Received June 8, 2020, accepted June 16, 2020, date of publication June 22, 2020, date of current version July 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003999

Screening of Pathogenic Genes for Colorectal Cancer and Deep Learning in the Diagnosis of Colorectal Cancer

YANKE LI, FUQIANG ZHANG, AND CHENGZHONG XING
Department of Anorectal Surgery, China Medical University First Hospital, Shenyang 110000, China
Corresponding author: Chengzhong Xing (xcz1966@126.com)

ABSTRACT Based on complex networks and machine learning methods, this paper studies the mining of colorectal cancer treatment genes, and innovatively combines a variety of feature extraction and comparative analysis methods, from gene network features, gene attribute features, network and attribute integration The three aspects of characteristics comprehensively excavate the genetic characteristics, and demonstrate the feasibility of the study through comparative analysis from different perspectives. Constructing a colorectal cancer gene network, analyzing the changes in the network structure during the development of colorectal cancer, and mining the network characteristics of genes are the first issues to be studied in this paper. The analysis of the network structure compares the changes in the network structure of the driver genes in the Normal network and the Tumor network and the edge mechanism of the driver genes, and the distribution of the eigenvalues of the driver genes and non-driver genes in the Tumor network. During the development of colorectal cancer, the network structure of the gene has changed significantly, and the prediction results based on the network structure show better prediction results than the non-network structure. These findings are feasible for the research direction of this paper. The argument is carried out, and the relevant analysis results are also given in the article. However, the research method of the thesis is based on network research, so comparing structural features with non-structural features only shows that structural features have a good classification ability, and cannot directly explain that modeling using gene networks is better than not using gene networks. Finally, based on the random forest, the optimized classification is improved to reveal the important factors affecting the diagnosis of colorectal cancer, and then to identify the true potential colorectal cancer driver genes, providing guidance for the clinical research of colorectal cancer and driving gene mining.

INDEX TERMS Colorectal cancer, gene screening, diagnosis, deep learning.

I. INTRODUCTION

Colorectal cancer is one of the most common causes of colorectal cancer death in human malignancies worldwide [1]–[5]. About 10-15% of patients with colorectal cancer have simultaneous lung metastases [6]–[8]. Despite the use of multiple chemotherapeutic agents such as 5-fluorouracil, oxaliplatin, and methotrexate in the treatment of colorectal cancer, the prognosis of most patients with colorectal cancer has not improved significantly [9]. In the past few years, radiotherapy has been used as an effective treatment for the treatment of colorectal cancer [10]–[12]. However, some patients receiving radiotherapy often have radiotherapy tolerance, which is the main factor for failure of radiotherapy and poor prognosis in patients with colorectal cancer [13]. Interestingly, it is believed that stem cells of colorectal cancer cells play an important role in the radiation resistance of colorectal cancer [14]. Therefore, it seems to be a promising strategy for the treatment of colorectal cancer by targeting factors such as radiation resistance to improve the efficacy of radiation therapy. Colonic polyps mainly refer to the benign bulge-like lesions of the mucosa in the intestinal cavity of the colorectal. It can be single or multiple in each intestinal segment. According to the type of pathology, it can be divided into inflammatory, adenomatous, hamartoma, etc. Adenomatous polyps can be divided into villous, dentate, and tubular adenoma [15]. Previous studies have shown that there is a close relationship between colon cancer and adenomatous polyps. Clinical research progress on risk factors related to the incidence of colon polyps. Studies have shown that

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

adenoma carcinogenesis is the main cause of colon cancer, which accounts for more than 50% of colon cancer; when the diameter of an adenoma exceeds 2 cm, the chance of cancelation will also exceed 50% [16]. With the morphological progress of adenoma to cancer, a series of genetic changes such as inactivation of tumor suppressor genes and activation of oncogenes will occur at each stage [17]. Since it is known that most colorectal adenoma develops into colorectal adenocarcinoma through the well-known adenoma sequence development process, the incidence of adult colorectal adenoma cannot be ignored. Related research indicates that the risk factors for adenoma in young adults include: men, smokers, drinking, obesity, etc. These clinical factors may be an indication for screening adenoma in young people [18]–[22].

Colorectal cancer is one of the most common gastrointestinal malignancies in the world today, and it has caused very serious harm to human health. Östlund *et al.* have confirmed through research that H19 is a miRNA675 precursor RNA, which shows up-regulated expression in colorectal cancer tissues and cells, and can also inhibit the activity of its downstream tumor suppressor gene (retinoblastoma, RB), thereby promoting tumor cell generation and proliferation [23]. Liu *et al.* found that H19 is a newly discovered important regulatory gene in the process of colorectal cancer epithelial mesenchymal transition [24]. It is highly expressed in mesenchymal cell-like tumor cells and original cancer tissues. The high expression of H19 can significantly promote epithelial mesenchymal transition process and tumor growth. According to research, H19 can also play the role of endogenous competitive RNA to inhibit the function of miR-138 and miR-200a, thereby eliminating their inhibition of the core imprints of mesenchymal cells such as ZEB1, ZEB2 and Vimentin [25]–[28]. There are many studies showing that H19 is overexpressed in colorectal cancer. However, Li *et al.* found that the low expression state of H19 can significantly promote the formation of colonic polyps in mice during the study of colon cancer models [29]. To a certain extent, it shows that H19 may have two different roles in oncogenes or tumor suppressor genes in colorectal cancer, but what role dominates in the different stages of colorectal cancer development requires our constant the study found [30].

Xu *et al.* developed the CHASM method to train a random forest to identify and prioritize missense mutations that are most likely to produce functional changes that enhance tumor cell proliferation [31]. Alexander *et al.* proposed a colorectal cancer driver annotation tool, which predicts missense mutations by training SVM [32]. Since the number of driver genes is much smaller than passenger genes, the method based on machine learning is full of challenges in constructing positive and negative samples [33]. However, through reasonable sample sampling, it is often possible to obtain better performance than other algorithms. Zhang *et al.* proposed a tree model-based prediction algorithm (20/20 +) and compared it with seven classic driver gene prediction algorithms [34]. The results show that 20/20 + performs best among the eight algorithms, indicating that machine learning models can predict driver genes well. Zhao *et al.* proposed the OncodriveCLUST method to predict driver genes by looking for categories generated at locations where the mutation rate is higher than the background mutation rate [35]. Xu *et al.* proposed the MutsigCV method, which uses gene expression data and DNA replication time data to establish patient-specific background mutation models to identify genes that are significantly mutated [36]. Brambila-Tapia *et al.* proposed the InVEx method to analyze potential exogenous driver genes whose mutation frequency is much higher than the background mutation rate by analyzing large-scale melanoma exome data [37]. However, most mutations in colorectal cancer occur relatively infrequently, and it is difficult to establish a reliable background mutation model, which limits the performance of methods based on mutation frequency. The method based on network analysis aims to mine important nodes in the network and identify them as driving genes [38]–[40]. Among them, Amer *et al.* proposed the DawnRank algorithm, which uses the PageRank algorithm to sort the genes in the gene interaction network to predict the driving genes [41]. Guo *et al.* proposed a single sample control strategy (SCS), using network control theory to find driving mutations that can regulate the network from normal state to disease state [42]. However, the gene interaction networks of such methods are generally downloaded from online databases, such as BioGrid [43] and HPRD [44], which usually contain many false positive data. Therefore, network-based methods require more precise gene interaction networks to improve prediction accuracy.

With the completion of the Human Genome Project and high-throughput sequencing technology [45]–[47], a large amount of colorectal cancer genomic data has been published, which directly introduces colorectal cancer related research into the fast-track development of high-speed development. Studies have found that carcinogens that can cause gene mutations in normal cells can promote the growth of colorectal cancer cells and malignant proliferation. Such carcinogens are called driver genes. The mutation of the driving gene is the most important factor in the development of colorectal cancer, and plays a major role in the development of colorectal cancer; correspondingly, the passenger gene refers to a gene that has a smaller effect on colorectal cancer, that is, it is associated with colorectal cancer cells. The growth and proliferation have little correlation. Therefore, the identification of colorectal cancer driver genes from all genes is of great significance. Not only that, the driver genes of different types of colorectal cancer are also different, and the selection of driver genes according to the type of colorectal cancer also helps to improve the efficiency of colorectal cancer prevention and treatment. The study of driver genes is of great significance to the diagnosis, prevention and treatment of colorectal cancer, the development of targeted drugs, and the understanding of pathological mechanisms. Machine learning methods are quite mature in the field of predicting driver genes, and have achieved very good results. However, in order to improve the prediction accuracy of the method, simply using various biological information features extracted from different data sets to train the classifier may not be the best way to integrate different types of data. Studies have shown that integrating different types of feature data (such as gene mutation and gene expression data) can more effectively improve disease prediction efficiency. Considering that the mining of key nodes in network analysis can be used to characterize the importance of nodes, if a gene interaction network for colorectal cancer is constructed for genetic data and combined with gene mutation information and expression, then the prediction accuracy of the algorithm should be better than The prediction accuracy obtained by concatenating pure biological information features is higher. Based on complex networks and machine learning methods, this paper studies the mining of colorectal cancer driver genes, and innovatively combines a variety of feature extraction and comparative analysis methods, respectively from three aspects: gene network features, gene attribute features, integrated features of networks and attributes. Dig out the genetic characteristics, and demonstrate the feasibility of the study through comparative analysis from different perspectives.

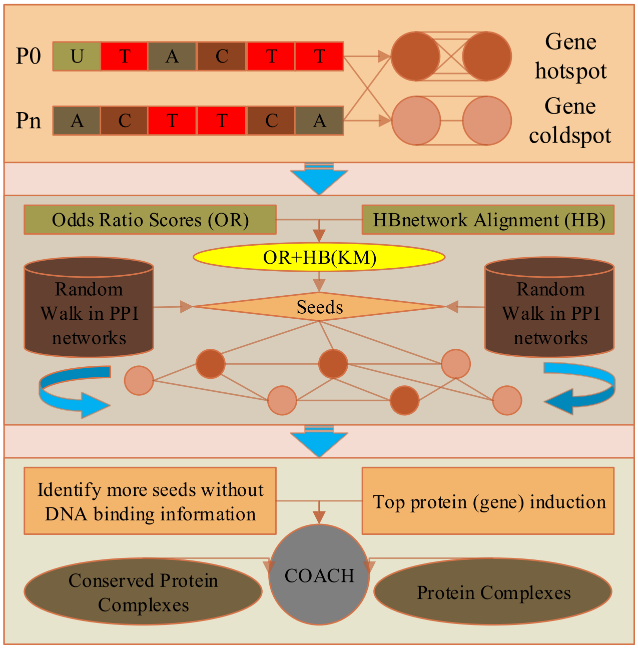
P0
U
T
A
C
T
T
Pn
A
C
T
T
C
A
Gene hotspot
Gene coldspot
Odds Ratio Scores (OR)
HBnetwork Alignment (HB)
OR+HB(KM)
Random Walk in PPI networks
Seeds
Random Walk in PPI networks
Identify more seeds without DNA binding information
Top protein (gene) induction
COACH
Conserved Protein Complexes
Protein Complexes

FIGURE 1. The method of restarting with symbolic random walk.

II. DEEP LEARNING ANALYSIS OF PATHOGENIC GENE SCREENING

A. GENE SCREENING WITH SYMBOLIC RANDOM WALK RESTART METHOD

The signed random walk restart algorithm is a model for personalized ranking of nodes in a signed network. Signed random walk restart is different from the traditional random walk-based method, because the traditional random walk-based method is only applicable to the network that is assumed to be a positive edge, and cannot effectively rank the nodes in the signed network [48], and Lack of the ability to consider complex edge relationships, the signed random walk restart algorithm makes up for this. The traditional walk mechanism based on the random walk method and the signed random walk restart method is shown in Figure 1. There is an interaction between directly connected genes and genes. This relationship may be a relationship that promotes expression, or a relationship that inhibits expression. Therefore, when constructing a colorectal cancer gene network, this special relationship between gene nodes can be distinguished by the marginal weight. If the genes show a mutual promotion relationship, then the marginal weight value is defined as +1; if the gene the mutual suppression relationship between them, then the weight value of the connected edge is defined as −1.

Mechanism of the interaction of genes in the syndrome gene network. Introducing the research idea of the symbolic random walk restart algorithm can well reveal the internal mechanism of genes and mine the features that can be used for model learning and training [49]. Considering that each gene node may promote expression with some neighbor nodes, and suppress expression with some neighbor nodes, so when calculating the node score, it is necessary to consider both the promotion relationship and inhibition relationship between genes, and define each gene The node's score M is equal to the inhibition expression score M^- and the promotion expression score M^+. The calculation method is illustrated in Figure 2 as an example.

This article takes node N as an example to introduce the calculation method of node N [50]. As can be seen from Figure 2, the joint edge weight of node i and node N is +1, the joint edge weight of node j and node N is −1, and the joint edge weight of node k and node N is also −1, Because the algorithm considers the influence of the edge weight, when the value passes through the edge with a weight value of −1, the sign of its value must be reversed; otherwise, the value remains unchanged when it passes through the edge with a weight value of +1. Therefore, the $M_N^+(t+1)$ score of node N is obtained by the weighted sum of the positive value of node i, the negative value of node j, and the negative value of node k; Similarly, the $M_N^-(t+1)$ score of node N It is

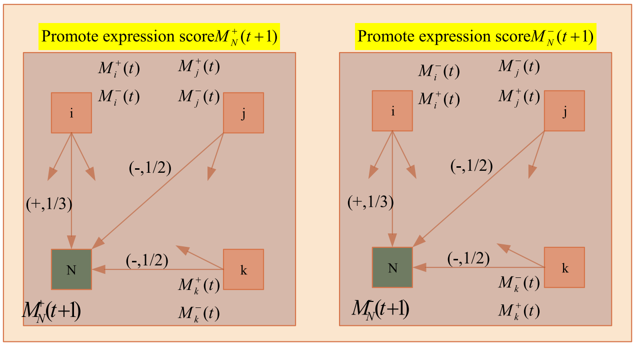

FIGURE 2. Definitions of M+ and M− in the random walk restart algorithm with sign.

obtained by the weighted sum of the negative value of node i, the positive value of node j, and the positive value of node k. Therefore, without considering the attenuation condition, the calculation formula of node N is as follows:

$$M_N^+(\mathrm{t}+1) = (1-c)(\frac{M_{\mathrm{i}}^+(t)}{3} + \frac{M_j^+(t)}{2} + \frac{M_k^+(t)}{2}) + c1^*(N+S) \quad (1)$$

$$M_N^-(\mathrm{t}+1) = (1-c)(\frac{M_{\mathrm{i}}^-(t)}{3} + \frac{M_j^-(t)}{2} + \frac{M_k^-(t)}{2}) \quad (2)$$

$$M_N = M_N^+(\mathrm{t}+1) + M_N^-(\mathrm{t}+1) \quad (3)$$

In the actual network, there is a problem of information attenuation during the transmission of information, so it is necessary to consider the situation of balanced attenuation. The paper gives the score M of each gene node, the suppression expression score M^- and the promotion expression score M^+. The definition is as follows:

$$M^- = (1-c)[G_-^T M^+ + \gamma G_+^T M^- + (1-\beta) G_-^T M^-] \quad (4)$$

$$M^+ = (1-c)[G_+^T M^+ + \beta G_-^T M^- + (1-\gamma) G_-^T M^-] + cq \quad (5)$$

$$M = M^- + M^+ \quad (6)$$

G is the adjacency matrix, T is the degree matrix, q is the start vector, and c is the restart probability; γ and β represent the balanced attenuation factors acting on the positive and negative weight edges after passing the negative weight edge, without considering the balanced attenuation time.

$$M^- = (1-c)(G_-^T M^+ + G_+^T M^-) \quad (7)$$

$$M^- = (1-c)(G_+^T M^+ + G_-^T M^-) + cq \quad (8)$$

In the research, this article uses the mutation frequency of the gene and the differential expression of the gene as initial values to conduct random walks [51]. The advantage of this is that the biological information of the genes is fused into the network, and the fusion characteristics that can characterize the genes are obtained.

Based on the above research, this paper also introduces the concept of network entropy in feature extraction. For gene i, the local network structure entropy $\mathrm{F_i}$ is defined as:

$$F_i = \frac{1}{\log k_i} \sum_{j \subseteq N_i} P_{ij} \log P_{ij} \quad (9)$$

B. DEEP LEARNING MODEL ANALYSIS

Use order statistics to score rank vectors: In most cases, it is to find genes that rank higher in many preference lists, thereby ignoring a small number of studies that do not provide information. Therefore, we can assume that all the normalized ranks that provide information come from a distribution [52]. For any normalized rank vector M, let $M(1) \leq M(2) \leq \ldots\ldots \leq M(n)$ be a reordering.

Then we estimate the probability of the rank vector ˆM (k) ≤ M (k) generated under the null model (that is, all rank orders rj are sampled from a uniform distribution). Let β, k and n (M) denote the probability of ˆM (k) ≤ M (k). Then under the zero model, the probability of order statistics ˆM (k) being less than or equal to x can be expressed as a binomial probability. As shown in Figure 3, it is a framework for deep learning.

Algorithms are a series of instruction sets for solving practical problems, and they are the strategy mechanism for solving problems. Similarly, whether an algorithm is good or bad is directly related to whether the problem can be solved reasonably, and the measurement of the algorithm quality needs some evaluation indicators, which is an indispensable work content in the modeling process. At present, the commonly used evaluation indicators to evaluate the performance of machine learning classification algorithms include AUC value, precision, and recall [53].

Whether it is the calculation of the AUC value, or the calculation of the precision rate and the recall rate, it is based on the comparison of the real category and the prediction result. By constructing the confusion matrix, this comparison result can be intuitively reflected, and the relevant evaluations are calculated on this basis index. The confusion matrix is shown in Table 1.

TABLE 1. Two-class confusion matrix.

Confusion matrix		Predictive value	
		positive	negative
Actual value	positive	TH	FR
	negative	FH	TR

In the analysis process, this paper defines the proportion of positive samples predicted correctly by the algorithm in the test set to all the positive samples as the precision rate, and the proportion of positive samples predicted correctly by the algorithm to the true samples is the recall rate (or Recall rate). The precision rate and the recall rate usually appear as a pair of confrontation evaluation indicators. In order to have a high precision rate and maintain a high recall rate, in actual operation, it is necessary to compromise on the selection of parameters to find the best result. The precision rate P and the recall rate M are defined as follows:

$$H = \frac{TH}{TH+FH} \quad (10)$$

$$R = \frac{TR}{TR+FR} \quad (11)$$

Among them, TH, FH, TR and FR are true positive, false positive, true negative and false negative, respectively. In this study, true positives indicate driver genes predicted as driver genes, false positives indicate passenger genes predicted as driver genes, true negatives indicate passenger genes predicted as passenger genes, and false negatives indicate passenger genes predicted as passenger genes.

For the differentially expressed genes obtained above, in order to explore their potential molecular mechanisms in the occurrence and development of colorectal cancer, detailed and complete bioinformatics annotations and descriptions

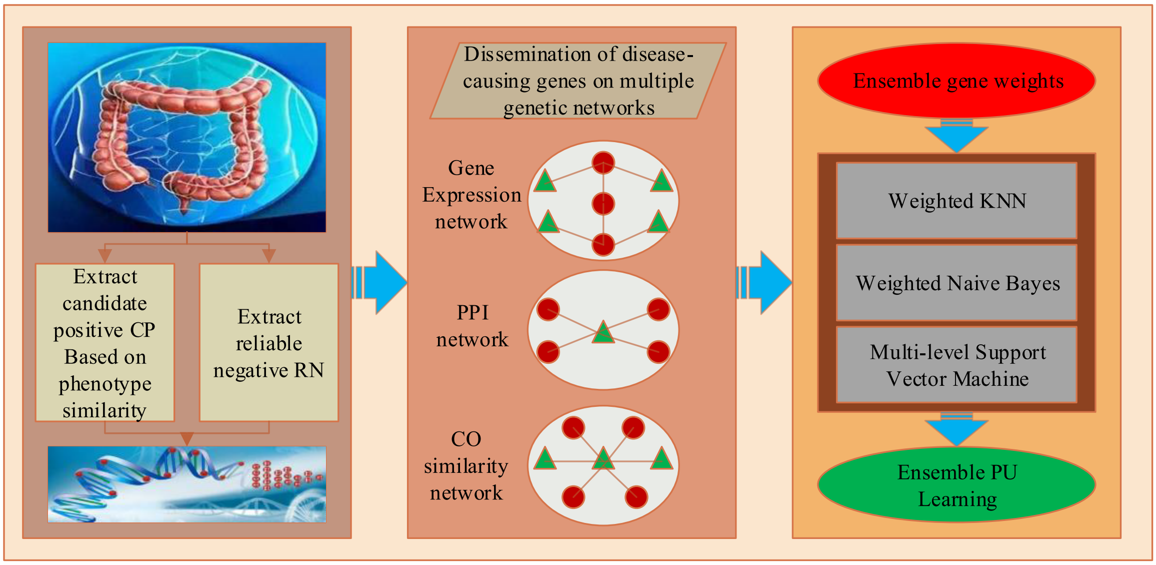
Extract candidate positive CP Based on phenotype similarity
Extract reliable negative RN
Dissemination of disease-causing genes on multiple genetic networks
Gene Expression network
PPI network
CO similarity network
Ensemble gene weights
Weighted KNN
Weighted Naive Bayes
Multi-level Support Vector Machine
Ensemble PU Learning

FIGURE 3. Deep learning framework.

are required. First, the use of gene function analysis (Gene Ontology, Gene Ontology, GO) and pathway analysis (Kyoto Encyclopedia of Genes and Genomes, KEGG). Gene Ontology (GO) is a method for systematically annotating the properties of species genes and their products [54]. It covers three aspects of biology: cellular components, molecular functions, and biological processes. 1) Cellular component (cellular component): each part of the cell and the extracellular environment; 2) molecular function (molecular function): can be described as molecular-level activity, such as catalytic or binding activity; 3) biological process (biological process): The process or collection of molecular events.

TABLE 2. Colorectal cancer data types and corresponding genes.

Types of cancer	Total number of genes	Number of known driver genes
BRCA	15523	313
COAD	15643	145
HNSC	15711	194
KIRC	15986	127
LUAD	15909	257
LUSC	16221	186
UCEC	16397	232

III. DIAGNOSTIC ANALYSIS OF COLORECTAL CANCER

A. DATA SOURCES AND DATA PREPROCESSING

The Cancer Genome Atlas (TCGA) is a powerful database that integrates a variety of colorectal cancer genome sequencing data and provides reliable and sufficient data resources for the rapid development of biomedical research [55]. Moreover, scientific research projects based on the rich data provided by the TCGA database have achieved rich results. This thesis is also based on research questions and research objects, and the related gene data is downloaded from the TCGA database for research. The data composition of different colorectal cancer types is not the same, not only in the difference in the number of genes, but also the number of known driver genes for different colorectal cancer. Table 2 shows the number of genes and known driver genes for different colorectal cancers in detail.

Based on the strict GEO data set screening conditions in the early stage, a total of 8 sets of colorectal cancer chip expression data sets that met the requirements were obtained. The total number of included samples was 706, including 493 colorectal cancer samples and 213 normal samples. These sample sizes are sufficient for subsequent bioinformatics-based differential expression gene screening. At the same time, this is also the current bioinformatics analysis of colorectal cancer based on the GEO database, which is included in the data mining research with the largest sample size.

Based on a series of data preprocessing, a box plot of the data distribution of each set of GEO data sets before and after normalization was drawn. The results are shown in Figure 4, and the expression distribution of each set of data was found. Before the quantiles are normalized, they appear disordered, but after normalization, they present a consistent data distribution, which is helpful for the use of subsequent analysis.

The work of the data pre-processing part is mainly to screen out the genes with significant differential expression, and to filter out the genes with insignificant differential expression at the same time, and use the selected genes for the subsequent construction of the colorectal cancer gene interaction network. According to the type of colorectal cancer, the samples were divided into a control group and a treatment group,

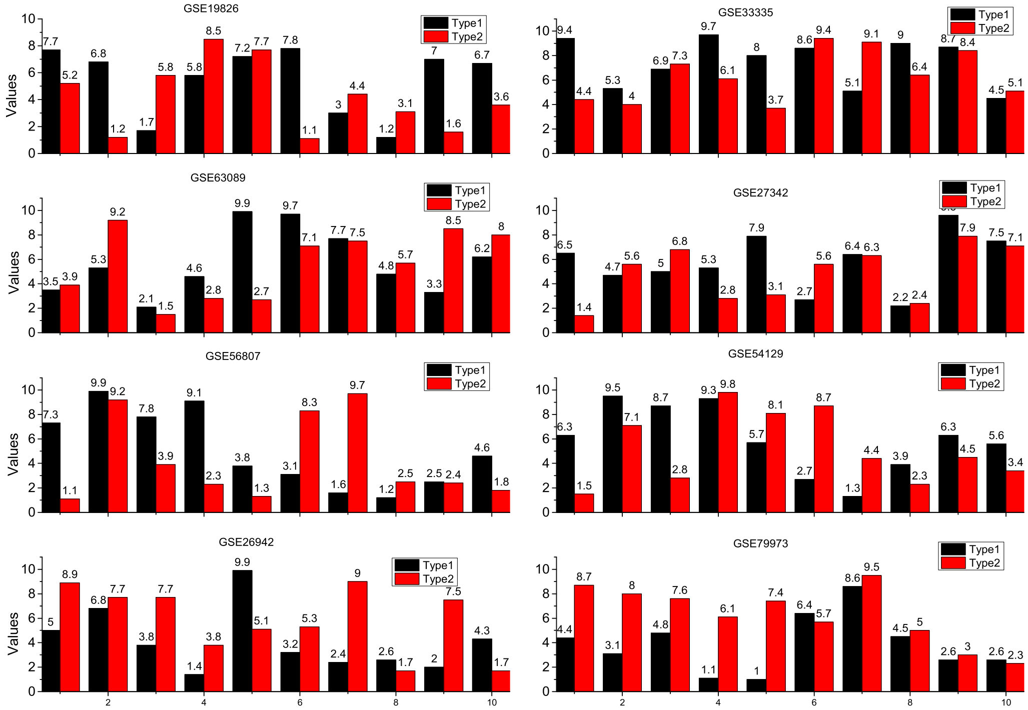

FIGURE 4. Differential gene screening data pretreatment for colorectal cancer.

and the differential expression analysis of genes was carried out by means of precise testing. The p-values corrected by Benjamin-Hochberg (BH) multiple test were corrected. The genes with a false detection rate (FDR) of less than 0.05 and a difference multiple (colorectal cancer sample and normal sample) of more than 4 times were used as differentially expressed genes for further research. The genes in the red and green parts are the differential genes screened, red indicates up-regulated gene expression, and green indicates down-regulated gene expression. Based on the screened differentially expressed genes, a colorectal cancer gene interaction network is constructed for subsequent feature mining related research.

Construct colorectal cancer gene network G = (V, E) based on the screened differential genes and gene relationship data, where V represents the node set, E represents the edge set, the node represents the gene, and the edge represents the protein between the two genes. There is an interaction relationship, and the weights of the edges are positive and negative, that is, the weights of the genes that promote mutual expression between genes are positive (+1), and the weights of the genes that mutually inhibit expression are negative (−1).

In fact, the network constructed here is actually composed of many subgraphs of uneven size, not a fully connected graph. The data analysis results also show that the proportion of the number of nodes in the largest connected subgraph of BRCA, COAD, HNSC, KIRC, LUAD, LUSC, and UCEC accounts for about 0.995, 0.992, 0.992, 0.995, 0.993, 0.994, and 0.988, respectively. And the genes in the largest connected subgraph of each colorectal cancer gene network include all known driver genes corresponding to colorectal cancer.

B. DIAGNOSTIC EVALUATION CRITERIA

In view of the particularity of the research data of the thesis, in addition to some known driver genes, other genes in the genetic data are strictly unknown genes in the strict sense. Unknown genes include potential driver genes and passenger genes, which need to be further studied and verified. Therefore, the ensuing problem is that there are only positive samples (known driver genes) and no negative samples (passenger genes) in predicting potential driver genes. That is, the negative samples in the training samples need to be constructed by themselves.

In order to solve the problem of no negative samples, this paper proposes a hypothesis during the research process, that is, each gene has a probability of being a driver gene, but the problem is that the gene is the probability of the driver gene. The probability is greater than that of a driver gene. In other words, the probability of people liking commodities must be greater than the probability of liking commodities. Based on this assumption, the next step is to screen for unknown genes. The steps for constructing a negative sample are as follows:

(1) Randomly sample a certain percentage of gene samples from unknown gene samples as a negative sample, and form a training set together with known colorectal cancer driver genes;

(2) Train a classification model based on the random forest algorithm, and predict all genes except the training set, and give the probability that each gene is a driving gene;

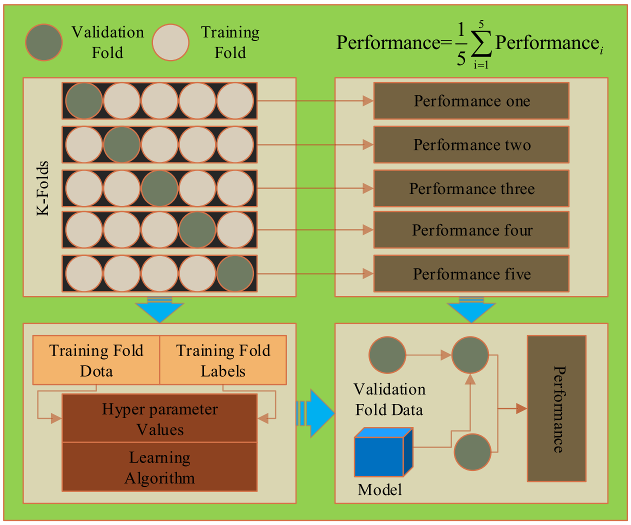

FIGURE 5. **Evaluation model.**

(3) Repeat steps (1) and (2) 100 times to obtain 100 predicted result sets;

(4) According to the principle that the probability data is small enough and the number of repetitions is as large as possible, some genes are selected as final negative samples for the 100 predicted result data;

(5) Combine the negative and positive samples to get a new data set.

By repeating the above 5 steps 7 times, 7 kinds of negative sample structure problems of colorectal cancer can be completed. However, due to the small number of colorectal cancer driver genes, the problem of imbalance in the sample ratio between the negative and positive samples of colorectal cancer still needs to be solved.

The index of algorithm evaluation is a very important index to measure the performance of the algorithm, and the performance of the algorithm can be intuitively distinguished from the performance gap between different algorithms through algorithm comparison. The algorithm proposed in this paper will be evaluated from the following two aspects:

Through the analysis of related research on the driving genes of colorectal cancer, it is not difficult to find that SVM as a classic algorithm is very widely used in the field of biomedicine, taking into account the uniqueness of the modeling in this paper (combining complex network analysis with machine learning) Different from other algorithms that do not consider the network structure, the SVM algorithm is a relatively suitable comparison algorithm, so the algorithm proposed in the paper will be compared with the SVM algorithm next. As shown in Figure 5, it is an evaluation model.

Except for 3 cases of colorectal cancer, the paired adenoma tissues have been exhausted, and the remaining 48 tissue samples were all exon captured and sequenced by Ouyi Biological Co Ltd. The average sequencing coverage was 150 × or more. The specific sequencing steps are as follows: DNA is extracted from tissue samples, and the library is constructed after passing the quality inspection. Qualified DNA tissue samples were processed by Covaris ultrasonic disrupter, which was randomly broken into 350 bp fragments, and then the library was constructed. The kit used was TruSeq DNA LT Sample Prep kit, and then the DNA fragments were completed through the following steps Construction of the library: end repair, add ployA tail, add sequencing adapter, purification, PCR amplification. After the library is qualified, the sequencer is used for double-end sequencing. After the sequencing data is off the machine, first filter the data, remove the low-quality data, and obtain Clean Reads. Then compare Clean Reads with the reference genome, use the GATK software to detect SNV and InDel sites according to the result of the comparison, and use CNVkit software to detect and annotate the CNV.

The comparison of network features aims to analyze the changes in the network structure characteristics of known driver genes of various types of colorectal cancer during the carcinogenesis from normal tissue state to colorectal cancer tissue state. On the one hand, it can be used to verify the research feasibility of integrating network structural features, and on the other hand, it can help to dig out some potential change mechanisms of the network in the process of cancelation, which has important guiding significance for revealing the occurrence and development of colorectal cancer from the network perspective.

First, construct the Normal gene network and the Tumor gene network separately according to the operation process; after constructing the network, this paper extracts the degree centrality, aggregation coefficient, near centrality of seven types of colorectal cancer genes in the network based on the Normal network and the Tumor network respectively. Intermediate centrality, feature vector centrality, K-shell value, and local structure entropy of the network. Then, compare the distribution of each network feature corresponding to the Normal network and the Tumor network. Among them, except for the centrality of the intermediary, the other six network characteristics have significantly increased changes. The reason is that the centrality of the intervening number reflects the importance of node i as a "bridge", which is determined by the number of shortest paths passing through node i and the total number of shortest paths. At the same time, in view of the particularity of the driver gene itself, whether it is in the Normal network or the Tumor network, its importance cannot be ignored, but the discussion in the two networks lacks a relative reference, that is, the centrality of the number The numerator and denominator values of the calculation formula may change in the same trend, so this comparative analysis is not suitable for the centrality of the intermediary.

IV. RESULTS ANALYSIS

A. FEATURE COMPARISON AND ANALYSIS

The paper does not compare the calculation results of the signed random walk restart algorithm in the section of network characteristics comparison and analysis. As shown

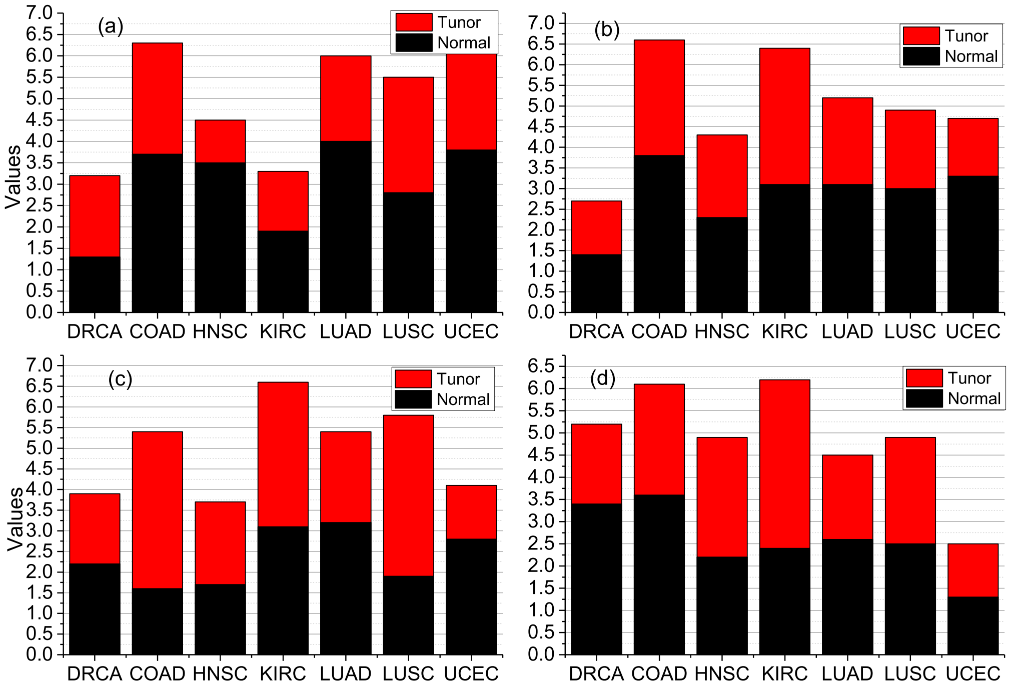

FIGURE 6. Comparison of analysis results of different network indicators in Normal network and Tumor network.

in Figure 6, the main reason is that the initial value of the random walk of the algorithm is selected by the mutation frequency and differential expression of the gene. In the research, we can only choose the Tumor network as the random walk network, so we cannot compare the difference between Normal and Tumor for the features mined by this algorithm. The evaluation of the feature extraction of the random walk restart algorithm with signs in the paper will be given later.

It can be clearly seen from Figure 6 that degree centrality, aggregation coefficient, near centrality, eigenvector neutrality and K-shell value all have significant changes, especially the index of the structural entropy of the local network changes most significantly before and after. This change in network structure not only echoes the literature research on entropy mentioned earlier, but also shows that the structural entropy of the local network is a characteristic representation that can well describe the changes in driving genes. The comparative analysis of the structural characteristics of the Normal network and the Tumor network once again demonstrates that it is evidence-based to mine the feature dimensions that can characterize genes from the perspective of network structure.

In addition to the comparative analysis of the network structure attributes, the paper also made an in-depth analysis and mining of the connection mechanism between the driver genes in the Normal network and the Tumor network, the purpose is to analyze the connection mechanism of the colorectal cancer genes, and tumor cancelation During the process, it is known that the driving genes are continuously changed. It has been pointed out that the number of known driver genes of various colorectal cancer types' accounts for a small proportion of the total number of genes corresponding to colorectal cancer, so only the mechanism of the connection of known driver genes is analyzed and attempts to find some valuable and valuable Refer to the conclusion of meaning.

For the analysis of the connection mechanism between known driver genes, this article mainly starts from two perspectives. First, it analyzes the connection situation of driver genes in the Tumor network, and then compares and analyzes the connection situation of the driver genes in the Normal network and the Tumor network. It was found that in different Tumor networks, the proportion of driver genes that form connected triads among known driver genes accounted for more than half of the total number of driver genes. Only colon adenocarcinoma (COAD) constitutes connected triads. The ratio is less than half, but it is basically close to half, which may be related to the small number of driver genes known in colon adenocarcinoma, which is the type of colorectal cancer with the smallest number of driver genes in the studied colorectal cancer. This discovery reveals that there is a high possibility that the driver gene will tend to be directly connected to the driver gene, because the number of driver genes is known to account for only about 1% of the number of genes in the constructed gene network, which is a high probability event under the condition of small probability. As shown in Figure 7. Among them, it represents the proportion of the known driver genes of the colorectal cancer that constitute the connected triplet, and the number of driver genes that have no

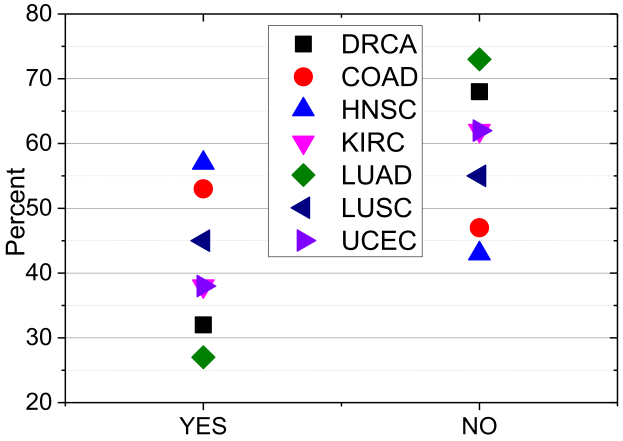

FIGURE 7. Proportion of connected triplets in the driving genes of colorectal cancer.

driver genes or only one driver gene in the neighboring nodes accounts for the proportion of the known driver genes of the colorectal cancer. Of course, if only the neighbor nodes of the driver gene are considered in the analysis, then the proportion of such driver genes will be higher.

Based on the above findings, this paper also conducted the same analysis on the Normal network and found that in the Normal network, the proportion of known driver genes that form a connected triplet is not as high as that in the Tumor network, but it is still quite high. The ratio is shown in Figure 8. Among them, the ordinate value (left) corresponding to the node of the solid red line is the proportion of orange in Figure 8. The blue dotted line represents the proportion of connected triples in the Normal network (blue line (Below the red line). This shows once again that whether in the Normal network or the Tumor network, the driver genes tend to be directly connected to the driver genes.

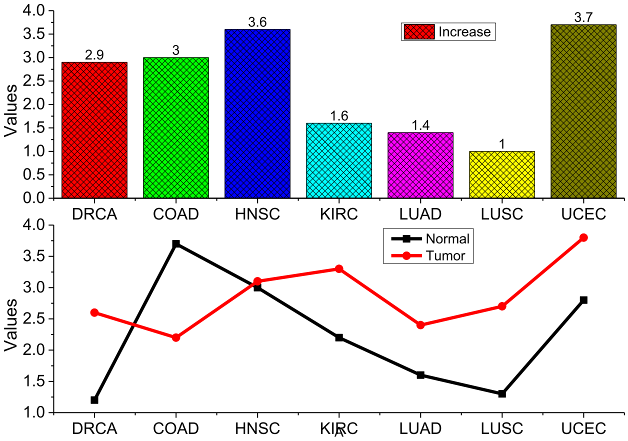

FIGURE 8. Proportion of connected triplets in the driving genes of Normal network and Tumor network.

On the basis of this analysis, the paper further compares the changes in the proportions of known triplets connected by the known driver genes in the Normal network and Tumor network. As shown in Figure 8, the abscissa corresponds to seven different types of colorectal cancer, the ordinate (left) represents the proportion of driver genes that form a connected triplet (broken line), and the ordinate (right) represents the Tumor network of different colorectal cancers The difference of the proportion corresponding to the Normal network. It was found that along with the process of cancelation, the proportion of connected triplets was significantly increased, among which renal clear cell carcinoma (KIRC) and endometrial cancer (UCEC) increased most obviously, indicating that the driver genes of colorectal cancer are not only inclined Because it is directly connected to the driver gene, and as the tumor deteriorates, this ratio shows a trend of continuous increase.

B. COMPARATIVE ANALYSIS OF DRIVER GENES AND NON-DRIVER GENES

The above is mainly a comparative analysis of the Normal network and Tumor network. This section will analyze the differences between the network structure characteristics of driver genes and non-driver genes (including passenger genes). Because the follow-up driver gene prediction is carried out with the Tumor network as the background network, the comparative analysis of driver genes and non-driver genes here is also based on the structural characteristics of the Tumor network.

This paper compares the seven network structure indicators of driver genes and non-driver genes, including degree centrality, aggregation coefficient, near centrality, intermediary centrality, feature vector centrality, K-shell value, and local network structure. As shown in Figure 9, it was found that the seven indicators showed significant differences among different types of colorectal cancer, and compared with the distribution of the characteristic values of non-driver genes, the driver genes had a higher characteristic distribution. Because the non-driver genes not only contain a large number of passenger genes, but also some drive genes waiting to be mined by the algorithm, the value of this part of potential drive genes will have a certain interference effect on the comparison of network structure characteristics, but in this case The distribution of driver genes and non-driver genes still shows a clear difference, so it can be well explained that there is a significant difference in the distribution of network structure characteristics between driver genes and passenger genes.

In Figure 9, each subgraph represents a network structure feature of a colorectal cancer. The box plot of each subgraph represents the distribution of driver genes and non-driver genes, and the red box plot represents the distribution of driver gene feature values. The blue boxplot represents the distribution of eigenvalues of non-driver genes. It can be clearly seen from the figure that the seven characteristic indicators analyzed have a good classification ability, and this result is also consistent with the conclusion of the analysis of feature importance.

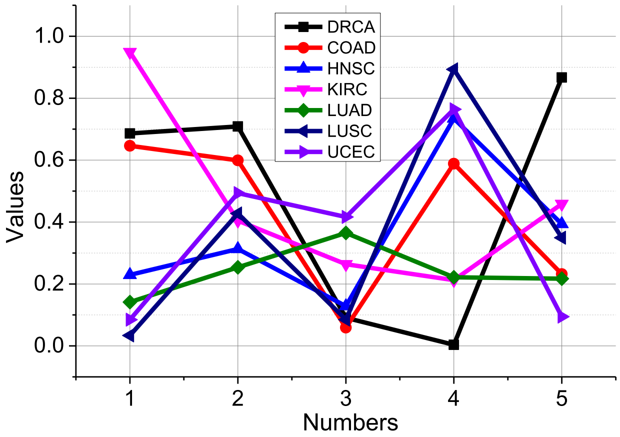

FIGURE 9. Analysis of network structure characteristics of driver genes and non-driver genes in Tumor network.

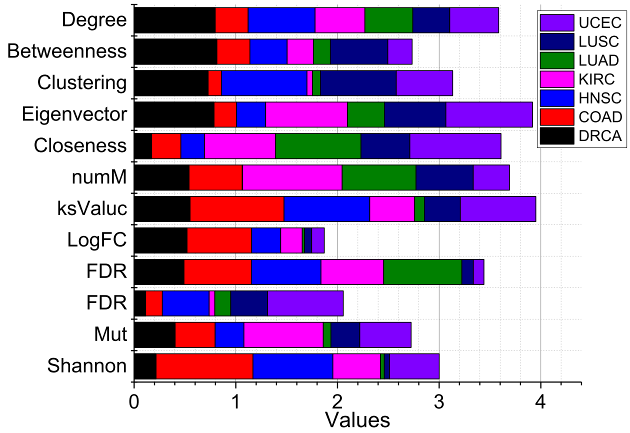

FIGURE 10. Results of single feature analysis of colorectal cancer.

In order to analyze the importance of the extracted features in the research process of the thesis, on the premise of ensuring that the model training method is unchanged, the method of splitting the features is adopted, that is, only one feature is selected at a time for model learning and training. Repeat this process until all the features are traversed, and the ROC curve and corresponding AUC score of the training model based on a single feature can be obtained, and the AUC score is used as an important measure of the corresponding single feature, as shown in Figure 10. In addition to the importance ranking of all features in Figure 10, it can also be clearly found that even in the prediction models of different colorectal cancer types, all individual features can maintain the AUC value greater than 0.5 (the red dotted line corresponds to the AUC The value is equal to 0.5), which can fully explain that the features extracted by feature engineering have a good classification ability. In addition, the degree of centrality, clustering coefficient, mutation frequency and K-shell value have the highest scores, indicating that these four features have relatively better classification capabilities and are more important for the prediction of colorectal cancer driver genes.

Among them, the ordinate represents the feature, the abscissa corresponds to the seven types of colorectal cancer, the abscissa of each sub-graph represents the importance score of the feature, that is, the importance of the feature corresponding to the histogram, the red dotted line is the reference line (AUC = 0.5).

The characteristics of genes are composed of structural and non-structural characteristics. Among them, the structural features are the features extracted based on the network, and the non-structural features refer to the attribute features of the gene, including the mutation frequency of the gene, logFC value and FDR value. In order to study the influence of structural features and non-structural features on the prediction results respectively, in this study, structural features and non-structural features were analyzed separately, and structural features and non-structural features were used for model learning and training respectively, and unknown genes Make predictions. For the two types of special prediction results, first, the genes predicted to be driver genes with a probability greater than 0.5 are sorted in descending order according to the probability, and different intervals are divided according to the threshold, and then the prediction results in different intervals are compared with the colorectal cancer gene census (Cancer Gene Census, CGC) database overlap number to compare the prediction results of structural features and non-structural features. The results correspond to the blue dotted line and orange dotted line in Figure 11, respectively.

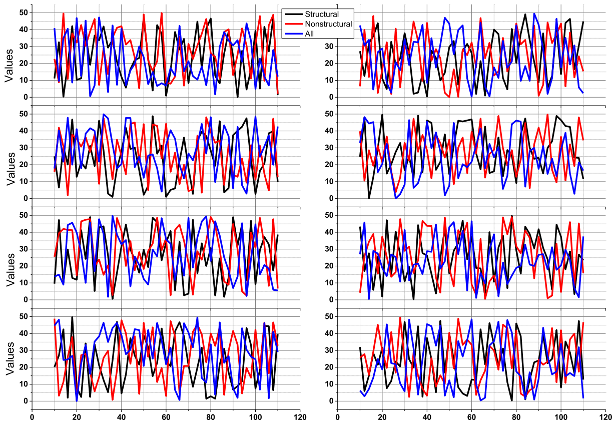

FIGURE 11. Analysis and comparison of type features.

Among them, the abscissa of each subgraph represents different thresholds, that is, the prediction results are divided in descending order of probability; the ordinate of the subgraph is the number of overlaps between the prediction results in the corresponding interval of the abscissa and the CGC database, and the overlap value is used as the structural feature And non-structural features affect the evaluation index of the algorithm. It can be found from the figure that except for UCEC, the overall prediction effect of structural features is better than non-structural features. On the one hand, it shows that structural features can characterize genes well. On the other hand, it also verifies the feasibility of the research direction of the paper. Mining gene characteristics from the perspective of network structure.

Based on the above analysis, it is found that during the development of colorectal cancer, the network structure of the gene has changed significantly, and the prediction results based on the network structure show better prediction results than the non-network structure. The feasibility of the research direction of the thesis is demonstrated, and the relevant analysis results are also given in the article. However, the research method of the thesis is based on network research, so comparing structural features with non-structural features only shows that structural features have a good classification ability, and cannot directly explain that modeling using gene networks is better than not using gene networks.

In the study, the structural features of the extracted gene network and the attribute features of the gene and the integrated features of the network and attribute were fused to optimize the predictive ability of colorectal cancer driver genes, and this fusion feature was used to train the model. Finally, compare the prediction results of the model using the fusion feature training with the prediction results of the model training with non-structural features, that is, the method using the gene network and the method not using the gene network, as shown in Figure 11, the red solid line The prediction result of the gene network, the yellow dotted line is the prediction result of the unused gene network. The results in Figure 11 show that the red solid line is basically above the yellow dashed line, indicating that the method using the gene network is superior to the method not using the gene network. Not only is that, but the overall result of the solid red line better than the dashed blue line. Therefore, after feature fusion, it can not only improve the algorithm's ability to predict colorectal cancer driver genes, but also make up for the shortcomings of structural features or non-structural features, and achieve the effect of complementary features.

C. DIAGNOSIS AND PREDICTION

There have been studies using this method to measure the prediction performance of the algorithm. According to the model, the probability that the unknown gene is the driving gene is given, and the probability is sorted in descending order according to the probability. The overlapping ratio between the top 10 genes and the CGC database is calculated. Then, compare with the 20/20 + algorithm, as shown in Figure 12. The results show that the overlapping ratio of the predicted results of the seven types of colorectal cancer and CGC can reach 40% or more, and the ratio of this result is quite high.

Among them, the abscissa represents the type of colorectal cancer, and each colorectal cancer only considers the top 10 prediction results; the ordinate represents the ratio of the number of overlaps between the prediction results and the CGC database in the prediction results; the red dotted line represents the 20/20 + algorithm prediction Of the average.

According to the results in Figure 13, it can be found that the prediction results achieved by the colorectal cancer driver gene mining algorithm proposed in this paper can be comparable to the most advanced 20/20 + algorithm. In order to further analyze whether other genes predicted to be driving genes have related research proofs, and whether the prediction results of the algorithm can reflect the current research trends, the paper analyzes the top 100 genes of each colorectal cancer prediction result, and the results are shown in Figure 13 As shown.

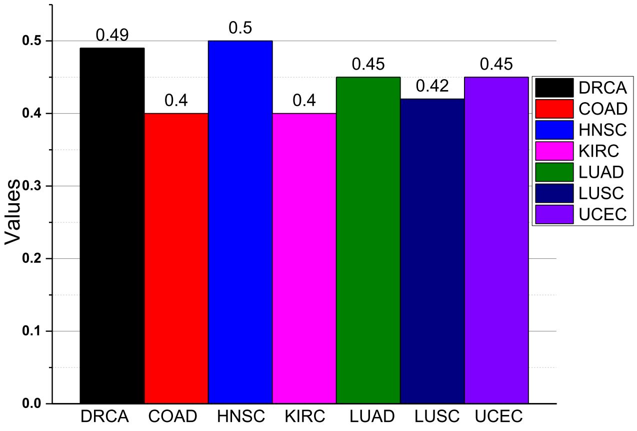

FIGURE 12. **Proportion of genes in the colorectal cancer driver gene prediction results in the CGC database.**

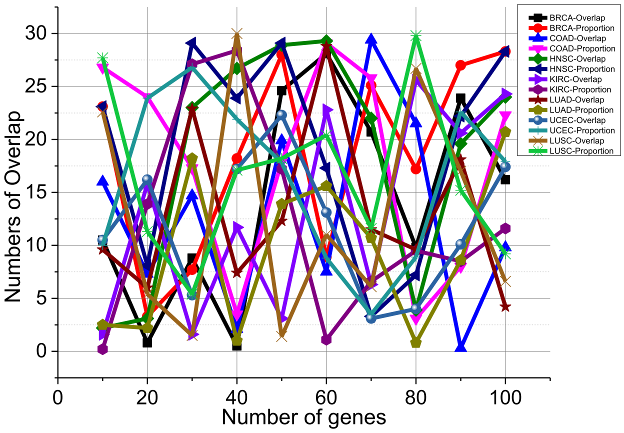

FIGURE 13. **The number and ratio of overlaps between the prediction results of colorectal cancer and the CGC database.**

In addition to comparing the algorithm prediction results with the CGC database above, this article also uses the NCBI (PubMed) database to search for relevant literature for unknown genes that do not appear in the CGC but have a higher prediction probability, to further evaluate the performance of the algorithm in this paper. There are many genes that have not been collected in the CGC database that have been or are being reported in the relevant literature, especially breast cancer and lung cancer (lung adenocarcinoma, lung squamous cell carcinoma). The reason is that lung cancer and breast cancer have always been the research hotspots in the field of biomedicine, and lung cancer and breast cancer are also the first onset of tumors in men and women, and are also the main tumor diseases in countries around the world. It is reasonable that cancer and lung cancer have better prediction effects.

This chapter is mainly based on complex network and machine learning methods to model the colorectal cancer driver gene mining algorithm, and to describe some details in the modeling process, such as missing data processing, standardization of features, negative sample construction problems, Data set sampling, and model training. Then, analyze and discuss the prediction results of the algorithm, including algorithm evaluation and feature analysis. Among them, algorithm evaluation is mainly to compare the performance of the algorithm proposed in the paper with the SVM and 20/20 + algorithms; feature analysis mainly includes feature importance analysis, comparative analysis of structural features and non-structural features, and the use of genetic networks and No comparative analysis of gene networks was used.

Through the research and analysis in this chapter, it is found that (1) the network characteristics of genes extracted from the paper, the attribute characteristics of genes, and the integrated characteristics of networks and attributes can all play a good role in classification. Among them, degree centrality, poly the class coefficient, mutation frequency and K-shell value contribute the most to gene classification. (2) The analysis of structural features and non-structural features found that structural features have a better classification effect than non-structural features, indicating that structural features have good classification capabilities, and also verified the network analysis in gene mining research. (3) The fusion feature of structural features and non-structural features can achieve better prediction results, indicating that the method using gene networks is significantly better than the method without using gene networks, and structural features improve the predictive ability of the algorithm, non-structural features. Then the prediction result of the optimization algorithm is optimized. (4) Through algorithm comparison, it is found that the NRFD algorithm proposed in this paper is superior to the SVM and 20/20 + algorithms, and the NRFD algorithm has better generalization ability, and can obtain better results in seven different types of colorectal cancer.

V. CONCLUSION

This paper is based on complex networks and machine learning methods. It mainly uses gene mutation data, gene expression data and gene correlation data to construct feature vectors of gene samples, and is used to learn classification models for pattern recognition and to mine seven colorectal cancer-related Potential driver genes. During the modeling process, detailed research was conducted on the complex structure of the network, feature mining, algorithm performance and prediction results. For the analysis of the results, this paper first compares the performance of the algorithm with the SVM algorithm through the ROC curve and AUC value, and then compares the genes predicted by the algorithm with the genes in the CGC database to verify the prediction results of the model, and in seven types of It has been verified in colorectal cancer. The main goal of the research on colorectal cancer driver gene mining algorithm based on complex network and machine learning method is to predict the potential colorectal cancer driver gene, and evaluate the model through the prediction result of the algorithm. In addition, this paper also analyzes the importance of individual features, the importance of structural features and non-structural features, as well as the methods using gene networks and methods not using gene networks. The purpose is not only to verify the usefulness of extracting features, but to mine It is an important indicator to measure the development and changes of colorectal cancer, and it also proves that using gene network methods to mine driver genes is significantly better than mining methods without using gene networks. In this paper, we use deep learning technology to study colorectal cancer pathogenic gene screening, propose a new method for colorectal cancer diagnosis, more scientific and accurate colorectal cancer pathogenic gene screening, provide technical and data guidance for colorectal cancer diagnosis, improve diagnosis efficiency, and provide other relevant Provide a reference method for diagnosing cases. The discovery of these conclusions has important guiding significance for the study of colorectal cancer driver genes.

REFERENCES

[1] I. Nazari, H. Tayara, and K. T. Chong, "Branch point selection in RNA splicing using deep learning," *IEEE Access*, vol. 7, pp. 1800–1807, 2019.

[2] M. Rath, S. E. Jenssen, K. Schwefel, S. Spiegler, D. Kleimeier, C. Sperling, L. Kaderali, and U. Felbor, "High-throughput sequencing of the entire genomic regions of CCM1/KRIT1, CCM2 and CCM3/PDCD10 to search for pathogenic deep-intronic splice mutations in cerebral cavernous malformations," *Eur. J. Med. Genet.*, vol. 60, no. 9, pp. 479–484, Sep. 2017.

[3] M. Chiara, I. Primon, L. Tarantini, L. Agnelli, V. Brancaleoni, F. Granata, V. Bollati, and E. Di Pierro, "Targeted resequencing of FECH locus reveals that a novel deep intronic pathogenic variant and eQTLs may cause erythropoietic protoporphyria (EPP) through a methylation-dependent mechanism," *Genet. Med.*, vol. 22, no. 1, pp. 35–43, Jan. 2020.

[4] S. Das, K. Fearnside, S. Sarker, J. K. Forwood, and S. R. Raidal, "A novel pathogenic aviadenovirus from red-bellied parrots (poicephalus rufiventris) unveils deep recombination events among avian host lineages," *Virology*, vol. 502, pp. 188–197, Feb. 2017.

[5] R. Vaz-Drago, N. Custódio, and M. Carmo-Fonseca, "Deep intronic mutations and human disease," *Human Genet.*, vol. 136, no. 9, pp. 1093–1111, Sep. 2017.

[6] C. L. Alston, M. T. Veling, J. Heidler, L. S. Taylor, J. T. Alaimo, A. Y. Sung, L. He, S. Hopton, A. Broomfield, J. Pavaine, J. Diaz, E. Leon, P. Wolf, R. McFarland, H. Prokisch, S. B. Wortmann, P. E. Bonnen, I. Wittig, D. J. Pagliarini, and R. W. Taylor, "Pathogenic bi-allelic mutations in NDUFAF8 cause leigh syndrome with an isolated complex i deficiency," *Amer. J. Human Genet.*, vol. 106, no. 1, pp. 92–101, Jan. 2020.

[7] S. Chakraborty, M. Britton, P. J. Martínez-García, and A. M. Dandekar, "Deep RNA-seq profile reveals biodiversity, plant–microbe interactions and a large family of NBS-LRR resistance genes in walnut (Juglans regia) tissues," *AMB Express*, vol. 6, no. 1, p. 12, Dec. 2016.

[8] H. L. Schulz, "Mutation spectrum of the ABCA4 gene in 335 stargardt disease patients from a multicenter german cohort-impact of selected deep intronic variants and common SNPs," *Investigative Ophthalmol. Vis. Sci.*, vol. 58, no. 1, pp. 394–403, 2017.

[9] K. M. Knapp, R. Sullivan, J. Murray, G. Gimenez, P. Arn, P. D'Souza, A. Gezdirici, W. G. Wilson, A. P. Jackson, C. Ferreira, and L. S. Bicknell, ''Linked-read genome sequencing identifies biallelic pathogenic variants in DONSON as a novel cause of meier-gorlin syndrome,'' *J. Med. Genet.*, vol. 57, no. 3, pp. 195–202, Mar. 2020.

[10] A. B. Gussow, S. Petrovski, Q. Wang, A. S. Allen, and D. B. Goldstein, ''The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes,'' *Genome Biol.*, vol. 17, no. 1, p. 9, Dec. 2016.

[11] G. Fu, Y. Wei, X. Wang, and L. Yu, ''Identification of candidate causal genes and their associated pathogenic mechanisms underlying teratozoospermia based on the spermatozoa transcript profiles,'' *Andrologia*, vol. 48, no. 5, pp. 576–583, Jun. 2016.

[12] H. Dubey, K. Kiran, R. Jaswal, P. Jain, A. M. Kayastha, S. C. Bhardwaj, T. K. Mondal, and T. R. Sharma, ''Discovery and profiling of small RNAs from puccinia triticina by deep sequencing and identification of their potential targets in wheat,'' *Funct. Integrative Genomics*, vol. 19, no. 3, pp. 391–407, May 2019.

[13] R. M. de Almeida, J. Tavares, S. Martins, T. Carvalho, F. J. Enguita, D. Brito, M. Carmo-Fonseca, and L. R. Lopes, ''Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients with hypertrophic cardiomyopathy,'' *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0182946.

[14] H. Guo, ''Novel pathogenic genes and rare mutations of congenital factor V deficiency: A case report,'' *Blood*, vol. 134, p. 4920, Oct. 2019.

[15] G. D. Fruscio, ''Are all the previously reported genetic variants in limb girdle muscular dystrophy genes pathogenic,'' *Eur. J. Hum. Genet.*, vol. 24, no. 1, pp. 73–77, 2016.

[16] Y. Ding, P. Liu, S. Zhang, L. Tao, and J. Han, ''Screening pathogenic genes in oral squamous cell carcinoma based on the mRNA expression microarray data,'' *Int. J. Mol. Med.*, vol. 10, pp. 3597–3603, Feb. 2018.

[17] X. Lin, Y. Liu, J. Deng, Y. Lyu, P. Qian, Y. Li, and S. Wang, ''Multiple advanced logic gates made of DNA-ag nanocluster and the application for intelligent detection of pathogenic bacterial genes,'' *Chem. Sci.*, vol. 9, no. 7, pp. 1774–1781, 2018.

[18] M. Duan, H. Li, J. Gu, X. Tuo, W. Sun, X. Qian, and X. Wang, ''Effects of biochar on reducing the abundance of oxytetracycline, antibiotic resistance genes, and human pathogenic bacteria in soil and lettuce,'' *Environ. Pollut.*, vol. 224, pp. 787–795, May 2017.

[19] B. Xia, Y. Li, J. Zhou, B. Tian, and L. Feng, ''Identification of potential pathogenic genes associated with osteoporosis,'' *Bone Joint Res.*, vol. 6, no. 12, pp. 640–648, Dec. 2017.

[20] F. Fakhouri, M. Fila, F. Provát, Y. Delmas, C. Barbet, V. Châtelet, C. Rafat, M. Cailliez, J. Hogan, A. Servais, A. Karras, R. Makdassi, F. Louillet, J.-P. Coindre, E. Rondeau, C. Loirat, and V. Frémeaux-Bacchi, ''Pathogenic variants in complement genes and risk of atypical hemolytic uremic syndrome relapse after Eculizumab discontinuation,'' *Clin. J. Amer. Soc. Nephrology*, vol. 12, no. 1, pp. 50–59, Jan. 2017.

[21] W. Xia, F. Liu, and D. Ma, ''Research progress in pathogenic genes of hereditary non-syndromic mid-frequency deafness,'' *Frontiers Med.*, vol. 10, no. 2, pp. 137–142, Jun. 2016.

[22] J. W. Byun, ''O-Serogroups, Virulence Genes of Pathogenic Escherichia Coli and Pulsed-Field Gel Electrophoresis (PFGE) Patterns of O149 Isolates from Diarrhoeic Piglets in Korea,'' *Veterinarni Medicina*, vol. 58, no. 9, pp. 468–476, 2018.

[23] C. Östlund, W. Chang, G. G. Gundersen, and H. J. Worman, ''Pathogenic mutations in genes encoding nuclear envelope proteins and defective nucleocytoplasmic connections,'' *Experim. Biol. Med.*, vol. 244, no. 15, pp. 1333–1344, Nov. 2019.

[24] X. Liu, W. Chen, W. Li, J. R. Priest, Y. Fu, K. Pang, B. Ma, B. Han, X. Liu, S. Hu, and Z. Zhou, ''Exome-based case-control analysis highlights the pathogenic role of ciliary genes in transposition of the great arteries,'' *Circulat. Res.*, vol. 126, no. 7, pp. 811–821, Mar. 2020.

[25] J. Feng and J. Xu, ''Identification of pathogenic genes and transcription factors in glaucoma,'' *Mol. Med. Rep.*, vol. 20, pp. 216–224, May 2019.

[26] J. Feng, Q. Zhou, W. Gao, Y. Wu, and R. Mu, ''Seeking for potential pathogenic genes of major depressive disorder in the gene expression omnibus database,'' *Asia–Pacific Psychiatry*, vol. 12, no. 1, pp. 617–684, Mar. 2020.

[27] M. Bauwens, ''ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: Novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants,'' *Genet. Med.*, vol. 21, no. 8, pp. 1761–1771, Aug. 2019.

[28] M. Ishii, Y. Matsumoto, I. Nakamura, and K. Sekimizu, ''Silkworm fungal infection model for identification of virulence genes in pathogenic fungus and screening of novel antifungal drugs,'' *Drug Discoveries Therapeutics*, vol. 11, no. 1, pp. 1–5, 2017.

[29] Z. Li, C. Zhou, L. Tan, P. Chen, Y. Cao, X. Li, J. Yan, H. Zeng, D.-W. Wang, and D.-W. Wang, ''A targeted sequencing approach to find novel pathogenic genes associated with sporadic aortic dissection,'' *Sci. China Life Sci.*, vol. 61, no. 12, pp. 1545–1553, Dec. 2018.

[30] L. Zhang, H. Qin, Z. Wu, W. Chen, and G. Zhang, ''Pathogenic genes related to the progression of actinic keratoses to cutaneous squamous cell carcinoma,'' *Int. J. Dermatol.*, vol. 57, no. 10, pp. 1208–1217, Oct. 2018.

[31] B. Xu, Y. Liu, S. Yu, L. Wang, L. Liu, H. Lin, Z. Yang, J. Wang, and F. Xia, ''Multipath2vec: Predicting pathogenic genes via heterogeneous network embedding,'' in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 951–956.

[32] W. G. Alexander, J. H. Wisecaver, A. Rokas, and C. T. Hittinger, ''Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides,'' *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 15, pp. 4116–4121, Apr. 2016.

[33] X.-L. Yue and Z.-Q. Gao, ''Identification of pathogenic genes of pterygium based on the gene expression omnibus database,'' *Int. J. Ophthalmol.*, vol. 12, no. 4, pp. 529–535, 2019.

[34] J. Zhang, R. Cheng, J. Liang, C. Ni, M. Li, and Z. Yao, ''Lentiginous phenotypes caused by diverse pathogenic genes (SASH1andPTPN11): Clinical and molecular discrimination,'' *Clin. Genet.*, vol. 90, no. 4, pp. 372–377, Oct. 2016.

[35] B. Zhao, M. Wang, J. Xu, M. Li, and Y. Yu, ''Identification of pathogenic genes and upstream regulators in age-related macular degeneration,'' *BMC Ophthalmol.*, vol. 17, no. 1, p. 102, Dec. 2017.

[36] S. Xu and P. Ning, ''Predicting pathogenic genes for primary myelofibrosis based on a system–network approach,'' *Mol. Med. Rep.*, vol. 17, pp. 186–192, Oct. 2017.

[37] A. J. L. Brambila-Tapia, A. C. Poot-Hernández, E. Perez-Rueda, and K. Rodríguez-Vázquez, ''Identification of DNA methyltransferase genes in human pathogenic bacteria by comparative genomics,'' *Indian J. Microbiol.*, vol. 56, no. 2, pp. 134–141, Jun. 2016.

[38] X.-Z. Liu, ''Simultaneous expression of two pathogenic genes in four chinese patients affected with inherited retinal dystrophy,'' *Int. J. Ophthalmol.*, vol. 13, no. 2, pp. 220–230, Feb. 2020.

[39] Z. Yue, H. Lin, M. Li, H. Wang, T. Liu, M. Hu, H. Chen, H. Tong, and L. Sun, ''Clinical and pathological features and varied mutational spectra of pathogenic genes in 55 chinese patients with nephronophthisis,'' *Clinica Chim. Acta*, vol. 506, pp. 136–144, Jul. 2020.

[40] T. Rodriguez, ''Phylogenetic considerations in the evolutionary development of aminoglycoside resistance genes in pathogenic bacteria,'' *J. Phylogenetics Evol. Biol.*, vol. 04, no. 01, pp. 4–24, 2016.

[41] M. M. Amer, H. M. Mekky, A. M. Amer, and H. S. Fedawy, ''Antimicrobial resistance genes in pathogenic escherichia coli isolated from diseased broiler chickens in egypt and their relationship with the phenotypic resistance characteristics,'' *Veterinary World*, vol. 11, no. 8, pp. 1082–1088, Aug. 2018.

[42] R. Aghdasi-Araghinezhad and K. Amini, ''Study of antibiotic resistance pattern and incidence of pathogenic genes of MgtC, Spi4R, AgfA, InvE/A and TtrC in salmonella Infantis isolated from clinical specimens,'' *KAUMS J.*, vol. 21, no. 5, pp. 443–449, 2017.

[43] Y. Zhang, T. Zhang, and Y. Chen, ''Comprehensive analysis of gene expression profiles and DNA methylome reveals oas1, ppie, Polr2g as pathogenic target genes of gestational diabetes mellitus,'' *Sci. Rep.*, vol. 8, no. 1, pp. 6–10, Dec. 2018.

[44] Y. Lei, P. Guo, J. An, C. Guo, F. Lu, and M. Liu, ''Identification of pathogenic genes and upstream regulators in allergic rhinitis,'' *Int. J. Pediatric Otorhinolaryngol.*, vol. 115, pp. 97–103, Dec. 2018.

[45] J. Feng and J. Xu, ''Identification of pathogenic genes and transcription factors in Osteosarcoma,'' *Mol. Med. Rep.*, vol. 20, pp. 3–8, May 2019.

[46] A. Backes, ''Expression analysis of cell wall-related genes in the plant pathogenic fungus drechslera teres,'' *Genes*, vol. 11, no. 3, pp. 300–351, 2020.

[47] K. Heidari, ''Prevalence of pathogenic genes CagA and VacA of helicobacter pylori isolated in patients with digestive disorders,'' *Iranian J. Med. Microbiol.*, vol. 13, no. 1, pp. 80–88, 2019.

[48] D. Zhang, Y. Zhou, D. Zhao, J. Zhu, Z. Yang, and M. Zhu, ''Complete genome sequence and pathogenic genes analysis of pectobacterium atroseptica JG10-08,'' *Genes Genomics*, vol. 39, no. 9, pp. 945–955, Sep. 2017.

[49] S. Zhang, Z. Tong, H. Yin, and Y. Feng, ''Pathogenic genes selection model of genetic disease based on network motifs slicing feedback,'' *Current Proteomics*, vol. 16, no. 5, pp. 392–401, Jul. 2019.

[50] S. Kalteh, ''Investigating the possibility of the listeria monocytogenes entering into a viable but non-culturable (VBNC) form and expression of the pathogenic genes during the frozen storage of (-18°C) rainbow trout fish nugget,'' *Iranian J. Med. Microbiol.*, vol. 13, no. 1, pp. 69–79, Mar. 2019.

[51] W. W. Deng, ''Whole Genome Sequencing Reveals the Distribution of Resistance and Virulence Genes of Pathogenic Escherichia Coli CCHTP from Giant Panda,'' *Hereditas*, vol. 41, no. 12, pp. 1138–1147, 2019.

[52] J. M. Rossato, B. G. Brito, R. K. T. Kobayashi, V. L. Koga, J. J. P. Sarmiento, G. Nakazato, L. F. D. Lopes, L. A. G. Balsan, T. T. Grassotti, and K. C. T. Brito, ''Antimicrobial resistance, diarrheagenic and avian pathogenic virulence genes in escherichia coli from poultry feed and the ingredients,'' *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, vol. 71, no. 6, pp. 1968–1976, Dec. 2019.

[53] D.-W. Lee, L.-J. Jun, and J.-B. Jeong, ''Distribution of tetracycline resistance genes in pathogenic bacteria isolated from cultured olive flounder (Paralichthys olivaceus) in jeju in 2016,'' *J. FISHRIES Mar. Sci. Edu.*, vol. 29, no. 3, pp. 834–846, Jun. 2017.

[54] O. Albarria, ''Molecular screening of sideropohore genes in extraintestinal pathogenic escherichia coli isolated from clinical and escherichia coli isolated food samples in turkey,'' *Pyrex J. Biomed. Res.*, vol. 5, no. 1, pp. 5–13, 2019.

[55] A. A. El Tawab, F. El-Hofy, A. Ammar, and N. A. Galil, ''Molecular screening of virulence genes in avian pathogenic esherichia coli,'' *Benha Veterinary Med. J.*, vol. 30, no. 1, pp. 137–149, Mar. 2016.

YANKE LI graduated from the Clinical Medicine of China Medical University, in 2009. He was with the Department of Anorectal Surgery, China Medical University First Hospital. His research interest includes colorectal oncology.

FUQIANG ZHANG graduated from the Clinical Medicine of Shenyang Medical College, in 2015. He studied at the Department of Anorectal Surgery, China Medical University First Hospital. His research interest includes colorectal oncology.

CHENGZHONG XING graduated from the Clinical Medicine of China Medical University, in 1987. He was with the Department of Anorectal Surgery, China Medical University First Hospital. His research interest includes colorectal oncology.