Clinics of Oncology

Research Article ISSN: 2640-1037 | Volume 7

A Machine-Learned Predictor of Colorectal Cancer Based on Metabolomics

Zhang C[1#], Bai Z[2#], Yin Q[1#], Qin J[1], Yang L[1], Zhao X[1], Ji J[1], Liu Y[1], Wang G[4*], Huang X[3*], and Wang Z[1*]

[1]Department of Medical Oncology, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[2]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[3]Department of Anesthesiology, Shanghai ChangZheng hospital, Second Affiliated Hospital of Naval Medical University, Shanghai, China

[4]Department of Critical Care Medicine, The Second People's Hospital of Dongying, Dongying, Shandong, China

[#]These Authors contributed equally to this work

[*]**Corresponding author:**

Zhongqi Wang,

Department of Medical Oncology, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, 725 Wanpingnan Road, Shanghai 200032, China

Xingshuai Huang,

Department of Anesthesiology, Shanghai Chang-Zheng hospital, Second Affiliated Hospital of Naval Medical University, 415 Fengyang Road, Shanghai, 200003, China

Guoying Wang,

Department of Critical Care Medicine, The Second People's Hospital of Dongying, Dongying, Shandong 257300, China

Received: 06 Apr 2024

Accepted: 20 May 2024

Published: 25 May 2024

J Short Name: COO

Citation:

Wang Z, Huang X, Wang G, A Machine-Learned Predictor of Colorectal Cancer Based on Metabolomics. Clin Onco. 2024; 7(10): 1-11

Keywords:

Colorectal cancer; Colorectum

1. Abstract

As for the increasing of prevalence of colorectal cancer (CRC) overworld, it is vital to decrease its morbidity and mortality with early detection. Tests being based on metabolomics are an creative ideal tool for CRC detection. Here, biomarkers to distinguish patients with CRC from those healthy controls (HC) have been explored and validated. Meanwhile, a total of 64 Colorectal cancer as well as healthy control samples were collected from patients who having CRC and HC. Untargeted colorectum metabolite profiling was conducted with Ultra High Performance Liquid Chromatography (UHPLC)-Q-TOF MS. With different kind of machine learning (ML) classifier models,we assessed the discrimination abilities of the quantified metabolites, including Random Foreset (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Support Vector Machine (SVM), The k-Nearest Neighbors (k-NN), Naïve Bayest (NB), and Logistic Regression. Data were divided into training (n =32) and validation datasets (n = 32) randomly. The clustering analysis showed a distinct consistency of aberrant metabolites between the two groups while approaches can be adjusted to tradeoff between sensitivity and specificity. To develop a more effective model, we also formed a series of logistic models with 18 significant metabolite, a combination of the top 3 metabolites, and 4 different reduced models, the results showed that introducing more variables may not add to the utility, the reduced logistic model (1,2-dioleoyl-sn-glycerol + Glyceryl tripalmitoleate, with AUC: 0.8491,) is the most efficient and effective machine learning model for our classification task. Moreover, we can also find a significant difference on the sensitivity between tumor markers (TMs) and ML models. Therefore, we report that colorectal metabolomics combined with ML demonstrate high accuracy and versatility in distinguishing CRC from those healthy controls.

2. Introduction

Although the diagnosis and management of cancer in the last decade have been advanced, CRC still represents a significant global hygeian burden. Overall, CRC ranks the third in cancer morbidity and second in mortality among whole kind of cancers all over the world [1,2]. The prevalence of CRC is intimately associated with the westernization in dietary and other health habits. At the same time, it is expected to increase further in developed countries which have a remarkable economic growth [3,4]. Apparently,

cancer detection is an important issue in CRC diagnosis and treatment,which should be resolved immediately.

Fecal occult blood test is the most commonly used in the tests of CRC screening. Although these tests have had a prominent contribution to the reducting the mortality rate associated with CRC [5,6], their limited sensitivity for early-stage precancerous lesions indicated the need for more improvement [7]. In addition, a large proportion of the at-risk population is still often diagnosed in advanced stages [8]. Currently blood-based biomarkers which is put into use for CRC such as CEA and cancer antigen 19-9 (CA19-9), are suitable for surveillance or prognostic indicator in CRC treatment but not for screening or diagnosing because of low sensitivity and specificity. Otherwise, the association with other types of gastrointestinal cancers, including gastric cancer, pancreatic cancer, or gynecological cancer such as ovarian cancer also have a infiluence on it[9]. Consequently developing a novel method that screens CRC more conveniantly with higher sensitivity and specificity is paramount.

Frequently mutated genes including BRAF, APC, KRAS, CTNNB1 and SMAD4 have been identified in association with CRC [10]. The epigenetic variation in CRC disorder the hyper-and hypomethylation, which inactivates the tumor suppressor genes and activates oncogenes and leads to epithelial cell growth resuilting in cancerous tumor formulation [11]. In addition to genetic changes, malignant cancers, including CRC, show drastic metabolic shifts. For instance, the glycolysis pathway can be actived by tumor cells with the ignorance of oxygen availability,which produces adenosine triphosphate (Warburg effect) [12]. In addition, the upregulation of oxidative phosphorylation has been found and reported in several cancers [13]. Glutamine is used as a carbon source alternative to glucose via the TCA in proliferating tumor cells to synthesize purines and pyrimidines[14]. In addition, holistic changes such as amino acid, pentose phosphate, urea cycle, polyamine, and nucleotide pathways in the metabolic pathways have been reported[15–17]. Therefore, the metabolites that reflect these metabolic aberrances associated with CRC have been analyzed to establish a novel set of biomarkers [18-24].

In order to enhance the discriminability of multiple biomarkers, ML is a cornerstone [25,26]. Nakajima et al previously have used an alternative decision tree (ADTree)–based prediction method to detect CRC [27], and metabolomics with this ML method showed a high discriminability for breast cancers [28].

In this study, we performed colorectal metabolomic profiling of colorectum collected from patients who have been diagnosed with CRC and HC. Six different kind of ML models were developed to determine the combination of metabolite concentrations that could discriminate between two groups among these models. More than 64 samples were examined, and the data were divided into two datasets. One dataset was used for the ML model development, and the other dataset was used to validate the ML model. Our approach has shown the screening potential of colorectal metabolomic profiles to detect CRC.

3. Materials and Methods

3.1. Subjects

This study was approved by the Ethics Committee of Longhua hospital affiliated to Shanghai University of Traditional Chinese Medicine and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants who agreed to serve as donors. Patients histopathologically diagnosed with colorectal adenocarcinoma were included while the patients with all other types of cancer (adenosquamous cell carcinoma, endocrine carcinoma, lymphoma, etc.) were excluded. The resected specimens were pathologically classified according to the 7th edition of the Union for International Cancer Control TNM Classification of Malignant Tumors [29].

3.2. Colorectum Collection

Colorectum samples were collected after colorectal surgery. Approximately 100mg of colorectal cancer tissues as well as corresponding paracancer tissues was collected and stored in 5 ml polypropylene tubes on ice in order to prevent the degeneration of metabolites. After that, the samples were immediately stored at a tamperature of −80°C.

3.3. Colorectum Preparation and LC-MS/MS Analysis

UHPLC-Q-TOF MS was used for nontargeted analyses of colorectal metabolites, UHPLC-Q-TOF MS analysis was performed using an UHPLC (1290 Infinity LC, Agilent Technologies) coupled to a quadrupole time-of-flight (AB Sciex TripleTOF 6600) in Shanghai Applied Protein Technology Co., Ltd.

Colorectum samples were analyzed via two methods. For HILIC separation, samples were analyzed using a 2.1 mm × 100 mm ACQUIY UPLC BEH Amide 1.7 μm column (waters, Ireland). In both ESI positive and negative modes, the mobile phase contained A=25 mM ammonium acetate and 25 mM ammonium hydroxide in water and B= acetonitrile. The gradient was 95% B for 0.5 min and was linearly reduced to 65% in 6.5 min, and then was reduced to 40% in 1 min and kept for 1 min, and then increased to 95% in 0.1 min, with a 3 min re-equilibration period employed.

The ESI source conditions were set as follows: Ion Source Gas1 (Gas1) as 60, Ion Source Gas2 (Gas2) as 60, curtain gas (CUR) as 30, source temperature: 600℃, IonSpray Voltage Floating (ISVF) ± 5500 V. In MS only acquisition, the instrument was set to acquire over the m/z range 60-1000 Da, and the accumulation time for TOF MS scan was set at 0.20 s/spectra. In auto MS/MS acquisition, the instrument was set to acquire over the m/z range 25-1000 Da, and the accumulation time for product ion scan was set at 0.05 s/spectra. The production scan is acquired using information dependent acquisition (IDA) with high sensitivity mode selected. The parameters were set as follows: the collision energy (CE) was fixed at 35 V with ± 15 eV; declustering potential (DP), 60 V (+)

and −60 V (−); exclude isotopes within 4 Da, candidate ions to monitor per cycle: 10.

3.4. Metabolomics Data Processing

The raw MS data were converted to MzXML files using ProteoWizard MSConvert before importing into freely available XCMS software. For peak picking, the following parameters were used: centWave m/z = 10 ppm, peakwidth = c (10, 60), prefilter = c (10, 100). For peak grouping, bw = 5, mzwid = 0.025, minfrac = 0.5 were used. CAMERA (Collection of Algorithms of MEtabolite pRofile Annotation) was sued for annotation of isotopes and adducts. In the extracted ion features, only the variables having more than 50% of the nonzero measurement values in at least one group were kept. Compound identification of metabolites was performed by comparing of accuracy m/z value (<10 ppm), and MS/MS spectra with an in-house database established with available authentic standards.

3.5. Statistical Methods

3.5.1. Data Introduction: In this experiment, 64 samples were collected, including metabolite data of 32 cancer cells and their corresponding normal tissues. A total of 22,604 detection materials were obtained, among which 1179 metabolites were characterized in HMDB or KEGG.

In addition to metabolomics data, clinical trial data, including age, sex, and smoking history, were added to the study. After we numeralize this information, we get a data set consisting entirely of numerical data, which we can use for machine learning.

3.5.2. Feature Selection: For the large number of metabolites, we need to screen out the differential metabolites that meet the requirements through statistical methods. The common method is to use Fold Change to find metabolites with large difference multiples. However, the metabolic data of cancer tissues obtained in this experiment appear in pairs with that of normal tissues, indicating that the simple difference of multiples may cause errors. Therefore, Wilcoxon Wigned Rank Test was also used for metabolites in the screening process. The metabolites with p value <0.05 were selected, and the results obtained in FC were intersected to obtain accurate differential metabolites.

3.5.3. Learning a Classifier: Machine learning can be divided into supervised machine learning, unsupervised machine learning, reinforcement learning, etc., according to whether the data is labeled or not. For generating classifiers, supervised machine learning is used. The basic principles of supervised machine learning including:

1. Input data: It starts with labeled training data, where inputs are paired with corresponding class labels or categories.

2. Training Phase: The classifier learns patterns from the labeled data to create a model that maps inputs to their respective classes. Various algorithms (e.g., decision trees, SVMs, neural networks) are used for this purpose.

3. Prediction: Once trained, the classifier can predict the class or category of new, unseen data based on the learned patterns.

In this article, we considered several classifiers, including Random Foreset (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Support Vector Machine (SVM), The k-Nearest Neighbors (k-NN), naïve Bayes (NB), and Logistics Regression, and try to develop the most suitable model for prediction.

For Random Forest, it starts by creating multiple bootstrap samples (random samples with replacement) from the original dataset. Each subset is used to build a decision tree. During the construction of each tree, nodes are split considering different subsets of features.

Lasso is an extension of linear regression. It aims to minimize this cost function, a combination of the mean squared error (MSE) and the L1 regularization term:

$$Cost = MSE + \lambda \sum_{i=1}^{n} |\theta_i|$$

As λ increases, more coefficients tend towards zero, resulting in automatic feature selection.

SVM aims to find the optimal hyperplane that separates data into different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points from each class. With kernel function, SVM can efficiently handle non-linearly separable data.

k-NN is an instance-based or lazy learning algorithm. It stores the entire training dataset and makes predictions based on similarity measures between instances. It approximates the target function locally around the data point being predicted by looking at its k-nearest neighbors.

Naïve Bayes is a probabilistic classifier based on Bayes' theorem with an assumption of independence between features. As we expect a binary classifier, Bernoulli Naïve Bayes is used in this article.

3.5.4. Software: All code was written in R (version 2.15.1). For SVM we used the e1071 (version 1.6) R library. For Naïve Bayes, we used the RWeka (version 0.4-12) library for R, which is an interface to the WEKA software. For KNN, and RBF SVM we used the caret R library (version 5.15-023). For LASSO, we used the glmnet R library (version 1.8). For C4.5 we used the C5.0 R library (version 0.1.0-15).

4. Results

4.1. Overview of Profiled Metabolites

To target potential biomarkers, we collected 64 groups of metabolomics data from 32 patients with colorectal cancer, including the metabolite levels from both cancerous and normal tissues. Table 1 summarizes the basic information of the patients.

As for the metabolomics data, we quantified a total of 22604 metabolites (11052 neg and 11152 pos), and 1179 of them have HMDB or KEGG ID, suggesting that they are characterized and

investigated. The result of unsupervised hierarchical clustering with heat map shown in Figure 1. We can see that the cancerous tissues and normal tissues are not fully separated, indicating the necessity of feature selection.

Table 1: Subject Information

n	32
Gender	
Male	21
Female	11
Age	
Mean	66.6875
SD	9.385619
Drink	2
Smoke	1
Anamnesis	
Hypertension	17
Diabetes	7
Hyperlipemia	3
Colitis	2
others	13
Lymphatic Metastasis	14
Distant Metastasis	1

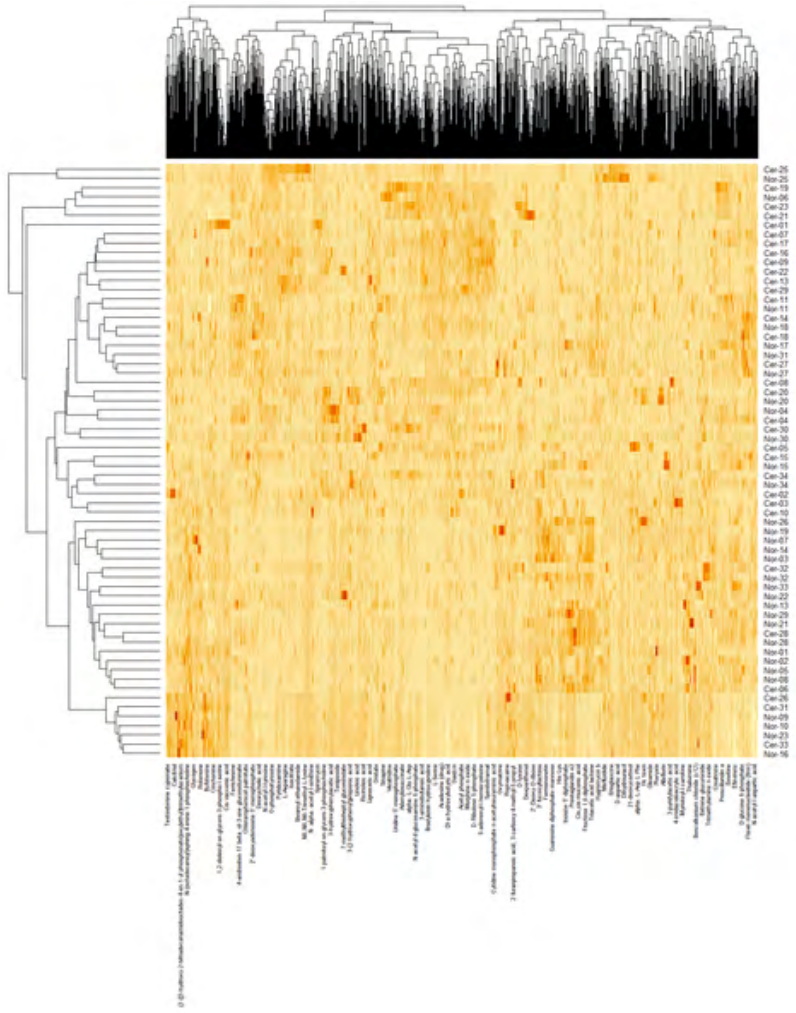

Figure 1: Heatmap for all metabolites

4.2. Feature Selection

To detect significant features in these metabolites, we applied Fold Change (average FC) and select ones that show an FC >4 or FC＜0.25. Since the data from cancerous and normal tissues appear in pairs, we also applied Wilcoxon Signed-Rank Test, and finally locked on 18 metabolites with significance, as showed in Figure 2 and Table 2.

These CRC metabolites showed higher concentrations than those of HC: Three lipids and lipid-like molecules, such as 4-Methylthio-2-oxobutanoic acid, 1,2-dioleoyl-sn-glycerol, Glyceryl tripalmitoleate; five organic acids, such as D-erythrose 4-phosphate, Leukotriene C4, 4-hydroxy-l-glutamic acid, Cystine, L-Aspartyl-L-phenylalanine; five Nucleosides, nucleotides and analogues, such as Udp-n-acetylglucosamine, Uridine 5'-diphosphogalactose, Adenylosuccinic acid, Nicotinate d-ribonucleotide, S-methyl-5'-thioadenosine were included. One phenylpropanoids (Phenyllactic acid), one organic nitrogen compounds (1-methylhistamine) were also included;

And several metabolites associated with CRC showed lower concentrations than those of HC, including one organic acids (Gamma-Glu-Cys), one organoheterocyclic compounds (1h-1,2,4-triazol-3-amine), and one Benzenoids (Benzalkonium chloride).

Then, the result of unsupervised hierarchical clustering after feature selection with heat map shown in Figure 3. We can find a better clustering effect than Figure 1.

Figure 2: Volcano Plot

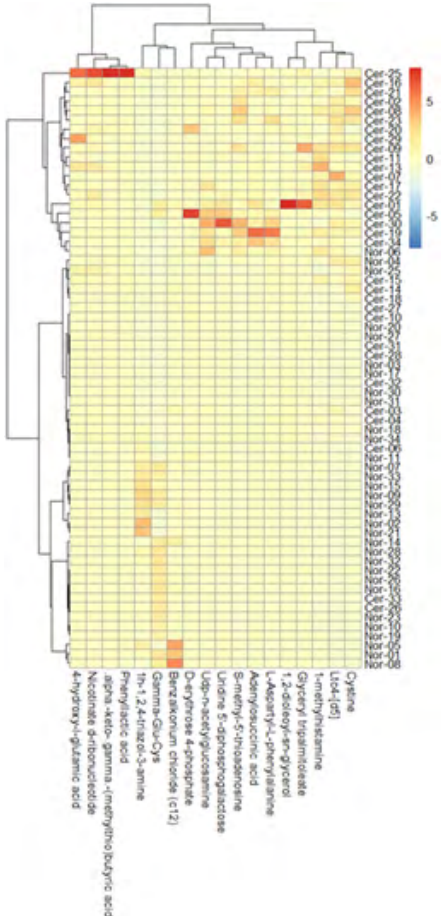

Figure 3: Heatmap for significant metabolics

Table 2: The different 18 metabolites between CRC and HC with significance

name	Superclass	FC	p-value
alpha.-keto-.gamma.-(methylthio)butyric acid (4-Methylthio-2-oxobutanoic acid)	Lipids and lipid-like molecules	5.23	0.00837
D-erythrose 4-phosphate	Organic oxygen compounds	4.23	0.04133
Gamma-Glu-Cys (gamma-Glutamylcysteine)	Organic acids	0.237	0.00051
Ltc4 (Leukotriene C4)	Organic acids	4.74	0.00004
Phenyllactic acid	Phenylpropanoids	4.994	0.0004
Udp-n-acetylglucosamine	Nucleosides, nucleotides, and analogues	4.408	0.00005
Uridine 5'-diphosphogalactose (UDP-galactose)	Nucleosides, nucleotides, and analogues	6.24	0.00066
1-methylhistamine	Organic nitrogen compounds	4.043	0.0002
1,2-dioleoyl-sn-glycerol (Diglyceride)	Lipids and lipid-like molecules	4.503	0.00018
1h-1,2,4-triazol-3-amine (Amitrole)	Organoheterocyclic compounds	0.244	0.0002
4-hydroxy-l-glutamic acid	Organic acids	4.13	0.00179
Adenylosuccinic acid	Nucleosides, nucleotides, and analogues	5.126	0.00018
Benzalkonium chloride (c12, Coniine)	Benzenoids	0.236	0.01323
Cystine	Organic acids	4.113	0.00031
Glyceryl tripalmitoleate (Triglyceride)	Lipids and lipid-like molecules	8.85	0.00004
L-Aspartyl-L-phenylalanine (Aspartic acid-phenylalanine dipeptide)	Organic acids	4.038	0.00106
Nicotinate d-ribonucleotide (Nicotinate ribonucleotide)	Nucleosides, nucleotides, and analogues	4.591	0.00003
S-methyl-5'-thioadenosine (Methylthioadenosine)	Nucleosides, nucleotides, and analogues	5.865	0.00031

4.3. Partial Least Squares-Discriminant Analysis

Partial least squares-discriminant analysis (PLS-DA) is a versatile algorithm that can be used for predictive and descriptive modelling as well as for discriminative variable selection [30]. PLS-DA can generate score plots and Variable importance projection (VIP) score plots [31].

It is widely used in metabolomics analysis and classification. We also applied it on our dataset, and evaluated the overall differences between cancerous and normal cells. Figure 4 shows the scores plot of PLS-DA analysis, distinguishing the two kinds according to their scores on the first two components. We did the analysis both on the whole dataset and on positive ions and negative ions separately. And then we found the metabolites with VIP > 1, which was certified as significant factor. Table 3 shows the VIP scores of selected metabolics.

According to the VIP scores we get, we can find that most of the metabolites we select from 3.1 (as showed in Table 3 with grey) show a VIP score >1, including 11 metabolites: 4-Methylthio-2-oxobutanoic acid, 1,2-dioleoyl-sn-glycerol, D-erythrose 4-phosphate, Cystine, Udp-n-acetylglucosamine, Uridine 5’-diphosphogalactose, S-methyl-5’-thioadenosine, Phenyllactic acid, 1-methylhistamine, Gamma-Glu-Cys, Benzalkonium chloride. So we continued to form our model with these metabolites.

Table 3: The VIP scores of selected metabolics

Name	data_VIP
alpha.-keto-.gamma.-(methylthio)butyric acid	0.651904
D-erythrose 4-phosphate	0.683736
Gamma-Glu-Cys	1.65263
Ltc4-[d5]	2.054495
Phenyllactic acid	0.664884
Udp-n-acetylglucosamine	1.097735
Uridine 5'-diphosphogalactose	1.007847
1-methylhistamine	1.705314
1,2-dioleoyl-sn-glycerol	0.683945
1h-1,2,4-triazol-3-amine	1.750107
4-hydroxy-l-glutamic acid	1.022703
Adenylosuccinic acid	1.312669
Benzalkonium chloride (c12)	0.943848
Cystine	1.86491
Glyceryl tripalmitoleate	1.21818
L-Aspartyl-L-phenylalanine	1.305451
Nicotinate d-ribonucleotide	1.204299
S-methyl-5'-thioadenosine	1.699929

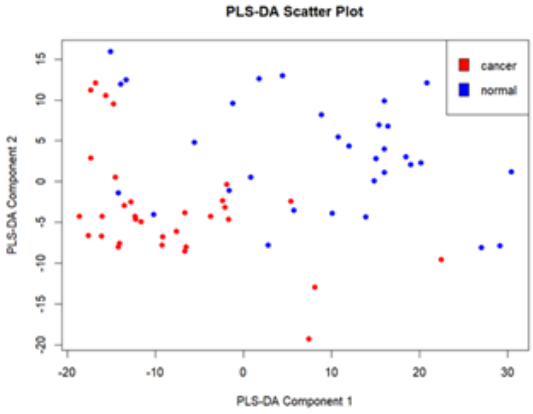

Figure 4: The scores plot of PLS-DA analysis

4.4. Machine Learning Models

With the 18 differential metabolites we find, we are prepared to create prediction models based on Machine Learning algorithm. The total 64 groups of data were randomly assigned to training (n = 32) and validation dataset (n = 32). We were going to form machine learning classifier models on the training set and test the effectiveness of our models on the validation set.

We applied multiple machine learning classifier models, including Random Foreset (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Support Vector Machine (SVM), The k-Nearest Neighbors (k-NN), Naïve Bayest (NB), and Logistic Regression, to find the most suitable model for our data. We first tested these models on the whole dataset and examine their effectiveness through their AUC values. The ROC curves are showed in Figure 5. Their performances on forecast accuracy, specificity.

The effectiveness of the models were not quite up to expectations. The result of Lasso regression reminded us that dimensionality reduction might lead to a more effective model.

To find appropriate covariates for our model, we formed a series of logistic models with only one metabolite. Since the model utilities are dependent on how we divide training set and validation set, we regenerate these models for 1000 times with different divide and calculate the average AUC, accuracy, specificity, sensitivity, in order to compare these models correctly. The result is showed in Table 4.

We can find that some metabolites outperform the others. For example, Ltc4-[d5] (AUC: 0.8191), 1,2-dioleoyl-sn-glycerol (AUC: 0.8166), and Glyceryl tripalmitoleate (AUC: 0.8143) are the top 3 metabolites. To develop a more effective model, we consider a combination of these top 3 metabolites. Table 5 shows the results of these models.

The results showed that introducing all of the variables into this model may not add to the utility. The highest AUC is the combination of "1,2-dioleoyl-sn-glycerol + Glyceryl tripalmitoleate", which is 0.8491. So we now have found the most efficient logistic model, as reduced logistic model.

In a similar way, we also formed many other reduced models based on different machine learning algorithm. The model comparison among those models are showed in Table 6. The model comparisons for RF, SVM, NB, Logistic with one metabolites.

From this we can conclude that the reduced logistic model is the most efficient and effective machine learning model for our classification task. The model can be described as:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where x_1 refers to 1,2-dioleoyl-sn-glycerol and x_2 refers to Glyceryl tripalmitoleate.

This model only contains 3 coefficients, which meets the requirments of 10 EPV (10 events per variable).

Table 4: Model compare for logistic models with one metabolites

	AUC	SD	Accuracy	SD	Specificity	SD	Sensitivity	SD
alpha.-keto-.gamma.-(methylthio)butyric acid	0.7516	0.0618	0.6371	0.0517	0.3893	0.1119	0.885	0.0978
D-erythrose 4-phosphate	0.6452	0.0721	0.5164	0.0305	0.3483	0.3957	0.6846	0.4107
Gamma-Glu-Cys	0.689	0.0749	0.601	0.0801	0.9003	0.1042	0.3018	0.2275
Ltc4-[d5]	0.8191	0.0549	0.7251	0.0506	0.5288	0.1215	0.9214	0.0766
Phenyllactic acid	0.7709	0.0592	0.6013	0.0552	0.3541	0.2133	0.8485	0.2077
Udp-n-acetylglucosamine	0.7814	0.0604	0.6086	0.0542	0.321	0.2002	0.8961	0.188
Uridine 5'-diphosphogalactose	0.6855	0.0717	0.5874	0.0426	0.2699	0.1291	0.9049	0.1306
1-methylhistamine	0.7093	0.0681	0.6858	0.0462	0.4218	0.1116	0.9498	0.0635
1,2-dioleoyl-sn-glycerol	0.8166	0.0556	0.7111	0.0574	0.5483	0.1594	0.874	0.0821
1h-1,2,4-triazol-3-amine	0.7777	0.0564	0.6473	0.0634	0.6308	0.1901	0.6638	0.2214
4-hydroxy-l-glutamic acid	0.6994	0.0674	0.5988	0.0426	0.2803	0.1544	0.9174	0.1374
Adenylosuccinic acid	0.7558	0.0661	0.6652	0.0489	0.3973	0.1222	0.9331	0.0569
Benzalkonium chloride (c12)	0.6201	0.0708	0.4985	0.028	0.8923	0.2732	0.1047	0.2658
Cystine	0.7125	0.0675	0.727	0.0475	0.5246	0.0949	0.9295	0.0552
Glyceryl tripalmitoleate	0.8143	0.0558	0.7317	0.051	0.5536	0.1056	0.9098	0.0742
L-Aspartyl-L-phenylalanine	0.738	0.0661	0.62	0.0562	0.3249	0.159	0.9151	0.0897
Nicotinate d-ribonucleotide	0.7551	0.0629	0.661	0.0571	0.4069	0.1614	0.9151	0.0865
S-methyl-5'-thioadenosine	0.7711	0.0611	0.6898	0.0504	0.4371	0.1294	0.9426	0.0743

Table 5: Model comparison for logistic models

	AUC	SD	Accuracy	SD	Specificity	SD	Sensitivity	SD
Ltc4-[d5] + 1,2-dioleoyl-sn-glycerol	0.8239	0.0591	0.7432	0.0548	0.6116	0.1299	0.8748	0.0882
Ltc4-[d5] + Glyceryl tripalmitoleate	0.8263	0.0619	0.7379	0.0541	0.5656	0.1083	0.9102	0.0862
1,2-dioleoyl-sn-glycerol + Glyceryl tripalmitoleate	0.8491	0.0559	0.7446	0.0528	0.6094	0.1297	0.8798	0.0991
Ltc4-[d5] + 1,2-dioleoyl-sn-glycerol + Glyceryl tripalmitoleate	0.8295	0.0691	0.7383	0.0565	0.6006	0.129	0.8761	0.098

Table 6: Model comparison for reduced models

	AUC	SD	Accuracy	SD	Specificity	SD	Sensitivity	SD
reduced_RF	0.8055	0.0497	0.7695	0.0488	0.7581	0.114	0.7808	0.1112
reduced_SVM	0.8168	0.0555	0.625	0	0.4375	0	0.8125	0
reduced_NB	0.8254	0.0531	0.7575	0.0533	0.6113	0.1164	0.9038	0.0872
reduced_Logistic	0.8491	0.0559	0.7446	0.0528	0.6094	0.1297	0.8798	0.0991

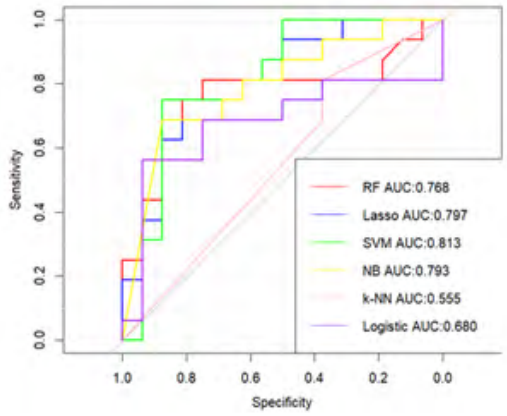

Figure 5: ROC Curves

4.5. Pathway Analysis

The overall differences in metabolomic profiles between HC and CRC were also evaluated by pathway analysis, while pathway of the most important 18 metabolites were marked in red (Figure 6).

Pathway analysis detected three significantly enriched pathways of the most important 18 metabolites, including (1)Protein digestion and absorbtion, (2)Biosynthesis of amino acids, and (3) Neuroactive ligand-receptor interaction. The Rich factor of pathway (1) was relatively high, while those of (2) and (3) were relatively small.

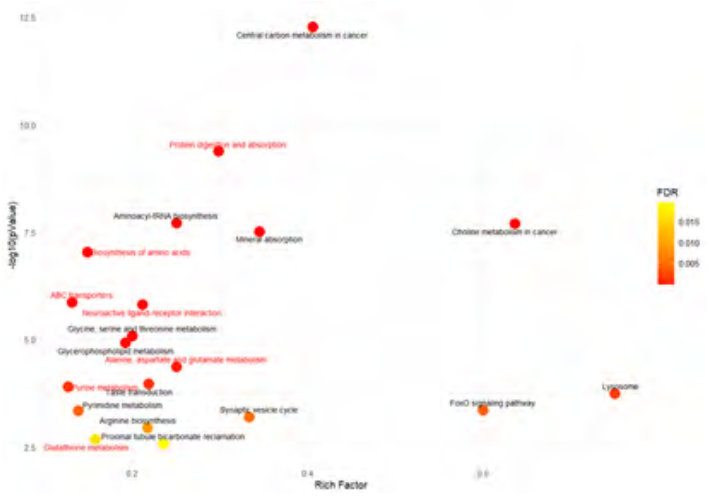

Figure 6: Pathway analysis

4.6. Comparisons with Tumor Markers

Carcinoembryonic antigen (CEA), CA724, CA125, CA153, CA50, CA242 and CA199 of the patients were measured. For tumor markers (TMs), the subjects showing a CEA＞5.0 ng/ml, CA724＞6.9 U/ml, CA125＞25 U/ml, CA50＞25 U/ml, CA242＞20 U/ml, or CA199＞30 U/ml were counted as positive. In the dataset, the sensitivity to CEA, CA724, CA125, CA153, CA50, CA242 and CA199 from CRC subjects were 14.89%, 8.51%, 6.38%, 2.13%, 2.13%, 8.51% and 10.64%, respectively. Those of reduced models RF, SVM, NB, and Logistic, the average sensitivity were 78.08%, 81.25%, 90.38%, and 87.98%, respectively. We can find a significant difference on the sensitivity between TMs and ML models.

5. Discussion

Mainly microRNA, most of microRNA are detected in plasma and stool by methylation and abnormal levels of circulating tumor DNA and noncoding RNA,which represent the recently developed liquid biopsies for CRCs [32]. In saliva samples, MiR-21 has shown the ability in the discrimination in CRC and HC [33]. A single marker showing high specificity for a disease is more beneficial for developing simple and reasonable assays compared with simultaneous analyses of multiple markers for the detection of diseases. The analysis of volatile compounds has also shown the potential of detection in CRC [34]. As a pity,the current study can only measure hydrophilic metabolites, and more comprehensive analyses should be explored to the accuration to the biomarkers of CRC.

In this study, we investigated the use of metabolomics to discover colorectum-based biomarkers and discriminate these biomarkers among CRC and HC. A total of 64 colorectum samples were collected from subjects with CRC (n =32) and HC (n =32).

As described in the heatmap (Figure 1), we quantified 1179 characterized from a total of 22604 metabolites under unsupervised hierarchical clustering, but the CRC tissues and normal tissues are not fully separated, indicating the necessity of feature selection. So we applied Fold Change (FC >4 or FC＜0.25) as well as Wilcoxon Signed-Rank Test and finally locked on 18 significant metabolites, as showed in Figure 2 and Table 2. We can find a better clustering effect than Figure 1 after feature selection with heat map shown in Figure 3.

We applied multiple machine learning classifier models, including Random Foreset (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Support Vector Machine (SVM), The k-Nearest Neighbors (k-NN), Naïve Bayest (NB), and Logistic Regression, to find the most suitable model for our data. We first tested these models on the whole dataset and examine their effectiveness through their AUC values. The ROC curves are showed in Figure 5. Their performances on forecast accuracy, specificity, sensitivity. The effectiveness of the models were not quite up to expectations. The result of Lasso regression reminded us that dimensionality reduction might lead to a more effective model.

A single marker may not be enough for a disease-specific index, but ML patterns that capture multiple metabolite would enhance the specificity. In order to find appropriate covariates for our model, we also formed a series of logistic models with these 18 significant metabolites, the result is showed in Table 4. The top 3 metabolites are Ltc4-[d5] (AUC: 0.8191), 1,2-dioleoyl-sn-glycerol (AUC: 0.8166), and Glyceryl tripalmitoleate (AUC: 0.8143). To develop a more effective model, we then formed a combination of these top 3 metabolites (Table 5) , it showed the highest AUC is 0.8491 of the group “1,2-dioleoyl-sn-glycerol + Glyceryl tripalmitoleate” . So we now have found the most efficient logistic model, as reduced logistic model.

Leukotriene C4, 1,2-dioleoyl-sn-glycerol, and Glyceryl tripalmitoleate were both involved in lipid metabolism, particularly in the synthesis and transformation of lipids. Leukotriene C4 is a product of arachidonic acid metabolism, while 1,2-dioleoyl-sn-glycerol and Glyceryl tripalmitoleate are lipid molecules associated with esterification reactions involving glycerol and fatty acids. It can be inferred that the Fatty acid metabolism and glutaminolysis, being characteristic of cancer metabolism, might underlie the observed characteristics. In addition, several lipids and lipid-like molecules, such as 4-Methylthio-2-oxobutanoic acid, 1,2-dioleoyl-sn-glycerol (Diglyceride), some of Organic acids, such as Cystine, and some Nucleosides, nucleotides, and analogues, such as Udp-n-acetylglucosamine, Uridine 5’-diphosphogalactose were also elevated in CRC. The intermediate metabolites associated with these energy and amino acid pathways were continually reported.

In a similar way, we also formed many other reduced models based on different ML algorithm. The model comparison among those models are showed in Table 6 for RF, SVM, NB, and we concluded that the reduced logistic model is the most efficient and effective machine learning model for our classification task.

We also compared the sensitivity of ML models with those of CEA, CA724, CA125, CA153, CA50, CA242 and CA199 in CRC data. Both ML models showed better sensitivity compared with these seven TM. The sensitivity of CEA, CA724, CA125, CA153, CA50, CA242 and CA199 from CRC subjects were 14.89%, 8.51%, 6.38%, 2.13%, 2.13%, 8.51% and 10.64%, respectively. While the sensitivity of four reduced models of RF, SVM, NB, and Logistic were 78.08%, 81.25%, 90.38%, and 87.98%, respectively. Therefore, we can find a significant difference on the sensitivity between TMs and ML models in our dataset, the complimentary use of ML models with TM may benefit the screening of CRC.

Several limitations need to be acknowledged. Firstly, the sample size is not relatively large enough and the proportion of these two groups does not reflect the actual prevalence of these diseases.

Secondly, the comparison with other diseases, using other cancer types in especial, was not performed. Thirdly, our approach in the current study showed CRC detection abilities; however, room to improve the sensitivity and specificity of CRC detection still exists.

In conclusion, we analyzed the colorectal metabolic profiles of CRC and HC. The data showed consistent profile patterns, including 1,2-dioleoyl-sn-glycerol and Glyceryl tripalmitoleate. The reduced logistic model successfully discriminated against these groups which have high sensitivity and specificity. In addition, the models showed higher sensitivity compared with CEA, CA724, CA125, CA153, CA50, CA242 and CA199. The models could give a great contribution to clinical screening for CRC in the future.

6. Acknowledgments

This work was supported by Shanghai Hospital Development Center (SHDC12020123, SHDC2020CR4050), "Science and Technology Innovation Action Plan" of Shanghai Municipal Science and Technology Commission (No.20Y21902300), Shanghai Municipal Health Commission, Health Youth Talent Project (2022YQ028), Shanghai "Science and Technology Innovation Action Plan" Medical Innovation Research Project (21MC1930500), National Thirteenth Five-Year Science and Technology Major Special Project for New Drug Innovation and Development (2017ZX09304001), Shanghai Municipal Health Commission (shslczdzk03701).

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71: 209-249.
2. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020; 70: 145-164.
3. Ahmed FE. Effect of diet, life style, and other environmental/chemopreventive factors on colorectal cancer development, and assessment of the risks. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 2004; 22: 91-147.
4. Randi G, Edefonti V, Ferraroni M, La Vecchia C, Decarli A. Dietary patterns and the risk of colorectal cancer and adenomas. Nutr Rev. 2010; 68: 389-408.
5. Hewitson P, Glasziou P, Watson E, Towler B, Irwig L. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. Am J Gastroenterol. 2008; 103: 1541-1549.
6. Simon JB. Fecal occult blood testing: clinical value and limitations. Gastroenterologist. 1998; 6: 66-78.
7. de Wijkerslooth TR, Stoop EM, Bossuyt PM, Meijer GA, van Ballegooijen M, van Roon AHC, et al. Immunochemical fecal occult blood testing is equally sensitive for proximal and distal advanced neoplasia. Am J Gastroenterol. 2012; 107: 1570-1578.
8. Maida M, Macaluso FS, Ianiro G, Mangiola F, Sinagra E, Hold G, et al. Screening of colorectal cancer: present and future. Expert Rev Anticancer Ther. 2017; 17: 1131-1146.
9. Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, Macdonald JS, et al. Asco 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol. 2006; 24: 5313-5327.
10. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487: 330-7.
11. Lao VV, Grady WM. Epigenetics and colorectal cancer. Nat Rev Gastroenterol Hepatol. 2011; 8: 686-700.
12. Warburg O. On the origin of cancer cells. Science. 1956; 123: 309-14.
13. Ashton TM, McKenna WG, Kunz-Schughart LA, Higgins GS. Oxidative phosphorylation as an emerging target in cancer therapy. Clin Cancer Res. 2018; 24: 2482-90.
14. Zhao Y, Zhao X, Chen V, Feng Y, Wang L, Croniger C, et al. Colorectal cancers utilize glutamine as an anaplerotic substrate of the tca cycle in vivo. Sci Rep. 2019; 9: 19180.
15. Hirayama A, Kami K, Sugimoto M, Sugawara M, Toki N, Onozuka H, et al. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. Cancer Res. 2009; 69: 4918-4925.
16. Satoh K, Yachida S, Sugimoto M, Oshima M, Nakagawa T, Akamoto S, et al. Global metabolic reprogramming of colorectal cancer occurs at adenoma stage and is induced by myc. Proc Natl Acad Sci USA. 2017; 114: E7697-E7706.
17. Tevini J, Eder SK, Huber-Schönauer U, Niederseer D, Strebinger G, Gostner JM, et al. Changing metabolic patterns along the colorectal adenoma-carcinoma sequence. J Clin Med. 2022; 11(3): 721.
18. Dalal N, Jalandra R, Sharma M, Prakash H, Makharia GK, Solanki PR, et al. Omics technologies for improved diagnosis and treatment of colorectal cancer: technical advancement and major perspectives. Biomed Pharmacother. 2020; 131: 110648.
19. Uchiyama K, Yagi N, Mizushima K, Higashimura Y, Hirai Y, Okayama T, et al. Serum metabolomics analysis for early detection of colorectal cancer. J Gastroenterol. 2017; 52: 677-694.
20. Sakurai T, Katsumata K, Udo R, Tago T, Kasahara K, Mazaki J, et al. Validation of urinary charged metabolite profiles in colorectal cancer using capillary electrophoresis-mass spectrometry. Metabolites. 2022; 12: 59.
21. Qiu Y, Cai G, Su M, Chen T, Liu Y, Xu Y, et al. Urinary metabonomic study on colorectal cancer. J Proteome Res. 2010; 9: 1627-1634.

22. Nakajima T, Katsumata K, Kuwabara H, Soya R, Enomoto M, Ishizaki T, et al. Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls. Int J Mol Sci. 2018; 19: 756. 23.
23. Leichtle AB, Nuoffer JM, Ceglarek U, Kase J, Conrad T, Witzigmann H, et al. Serum amino acid profiles and their alterations in colorectal cancer. Metabolomics. 2012; 8: 643-653.
24. Farshidfar F, Weljie AM, Kopciuk KA, Hilsden R, McGregor SE, Buie WD, et al. A validated metabolomic signature for colorectal cancer: exploration of the clinical value of metabolomics. Br J Cancer. 2016; 115: 848-857.
25. Takada M, Sugimoto M, Naito Y, Moon HG, Han W, Noh DY, et al. Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. BMC Med Inform Decis Mak. 2012; 12: 54.
26. Takada M, Sugimoto M, Ohno S, Kuroi K, Sato N, Bando H, et al. Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique. Breast Cancer Res Treat. 2012; 134: 661-670.
27. Nakajima T, Katsumata K, Kuwabara H, Soya R, Enomoto M, Ishizaki T, et al. Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls. Int J Mol Sci. 2018; 19: 756.
28. Murata T, Yanagisawa T, Kurihara T, Kaneko M, Ota S, Enomoto A, et al. Salivary metabolomics with alternative decision tree-based machine learning methods for breast cancer discrimination. Breast Cancer Res Treat. 2019; 177: 591-601.
29. Sobin LH, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours. 7th ed. Wiley-Blackwell; 2009.
30. Lee LC, Liong CY, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst. 2018; 143(15): 3526-39.
31. Xia J, Psychogios N, Young N, Wishart DS. Metaboanalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009; 37: W652-60.
32. Palanca-Ballester C, Rodriguez-Casanova A, Torres S, Calabuig-Fariñas S, Exposito F, Serrano D, et al. Cancer epigenetic biomarkers in liquid biopsy for high incidence malignancies. Cancers. 2021; 13: 3016.
33. Sazanov AA, Kiselyova E, Zakharenko A, Romanov MN, Zaraysky M. Plasma and saliva mir-21 expression in colorectal cancer patients. J Appl Genet. 2017; 58: 231-237.
34. Bel'skaya LV, Sarf EA, Shalygin SP, Postnova TV, Kosenok VK. Identification of salivary volatile organic compounds as potential markers of stomach and colorectal cancer: a pilot study. J Oral Biosci. 2020; 62: 212-221.