Article

A Novel Prediction Scheme for Risk Factors of Second Colorectal Cancer in Patients with Colorectal Cancer

Wen-Chien Ting[1,2], Horng-Rong Chang[3,4*], Chi-Chang Chang[5*], Chi-Jie Lu[6,*]

[1] Division of Colorectal Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taiwan
[2] Institute of Medicine, Chung Shan Medical University, Taiwan
[3] Division of Nephrology, Department of Internal medicine, Chung Shan Medical University Hospital, Taiwan
[4] School of Medicine, Chung Shan Medical University, Taiwan
[5] School of Medical Informatics, Chung Shan Medical University & IT office, Chung Shan Medical University Hospital, Taiwan
[6] Graduate institute of Business Administration and Department of Information Management, Fu-Jen Catholic University, Taiwan

*Correspondence: chrcsmu@gmail.com; threec@csmu.edu.tw; 055099@mail.fju.edu.tw; Tel.: +886-4-24730022 12218

Abstract: In Taiwan, colorectal cancer is ranked second and third in terms of mortality and cancer incidence, respectively. In addition, medical expenditures related to colorectal cancer are considered to be the third highest. While advances in treatment strategies have provided cancer patients with longer survival, potentially harmful second primary cancers can occur. Therefore, second primary colorectal cancer analysis is an important issue with regard to clinical management. In this study, a novel predictive scheme was developed for predicting the risk factors associated with second colorectal cancer in patients with colorectal cancer by integrating five data mining classification techniques, including support vector machine, random forest, multivariate adaptive regression splines, extreme learning machine, and extreme gradient boosting. In total, 4,287 patients in the datasets provided by three hospital tumor registries were used. Our empirical results revealed that this proposed predictive scheme provided promising classification results and the identification of important risk factors for predicting second colorectal cancer based on accuracy, sensitivity, specificity, and area under the curve metrics. Collectively, our clinical findings suggested that the most important risk factors were the combined stage, age at diagnosis, BMI, surgical margins of the primary site, tumor size, sex, regional lymph nodes positive, grade/differentiation, primary site, and drinking behavior. Accordingly, these risk factors should be monitored for the early detection of second primary tumors in order to improve treatment and intervention strategies.

Keywords: risk factors, second primary cancer (SPC), colorectal cancer, classification techniques, extreme gradient boosting

1. Introduction

Worldwide, colorectal cancer is considered one of the top three causes of cancer-related deaths in developed countries (Zinatizadeh et al., 2018). In Taiwan, it is also a leading cause of death, ranking second and third in terms of incidence and mortality, respectively. However, due to the success of cancer screening in Taiwan, the early detection and diagnosis of malignant tumors have become feasible. In addition, due to advances in therapeutic instruments and techniques, such as three-dimensional spatial conformal radiation therapy, intensity-modulated radiation therapy, and proximity radiation therapy, cancer patients have longer survival. However, there is a risk of the occurrence of potentially harmful second primary cancers (SPCs; Sakellakis et al., 2014; Santangelo, 2015; Xu et al., 2016).

Five-year cancer survival rates have historically been an important indicator of clinical treatment. Recently, the overall cancer survival rate has increased to 66.5% in the United States (Mahmoud et al., 2016). In Taiwan, excluding the low survival rates of lung, liver, and gastric cancers, the survival rate of other cancers has also increased significantly. However, one of the most difficult clinical issues for cancer survivors is the occurrence of multiple primary malignant neoplasms (MPMNs). Multiple malignancies are characterized as two or more independent primary malignancies diagnosed in different tissues/organs in the same individual (Li et al., 2015). In general, MPMNs are most present in double cancers. According to the literature, the incidence of second primary malignant tumors in patients with malignant tumors is six times higher than that in healthy people. Second primary malignant tumors occur most often within 3 years of the first tumor treatment, with the shorter the interval between the first cancer and the SPC, the worse the prognosis (Wu et al., 2014). The prevention of MPMNs has always been a significant problem faced by both doctors and patients. The high prevalence age range for MPMNs is 50–59 years, with most patients over 50 years (Sakellakis, 2014).

The first research report on MPMNs was published by Warren and Gates in 1932. According to their definition, MPMNs should have first and second malignant tumors, there should be at least 2 cm between the two tumors, they should be excluded from metastatic tumors within 5 years, and occur at a time more than 3 years from the primary tumor (Meng et al., 2017). The definition of SPC (synchronous vs metachronous) is based on the diagnosed time of the first primary cancer. Accordingly, primary cancers found within 6 months of the first diagnosis are considered to be synchronous, whereas metachronous cancers refer to a primary cancer discovered 6 months after the first diagnosis (Huang et al., 2015). Figure 1 shows the trajectory of cancer treatment, where the patient is diagnosed and staged first, followed by the targeted therapy and palliative treatment. The treatment target can be divided into cancer-free survival and chronic comorbid management. The latter can result in treatment failure, leading to palliative treatment, and in more severe cases, to an SPC (Patricia et al., 2015).

In Taiwan, the incidence of MPMNs in rapidly increasing. According to the guidelines of the Institute of Medicine's prevention and treatment recommendations for multiple malignancies, "Based on the cancer-registered population, it is imperative to use the empirical medical perspective and systematic analysis of therapeutic techniques to further develop clinical treatment guidelines for multiple malignancies (MPMNs)" (Vogt et al., 2017).

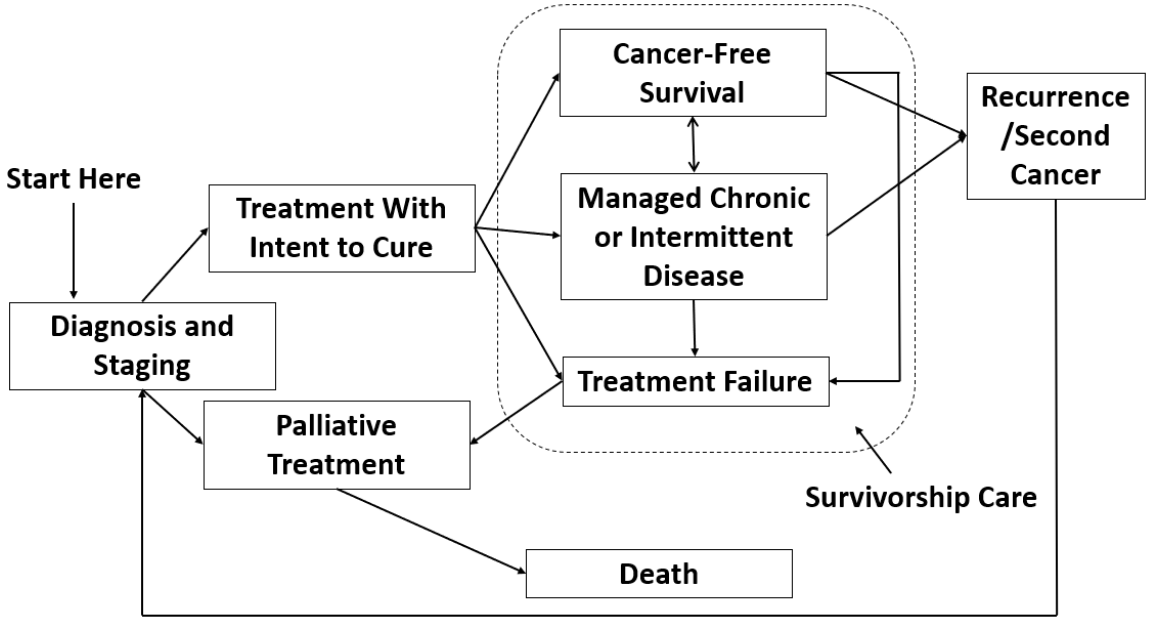


Figure 1. Cancer Care Trajectory
(Modified from source: Patricia et al., 2015)

With recent developments in information technology, data classification methods represent an important research field. Data mining technologies have also become useful tools to support clinical diagnostic guidelines. Machine learning is used to analyze important information hidden in the vast amount of data stored in databases. For example, breast cancer (Chang et al., 2019), ovarian cancer (Tseng et al., 2017), and colorectal cancer (Ting et al., 2018) have achieved good performances using these techniques.

Over the last two decades, cancer registration databases have been used to store records related to the treatment of colorectal cancer patients. Indeed, a vast network of useful information is hidden in these collected datasets. Although traditional data query and statistical functions can be utilized, it is not easy to find unknown information features in practice and information about their potential value cannot be directly observed from the dataset. As such, how to explore hidden, unknown, and valuable information from SPC databases through specific procedures and methods is an important research topic that aims to improve prevention and treatment strategies for colorectal cancer survivors.

In this study, we used machine learning techniques to develop a predictive model of colorectal cancer and an analyzing model of SPC. These classification techniques can be used to identify various analyzable risk factors and clinical features within SPC, providing decision support for clinical treatment.

2. Methods

2.1 MARS

Multivariate adaptive regression splines (MARS) is a flexible procedure used to find optimal variable transformations and interactions. It can be used to identify model relationships that are nearly additive or that involve interactions with fewer variables. MARS is a nonparametric statistical method based on a divide-and-conquer strategy for partitioning training datasets into separate groups, each of which gets its own regression equation. The non-linearity of the MARS model is approximated via the use of separate linear regression slopes in distinct intervals of the independent variable space.

The MARS function is a weighted sum of the basis functions (BFs), which are splines piecewise polynomial functions. It can be represented using the following equation [Friedman 1991]:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \, B_m(x) \quad (1)$$

where α_0 and α_m are constant coefficients that can be estimated using the least-squares method. M is the number of basis functions. $B_m(x)$ represents the basis functions. The hinge functions, $max\,(0, x-k)$ *or* $max\,(0, k-x)$, with a knot defined at value t are used in MARS modeling. In addition, MARS automatically selects the variables and values of those variables for knots of the hinge functions based on generalized cross-validation criterion (Zhang and Goh 2016).

2.2 RF

Random forest (RF) is an ensemble classification method based on statistical learning theory that combines several individual classification trees [Breiman, 2001, Yuk et al. 2018]. RF is a supervised machine learning algorithm that considers the unweighted majority of the class votes. First, various random samples of variables are selected as the training dataset using the bagging procedure, which is a meta-algorithm that uses random sampling with replacement to synchronously reduce variance and elude over-fitting. Classification trees using selected samples are then built into the training process. A large number of classification trees are then used to form a RF from the selected samples. Classification and regression tree (CART) is typically the classification method used for RF modeling. Finally, all classification trees are combined and the final classification results are obtained by voting on each class and then choosing the winner class in terms of the number of votes. RF performance is measured by a metric called 'out of bag' error, which is calculated as the average of the rate of error for each weak learner. In RF, each individual tree is explored in a particular way. The most important variable randomly chosen is used as a node and each tree is developed to its maximum expansion (Breiman, 2001).

2.3 SVM

Support vector machine (SVM) is a machine learning algorithm based on the structural risk minimization principle for estimating a function by minimizing the upper bound of the generalization error (Vapnik 2000). In modeling an SVM model, one can initially use the kernel function to, either linearly or non-linearly, map the input vectors into one feature space. Then, within the feature space, the SVM attempts to seek an optimized linear division to construct a hyperplane that separates the classes. In order to optimize the hyperplane, SVM solves the optimization problem using the following equation (Vapnik 2000):

$$Min\ \phi(x) = \frac{1}{2}\|w\|^2$$

$$Subject\ to\ \ y_i(w^T x_i + b) \geq 1\,, i = 1,2,\ldots,N \quad (2)$$

where $x_i \in R^d$ is the input variable, $y_i \in \{-1,1\}$ is the known target variable, N is the number of sample observations, d is the dimension of each observation, w is the vector of the hyperplane, and b is a bias term.

In order to solve eq. (2), the Lagrange method is used to transform the optimization problem into a dual problem. The penalty factor is used as a tuning parameter in the transformed dual problem to control the trade-off between maximizing the margin and the classification error. In general, SVM does not find the linear separate hyperplane for all application data. For non-linear data, it must transform the original data to a higher dimension of linearity separately as the best solution. The higher dimension is called the feature space and it improves the data separated by classification. The common kernel functions are linear, polynomial, radial basis function, and sigmoid. Although several choices for the kernel function are available, the most widely used is the radial basis function kernel (Tseng et al. 2017; Li et al. 2018).

2.4 ELM

Extreme learning machine (ELM) is a single hidden layer feed-forward neural-network (SLFN) that randomly selects the input weights and analytically determines the output weights of the SLFN (Huang et al. 2006). The modeling time of ELM is faster than traditional feedforward network learning algorithms such as the back-propagation (BP) algorithm. It also avoids many difficulties present in gradient-based methods such as the stopping criteria, learning rate, learning epochs, local minimal, and over tuning issues.

In SLFNs, N represents the arbitrary distinct samples (x_i, y_i), using ρ hidden neurons and the activation function vector θ(x), and approximates N samples with zero error, written as:

$$\boldsymbol{HA} = \boldsymbol{Y} \tag{3}$$

where $\boldsymbol{H}_{N\times\rho} = \left[\theta\left(w_i x_j + b_i\right)\right]$ is the hidden layer output matrix of the neural network and the *i*-th column of $\boldsymbol{H}$; $\boldsymbol{A}$ is the matrix of the output weights; w_i is the weight vector connecting the *i-th* hidden node and the input nodes; b_i is the threshold (bias) of the *i-th* hidden node; and $\boldsymbol{Y}$ is the matrix of the targets.

Huang et al. (2006) demonstrated that the input weights and hidden layer biases can be randomly generated in the ELM algorithm, and the output weights can be determined as simply as finding the least-square solution to a given linear system. Accordingly, the minimum norm least-square solution to the linear system is $\widehat{\mathbf{A}} = \widetilde{\boldsymbol{H}}\mathbf{Y}$, where $\widetilde{\boldsymbol{H}}$ is the Moore-Penrose generalized inverse of the matrix $\boldsymbol{H}$. The minimum norm least-square solution is unique and has the smallest norm among all least-square solutions (Huang et al., 2006).

2.5 XGboost

XGBoost belongs to the group of widely used tree learning algorithms. It is a supervised learning algorithm based on a scalable end-to-end gradient tree boosting system (Chen & Guestrin 2016). Boosting refers to the ensemble learning technique of building many models sequentially, with each new model attempting to correct for the imperfections or inadequacies in the previous model. In other words, in gradient boosting, a new weak learner is constructed to be maximally correlated with the negative gradient of the loss function associated with the whole assembly for each iteration [Natekin and Knoll 2013].

XGBoost is the implementation of a generalized gradient boosting decision tree that uses a new distributed algorithm for tree searching, which speeds up tree construction. XGBoost includes a regularization term that is used to alleviate overfitting, as well as support for arbitrary differentiable loss functions (Torlay et al. 2017). The objective function of Xgboost consists of two parts, namely, a loss function over the training set and a regularization term that penalizes the complexity of the model as follows (Mitchell and Frank 2017):

$$\text{Objective} = \sum_i \boldsymbol{L}(y_i, \hat{y}_i) + \sum_k \boldsymbol{\Omega}(t_k) \tag{4}$$

where $\boldsymbol{L}(y_i, \hat{y}_i)$ can be any convex differentiable loss function that measures the difference between the prediction and the true label for a given training instance. $\boldsymbol{\Omega}(t_k)$ describes the complexity of the tree f_k and is defined in the XGBoost algorithm as:

$$\boldsymbol{\Omega}(t_k) = \gamma\boldsymbol{T} + \frac{1}{2}\lambda\omega^2 \tag{5}$$

where $\boldsymbol{T}$ is the number of leaves on tree t_k and ω is the weight of the leaves. When $\boldsymbol{\Omega}(t_k)$ is included in the objective function, it is forced to optimize for a less complex tree, which simultaneously minimizes $\boldsymbol{L}(y_i, \hat{y}_i)$. This helps to alleviate any overfitting issues. $\gamma\boldsymbol{T}$ provides a

constant penalty for each additional tree leaf and $\lambda\omega^2$ penalizes for extreme weights. γ and λ are user configurable parameters (Mitchell and Frank 2017).

3 Proposed Prediction Scheme

In this study, the five data mining classification techniques described above were integrated to propose a scheme for predicting SPC in colorectal cancer patients. The flowchart of the proposed scheme is shown in Figure 1.

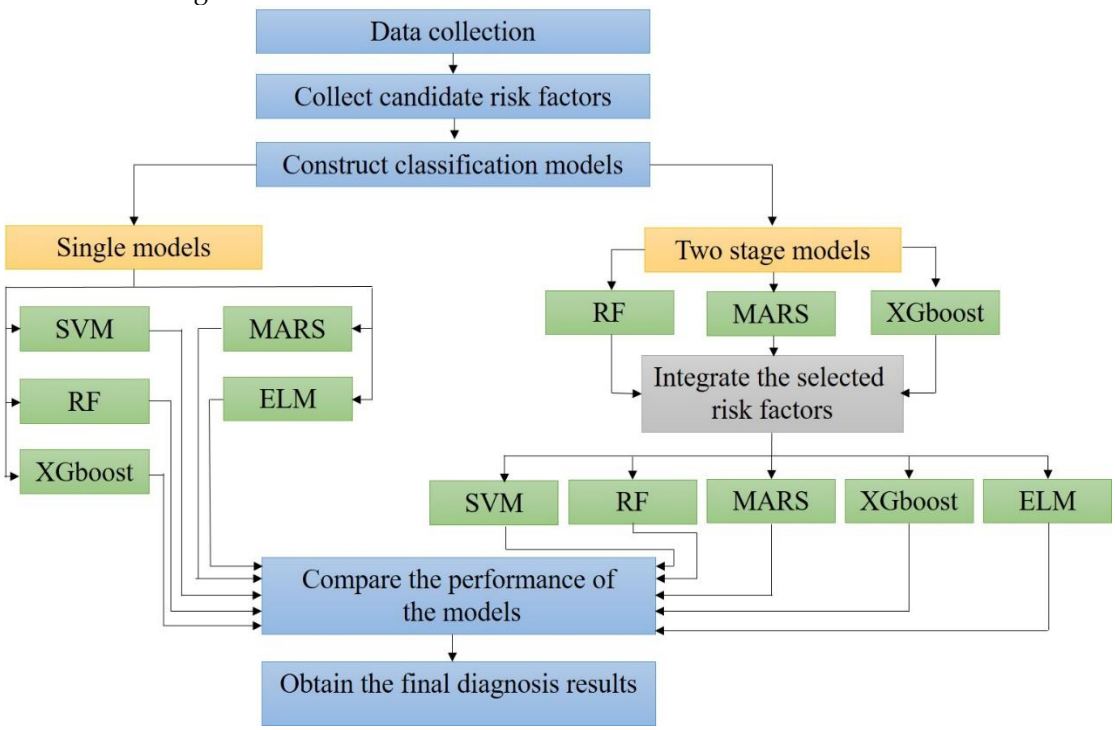


Figure 1. The proposed scheme for risk factor prediction

The first step of the proposed scheme was to collect the data. The second step was to collect candidate risk factors as predictor variables. As shown in Table 1, the 14 risk factors for SPC in colorectal cancer patients are represented as X1 to X14. The target variable is SPC or not (Y).

Table 1. The fourteen candidate risk factors for SPC in colorectal cancer patients

Variables	Description
X1. Sex	Male/female
X2. Age at diagnosis	Age at diagnosis
X3. Primary site	Colon/rectal
X4. Grade/differentiation	Distinguish by differentiation
X5. Tumor size	Distinguish by unit size
X6. Regional lymph nodes positive	Differentiated by lymphoid number
X7. Combined stage	Sorted out by clinical stage and pathologic stage
X8. Surgical margins of the primary site	Residual/no residual
X9. Radiation therapy/no radiation therapy	Radiation therapy/no radiation therapy
X10. Chemotherapy/no chemotherapy	Chemotherapy/no chemotherapy

X11. BMI	BMI
X12. Smoking behavior	Smoking behavior/no smoking behavior
X13. Betel nut chewing	Betel nut chewing/no betel nut chewing
X14. Drinking	Drinking/no drinking
Y: SPC	1: No, 2: yes

In the third step, we constructed classification models for predicting SPC in colorectal cancer patients. In building the classification models, we used two types of modeling processes. One was a single model and the other was a two-stage model. In modeling the single models, the entire 14 risk factors were directly used as predictors for SVM, RF, MARS, ELM, and XGboost for constructing five single classification models. These were termed single SVM (S-SVM), single RF (S-RF), single MARS (S-MARS), single ELM (S-ELM), and single XGboost (S-XGboost) models.

The two-stage model integrating the feature selection method and classifier were used in the third step of the proposed scheme as important disease risk factors are often fundamental indicators that provide useful information for modeling effective disease predictions. In modeling the two-stage model, a feature selection method was first used to select the important risk factors. Among the five data mining methods, only RF, MARS, and XGboost can be used to select important risk factors based on their fundamental algorithms, thus these were used as the three feature selection methods to identify and rank important risk factors for predicting SPC in colorectal cancer patients. Each feature selection method generated one set of important risk factors. Using only one feature selection technique may not provide stable and effective selection results. A simple average rank method was used to combine the risk factor selection results of the three methods.

Table 2 shows the selected and ranked risk factors using the RF, MARS, and XGboost methods. Note that a risk factor with a rank of 1 indicates that it is the most important risk factor, while that with a rank of 14 indicates that it is a risk factor not selected by the method. For each risk factor, the average rank was obtained by calculating the average value of its rankings in the RF, MARS, and XGboost methods. Table 2 shows also the average rank of every risk factor. The ranked overall variable importance of all the risk factors is shown in Figure 2. It can be observed that X7, with an average rank of 1, is the most important risk factor, followed by X2 and X11.

Table 2. The selected and ranked risk factors using the RF, MARS, and XGboost methods

Risk factors	RF	MARS	XGboost	Average Rank
X1	10	2	4	5.3
X2	2	3	2	2.3
X3	11	5	11	9.0
X4	6	14	5	8.3
X5	5	8	3	5.3
X6	7	9	9	8.3
X7	1	1	1	1.0
X8	4	4	8	5.3
X9	14	9	14	12.3
X10	13	14	13	13.3
X11	3	6	6	5.0
X12	12	7	12	10.3
X13	9	14	10	11.0
X14	8	14	7	9.7

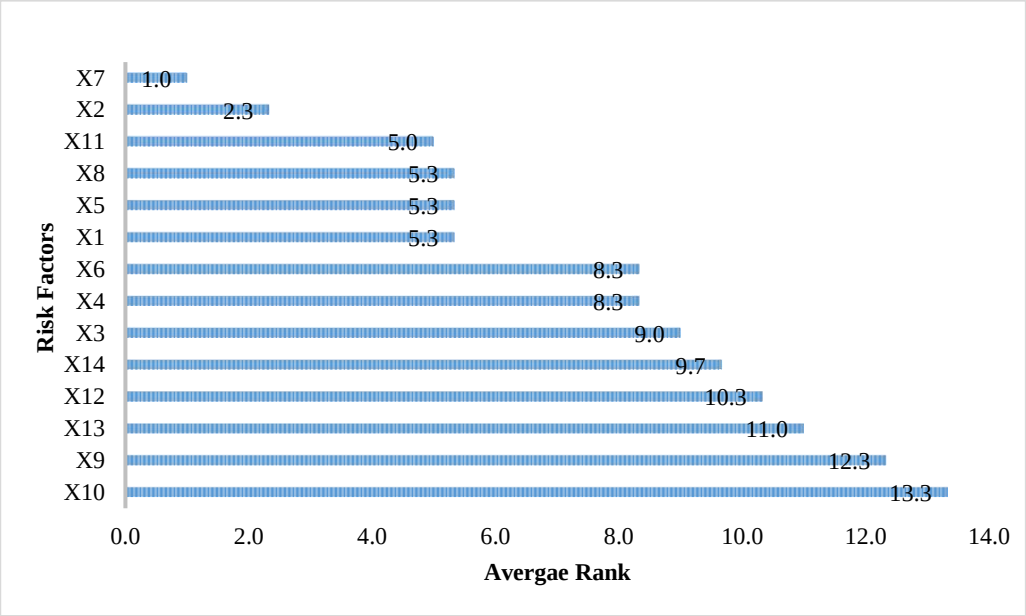


Figure 2. The ranking of all risk factors

In the modeling process of the two-stage method, after obtaining the average rank of each risk factor, the overall important risk factors should be identified before constructing a classification model. In this study, an average rank value less than 10 was used as the criteria for selecting the overall important risk factors. These criteria were determined by the suggestion of clinical physicians. Based on these criteria, it can be observed from Figure 2 that the 10 risk factors, including X7 (combined stage), X2 (age at diagnosis), X11 (BMI), X8 (surgical margins of the primary site), X5 (tumor size), X1 (sex), X6 (regional lymph nodes positive), X4 (grade/differentiation), X3 (primary site), and X14 (drinking) were selected as the important risk factors.

In the final stage of the two-stage method, the identified 10 overall important risk factors were served as the input variables for the SVM, RF, MARS, ELM, and XGboost methods in order to predict SPC in colorectal cancer patients. The five two-stage methods were termed A-SVM, A-RF, A-MARS, A-ELM, and A-XGboost, respectively.

In the fourth step of the proposed scheme, after obtaining the classification results from the five single methods and the five two-stage methods, we used accuracy, sensitivity, specificity, and area under the curve (AUC) parameters as classification accuracy metrics to compare the performance of the ten models.

In the final step, after comparing the classification performance of the S-SVM, S-RF, S-MARS, S-ELM, S-XGboost, A-SVM, A-RF, A-MARS, A-ELM, and A-XGboost models, we obtained the final diagnosis results and identified the important risk factors for predicting SPC in colorectal cancer patients.

4. Empirical Results

In this study, colorectal cancer datasets provided by three hospital cancer registries were used to verify the proposed medical diagnostic scheme for predicting the occurrence of SPC in colorectal cancer patients. Each patient in the dataset had 14 predictor variables, with one response variable indicating SPC or not. Excluding incomplete records, there were a total of 4,287 patients in the dataset. The 10-fold cross-validation method was used in this study for evaluating the performance of the proposed scheme.

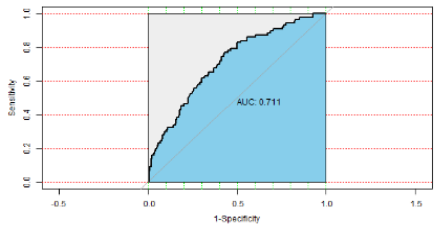
For modeling the ten models, including the S-SVM, S-RF, S-MARS, S-ELM, S-XGboost, A-SVM, A-RF, A-MARS, A-ELM, and A-XGboost models, for their predictive ability for the risk of SPC in

colorectal cancer patients, the software R (version 3.6.1) was employed. Each method used a different R package for analysis. This study used a 10-fold cross-validation procedure for training and testing the performance of the ten models.

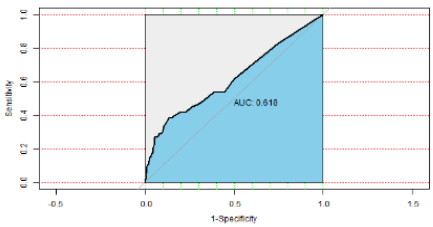
Using the process detailed in Section 3, Table 3 shows the classification results of the five single methods, including the S-SVM, S-RF, S-MARS, S-ELM, and S-XGboost models. From Table 3, it can be observed that the AUC values of the S-SVM, S-RF, S-MARS, S-ELM, and S-XGboost models were 0.711, 0.618, 0.640, 0.710, and 0.550, respectively. The single SVM model provided the highest AUC value, followed by the single XGboost model with a slightly smaller AUC value. However, it also can be seen from Table 3 that the accuracy value of the S-XGboost model was 0.641, which is significantly greater than that of the single SVM model at 0.408. Figure 3 shows the ROC curves of the five single classification methods for the occurrence of SPC in colorectal cancer patients. Thus, among the five single classification methods, the single XGboost model provided the best classification results.

Table 3. Classification results of the five single methods

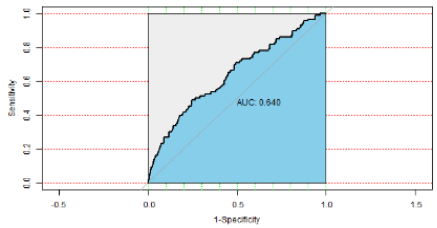
Methods	Accuracy	Sensitivity	Specificity	AUC
S-SVM	0.408	0.233	0.428	0.711
S-RF	0.819	0.384	0.868	0.618
S-MARS	0.727	0.488	0.754	0.640
S-XGboost	0.641	0.709	0.633	0.710
S-ELM	0.483	0.361	0.496	0.550



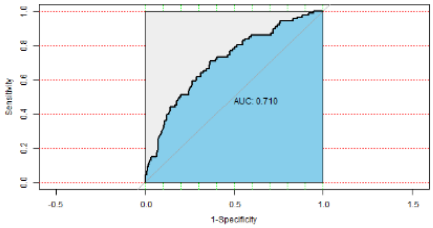
(a) S-SVM



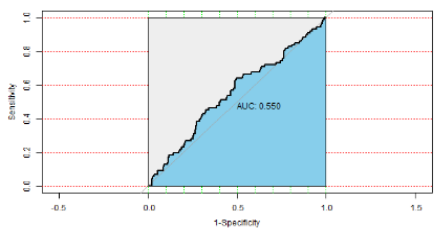
(b) S-RF



(c) S-MARS



(d) S-XGboost



(e) S-ELM

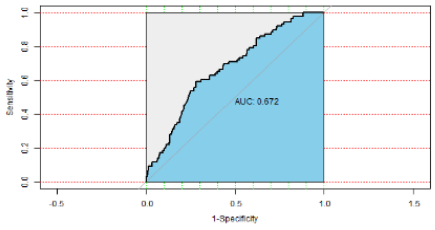
Figure 3. ROC curves of the five single methods

As aforementioned, the 10 risk factors, including X7, X2, X11, X8, X5, X1, X6, X4, X3, and X14, were selected as the important risk factors and then served as the critical predictor variables for constructing the five two-stage methods, including the A-SVM, A-RF, A-MARS, A-ELM, and A-XGboost models.

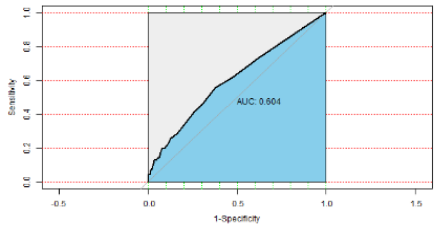
Table 4 shows the classification accuracy matrices of the five two-stage methods. As depicted in Table 4, it can be observed that the A-XGboost method generated the highest AUC value at 0.714, with a sensitivity value of 0.767, compared with the competing models. Figure 4 displays the ROC curves of the five two-stage methods. From Table 4 and Figure 4, it can be observed that the A-XGboost method generated the best performance for predicting the occurrence of SPC in colorectal cancer patients and is the best method among the five two-stage models.

Table 4. Classification results of the five two-stage methods

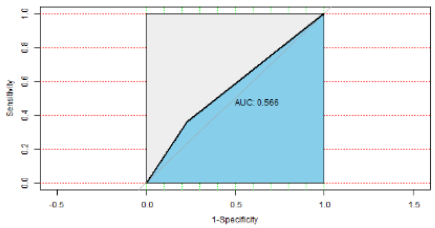
Methods	Accuracy	Sensitivity	Specificity	AUC
A-SVM	0.294	0.407	0.281	0.672
A-RF	0.615	0.558	0.622	0.604
A-MARS	0.731	0.361	0.772	0.566
A-XGboost	0.611	0.767	0.593	0.714
A-ELM	0.425	0.442	0.424	0.546



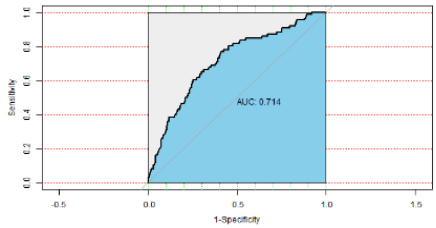
(a) A-SVM



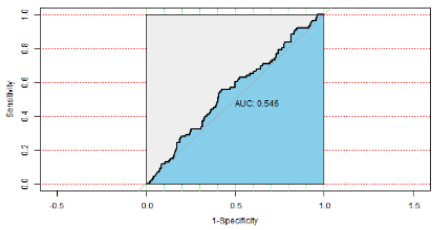
(b) A-RF



(c) A-MARS



(d) A-XGboost



(e) A-ELM

Figure 4. ROC curves of the five two-stage methods

For comparing the classification performance between the five single methods and the five two-stage models, Figure 5 depicts the AUC values of the ten models in decreasing order. It can be observed from Figure 5 that the A-XGboost model generated the best AUC value, followed by the S-SVM and S-XGboost models. These results indicated that the A-XGboost method is a good alternative for constructing a classification model for diagnosing the occurrence of SPC in colorectal cancer. Moreover, the A-XGboost method can be used to select important risk factors that are more influential on patients with SPC of colorectal cancer.

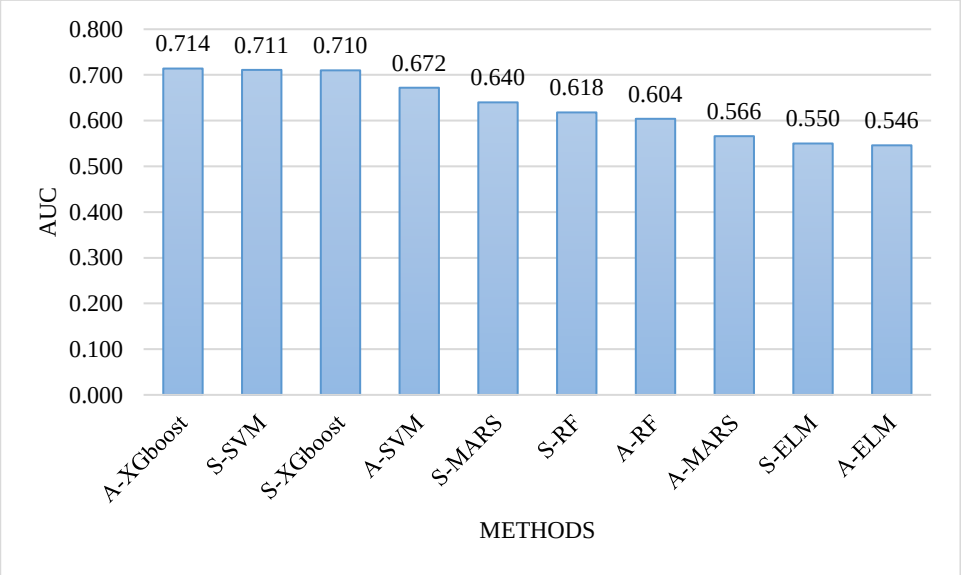


Figure 5. Comparison of the AUC values of the five classifiers with and without using the proposed scheme

5. Discussion and Conclusions

In this study, 10 important risk factors, including the combined stage, age at diagnosis, BMI, surgical margins of the primary site, tumor size, sex, regional lymph nodes positive, grade/differentiation, primary site, and drinking behavior, were selected by the A-XGboost model, which provided the best classification performance among the ten models constructed in this study.

Colorectal cancer ranks second and third in terms of mortality and incidence, respectively, in Taiwan. It is also the third highest cancer in terms of medical expenditure. While patient survival has improved, the occurrence of second primary cancers in colorectal cancer patients has become an important issue for clinical management. To address this issue, data from the cancer registry can be used to better understand the disease and maximize the prevention of SPC. Important issues for future research include predictive models (radiotherapy and chemotherapy) and their association with SPC, as well as a better understanding of the interactions with other genetic factors. Further discussion with patients after diagnosis should help determine the optimal duration of monitoring and follow-up.

References

1. Zinatizadeh N, Khalili F, Fallah P, Farid M, Geravand M, Yaslianifard S. (2018). Potential preventive effect of lactobacillus acidophilus and lactobacillus plantarum in patients with polyps or colorectal cancer, Arq Gastroenterol. 2018 Oct-Dec, 55(4), 407-411.
2. Sakellakis M. (2014). Multiple primary malignancies: a report of two cases. Chinese Journal of Cancer Research, 26.2, 215-8.
3. Santangelo ML. (2015). Immunosuppression and Multiple Primary Malignancies in Kidney-Transplanted Patients: A Single-Institute Study. BioMed Research International. 2015, 183-523.
4. Xu W. (2016). Multiple Primary Malignancies in Patients with Hepatocellular Carcinoma: A Largest Series With 26-Year Follow-Up. Medicine, 95(17), e3491.
5. Mahmoud O, Dosch A, Kwon D, Pitcher JD, Conway S, Benedetto P, Fernandez G, Trent J, Temple HT, Wolfson AH. (2016). The Impact of Perioperative Chemotherapy Timing in Conjunction With Postoperative External-Beam Radiation Therapy on Extremity Soft-Tissue Sarcomas Outcome, Am J Clin Oncol. 2016 Oct, 39(5), 528-34.
6. Li F. (2015). Multiple primary malignancies involving lung cancer. BMC Cancer 15, 696.
7. Wu LL, Gu KS (2014). Clinical retrospective analysis of cases with multiple primary malignant neoplasms. Genetics And Molecular Research, 13(4), 9271-84.
8. Meng LV, Zhang X, Shen y, Wang F, Yang J, Wang B, Chen Z, Li P, Zhang X, Li S, Yang J. (2017) Clinical analysis and prognosis of synchronous and metachronous multiple primary malignant tumors, Medicine (Baltimore). 2017 Apr, 96(17), e6799.
9. Huang CS, Yang SH, Lin CC, Lan YT, Chang SC, Wang HS, Chen WS, Lin TC, Lin JK, Jiang JK. (2015). Synchronous and Metachronous Colorectal Cancers: Distinct Disease Entities or Different Disease Courses? Hepato-gastroenterology, 2015 Jun, 62(140), 838-42.
10. Patricia AG, Jacqueline C, Erin EH. (2015). Ensuring quality care for cancer survivors: Implementing the survivorship care plan. Seminars in Oncology Nursing, 24(3), 208-217.
11. Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, Omlin A.(2017).Multiple primary tumours: challenges and approaches, a review, ESMO Open. 2017 May 2, 2(2), e000172.
12. Chang CC and Ssu-HC (2019). Developing a Novel Machine Learning-based Classification Scheme for Predicting SPCs in Women with Breast Cancer, Frontiers in Genetics. 2019, https://doi.org/10.3389/fgene.2019.00848
13. Tseng CJ, Chang CC, Lu CJ, Chen GD (2015). Application of Machine Learning to Predict the Recurrence-Proneness for Cervical Cancer, Neural Computing and Applications, 24, 1311-1316.
14. Ting WC, Lu YC, Lu CJ, Chalong C, Chang CC (2018). Recurrence Impact of Primary Site and Pathologic Stage in Patients Diagnosed with Colorectal Cancer, Journal of Quality, 25(3), 166-184.
15. J.H. Friedman, Multivariate adaptive regression splines, The Annals of Statistics, 19 (1991), pp. 1-141
16. Zhang, W., & Goh, A. T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geoscience Frontiers, 7(1), 45-52.
17. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
18. Yuk, E., Park, S., Park, C. S., & Baek, J. G. (2018). Feature-learning-based printed circuit board inspection via speeded-up robust features and random forest. *Applied Sciences*, *8*(6), 932.
19. Vapnik, V.N. (2000) The Nature of Statistical Learning Theory. Springer, Berlin
20. Tseng, C. J., Lu, C. J., Chang, C. C., Chen, G. D., & Cheewakriangkrai, C. (2017). Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artificial intelligence in medicine*, *78*, 47-54.
21. Li, T., Gao, M., Song, R., Yin, Q., & Chen, Y. (2018). Support Vector Machine Classifier for Accurate Identification of piRNA. *Applied Sciences*, *8*(11), 2204.
22. Huang, G.R., Zhu, Q.Y., Siew, C.X. (2006) Extreme learning machine: theory and applications. Neurocomputing 70: 489-501
23. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.
24. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front neurorobotics 7:21
25. Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain informatics, 4(3), 159.

401 26. Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. PeerJ
402 Computer Science, 3, e127.
403 27. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 2006; 6(3):
404 21–45.
405