Bacterial signatures for diagnosis of colorectal cancer by machine learning

Kexin Yuan [1,†,‡*], Xuexinyi Chen [2,‡] , Xinru Zhu [3,],Junlu Wang[4,] ,Tianzi Li[5,],Yun Li[4,]

[1] Qian Weichang College, Shanghai University, Shanghai 200444, China
[2] School of Chinese Medicine·School of Integrative Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China
[3] Medical School·Integrated Medical School, Nanjing University of Chinese Medicine, Nanjing 210023, China
[4] College of Pharmacy, Nanjing University of Chinese Medicine, Nanjing 210023, China
[5] College of Life Sciences, Inner Mongolia University, Hohhot 010020, China
* Correspondence: kiyoblingrank@163.com; Tel.: +86-1805605684
† Current address: Shanghai University, Shanghai 200444, China
‡ These authors contributed equally to this work.

Abstract: Invasive methods such as colonoscopy are more commonly used in colorectal cancer (CRC) screening and diagnosis, but these methods are not easily accepted and have limitations. In this paper, we aim to exploit the close relationship between intestinal flora and the development of CRC. A T-test was used to screen and compare the intestinal flora of healthy individuals and patients, and strains with significant differences were selected as characteristic ones. In addition, three AI learning models, Random forest (RF), K-Nearest Neighbor (KNN), and Back propagation neural network (BPNN), were used to build a colorectal cancer diagnosis model based on intestinal flora. Overall, the investigation carried out by us has revealed six highly divergent species between healthy individuals and patients from t-tests and key species associated with CRC. The results were validated against each other, confirming the reliability of the obtained key strains, and providing a new idea for the clinical diagnosis of CRC.

Keywords: colorectal cancer; bacterial signatures; machine learning; clinical diagnosis.

1 Project Summary

Colorectal cancer is a common malignant tumour of the gastrointestinal tract that occurs in the colon [1]. It has been reported that 95% of all diseases are related to intestinal flora [2]. Under normal conditions, the intestinal flora can maintain a dynamic ecological balance with the host and the external environment [3,4].

In this project, we screened and compared the intestinal flora of healthy individuals and patients using the t-test to screen out strains with significant characteristic differences. In addition, three AI learning models, Random forest (RF), K-Nearest Neighbor (KNN), and Back propagation neural network (BPNN), were used to build a colorectal cancer diagnosis model based on intestinal flora, and their accuracy was evaluated. The results showed that RF had the highest accuracy in determining disease or not (0.792), which was higher than that of BPNN (0.667) and KNN (0.542). The colorectal cancer diagnosis model established with RF was better fitted.

The model can be used to assist in the diagnosis of colorectal cancer and to further explore applications of artificial intelligence, such as the use of random forest algorithms to predict the anti-cancer effects of colorectal cancer drugs.

2 Project Results

The results are divided into three sections: data, algorithm, and discussion. The data part includes the selection of the database, the process of data pre-processing and the data visualization methods, and the presentation of the results. The algorithm part includes the selection of the algorithm, the results of the algorithm, and its results analysis and evaluation. The discussion part includes a discussion of AI applications, deficiencies, and future prospects.

2.1 Data processing and visualization

2.1.1 Data set

The data for this study were obtained from the GMrepo Gut Genomics online repository (https://gmrepo.humangut.info/home)[5]. A total of 79 data items were collected, including indicators of intestinal flora, viruses, and archaea that may influence the development of colorectal cancer. As the data were concentrated in Japan and the subjects were mostly between 50 and 80 years of age with a body mass index concentrated between 20 and 25, the effect of age and weight on colorectal cancer was not considered for the time being.

2.1.2 Data pre-processing

(i) Data screening and supplementation

Firstly, we deleted the viral and archaea parts of the database and studied only the intestinal flora. At the same time, in order to observe whether age, body mass index, and sex had any effect on the occurrence of colorectal cancer, we added the relevant information to each sample for subsequent analysis.

(ii) Non-numerical data coding

Information such as gender and whether or not they had colorectal cancer was not coded in the original database, making the information impossible to analyse. We coded the information on gender and whether or not they had colorectal cancer, using 0 / 1 to distinguish between different genders and whether or not they had colorectal cancer, for subsequent analysis.

2.1.3 Visualisation of results

After data pre-processing, in order to visualize the differences between the intestinal flora of healthy people and colorectal cancer patients, we used a line graph to visualize the flora abundance data, and the results are shown in Figure 1. We calculated the mean abundance of the six flora in the healthy population and in patients with colorectal cancer, and the slope of the line shows the difference in abundance between the two groups.

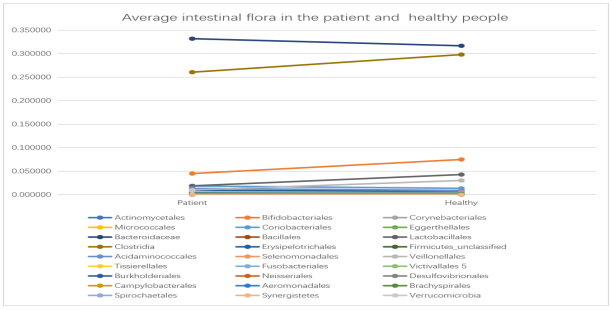


Figure 1: Average intestinal flora in patients and healthy people

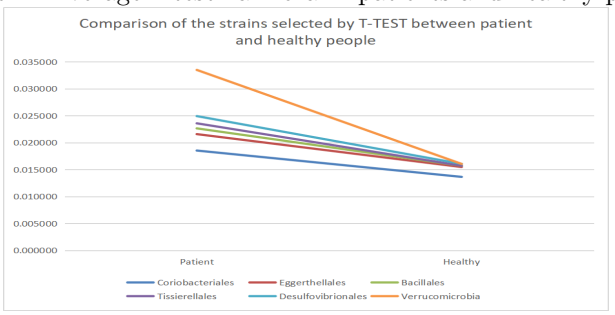


Figure 2: Comparison of the strains selected by the T-TEST between patients and healthy people

Based on the t-test, we made violin plots of the resulting differential strains to better show the distribution of the differential strains between the two populations, as shown in Figure 3.
The distribution of Desulfovibrionales in the healthy population is shown in the lower half of the graph, indicating that the distribution of this organism in the intestine of the healthy population is low in proportion to the overall distribution. The graph of the distribution of Desulfovibrionales in colorectal cancer patients is longer than that of the healthy population, and the difference between the maximum and minimum values is large, indicating that the distribution of Desulfovibrionales in the intestine of colorectal cancer patients is not uniform. The part of the graph that is higher than that of the healthy population indicates that some colorectal cancer patients are over-represented and significantly different from the healthy population, so we speculate that this strain may be used as an indicator for colorectal cancer. Other violin diagrams can be used as examples for related analyses.
The ability of certain strains of bacteria to diagnose colorectal cancer has been confirmed in some studies, but the accuracy of the diagnosis of colorectal cancer is yet to be confirmed, as they are mostly single strains of bacteria [6].

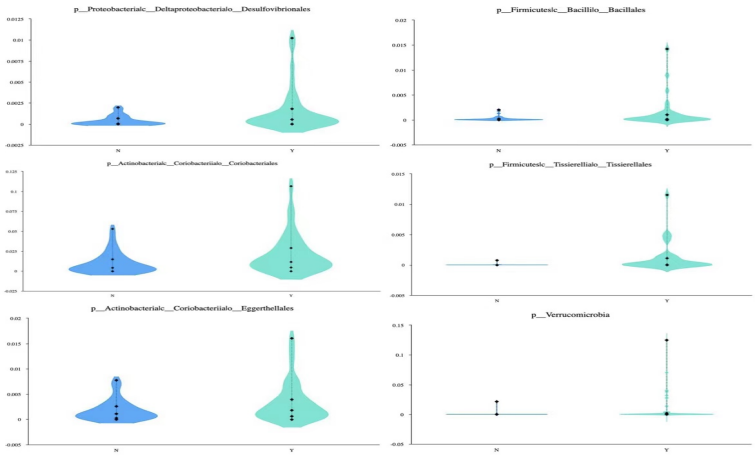


Figure 3: Violin diagram of six differential strains

2.2 Algorithm and Modeling

2.2.1 Modeling

In building disease prediction models, specific machine learning algorithms need to be compared and selected for specific problem scenarios and different performance requirements. However, there has been a lack of general guidelines for algorithm selection, and experimental comparisons are usually required to select the optimal algorithm for the current problem [7]. Therefore, we selected three classical machine learning algorithms, including Random Forest, K-Nearest Neighbor, and Back Propagation, for the prediction of CRC. We used 70% of the dataset's data volume as each algorithm's training set and the remaining 30% as the test set.

2.2.2 Model evaluation

(i)Model evaluation indexes

Accuracy, precision, recall, and f1 were used to evaluate the performance of each model. The number of positive samples is denoted as P, the number of negative samples is denoted as N, and the number of positive cases correctly predicted is denoted as TP, the number of negative samples predicted as positive samples are denoted as FP, the number of positive samples predicted as negative samples are denoted as FN, and the number of negative samples correctly predicted is denoted as TN [8]. The calculation formula for each indicator is as follows:

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{Recall}=\frac{TP}{TP+FP}$$

$$\text{Precision}=\frac{TP}{FN+TP}$$

$$\text{F1}=\frac{2TP}{2TP+FN+FP}$$

The confusion matrix (Table 1) is widely used in the performance evaluation of gut flora-based disease prediction models [9].

Predict	Actual (True)	Actual (False)
Predict (True)	TP (True positive)	FP (False positive)
Predict (False)	FN (False negative)	TN (True negative)

Table 1 Confusion matrix

(ii) Analysis of results

Model	accuracy	precision	recall	f1
RF	0.792	0.875	0.636	0.737
KNN	0.542	0.542	0.563	0.499
BP	0.667	0.667	0.688	0.657

Table 2 Initial Model Evaluation Results

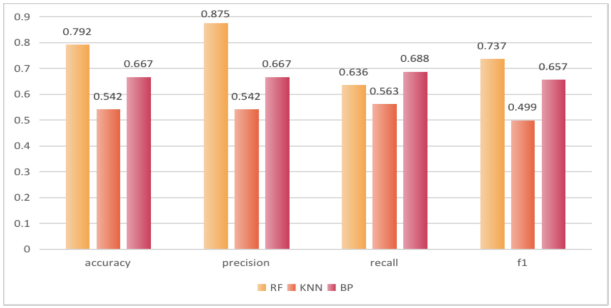


Figure 4: Histogram of initial Model Evaluation Results

In this project, we wanted to minimize the number of missed diagnoses for predicting colorectal cancer, including other medical diagnoses. Therefore, accuracy is the most important performance index to evaluate the quality of the model in this experiment.

Algorithm	Accuracy
RF	0.792
KNN	0.542
BP	0.667

Table 3 Comparison of classification accuracy results

Figure 5: Histogram of accuracy

Based on the classification accuracy, we found that the Random Forest algorithm had the best performance compared with other algorithms applied to this dataset.

(iii) Further analysis

We selected classification prediction accuracy for the model output performance evaluation index. According to the performance of the test set of each algorithm, a confusion matrix was established to evaluate the model's actual classification and prediction ability. The confusion matrix was drawn up using a heat map, and the heat map was set to a darker color for larger values to clearly see the predicted effect of each class.

Figure 6 shows the confusion matrix constructed for the three algorithms, and their performances on the test set are compared. Among them, the corresponding accuracy rate of RF is the highest, the number of incorrectly predicted samples is 6, the precision rate is large, and the recall rate has high credibility. It has a powerful capability of distinguishing between positive and negative samples, as well as the recognition of positive samples. The ability of BP and KNN to identify positive samples and distinguish the difference between positive and negative samples is poor, and the number of incorrectly predicted samples is 9 and 11.

The darkest color on the diagonal of the confusion matrix of RF indicates that the prediction is more correct.

The figure shows the confusion matrix for the test set of RF. 8, in the upper left corner of the figure, represents 8 subjects predicted to be patients, 4 in the lower left corner represents 4 subjects misidentified as healthy, 2 in the upper right corner are represented as healthy people misidentified as patients and 10 in the

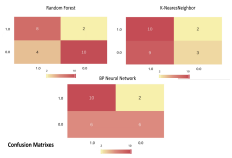


Figure 6: Confusion matrices of three models

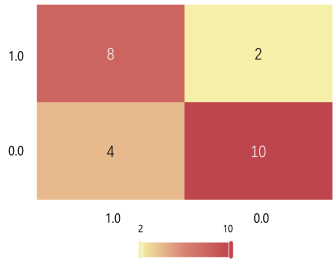


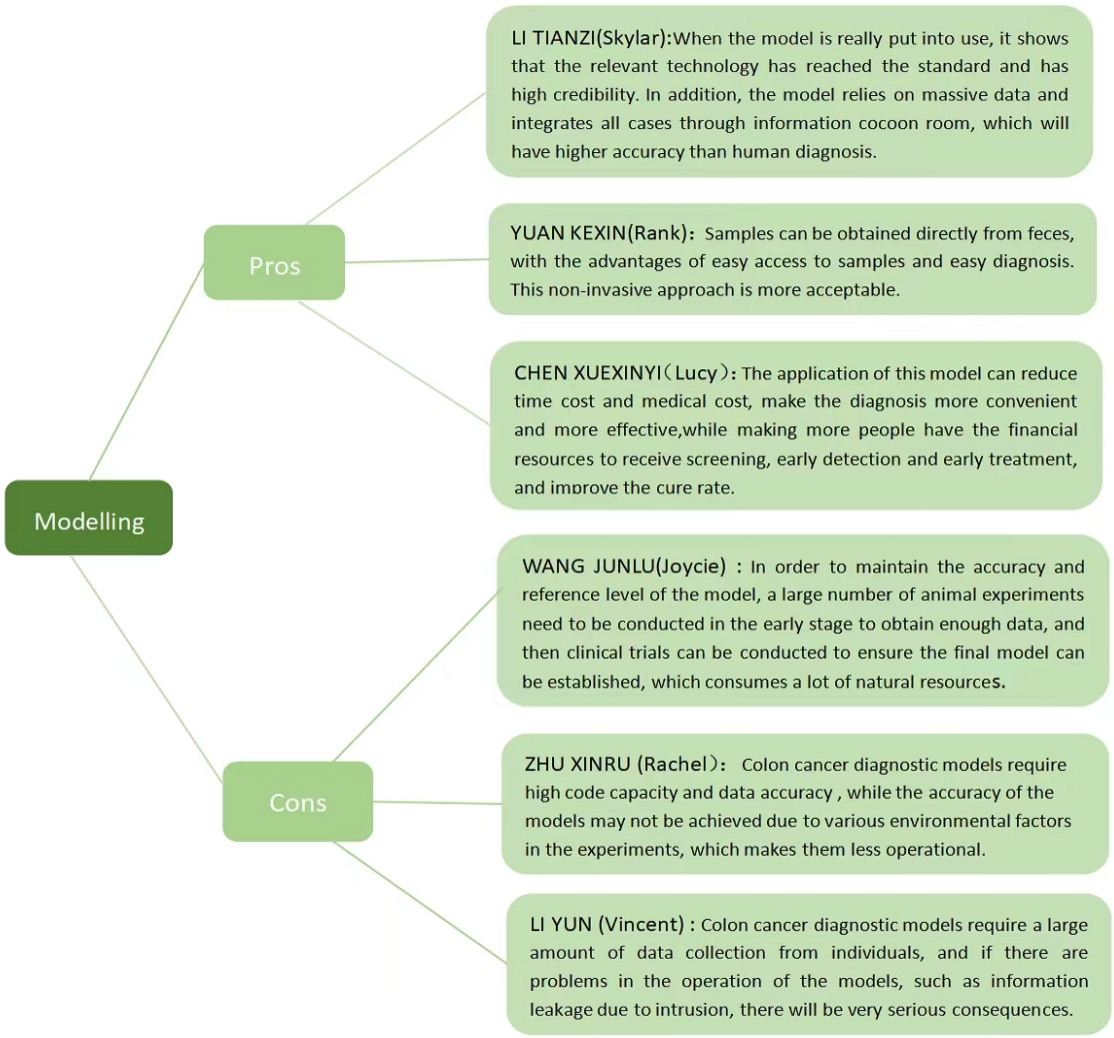
Figure 7: Confusion matrices of RF

lower right corner are represented as healthy people misidentified as unhealthy.

2.3 Discussion

2.3.1 Discussion of AI application

In the future, after further refinement, the diagnostic model for colorectal cancer that we have developed can be put into clinical use and popularized. However, whether people will be willing and able to accept the model diagnosis is still a matter of debate. We discuss his question later.


Modelling
Pros
LI TIANZI(Skylar):When the model is really put into use, it shows that the relevant technology has reached the standard and has high credibility. In addition, the model relies on massive data and integrates all cases through information cocoon room, which will have higher accuracy than human diagnosis.
YUAN KEXIN(Rank): Samples can be obtained directly from feces, with the advantages of easy access to samples and easy diagnosis. This non-invasive approach is more acceptable.
CHEN XUEXINYI（Lucy）: The application of this model can reduce time cost and medical cost, make the diagnosis more convenient and more effective,while making more people have the financial resources to receive screening, early detection and early treatment, and improve the cure rate.
Cons
WANG JUNLU(Joycie) : In order to maintain the accuracy and reference level of the model, a large number of animal experiments need to be conducted in the early stage to obtain enough data, and then clinical trials can be conducted to ensure the final model can be established, which consumes a lot of natural resources.
ZHU XINRU (Rachel)： Colon cancer diagnostic models require high code capacity and data accuracy , while the accuracy of the models may not be achieved due to various environmental factors in the experiments, which makes them less operational.
LI YUN (Vincent) : Colon cancer diagnostic models require a large amount of data collection from individuals, and if there are problems in the operation of the models, such as information leakage due to intrusion, there will be very serious consequences.

2.3.2 Deficiencies and prospects

(i) Deficiencies in the model

1. Many indicators of intestinal flora were measured, while there were fewer healthy people and patients, which may cause prediction bias due to the small sample size.
2. The samples were all from the same hospital in Japan, and the composition of intestinal flora may be different among people in different regions, thus leading to problems of generalizability.

(ii) Prospects

1. It has been reported that 95% of all diseases are related to intestinal flora [10]. Thus, we expect to extend this model to detect whether intestinal flora has an effect on other cancers including brain cancer and breast cancer in the future, and to establish an AI model to determine whether patients have other types of cancers based on differences in intestinal flora strains.
2. The prediction accuracy of the model is 70% to 80%, and there is room for further improvement. We expect to further improve the accuracy of the model by expanding the database, improving the model, and replacing it with a more suitable model in the future so that the accuracy of the model can reach 90%.

3 Project Skills

The project skills section focuses on project management, the individual contribution to the project, and the challenges and solutions encountered during project progression.

3.1 Project Management and Contributions

The first diagram presents the main project nodes, which divide the project into seven parts: Pre-Planning, Project Preparation, Data Collection, Data Collecting, Algorithms and Modeling, Model Evaluation and Summary, as well as the timing and main content of each part. The last column of the chart shows the group leader responsible for each part of the content in the organization of the distribution of work, with each part consisting of 2-3 people working together to complete the task. This chart gives a clear picture of the group's project planning.

Project Timeline Schedule					
Project Node Name	Start Time	End Time	Elapsed Time	Specific Notes	Key Person Responsible
Pre-Planning	2023/1/16	2023/1/25	9	Do background research on colorectal cancer Define the project goal of using AI to build a diagnostic model of the bowel	WANG JUNLLU LI TIANZI
Project Preparation	2023/1/16	2023/3/2	45	Develop an overall project plan Designate the division of labour within the group Determine when the weekly group meetings will take place	YUAN KEXIN
Data Collection	2023/1/26	2023/2/5	10	In vivo strain data from GMrepo repository for healthy and colorectal cancer patients	ZHU XINRU
Phase 1: Data Processing	2023/2/6	2023/2/10	4	Collation of data Selection of key strains by t-test	CHEN XUEXINYI
Phase 2: Algorithms and Modelling	2023/2/11	2023/2/17	6	Random Forest algorithm (accuracy: 0.792) K-Nearest Neighbor (accuracy: 0.542) BP neural network algorithm (accuracy: 0.667)	LI YUN
Model Evaluation	2023/2/11	2023/2/25	14	Comparing the accuracy of models built with various algorithms The Random Forest algorithm was found to be the most accurate	YUAN KEXIN
Summary and Reflections	2023/2/26	2023/3/2	4	Thinking about the use of AI technology under colorectal cancer treatment Summarising project completion Summarising project shortcomings	ALL

DATE	1.26-1.29	1.30-2.1	2.2-2.5	2.6-2.10	2.11-2.14	2.15-2.17	2.18-2.21	2.22-2.25	2.26-3.2
Data Collection									
Phase 1: Data Processing									
Phase 2: Algorithms and Modelling									
Model Evaluation									
Summary and Reflections									

The second chart is a Gantt chart that divides the duration of the 5 main sections from data collection to the project summary, which is also the core of the project. The overall project progression can be seen throughout the table: the sections marked in yellow indicate the duration of each phase, from which the overall progress and time progression of the project can be clearly seen.

3.2 Challenges and solutions

The graph is on page 13.

Challenges, Process, and Solutions		
During the first Panel Discussion, Motivationit was a challenge to understand the scope of the project, which was a time-consuming exercise.	We redefined the project, subdivided the overall task, held frequent group meetings to share progress points and give updates.	With a fuller understanding of the project, the team began to work efficiently and learned how to form relevant questions.
Zero basis to AI	None of our team members had studied machine learning and AI before, and the learning curve was steep as to how to properly apply AI to the project.	Through the first three weeks of literature reading and learning about related machine learning, we finally had an initial clear plan in the fourth week. Then we completed most of the data collection and processing tasks and started to experiment with visualization and algorithms, all the members participated fully.
AI application	We tried to start from a general direction AI and biomedical perspective, but the results were not comprehensive enough.	We re-researched information in an AI perspective related to our topic, and used a tree diagram to sort out AI applications.
Visualization	Our initial visualization was too complicated and not clear enough for audiences to understand.	We tried to reduce the color, reduced the information in the graphic, and labeled the key parts on the slides.
Engaging with the algorithm was difficult	In the first four weeks, the progress of the algorithm group was the slowest. Python kept reporting errors, and the results came out with very low accuracy.	We consulted relevant teachers and classmates, reinstalled and debugged the python environment; we processed the vacancy values, coded the non-numerical data, modified the model. Finally, the accuracy improved.
Time-pressure	From the fifth to the seventh week, five of the six teammates needed to take their final examination. It was hard to get everyone together for group meetings and talk about the project.	We documented each group meeting in detail and update the notes on the trello board to make sure everyone knew the progress of the project.

4 Patents

Author Contributions: Conceptualization: K,Y., R,Z., X, C., J, W., T, L. and Y, L.; literature search: J, W. and T, L.; data processing and visualization: R, Z. and X, C.; algorithms and modeling: K, Y. and Y, L.; writing and reviewing : K,Y., R,Z., X, C., J, W., T, L. and Y, L.; overleaf and editing: J, W. and X, C. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Institutional Review : Biotechnology Engineering and Healthcare Technology, Cambridge Programme 2023
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.
Data Availability Statement: please contact the corresponding author(s) for all reasonable requests for access to the data.
Acknowledgments : We would like to give our thanks to our faculty professors and teachers at Cambridge University for their guidance.
Conflicts of Interest : The authors declare no conflict of interest.

5 References

1. Dekker, E.; Tanis, P. J.; Vleugels, J. L. A.; Kasi, P. M.; Wallace, M. B., Colorectal cancer. Lancet 2019, 394 (10207), 1467-1480.
2. Weitz, J.; Koch, M.; Debus, J.; Hohler, T.; Galle, P. R.; Buchler, M. W., Colorectal cancer. Lancet 2005, 365 (9454), 153-165.
3. Kahrstrom, C. T.; Pariente, N.; Weiss, U., Intestinal microbiota in health and disease. Nature 2016, 535 (7610), 47.
4. Jin, M.; Qian, Z.; Yin, J.; Xu, W.; Zhou, X., The role of intestinal microbiota in cardiovascular disease. J Cell Mol Med 2019, 23 (4), 2343-2350.
5. Dai, D.; Zhu, J. Y.; Sun, C. Q.; Li, M.; Liu, J. X.; Wu, S. C.; Ning, K.; He, L. J.; Zhao, X. M.; Chen, W. H., GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic Acids Res 2022, 50 (D1), D777-D784.
6. Eklof, V.; Lofgren-Burstrom, A.; Zingmark, C.; Edin, S.; Larsson, P.; Karling, P.; Alexeyev, O.; Rutegard, J.; Wikberg, M. L.; Palmqvist, R., Cancer-associated fecal microbial markers in colorectal cancer detection. International Journal of Cancer 2017, 141 (12), 2528-2536.
7. Uddin, S.; Khan, A.; Hossain, M. E.; Moni, M. A., Comparing different supervised machine learning algorithms for disease prediction. Bmc Med Inform Decis 2019, 19 (1).
8. Liu, Z. Q.; Guo, C. G.; Dang, Q.; Wang, L. B.; Liu, L.; Weng, S. Y.; Xu, H.; Lu, T. Y.; Sun, Z. Q.; Han, X. W., Integrative analysis from multi-center studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer. EBioMedicine 2022, 75.
9. Pasolli, E.; Truong, D. T.; Malik, F.; Waldron, L.; Segata, N., Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. Plos Comput Biol 2016, 12 (7).
10. Heinimann, K., [Hereditary Colorectal Cancer: Clinics, Diagnostics and Management]. Ther Umsch 2018, 75 (10), 601-606.