results of this experiment provide a favorable basis for the in-depth application of artificial intelligence methods in the diagnosis of medical cancers and provide a basis for the classification of early and late stage of cancers.

To investigate whether the genes screened by machine learning and raw letter analysis are more meaningful, we enriched 12 genes CLEC3B, PLP1, GCNT2, MAMDC2, GLP2R, LYVE1, TMEM100, SCARA5, ADH1B, GREM2, JCHAIN and GLDN that overlap between the Lasso algorithm and raw letter analysis. The analysis showed that these genes are focused on the cellular response to growth factor stimulation and the transmembrane receptor protein serine/threonine kinase signaling pathway. Growth factor is a peptide that regulates cell growth and other cell functions by binding to specific, high-affinity cell membrane receptors. Hong Lun et al. showed that both growth factors EGF or TGF-βl induced epithelial-mesenchymal transition (EMT) in colon cancer cells, and the combined induction of both had a synergistic effect in enhancing the EMT phenomenon [37]. Receptor serine/threonine kinases primarily phosphorylate serine or threonine in downstream signaling proteins to transmit signals from outside the cell into the cell, which in turn affects gene transcription to achieve a variety of biological functions.

There were 8 protein clusters in the 289-gene PPI network analysis of the protein network, with MCODE1 containing 2 prognostic genes and MCODE2, MCODE3, MCODE4 and MCODE5 each containing one gene associated with prognosis. MCODE1 contains mainly hormone metabolism, cellular hormone metabolic and ethanol metabolic processes, MCODE2 contains G alpha (i) signaling events, Class A/1 and GPCR ligand binding, MCODE3 contains ADORA2B-mediated anti-inflammatory cytokines production, G alpha (s) signaling events and Anti-inflammatory response favouring Leishmania parasite infection, MCODE4 contains pathway regulation of phosphorylation of restricted SMAD proteins, and MCODE5 contains PPAR-alpha pathway, long-chain

fatty acid transport and PPAR signaling pathway. Each component network contains prognostic genes that are involved in a number of pathways associated with cancer. The relationship between the genes contained in each of their components deserves to be explored more closely to provide a greater basis for prognostic improvement and treatment strategies.

Although the experimentally established model has good results for the diagnosis of colon cancer, there are still some limitations. The first point is that the factors affecting the classification of the three machine learning methods are not further explored; the second point is that the staging diagnosis accuracy of the model is low and the model is not further improved. All these aspects have some negative impact on the diagnosis and treatment of colon cancer in practice. For the above limitations, we will continue to explore them in the subsequent studies.

5. Conclusion

In the diagnosis of colon cancer, the average accuracy of RF model in this experiment reached 99.81%, PR reached 99.88%, recall rate 99.50%, F1 value 99.68%, and specificity and ROC_auc reached 100%. The RF model achieved an average accuracy of 91.52%, PR of 86.94%, recall of 73.04%, F1 value of 76.79%, specificity of 98.16%, and ROC_auc of 82.37% in the diagnosis of first three stages of colon cancer with stage IV staging. By survival analysis of LASSO-extracted genes with PPI-screened genes, we also identified GCNT2, GLDN, SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5, and CPT2 associated with colon cancer prognosis.

Many studies in recent years have shown that in many cancers, machine learning approaches can screen for new marker genes, which will drive more cancer treatment modalities and provide the basis for targeted therapies. There is great potential for machine learning in cancer staging diagnosis, and as models are enhanced and refined, this approach can provide more accurate diagnostic staging for early and late stage treatment.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be foundhere: TCGA (https://portal.gdc.cancer.gov/repository).

Authors' contributions

Study design: YS, ChenCheng, WG, XT, YL. Data acquisition: YS, WG, XT, RG. Statistical analysis: YS, ChenCheng, WG, XT, ChenChen, DJ, RG. Manuscript preparation: YS, ChenCheng, WG, XT, ChenChen, TL, YL. Manuscript editing: YS, ChenCheng, XT, YL, RG. Manuscript review: YS, ChenCheng, WG, XT, ChenChen, DJ, TL, YL, RG.

Declaration of competing interest

The authors declare that they have no conflicting financial interests.

Acknowledgment

This work supported by the Clinical Research Center of Breast Tumor and Thyroid Tumor in Xinjiang Autonomous Region, the Special Project of Tianshan Innovation Team in Xinjiang Uygur Autonomous Region (2020D14031) and Tianshan Youth Project in Xinjiang Uygur Autonomous Region (2019Q043) .

References

[1] R.L. Siegel, K.D. Miller, A. Goding Sauer, S.A. Fedewa, L.F. Butterly, J.C. Anderson, A. Jemal, Colorectal cancer statistics, 2020, CA A Cancer J. Clin. 70 (3) (2020) 145–164, https://doi.org/10.3322/caac.21601.

[2] Statistics | Cancer.Net. [(accessed on 27 November 2021)]; Available online: https://www.cancer.net/cancer-types/colorectal-cancer/statistics.

[3] J.N. Li, S.Y. Yuan, Fecal occult blood test in colorectal caPhotodiagnosis Photodyn. Ther.ncer screening, Journal of digestive diseases 20 (2) (2019) 62–64, https://doi.org/10.1111/1751-2980.12712.

[4] H. Goyal, R. Mann, Z. Gandhi, et al., Scope of artificial intelligence in screening and diagnosis of colorectal cancer[J], J. Clin. Med. 9 (3313) (2020), https://doi.org/10.3390/jcm9103313.

[5] L. Krakowczyk, J.K. Strzelczyk, Epigenetic modification of gene expression in colorectal carcinogenesis, Wspólczesna Onkol. 11 (6) (2007) 289.

[6] A. Horwich, G. Ross, Circulating tumor markers, in: Principles of Molecular Oncology, Humana Press, Totowa, NJ, 2004, pp. 233–246.

[7] Staged | Cancer.Net. [(accessed on 27 November 2021)]; Available online:https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.

[8] Stage, T., Stage, N., & Stage, M. Carcinoma in Situ Corresponds to the TNM Classification. Laryngeal Cancer: Stages. M-distant metastases.

[9] Stages of Cancer | Cancer.Net. [(accessed on 27 November 2021)]; Availableonline: https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer.

[10] Cancer Survival Rates. [(accessed on 27 November 2021)]; Available online: https://cancersurvivalrates.com/?type=colon&role=patient.

[11] L.F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F.M. Sánchez-Margallo, J.B. Pagador, Deep learning to find colorectal polyps in colonoscopy: a systematic literature review, Artif. Intell. Med. (2020) 101923, https://doi.org/10.1016/j.artmed.2020.101923.

[12] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[13] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 1999.

[14] J. Shawe-Taylor, N. Cristianini, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, vol. 204, 2000.

[15] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998.

[16] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[17] A. Trabelsi, Z. Elouedi, E. Lefevre, Decision tree classifiers for evidential attribute values and class labels, Fuzzy Set Syst. 366 (2019) 46–62, https://doi.org/10.1016/j.fss.2018.11.006.

[18] M. Fratello, R. Tagliaferri, Decision trees and random forests, in: Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 374–383.

[19] X. Xie, C. Chen, T. Sun, G. Mamati, X. Wan, W. Zhang, G. Wu, Rapid, non-invasive screening of keratitis based on Raman spectroscopy combined with multivariate statistical analysis, Photodiagnosis Photodyn. Ther. 31 (2020) 101932, https://doi.org/10.1016/j.pdpdt.2020.101932.

[20] Fangfang Chen, Chunzhi Meng, Hanwen Qu, Chen Cheng, Chen Chen, Bo Yang, Rui Gao, Xiaoyi Lv, Human serum mid-infrared spectroscopy combined with machine learning algorithms for rapid detection of gliomas, Photodiagn. Photodynamic Ther. 35 (2021) 102308, https://doi.org/10.1016/j.pdpdt.2021.102308.

[21] S. Jubair, A. Alkhateeb, A.A. Tabl, et al., A novel approach to identify subtype-specific network biomarkers of breast cancer survivability, Netw Model Anal. Health Inform. Bioinfo. 9 (2020) 43, https://doi.org/10.1007/s13721-020-00249-4.

[22] S. Li, F. Han, N. Qi, et al., Determination of a six-gene prognostic model for cervical cancer based on WGCNA combined with LASSO and Cox-PH analysis, World J. Surg. Oncol. 19 (2021) 277, https://doi.org/10.1186/s12957-021-02384-2.

[23] J. Li, D. Zhou, W. Qiu, Y. Shi, J.J. Yang, S. Chen, H. Pan, Application of weighted gene Co-expression network analysis for data from paired design, Sci. Rep. 8 (1) (2018) 622, https://doi.org/10.1038/s41598-017-18705-z.

[24] C.G. Saris, S. Horvath, P.W. van Vught, M.A. van Es, H.M. Blauw, T.F. Fuller, R. A. Ophoff, Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients, BMC Genom. 10 (2009) 405, https://doi.org/10.1186/1471-2164-10-405.

[25] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, H. Liang, Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types, Nat. Commun. 5 (2014) 3231, https://doi.org/10.1038/ncomms4231.

[26] Y. Di, D. Chen, W. Yu, L. Yan, Bladder cancer stage-associated hub genes revealed by WGCNA co-expression network analysis, Hereditas 156 (2019) 7, https://doi.org/10.1186/s41065-019-0083-y.

[27] R. Jia, H. Zhao, M. Jia, Identification of co-expression modules and potential biomarkers of breast cancer by WGCNA, Gene 750 (2020) 144757, https://doi.org/10.1016/j.gene.2020.144757.

[28] H. Jiang, S. Luo, Y. Dong, Simultaneous feature selection and clustering based on square root optimization[J], Eur. J. Oper. Res. (2020), https://doi.org/10.1016/j.ejor.2020.06.045.

[29] O. Queen, S.J. Emrich, LASSO-based feature selection for improved microbial and microbiome classification, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 2021, pp. 2301–2308, https://doi.org/10.1109/BIBM52615.2021.9669485.

[30] N.S. Maurya, S. Kushwaha, A. Chawade, et al., Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer, Sci. Rep. 11 (2021) 14304, https://doi.org/10.1038/s41598-021-92692-0.

[31] Stages of Cancer | Webmd.Com. [(accessed on 5 December 2021)]; Available online: https://www.webmd.com/cancer/cancer-stages.

[32] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140, https://doi.org/10.1093/bioinformatics/btp616.
[33] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004), https://doi.org/10.2202/1544-6115.1027. Article3.
[34] M.M. Koo, R. Swann, S. McPhail, G.A. Abel, L. Elliss-Brookes, G.P. Rubin, G. Lyratzopoulos, Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study, Lancet Oncol 21 (1) (2020) 73–79, https://doi.org/10.1016/S1470-2045(19)30595-9.
[35] G. Norcic, Liquid biopsy in colorectal cancer-current status and potential clinical applications, Micromachines 9 (6) (2018) 300, https://doi.org/10.1016/j.tranon.2020.100907. Published 2018 Jun 15.
[36] Y. Xie, W.Y. Meng, R.Z. Li, Y.W. Wang, X. Qian, C. Chan, E.L.H. Leung, Early lung cancer diagnostic biomarker discovery by machine learning methods[J], Translational Oncol. 14 (1) (2021) 100907, https://doi.org/10.1016/j.tranon.2020.100907.
[37] L. Hong, Study of Growth Factor-Induced Epithelial-Mesenchymal Transition in Human Colon Cancer Cells [D], Central South University, 2011 (in Chinese), https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1011180350.nh.