Hindawi
Complexity
Volume 2017, Article ID 8917258, 14 pages
https://doi.org/10.1155/2017/8917258

Research Article

Building Up a Robust Risk Mathematical Platform to Predict Colorectal Cancer

Le Zhang,[1,2,3] Chunqiu Zheng,[2] Tian Li,[2] Lei Xing,[4] Han Zeng,[2] Tingting Li,[3] Huan Yang,[5] Jia Cao,[5] Badong Chen,[4] and Ziyuan Zhou[6]

[1]*College of Computer Science, Sichuan University, Chengdu 610065, China*
[2]*College of Computer and Information Science, Southwest University, Chongqing 400715, China*
[3]*College of Mathematics and Statistics, Southwest University, Chongqing 400715, China*
[4]*School of Electronic and Information Engineering, Xi'an Jiaotong University, 28 Xianning West Road, Beilin District, Xi'an 710049, China*
[5]*Toxicology Institute, College of Preventive Medicine, Third Military Medical University, 30 Gaotanyan Street, Shapingba District, Chongqing 400038, China*
[6]*Department of Environment Health, College of Preventive Medicine, Third Military Medical University, 30 Gaotanyan Street, Shapingba District, Chongqing 400038, China*

Correspondence should be addressed to Badong Chen; chenbd@mail.xjtu.edu.cn and Ziyuan Zhou; ziyuanzhou@tmmu.edu.cn

Received 30 April 2017; Revised 10 July 2017; Accepted 17 August 2017; Published 16 October 2017

Academic Editor: Fang-Xiang Wu

Colorectal cancer (CRC), as a result of a multistep process and under multiple factors, is one of the most common life-threatening cancers worldwide. To identify the "high risk" populations is critical for early diagnosis and improvement of overall survival rate. Of the complicated genetic and environmental factors, which group is mostly concerning colorectal carcinogenesis remains contentious. For this reason, this study collects relatively complete information of genetic variations and environmental exposure for both CRC patients and cancer-free controls; a multimethod ensemble model for CRC-risk prediction is developed by employing such big data to train and test the model. Our results demonstrate that (1) the explored genetic and environmental biomarkers are validated to connect to the CRC by biological function- or population-based evidences, (2) the model can efficiently predict the risk of CRC after parameter optimization by the big CRC-related data, and (3) our innovated heterogeneous ensemble learning model (HELM) and generalized kernel recursive maximum correntropy (GKRMC) algorithm have high prediction power. Finally, we discuss why the HELM and GKRMC can outperform the classical regression algorithms and related subjects for future study.

1. Introduction

During past decades, new strategies are developed to decrease the incidence and to improve the prognosis of colorectal cancer (CRC), from popularizing regular screening in individuals older than 50 years for prevention to taking some new technologies like laparoscopic surgery, neoadjuvant chemotherapies, and bio-targeted therapy into consideration for more precise and individualized treatment. However, CRC is still one of the important contributors to cancer worldwide [1–7]. CRC ranks 4 in cancer incidences and accounts for approximately 8–10% cancer-related death [8], and the 5-year survival rate (40–50%) is still not as satisfied as expected. CRC is now recognized as a result of multistep process under very complicated gene-environment interactions; either genetic variation and environmental factors or dietary pattern and unfavorable lifestyle may jointly play the important roles in colorectal neoplasia [9–12]. Accordingly, to efficiently identify CRC-risk factors is the first step for prevention and early diagnosis which is critical for decreasing CRC morbidity and mortality [13, 14]. Based on this hypothesis, a consortium that includes institutions from South Korea, Japan, and China cooperatively performs a multicenter case-control study (KOJACH study) during 2000–2004 to explore the CRC-risk factors in East Asia populations [15–18]. In this cooperative study, information of family history, life styles,

food, nutrition intakes, and single nucleotide polymorphisms (SNPs) of each participant is collected for both CRC cases and cancer-free controls. Then this study plans to develop such a CRC predictive model that can not only investigate which potential risk factors have the significant impact on the occurrence of CRC regarding the collected data but also efficiently and reliably predict the risk of CRC before being diagnosed as early as possible.

There are some mathematical models already developed and used to process different type of data for CRC occurrence prediction. For low dimensional data, Wu et al. [19] and Huang et al. [20] propose the logistic regression and the greedy Bayesian model. To process high dimensional dichotomous data, Hahn and his colleagues [21–23] propose to use multifactor dimensionality reduction (MDR) method for mapping them into the low dimensional space and Li et al. [24] propose a novel forward U test to estimate the possibility of the risk of CRC. In addition, Andrew et al. [25], Meredith et al. [26], and Rutledge et al. [27] employ the linear regression models to predict the occurrence of CRC. However, these previous models cannot simultaneously process our big high dimensional CRC data with both continuous and discrete data type to obtain enough high predictive accuracy.

For this reason, to avoid the shortcomings of the previous research when they are used for such complicated data collected in the KOJACH study as mentioned above, we propose a robust CRC cancer predictive model based on our latest study [28] with the following three innovations. Firstly, we use a common standard to collect clinical CRC data with information of genetic variations and environmental exposure [29], since the quickly collected high dimensional data not only have the large volume including 369 CRC patients and 929 cancer-free controls, but also have 305 data types. Secondly, the biological classification, dimensionality reduction, and regression analysis stages are integrated into the CRC predictive model to make it robust and reliable. Thirdly, both heterogeneous ensemble learning model (HELM) and a generalized kernel recursive maximum correntropy (GKRMC) algorithm are developed to increase the predictive accuracy of the model.

The research results indicate that (1) both genetic and environmental related factors play the significant role in the occurrence of CRC; (2) CRC risk can be accurately and efficiently identified with this model by using these explored biomarkers as the classifiers; and (3) our innovated HELM and GKRMC have higher predictive power than the classical regression algorithms.

Finally, we analyze the outperformance reasons for both HELM and GKRMC algorithm and discuss the future study for the CRC predictive model.

2. Materials and Methods

The data used in this study is from the hospital-based case-control study of colorectal cancer in Chongqing, China, by the Department of Toxicology at the Third Military Medical University [18]. The clinical case data is comprised of 369 pathologically diagnosed colorectal cancer patients. The control data consists of 929 cancer-free patients with frequency matched by age, gender, and birthplace. All controls are selected from the orthopedics and general surgery department of the same hospitals and those who have cancer history or any cancer-related diseases are excluded. All recruitments sign a written informed consent.

Food intake is evaluated by our previously developed Semi-Quantitative Food Frequency Questionnaire [30]. The SNP information of full-length genes plus 2,000 bp in the upper stream of each candidate gene is obtained from the HapMap [31]. After setting the minor allele frequency at 0.01 [32], the Haploview software [33] is used to screen the tag SNPs and only one SNP is selected in each of linkage disequilibrium blocks. As a result, there is a total of 46 tag SNPs from the 127 reported SNPs of the three key alcohol-metabolism genes (ADH1B, ALDH2, and CYP2E1) [34–36]. DNA is extracted from 2.5 mL whole blood according to the manufacturer's instructions of Promega DNA Purification Wizard kit. The DNA purification and Polymerase Chain Reactions (PCR) are done by Eppendorf 5333 Mastercycler. Genotyping of the selected TagSNPs is done by ABI 3130xl Gene Analyzer. This study protocol is approved by the Third Military Medical University Ethics Committee.

The items in the dataset include general information (such as gender and age), polymorphism distribution of genes related to ethanol metabolism (the distribution of homozygotes and heterozygotes of gene loci), and demographic characteristics, food, and lifestyle habits (smoking and alcohol consumption). To avoid any bias, a standard questionnaire is generated in which each survey item has a specific definition. The examination is carried out as a face-to-face query. Several survey items, such as the amount of alcohol and cigarettes consumed, are quantitatively estimated. Using age 60 as the demarcation point, the surveyed patients are divided into the elderly group and the young/middle-aged group. Alcohol consumption is divided into healthy drinking (including people who do not drink and people who drink no more than 15 g per day) and nonhealthy drinking (including people who drink more than 15 g per day). Based on smoking habits, the participants are divided into nonsmokers and smokers (including those who had quit smoking).

This study employs these data to build the predictive CRC model with biological classification, dimensionality reduction, and regression analysis stages, which will be illustrated in detail in the next section.

2.1. Biological Classification. The biological classification is carried out from the perspective of medical science to divide the original dataset into four subclasses, which are as follows: (1) polymorphism distribution of genes related to ethanol metabolism: the data of the SNPs are listed in Supplementary S1 in Supplementary Material available online at https://doi.org/10.1155/2017/8917258; (2) demographic characteristics information: the data of the demographic characteristics are listed in Supplementary S2; (3) lifestyle habits: the data of the lifestyles are listed in Supplementary S3; (4) food: the data of the foods are listed in Supplementary S4.

2.2. Dimensionality Reduction for the Original Data. This study employs three broadly used dimensionality reduction

methods, namely, principal component analysis, entropy of information, and relief method to obtain the mutually explored biomarkers for each subclass.

(1) Sparse Principal Component Analysis (SPCA) Method. Principal component analysis (PCA) [37–39] is a dimensionality reduction technique to ease complexity in multivariate data analyses by replacing the original variables with a small group of principal components. SPCA uses the Lasso [40] to produce modified principal components with sparse loadings. PCs are the uncorrelated linear combinations of original variables ranked by their variances in the descending order:

$$\begin{aligned} &\mathrm{PC}_i = l_{1i}X_1 + l_{2i}X_2 + \cdots + l_{mi}X_m \\ &\max \quad (\mathrm{var}(\mathrm{PC}_i)) \\ &\text{s.t.} \quad \sum_{j=1}^{m} l_{ji}^2 = 1, \\ &\qquad \sum_{j=1}^{m} l_{ji} \cdot l_{jk} = 0, \\ &\qquad\qquad 0 \le k < i, \end{aligned} \tag{1}$$

where $X_1, X_2, \ldots, X_m$ are the original variables and $l_{1i}, l_{2i}, \ldots, l_{mi}$ are the coefficients of principal components PC_i corresponding to the original variables estimated by the R-system packages.

(2) Entropy Method. Entropy measures the uncertainty associated with a random variable [41–43] as

$$H(X) = -E\left[\log_p(X)\right] = -\sum_{x\in\chi} p(x)\log p(x), \tag{2}$$

where $p(x) = P(X = x), x \in \chi$, is the probability mass function of the random variable X and χ is a finite set (e.g., $\{1, 2, \ldots, n\}$) or an enumerable infinite set (e.g., $\{1, 2, \ldots\}$). High entropy $H(X)$ indicates high uncertainty about the random variable X.

(3) Relief Method. Relief algorithm [44] is applied to classification of two kinds of data. Relief is a kind of feature weighting algorithm, which gives different weights according to the relevance of features and categories. Also, the relevance of features and categories in relief algorithms is based on the ability of features to distinguish between close samples. Relief algorithm process is as follows:

$$w_i = w_i + \left|x^{(i)} - \mathrm{NM}^{(i)}(x)\right| + \left|x^{(i)} - \mathrm{NH}^{(i)}(x)\right|, \quad \text{for } i = 1:T. \tag{3}$$

The key idea of relief is to iteratively estimate feature weights according to their ability to discriminate between neighboring patterns. In each of the iterations, a pattern x is randomly selected and then two nearest neighbors of x are found, one from the same class (termed the nearest hit or NH) and the other from a different class (termed the nearest miss or NM). w_i represents the weight of the ith feature.

2.3. Regression Analysis. After biological classification and data dimensional reduction stages, we used the logistic regression (LR), support vector machine (SVM), heterogeneous ensemble learning model (HELM), kernel recursive lease squares (KRLS) [45], and our innovated generalized kernel recursive maximum correntropy (GKRMC) algorithm to build up the predictive regression model.

(1) Logistic Regression. The logistic regression (LR) [46, 47] (see (4)) can be considered as a type of semilinear regression (Huang et al., 2006), which assumes that dependent variable has 0 and 1 states.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{4}$$

where $x_1, x_2, \ldots, x_k$ are covariates and $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown coefficients for the covariates and p is the probability of the dependent variable equaling a "success" or "case."

(2) Support Vector Machine. Support vector machine (SVM) [48] is a machine learning method proposed by Vapnik in the early 1990s and successively extended by other researchers. The general form of the equation of the separating line is given as

$$f(x) = (W \cdot X) + b, \tag{5}$$

where $(W \cdot X)$ represents the inner product of the vector W and the X vector. If the linear discriminator function is normalized so that all samples meet $|f(x) \ge 1|$, then the margin between the classification face $(W \cdot X) + b = 1$ and $(W \cdot X) + b = -1$ is $2\|W\|$ (namely, the classification interval).

Minimizing the distance $2/\|W\|$, it is equivalent to maximizing $1/2||W||^2$, and then we can get the optimal classification face. Thus, the problem of seeking the optimal classification face is transformed into the following optimization problem:

$$\min \quad \frac{1}{2}w'w + c\sum_{i=1:N}\xi_i. \tag{6}$$

(3) Heterogeneous Ensemble Learning Model (HELM). Ensemble learning [49] employs multiple learners to solve a problem. The generalization ability of an ensemble is usually significantly better than that of a single learner [50]. The adaboost algorithm [51] is a type of ensemble learning. Based on previous studies, most of the ensemble learning algorithms are the integration of several of the same (homomorphic ensemble) or different (anomaly ensemble) weak classifiers. Here we propose such a HELM algorithm based on the adaboost algorithm that integrates the advantages of both homomorphic and anomaly ensemble. HELM algorithm process is illustrated in Figure 1.

Input. Sample set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_n is the examples and $y_n \in \{0, 1\}$ is the label; weak classifier $\mathscr{L} \in \{\mathscr{L}_1 = \mathrm{svm}, \mathscr{L}_2 = \text{logistic regression}, \mathscr{L}_3 = \mathrm{KRLS}\}$. T is the iteration number.

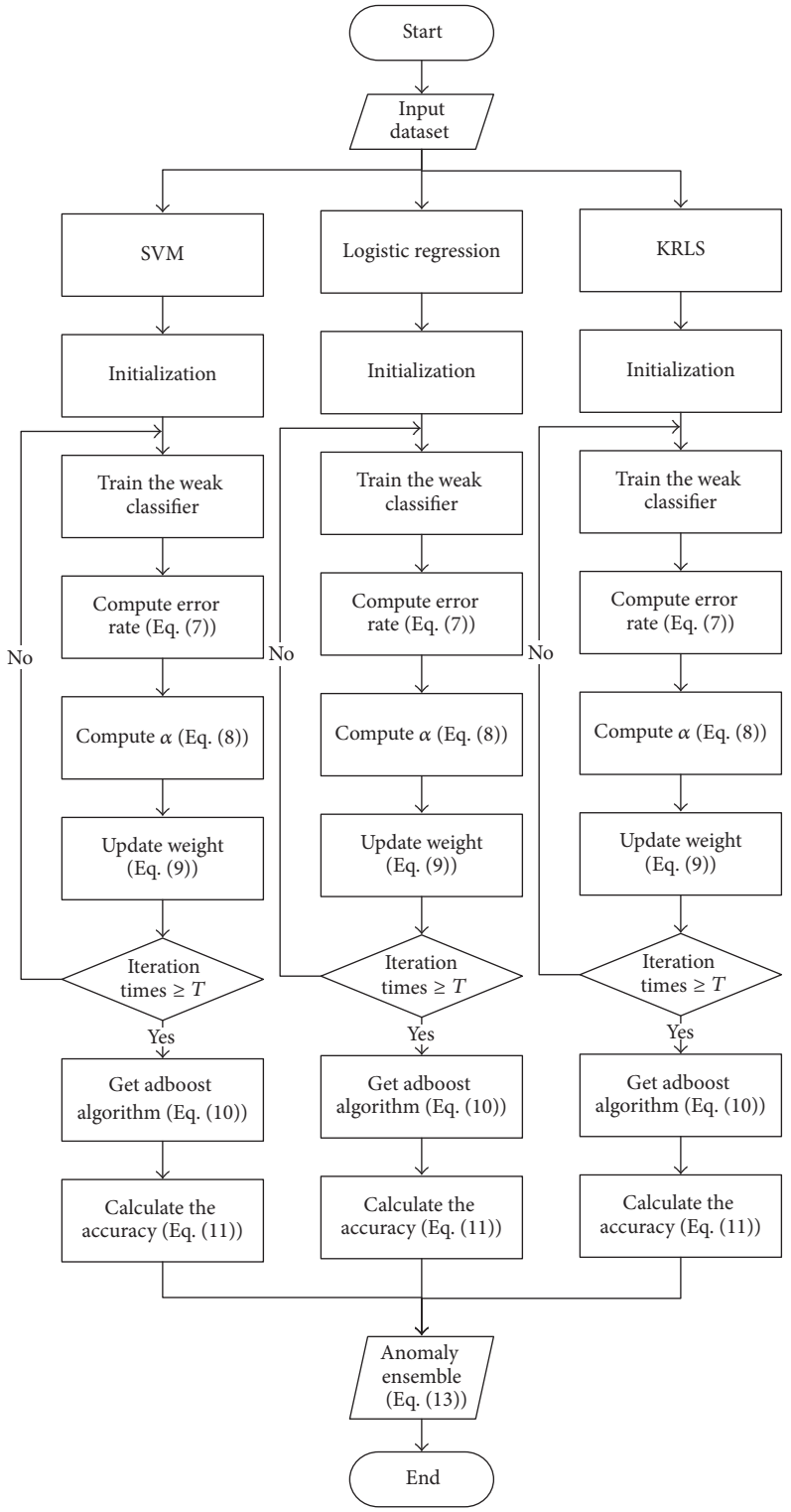

Figure 1: Workflow of HELM algorithm.

Process

(1) For $m = 1, \ldots, \mathscr{L}$,

(2) initialize the weight distribution $D_1(i) = 1/n$ (n is the number of examples; i is the index of the example),

(3) for $t = 1, \ldots, T$

(4) based on the sample distribution D_t and $\mathscr{L}_m$, we train the weak classifier h_t,

(5) compute the error (ε_t) for h_t

$$\varepsilon_t = \frac{\text{number of incorrectly classified example}}{\text{total number of examples}}, \tag{7}$$

(6) compute the weight (α_t) for h_t

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}, \tag{8}$$

(7) update the weight for each sample

$$D_{t+1}(i) = \frac{D_t(i)}{\text{sum}(D)} \begin{cases} \exp(-\alpha_t), & \text{if } h_t(x_i) = y_i, \\ \exp(\alpha_t), & \text{if } h_t(x_i) \neq y_i, \end{cases} \tag{9}$$

(8) end,

(9) obtain the ensemble learning classifier H_m by adboost algorithm [49, 50]

$$H_m(x) = \text{sign}(f(x)) = \text{sign}\sum_{t=1}^{T}\alpha_t h_t(x), \tag{10}$$

(10) calculate the accuracy of H_m

$$P_{H_m} = \frac{\text{number of correctly classified example}}{\text{total number of examples}}, \tag{11}$$

(11) end,

(12) assign a weight w_{H_m} to each H_m

$$w_{H_m} = \frac{P_{H_m}}{P_{H_1} + P_{H_2} + P_{H_3}}. \tag{12}$$

Output. Anomaly ensemble:

$$\text{HELM}(x) = \text{sign}\sum_{m=1}^{3} w_{H_m} H_m(x). \tag{13}$$

(4) Generalized Kernel Recursive Maximum Correntropy (GKRMC) Algorithm. It is well known that linear regression models can quickly estimate the occurrence rate of CRC. Nonetheless, using nonlinear model should sacrifice the computing cost to obtain the high predictive accuracy. Regarding the nature of our collected data, this study developed a nonlinear regression algorithm, GKRMC (Pseudocode 1), which can significantly increase the predictive accuracy with a reasonable computing cost. GKRMC is based on the kernel recursive least squares (KRLS) algorithm [45, 52–55] and the novel concept of the generalized correntropy [56]. Equation (14) gives the corresponding weighted and regularized cost function.

$$J = \max_{\Omega}\sum_{i=1}^{j}\beta^{i-j}G_{\alpha,\beta}\left(d_j - \Omega^T\varphi_j\right) - \frac{1}{2}\beta^i\gamma_2\left\|\Omega\right\|^2, \tag{14}$$

where $G_{\alpha,\beta}(\varepsilon) = (\alpha/2\beta\Gamma(1/\alpha))\exp(-|\varepsilon/\beta|^\alpha) = \gamma_{\alpha.\beta}\exp(-\lambda|\varepsilon|^\alpha)$, $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ is the shape parameter, β is the forgetting factor and it is set to 1, φ_i stands for $\varphi(u_i)$, with φ being the nonlinear mapping induced by a Mercer kernel, γ_2 is the regularization factor, i, j denote the numerical order of the samples, and $\gamma_{\alpha,\beta} = \alpha/(2\beta\Gamma(1/\alpha))$ is the normalization constant. Setting its gradient with respect to Ω equal to zero, one can obtain the solution as

$$\Omega_i = \left(\Phi_i B_i \Phi_i^T + \gamma_2\beta^i{\sigma_1}^\alpha\mathbf{I}\right)^{-1}\Phi_i B_i d_i, \tag{15}$$

where $\Phi_i = [\varphi_1, \varphi_2, \ldots, \varphi_i]$, $\sigma_1 = \beta^{\alpha/2}$ and $\mathbf{I}$ is an identity matrix.

$$B_i = \text{diag}\begin{bmatrix} \beta^{i-1}\left(d_1 - \Omega^T\varphi_1\right)^{\alpha-2} \times \left(\dfrac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\dfrac{d_1 - \Omega^T\varphi_1}{\sigma_1}\right|^\alpha\right) \\ \beta^{i-2}\left(d_2 - \Omega^T\varphi_2\right)^{\alpha-2} \times \left(\dfrac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\dfrac{d_2 - \Omega^T\varphi_2}{\sigma_1}\right|^\alpha\right) \\ \vdots \\ \left(d_1 - \Omega^T\varphi_1\right)^{\alpha-2} \times \left(\dfrac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\dfrac{d_i - \Omega^T\varphi_i}{\sigma_1}\right|^\alpha\right) \end{bmatrix}. \tag{16}$$

Generalized Kernel Recursive Maximum Correntropy
Initialization:
$\mathbf{Q}_1 = (\gamma_2 \beta {\sigma_1}^{\alpha} + G_{\sigma_2}(\mathbf{u}_1 - \mathbf{u}_1))^{-1}$
$\mathbf{a}_1 = \mathbf{Q}_1 d_1$
Computation:
Iterate for $i > 1$:
$\mathbf{h}_i = [G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_1), \ldots, G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_{i-1})]^T$
$y_i = \mathbf{h}_i^T \mathbf{a}_{i-1}$
$e_i = d_i - y_i$
$\mathbf{z}_i = \mathbf{Q}_{i-1}\mathbf{h}_i$
$\theta_i = (\exp(-{e_i}^{\alpha}/2{\sigma_1}^2))^{-1}$
$r_i = \gamma_2 \beta^i {\sigma_1}^{\alpha}\theta_i + G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_i) - \mathbf{z}_i^T \mathbf{h}_i$
$\mathbf{Q}_i = r_i^{-1} \begin{bmatrix} \mathbf{Q}_{i-1} r_i + \mathbf{z}_i \mathbf{z}_i^T & -\mathbf{z}_i \\ -\mathbf{z}_i^T & 1 \end{bmatrix}$
$\mathbf{a}_i = \begin{bmatrix} \mathbf{a}_{i-1} - \mathbf{z}_i r_i^{-1} e_i \\ r_i^{-1} e_i \end{bmatrix}$

Pseudocode 1: Pseudocode of GKRMC.

Using the matrix inversion lemma [54], we have

$$\left(\Phi_i B_i \Phi_i^T + \gamma_2 \beta^i {\sigma_1}^{\alpha} \mathbf{I}\right)^{-1} \Phi_i B_i = \Phi_i \left(\Phi_i^T \Phi_i + \gamma_2 \beta^i {\sigma_1}^{\alpha} B_i^{-1}\right)^{-1}. \tag{17}$$

Substituting (17) into (15) yields

$$\Omega_i = \Phi_i \left(\Phi_i^T \Phi_i + \gamma_2 \beta^i {\sigma_1}^{\alpha} B_i^{-1}\right)^{-1} d_i. \tag{18}$$

The weight vector can be expressed explicitly as a linear combination of the transformed data; that is, $\Omega_i = \Phi_i a_i$, where the coefficients vector $a_i = (\Phi_i^T \Phi_i + \gamma_2 \beta^i \sigma_1^{\alpha} B_i^{-1})^{-1} d_i$ can be computed using the kernel trick. Denote $Q_i = (\Phi_i^T \Phi_i + \gamma_2 \beta^i \sigma_1^{\alpha} B_i^{-1})^{-1}$; we have

$$Q_i = \begin{bmatrix} \Phi_{i-1}^T \Phi_{i-1} + \gamma_2 \beta^i \sigma_1^{\alpha} B_{i-1}^{-1} & \Phi_{i-1}^T \varphi_i \\ \varphi_i^T \Phi_{i-1} & \varphi_i^T \varphi_i + \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i \end{bmatrix}^{-1}, \tag{19}$$

where $\theta_i = (d_i - \Omega^T \varphi_i)^{\alpha-2} \times (\alpha^2 / 2\sigma_1 \Gamma(1/\alpha)) \times \exp(-|(d_i - \Omega^T \varphi_i)/\sigma_1|^{\alpha})$. It is easy to observe that

$$Q_i^{-1} = \begin{bmatrix} Q_{i-1}^{-1} & h_i \\ h_i^T & \varphi_i^T \varphi_i + \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i \end{bmatrix}, \tag{20}$$

where $h_i = \Phi_{i-1}^T \varphi_i$. Using the block matrix inversion identity, we can derive

$$Q_i^{-1} = \begin{bmatrix} Q_{i-1} r_i + z_i z_i^T & -z_i \\ -z_i^T & 1 \end{bmatrix}, \tag{21}$$

where $z_i = Q_{i-1} h_i$ and

$$r_i = \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i + \varphi_i^T \varphi_i - z_i^T h_i. \tag{22}$$

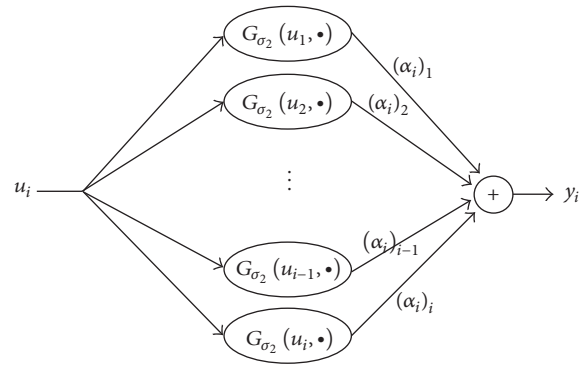

Figure 2: Network topology of GKRMC at ith iteration.

So,

$$a_i = Q_i d_i = r_i^{-1} \begin{bmatrix} Q_{i-1} r_i + z_i z_i^T & -z_i \\ -z_i^T & 1 \end{bmatrix} \begin{bmatrix} d_{i-1} \\ d_i \end{bmatrix} = \begin{bmatrix} a_{i-1} - z_i r_i^{-1} e_i \\ r_i^{-1} e_i \end{bmatrix}, \tag{23}$$
$$e_i = d_i - \Omega^T \varphi_i.$$

Then we obtain the GKRMC algorithm, in which the coefficients update follows (23) and r_i is computed by (22). This study uses $G_{\sigma_2}(\cdot)$ to denote the Gaussian kernel for RKHS [57], with σ_2 being the kernel size. The GKRMC produces a RBF [58] type network, which is a linear combination of the kernel functions (Figure 2). a_i denotes the coefficient vector of the network at iteration i and $(a_i)_j$ denotes the jth scalar in a_i.

3. Results

3.1. The Results of the Biological Classification. In past decades, a number of candidate factors implicated in CRC risk are proposed by epidemiology studies, which can be divided into two groups in total, genetic factors and non-genetic factors. The genetic factors' group consists of many SNPs, and the nongenetic factors' group is comprised of several kinds of environment factors. According to the biological characteristics and the manner that human beings are exposed to environmental factors in whole lifetime, the raw big CRC-related genetic and environmental data can be classified into four biological categories: SNPs, demographic characteristics, lifestyles, and foods as in Table 1.

3.2. Results of Original Data Dimensionality Reduction. To process the dataset of SNPs, demographic characteristics, lifestyle and food, SPCA, and entropy and relief methods are employed, respectively.

Table 2 shows the principal components for the SNPs, demographic characteristics, and lifestyle and food by SPCA method, respectively. The result of the SPCA is listed in Supplementary S5.

Table 1: Results of biological classification.

Categories	Illustration
SNPs	Polymorphism distribution of genes
Demographic characteristics	Including factors like age, sex, body weight, income levels, and educations, which represents the individually biological or social-psychological features
Lifestyles	Behavioral factors, such as smoking and alcohol drinking
Foods	The amount of food intake

Table 2: The results by SPCA method.

SNPs	rs10046, rs10505477, rs1152579, rs1229984, rs1255998, rs1256030, rs1256049, rs1271572, rs12953717, rs1329149, rs16941669, rs17033, rs1801132, rs2075633, rs2077647, rs3798758, rs3820033, rs4767939, rs4767944, rs4939827, rs676387, rs6905370, rs6983267, rs7296651, rs7837688, rs827421, rs886205, rs928554, rs9322354, rs9340799
Demographic characteristics	Cholesterol, blood triglyceride, psychological trauma, depression, age, exercise, BMI, physical activity, activity, marriage status, emotion status
Lifestyles	Smoking, drinking, coffee consumption, drinking and smoking in the same time point, tea consumption
Foods	Grains, melons, bean products, roots, vegetables, fruits, eggs and milk, mushrooms, oil, seasoning, meat, seafood, pickles

We consider that the features with high weight will result in the colorectal cancer when the relief algorithm is applied to extract key features from the dataset. The result of relief algorithm is shown in Figure 3. In the upper part of Figures 3(a), 3(b), 3(c), and 3(d), the horizontal axis shows the feature numerical number and the vertical axis shows the feature weight. In the lower part of Figures 3(a), 3(b), 3(c), and 3(d), the horizontal axis shows the feature weight and the vertical axis shows the feature value, while the bars in Figure 3 represent the numbers of the features according to the feature weight.

Table 3 shows the results of dimensionality reduction by entropy method for the SNPs, demographic characteristics, and lifestyle and food, respectively. The entropy $H(X)$ in (2) is for data dimensionality reduction.

Regarding the results of Figure 3, Table 4 shows the common factors for the SNPs, demographic characteristics, and lifestyle and food by relief method, respectively.

Figure 4 shows the interaction results for the three dimensionality reduction methods. Figure 4(a) indicates that rs1256030 is the mutually explored biomarker by SPCA, entropy, and relief; rs10046, rs1152579, rs676387, rs6905370, rs928554, and rs6983267 are the mutually explored biomarkers by SPCA and entropy and rs4939827, rs4767944, rs1801132, rs4767939, rs10505477, rs3798758, and rs2075633 are the mutually explored biomarker by SPCA and relief.

Table 3: The results by entropy method.

SNPs	rs6983267, rs1256030, rs10046, rs928554, rs1152579, rs690537, rs676387
Demographic characteristics	Age, BMI, blood triglyceride, depression, mental stress, psychological trauma
Lifestyles	Drinking and smoking in the same time point, drinking
Foods	Vegetables, nuts, mushrooms, seasoning, pickles, grains

Table 4: The results by relief method.

SNPs	rs10505477, rs1256030, rs1801132, rs2071454, rs2075633, rs2228480, rs2249695, rs2486758, rs3798758, rs4767939, rs4767944, rs4939827
Demographic characteristics	Age, BMI, physical activity, activity, family number, emotion status, temperament, mental stress, psychological trauma, depression, cholesterol
Lifestyles	Drinking, tea consumption, drinking and smoking in the same time point
Foods	Nuts, vegetables, meat, eggs and milk, seafood

Figure 4(b) indicates that age, depression, blood triglyceride, and BMI are the mutually explored biomarkers by SPCA, entropy, and relief; blood triglyceride is the mutually explored biomarker by SPCA and entropy; cholesterol, activity, emotion status, and physical activity are the mutually explored biomarkers by SPCA and relief and mental stress is the mutually explored biomarkers by entropy and relief.

Figure 4(c) indicates that drinking and drinking and smoking in same time point are the mutually explored biomarkers by SPCA, entropy, and relief; tea consumption is the mutually explored biomarker by SPCA and relief.

Figure 4(d) indicates that vegetables are the mutually explored biomarkers by SPCA, entropy, and relief; mushrooms, seasoning, pickles, and grains are the mutually explored biomarkers by SPCA and entropy; eggs and milk, meat, and seafood are the mutually explored biomarkers by SPCA and relief and nuts is the mutually explored biomarker by entropy and relief.

We have 36 features mutually explored by every two of the SPCA, entropy, and relief methods.

By U test [59], Table 5 shows that 13 out of 36 features have small p value.

Table 6 shows that 13 features with small p value are important biomarkers.

3.3. Results of Regression. According to the dimensionality reduction analysis, there are 13 biomarkers selected as the classifier for these four biological datasets. Next, we employ LR, SVM, KRLS, HELM, and GKRMC algorithm to build up the predictive cancer model based on these selected classifiers.

Table 7 presents four measures (accuracy, sensitivity, specificity, and precision) to assess how good or how "accurate" the classifier is.

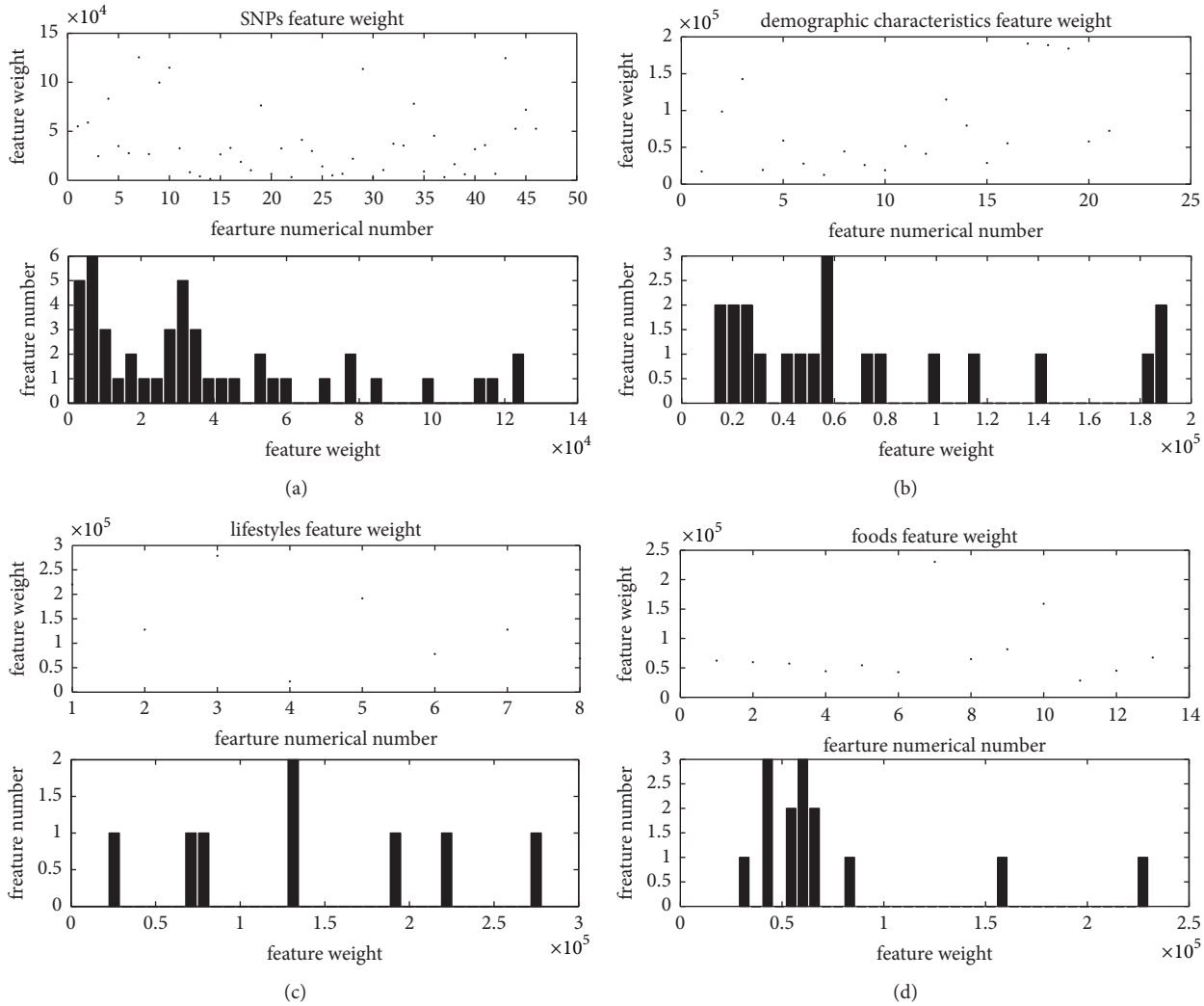

Figure 3: Feature selection by relief algorithm: (a) SNPs feature (note: the feature numerical number in the upper figure is regarding Supplementary S1 from columns B(1) to AU(46)), (b) demographic characteristics feature (note: the Feature numerical number in the upper figure is regarding Supplementary S2 from columns A(1) to U(21)), (c) lifestyle feature (note: the feature numerical number in the upper figure is regarding Supplementary S3 from columns B(1) to I(8)), and (d) food feature (note: the feature numerical number in the upper figure is regarding Supplementary S4 from columns B(1) to O(14)).

There are 1298 cases-control samples, 369 of which are case and 929 of which are control. Cross validation [60] method randomly chooses 75% of samples (973 samples) as the training dataset and the rest (325 samples) are used for testing dataset. Since cross validation introduces the random effect, we have to repeat the experiment 10 times. Figure 5 shows that GKRMC always has the greatest sensitive, precision, and accuracy values as well as greater specificity value compared to KRLS. Moreover, Table 8 lists the average value and standard deviation of the classification measurement for each algorithm.

4. Discussion and Conclusion

For CRC tumorigenesis, both genetic and environmental factors, as well as their interaction, playing important role in CRC risk is already the common view of most previously studies [61], but to figure out how to predict the occurrence of CRC by using the risk factors is still a challenge today. In the present study, we use big data of 1298 samples from a CRC case-control study in which relatively complete information of genetic and demographic characteristics and life style and food intake is simultaneously collected; furthermore, we expect to develop such a CRC-risk predictive model that not only can explore which risk factors included in the collected big dataset have significant impact on the occurrence of CRC, but also can accurately predict the occurrence of CRC as early as possible.

Such big datasets are classified into four different categories in the biological classification stage. And 13 of all explored potential biomarkers consisting of 4 SNPs, 6 demographic characteristics, 1 lifestyle factor, and 2 foods are screened out in data dimensionality reduction stage.

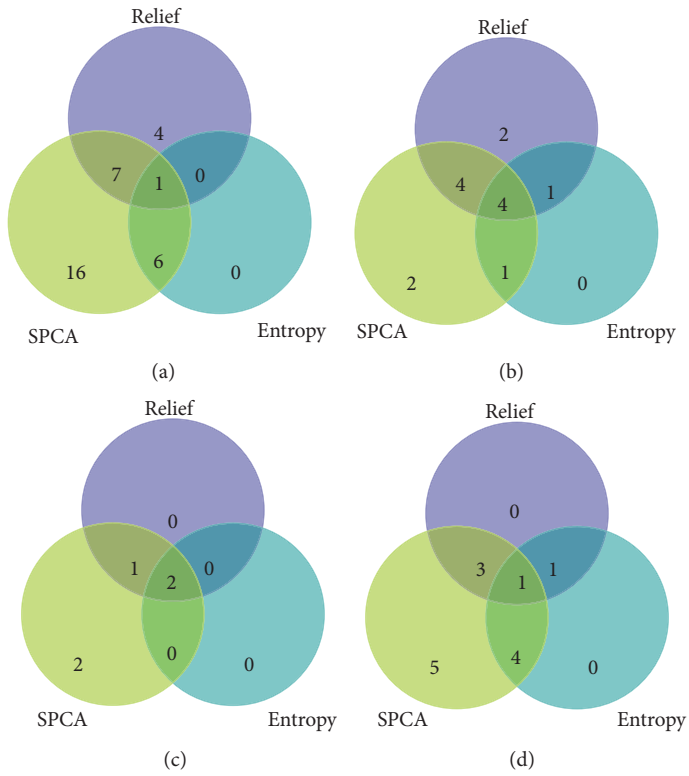

Figure 4: Venn plots of (a) SNPs, (b) demographic characteristics, (c) lifestyle, and (d) food.

Table 5: p value of 13 important biomarkers.

Biomarkers	p value
rs10046	0.0172
rs1256030	0.0004
rs6766387	0.0015
rs6983267	0.0000
age	0.0152
BMI	0.0019
Physical activity	0.0030
Emotion status	0.0247
Mental stress	0.0213
Cholesterol	0.0000
Drinking and smoking in the same time point	0.0000
Vegetables	0.0000
Seafood	0.0023

Table 6: Mutually explored biomarkers.

SNPS	rs10046, rs1256030, rs676387, 6983267
Demographic characteristics	Age, BMI, physical activity, emotion status, mental stress, cholesterol
Lifestyle	Drinking and smoking in the same time point
Foods	Vegetables, seafood

Unlike pure mathematical formulae, the biological rationality of such model depends on whether the selected biomarkers can be biologically explained as validated etiology of colorectal cancer supported by either population-based association study or biological function-based mechanisms experimental study. And then, these explored biomarkers can be used as the classifiers for the predictive model to access the risk of colorectal cancer in the regression analysis stage.

In fact, results from substantial epidemiology studies focusing on CRC risk/protective factors provide evidences for the associations between each category and risk of CRC. For the genetic variations, at least 2 (rs10046, rs6983267) of the 4 currently selected SNPs listed in Table 5 were reported to have significant association with CRC risk in either genome-wide association studies or candidate gene based study [59, 62]. Particularly, SNP rs6983267 is one of the most significant variations associated with increasing CRC risk in Caucasians, Asians, and Africans [63]. Regarding the other two selected SNPs (rs1256030, rs676387) located, respectively, in estrogen receptor beta gene (ESR2) and 17 β-hydroxysteroid dehydrogenases gene (HSD17B1) (both are estrogen metabolism pathway genes), though there is no direct evidence supporting their association with CRC, they both are found significantly associated with cancers such as liver and ovarian cancers [64, 65]. Moreover, considerable evidence from epidemiological and metabolic studies support that the estrogen metabolism pathway genes undoubtedly play an important role in CRC and other cancers [66], which implies the potential that the two SNPs may affect the susceptibility of CRC.

For demographic factors, almost all the 6 selected factors have been reported to be the unfavorable factors for CRC risk in a bunch of previous studies [67, 68].

Table 7: The definition of the classification measurement.

Measure	Formula	Illustration
Sensitivity	$\frac{TP}{P}$	TP: the number of true positives P: the number of positives
Specificity	$\frac{TN}{N}$	TN: the number of true negatives N: the number of negatives
Precision	$\frac{TP}{TP + FP}$	TP: the number of true positives FP: the number of false positives
Accuracy	$\frac{TP + TN}{P + N}$	TP: the number of true positives TN: the number of true negatives P: the number of positives N: the number of negatives

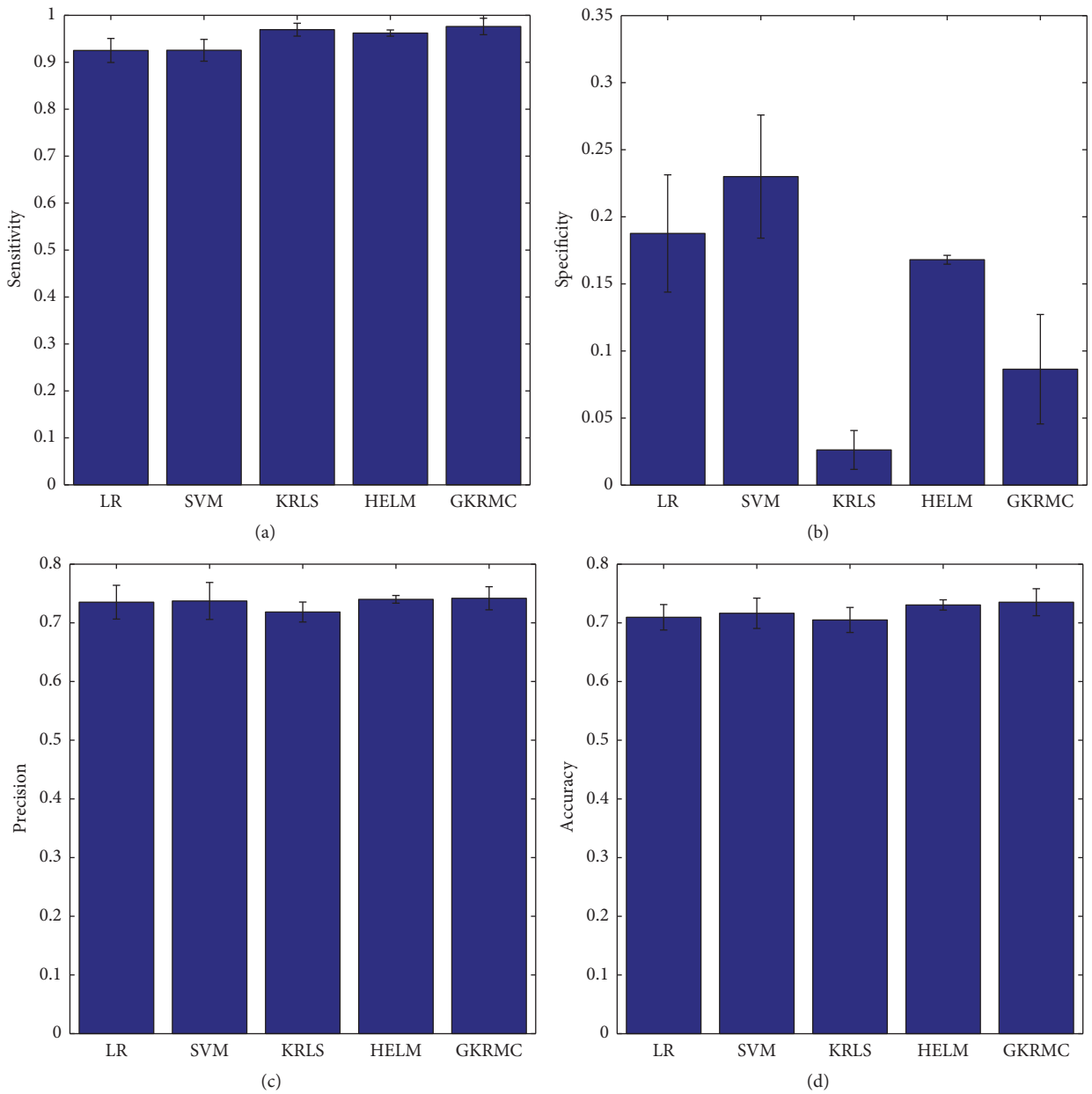

Figure 5: Predictive performance for the LR, SVM, KRLS, HELM, and GKRMC.

Table 8: The mutually explored biomarkers.

	LR	SVM	KRLS	HELM	GKRMC
Sensitivity	0.9251 ± 0.0256	0.9255 ± 0.0233	0.9694 ± 0.0137	0.9621 ± 0.0066	0.9762 ± 0.0175
Specificity	0.1876 ± 0.0437	0.2300 ± 0.0459	0.0262 ± 0.0145	0.1680 ± 0.0033	0.0864 ± 0.0408
Precision	0.7351 ± 0.0288	0.7372 ± 0.0315	0.7184 ± 0.0170	0.7400 ± 0.0066	0.7418 ± 0.0197
Accuracy	0.7095 ± 0.0217	0.7163 ± 0.0258	0.7049 ± 0.0213	0.7305 ± 0.0087	0.7351 ± 0.0230

For lifestyles, alcohol drinking and smoking are proved as two significant risk factors of CRC [18, 68]. Alcohol drinking, in a dose-response manner, evidently contributes to the increase of CRC risk. Meanwhile, obvious positive associations between CRC risk and cigarette smoking are observed in most measures [69].

For food, extensive epidemiologic and experimental studies confirm their important roles in the development of CRC. For example, higher consumption of vegetables and seafood is always associated with relatively lower CRC risk due to their relatively high content of antioxidant nutrients such as dietary fiber, vitamins, and long-chain unsaturated fatty acids [70–73]. On the contrary, the excessive consumption of smoked/salted/processed meat is linked to higher risk of colorectal neoplasia [73].

In general, it is demonstrated that the 13 currently explored biomarkers can be used as the classifiers in the regression analysis stage, which is supported by these manually reviewed experimental evidences [59, 63, 67, 69–71].

Although LR and SVM may perform very well for linear systems, their performance will get worse when applied to the nonlinear and non-Gaussian situations [74], which is rather common in real world applications. Therefore, we suggest using nonlinear regression algorithm to process our dataset, which is comprised of continuous and discrete data with multivariate data type. However, using classical nonlinear algorithm such as KRLS will suffer from outliers.

To overcome the shortcoming of both linear and conventional nonlinear regression algorithms, this study proposes an ensemble learning model (HELM) and a generalized kernel recursive maximum correntropy (GKRMC) algorithm to increase the predictive power of the model. Next, we analyze the reason why HELM and GKRMC can outperform LR, SVM, and KRLS algorithms.

HELM is an ensemble learning algorithm, which integrates linear and nonlinear classifiers to classify the data points. Based on the previous study [75], the diversity of weak classifiers is one of the evaluation criteria for ensemble algorithm. HELM includes both linear (SVM and logistic regression) and nonlinear (KRLS) classifiers and its superior performance has been shown in Figure 5.

The cost function of GKRMC (see (14)) is so robust that is not sensitive to large outliers as KRLS, since an exponentially weighted mechanism $G_{\alpha,\beta}(d_i - \Omega^T \varphi_i)$ of (14) can assign greater weight to the samples with smaller error but not to the samples with greater error. Since the big dataset usually consists of outliers [29, 76], GKRMC can achieve the higher predictive accuracy with the less standard deviation (Table 8) than KRLS. As mentioned before, the predictive power of GKRMC should be better than LR, SVM, and KRLS due to the nature of nonlinear regression (Figure 5).

In conclusion, this study proposes a robust CRC-risk predictive model to analyze the big data with information of genetic variations and environmental exposure for the CRC patients and cancer-free controls. The research results indicate that both genetic and environmental related factors explored by our model play the significant roles in the occurrence of CRC and the innovated HELM and GKRMC can increase the predictive power of the model.

However, this novel predictive model is the first step in predicting the risk of CRC tumor growth. Except for the environment factors and SNPs involved in the current model, if other factors such as pathway-pathway and pathway-environment interactions are included, there will be a higher chance to find a set of variations which may be integrative biomarkers, as proved in other researches [77, 78]. A limitation of our study is that there is only a finite number of tag SNPs located in a relatively small number of genes, which results in the nonuse of employing pathway interaction into model construction. Also, how to improve the GKRMC's specificity is an important topic for future study, which will further improve the whole system's performance. While extensions will be necessary to account in greater detail for the complexity of the CRC involved, we believe that if properly combined with more experimental data such as RNA sequence analysis and recent modeling techniques[79–86], advanced in silico platforms such as this one will evolve into powerful integrative research platforms that improve our understanding of CRC tumorigenesis.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the General Program from National Natural Science Foundation of China (nos. 81273156, 30771841, 61372138, and 61372152), Chongqing Excellent Youth Award and the Chinese Recruitment Program of Global Youth Experts, and the Fundamental Research Funding of the Chinese Central Universities (nos. XDJK2014B012 and XDJK2016A00).

References

[1] M. A. Arafa, M. I. Waly, J. Sahar, A. K. Ahmed, and S. Sunny, "Dietary and lifestyle characteristics of colorectal cancer in

Jordan: a case-control study," *Asian Pacific Journal of Cancer Prevention*, vol. 12, no. 8, pp. 1931–1936, 2011.

[2] M. M. Center, A. Jemal, and E. Ward, "International trends in colorectal cancer incidence rates," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 18, pp. 1688–1694, 2009.

[3] M. Li and J. Gu, "Changing patterns of colorectal cancer in China over a period of 20 years," *World Journal of Gastroenterology*, vol. 11, no. 30, pp. 4685–4688, 2005.

[4] S. Liu, R. Zheng, M. Zhang, S. Zhang, X. Sun, and W. Chen, "Incidence and mortality of colorectal cancer in China, 2011," *Journal of Thoracic Disease*, vol. 5, pp. 330–336, 2014.

[5] D. M. Parkin, C. A. Stiller, and J. Nectoux, "International variations in the incidence of childhood bone tumours," *International Journal of Cancer*, vol. 53, no. 3, pp. 371–376, 1993.

[6] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, and J. Lortet-Tieulent, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[7] Y. Zhao, X. Deng, Z. Wang, Q. Wang, and Y. Liu, "Genetic polymorphisms of DNA repair genes XRCC1 and XRCC3 and risk of colorectal cancer in chinese population," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 2, pp. 665–669, 2012.

[8] J. He and W. Chen, "Chinese cancer registry annual report 2012," Press of Military Medical Sciences, Beijing, China, pp. 68–71, 2012.

[9] S. D. Markowitz and M. M. Bertagnolli, "Molecular basis of colorectal cancer," *The New England Journal of Medicine*, vol. 361, no. 25, pp. 2404–2460, 2009.

[10] M. Van Engeland, S. Derks, K. M. Smits, G. A. Meijer, and J. G. Herman, "Colorectal cancer epigenetics: Complex simplicity," *Journal of Clinical Oncology*, vol. 29, no. 10, pp. 1382–1391, 2011.

[11] A. A. Ghazarian, N. I. Simonds, K. Bennett et al., "A review of NCI's extramural grant portfolio: Identifying opportunities for future research in genes and environment in cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 22, no. 4, pp. 501–507, 2013.

[12] E. J. Kuipers, W. M. Grady, D. Lieberman et al., "Colorectal cancer," *Nature Reviews Disease Primers*, vol. 1, pp. 1–25, 2015.

[13] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *The Lancet*, vol. 383, no. 9927, pp. 1490–1502, 2014.

[14] V. T. DeVita Jr. and S. A. Rosenberg, "Two hundred years of cancer research," *New England Journal of Medicine*, vol. 366, no. 23, pp. 2207–2214, 2012.

[15] J. Kim, Y. A. Cho, D.-H. Kim et al., "Dietary intake of folate and alcohol, MTHFR C677T polymorphism, and colorectal cancer risk in Korea," *American Journal of Clinical Nutrition*, vol. 95, no. 2, pp. 405–412, 2012.

[16] K. Matsuo, T. Suzuki, H. Ito et al., "Association between an 8q24 locus and the risk of colorectal cancer in Japanese," *BMC Cancer*, vol. 9, article no. 379, 2009.

[17] K. Wakai, K. Hirose, K. Matsuo et al., "Dietary risk factors for colon and rectal cancers: a comparative case-control study," *Journal of Epidemiology*, vol. 16, no. 3, pp. 125–135, 2006.

[18] H. Yang, Y. Zhou, Z. Zhou et al., "A Novel Polymorphism rs1329149 of CYP2E1 and a Known Polymorphism rs671 of ALDH2 of Alcohol Metabolizing Enzymes Are Associated with Colorectal Cancer in a Southwestern Chinese Population," *Cancer Epidemiology Biomarkers and Prevention*, vol. 18, no. 9, pp. 2522–2527, 2009.

[19] Y.-Z. Wu, H. Yang, L. Zhang et al., "Application of crossover analysis-logistic regression in the assessment of gene-environmental interactions for colorectal cancer," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 5, pp. 2031–2037, 2012.

[20] H. Huang, P. Chanda, A. Alonso, J. S. Bader, and D. E. Arking, "Gene-based tests of association," *PLoS genetics*, vol. 7, no. 7, Article ID e1002177, 2011.

[21] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, no. 3, pp. 376–382, 2003.

[22] J. H. Moore, J. C. Gilbert, C.-T. Tsai et al., "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of Theoretical Biology*, vol. 241, no. 2, pp. 252–261, 2006.

[23] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[24] M. Li, C. Ye, W. Fu, R. C. Elston, and Q. Lu, "Detecting genetic interactions for quantitative traits with *U*-statistics," *Genetic Epidemiology*, vol. 35, no. 6, pp. 457–468, 2011.

[25] A. S. Andrew, H. H. Nelson, K. T. Kelsey et al., "Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility," *Carcinogenesis*, vol. 27, no. 5, pp. 1030–1037, 2006.

[26] W. Meredith, R. Rutledge, S. M. Fakhry, S. Emery, and S. Kromhout-Schiro, "The conundrum of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the Glasgow verbal score from the Glasgow eye and motor scores," *Journal of Trauma*, vol. 44, no. 5, pp. 839–845, 1998.

[27] R. Rutledge, C. W. Lentz, S. Fakhry, and J. Hunt, "Appropriate use of the glasgow coma scale in intubated patients: a linear regression prediction of the glasgow verbal score from the glasgow eye and motor scores," *The Journal of Trauma*, vol. 41, no. 3, pp. 514–522, 1996.

[28] C. Zheng, L. Xing, T. Li, H. Yang, J. Cao et al., "Developing a robust colorectal cancer (CRC) risk predictive model with the big genetic and environment related CRC data," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1885–1893, IEEE, Shenzhen, China, 2016.

[29] C. A. Mattmann, "Computing: A vision for data science," *Nature*, vol. 493, no. 7433, pp. 473–475, 2013.

[30] Z. Y. Zhou, B. Q. Takezaki TMo, H. M. Sun, W. C. Wang, L. P. Sun, and S. X. Liu, "Development of a semi-quantitative food frequency questionnaire to determine variation in nutrient intakes between urban and rural areas of Chongqing, China," *Asia Pacific Journal of Clinical Nutrition*, vol. 13, pp. 273–283, 2004.

[31] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The International HapMap Project Web site," *Genome Research*, vol. 15, no. 11, pp. 1592-1593, 2005.

[32] O. De la Cruz and P. Raska, "Population structure at different minor allele frequency levels," *BMC Proceedings*, vol. 8, pp. 1–5, 2014.

[33] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[34] C. J. Mulligan, R. W. Robin, M. V. Osier et al., "Allelic variation at alcohol metabolism genes (ADH1B, ADH1C, ALDH2) and alcohol dependence in an American Indian population," *Human Genetics*, vol. 113, no. 4, pp. 325–336, 2003.

[35] M. Crous-Bou, G. Rennert, D. Cuadras et al., "Polymorphisms in alcohol metabolism genes ADH1B and ALDH2, alcohol consumption and colorectal cancer," *PLoS ONE*, vol. 8, no. 11, Article ID e80158, 2013.

[36] T. S. Kang, S. W. Woo, H. J. Park, Y. Lee, and J. Roh, "Comparison of genetic polymorphisms of CYP2E1, ADH2, and ALDH2 genes involved in alcohol metabolism in Koreans and four other ethnic groups," *Journal of Clinical Pharmacy and Therapeutics*, vol. 34, no. 2, pp. 225–230, 2009.

[37] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[38] I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2nd edition, 2010.

[39] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[40] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational Graphical Statistics*, p. 2007, 2012.

[41] T. M. Cover and J. A. Thomas, *Thomas JA. Elements of Information Theory*, Cognitive Science - A Multidisciplinary Journal, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2005.

[42] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[43] Z.-H. Zou, Y. Yun, and J.-N. Sun, "Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment," *Journal of Environmental Sciences* , vol. 18, no. 5, pp. 1020–1023, 2006.

[44] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1035–1051, 2007.

[45] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[46] K. Koh, S. J. Kim, and S. P. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[47] J. Pearce and S. Ferrier, "Evaluating the predictive performance of habitat models developed using logistic regression," *Ecological Modelling*, vol. 133, no. 3, pp. 225–245, 2000.

[48] A. Al-Anazi and I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs," *Engineering Geology*, vol. 114, no. 3-4, pp. 267–277, 2010.

[49] T. G. Dietterich, "Machine-learning research: four current directions," *AI Magazine*, vol. 18, no. 4, pp. 97–136, 2000.

[50] X. Wu, V. Kumar, J. R. Quinlan et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[51] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, vol. 14, Springer, Berlin, Germany, 1995.

[52] W. Liu, P. P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 543–554, 2008.

[53] B. Chen, S. Zhao, P. Zhu, and J. C. Príncipe, "Quantized kernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 22–32, 2012.

[54] B. Chen, S. Zhao, P. Zhu, and J. C. Principe, "Quantized kernel recursive least squares algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1484–1491, 2013.

[55] W. Liu, J. C. Principe, and S. Haykin, "Kernel Adaptive Filtering: A Comprehensive Introduction," in *Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2010.

[56] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Príncipe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.

[57] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Space in Probability and Statistics*, Kluwer Academic, Boston, Mass, USA, 2004.

[58] M. J. L. Orr, "Introduction to Radial Basis Function Networks," *Journal of Vitamin Research*, vol. 4, pp. 2797–2800, 1967.

[59] N. H. Ramzi, J. K. Chahil, S. H. Lye et al., "Role of genetic & environment risk factors in the aetiology of colorectal cancer in Malaysia," *Indian Journal of Medical Research*, vol. 139, pp. 873–882, 2014.

[60] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1137-43, International Joint Conference on Artificial Intelligence, Stanford, Calif, USA, 2001.

[61] H. Mahdi, B. A. Fisher, H. Källberg et al., "Specific interaction between genotype, smoking and autoimmunity to citrullinated α-enolase in the etiology of rheumatoid arthritis," *Nature Genetics*, vol. 41, no. 12, pp. 1319–1324, 2009.

[62] R. Cui, Y. Okada, S. G. Jang et al., "Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population," *Gut*, vol. 60, no. 6, pp. 799–805, 2011.

[63] C. A. Haiman, L. le Marchand, J. Yamamato et al., "A common genetic risk factor for colorectal and prostate cancer," *Nature Genetics*, vol. 39, no. 8, pp. 954–956, 2007.

[64] G. Lurie, L. R. Wilkens, P. J. Thompson et al., "Genetic polymorphisms in the estrogen receptor beta (ESR2) gene and the risk of epithelial ovarian carcinoma," *Cancer Causes and Control*, vol. 20, no. 1, pp. 47–55, 2009.

[65] L. S. Zhang, F. Yuan, X. Guan et al., "Association of genetic polymorphisms in HSD17B1, HSD17B2 and SHBG genes with hepatocellular carcinoma risk," *Pathology and Oncology Research*, vol. 20, no. 3, pp. 661–666, 2014.

[66] A. Barzi, A. M. Lenz, M. J. Labonte, and H.-J. Lenz, "Molecular pathways: Estrogen pathway in colorectal cancer," *Clinical Cancer Research*, vol. 19, no. 21, pp. 5842–5848, 2013.

[67] T. E. Robsahm, B. Aagnes, A. Hjartåker, H. Langseth, F. I. Bray, and I. K. Larsen, "Body mass index, physical activity, and colorectal cancer by anatomical subsites: A systematic review and meta-analysis of cohort studies," *European Journal of Cancer Prevention*, vol. 22, no. 6, pp. 492–505, 2013.

[68] Z.-Y. Zhou, H. Yang, J. Cao, K. Tajima, K. Matsuo, and W.-C. Wang, *Dietary Risks: Folate, Alcohol and Gene Polymorphisms*, INTECH Open Access, 2012.

[69] H. Raskov, H. C. Pommergaard, J. Burcharth, and J. Rosenberg, "Colorectal carcinogenesis—update and perspectives," *World Journal of Gastroenterology*, vol. 20, no. 48, pp. 18151–18164, 2014.

[70] M. Song, W. S. Garrett, and A. T. Chan, "Nutrients, foods, and colorectal cancer prevention," *Gastroenterology*, vol. 148, no. 6, pp. 1244–1260, 2015.

[71] Q. Ben, J. Zhong, J. Liu et al., "Association between consumption of fruits and vegetables and risk of colorectal adenoma a prisma-compliant meta-Analysis of observational studies," *Medicine (United States)*, vol. 94, no. 42, p. e1599, 2015.

[72] S. S. Young, "Re: Low-fat dietary pattern and cancer incidence in the Women's Health Initiative Dietary Modification Randomized Controlled Trial," *Journal of the National Cancer Institute*, vol. 100, no. 4, p. 284, 2008.

[73] E. D. Kantor, J. W. Lampe, U. Peters, T. L. Vaughan, and E. White, "Long-chain omega-3 polyunsaturated fatty acid intake and risk of colorectal cancer," *Nutrition and Cancer*, vol. 66, no. 4, pp. 716–727, 2014.

[74] W. Linde, "Stable non-gaussian random processes: stochastic models with infinite variance," *Bulletin of the London Mathematical Society*, vol. 28, no. 430, 1994.

[75] E. C. Ensembles, E. S. O. Hyperspectrale, S. Yu, and S. Cao, "Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data," 2003.

[76] L. V. Subramaniam, "Big Data and Veracity Challenges," 2014, http://wwwisicalacin/~acmsc/TMW2014/LVSpdf.

[77] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, article 126, 2012.

[78] R. Visakh and K. A. Abdul Nazeer, "Identifying epigenetically dysregulated pathways from pathway–pathway interaction networks," *Computers in Biology and Medicine*, vol. 76, pp. 160–167, 2016.

[79] B. Jiang, W. Dai, A. Khaliq, M. Carey, X. Zhou, and L. Zhang, "Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation," *Mathematics and Computers in Simulation*, vol. 109, pp. 1–19, 2015.

[80] B. Jiang, A. Struthers, Z. Sun et al., "Employing graphics processing unit technology, alternating direction implicit method and domain decomposition to speed up the numerical diffusion solver for the biomedical engineering research," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 27, no. 11, pp. 1829–1849, 2011.

[81] H. Peng, T. Peng, J. Wen et al., "Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach," *Bioinformatics*, vol. 30, no. 13, pp. 1899–1907, 2014.

[82] Y. Xia, C. Yang, N. Hu et al., "Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model," *BMC Genomics*, vol. 18, article no. 950, 2017.

[83] L. Zhang, B. Jiang, Y. Wu et al., "Developing a multiscale, multi-resolution agent-based brain tumor model by graphics processing units," *Theoretical Biology and Medical Modelling*, vol. 8, no. 1, article no. 46, 2011.

[84] L. Zhang, M. Qiao, H. Gao et al., "Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation," *Nanoscale*, vol. 8, no. 31, pp. 14877–14887, 2016.

[85] L. Zhang, Y. Xue, B. Jiang et al., "Multiscale agent-based modelling of ovarian cancer progression under the stimulation of the STAT 3 pathway," *International Journal of Data Mining and Bioinformatics*, vol. 9, no. 3, pp. 235–253, 2014.

[86] L. Zhang and S. Zhang, "Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: 'Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition' by Qian Wang et al," *Physics of Life Reviews*, vol. 20, pp. 140–142, 2017.