BMC Cancer

RESEARCH

Open Access

Machine learning-based colorectal cancer prediction using global dietary data

Hanif Abdul Rahman[1,2*], Mohammad Ashraf Ottom[1,3] and Ivo D. Dinov[1]

Abstract

Background Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide. Active health screening for CRC yielded detection of an increasingly younger adults. However, current machine learning algorithms that are trained using older adults and smaller datasets, may not perform well in practice for large populations.

Aim To evaluate machine learning algorithms using large datasets accounting for both younger and older adults from multiple regions and diverse sociodemographics.

Methods A large dataset including 109,343 participants in a dietary-based colorectal cancer ase study from Canada, India, Italy, South Korea, Mexico, Sweden, and the United States was collected by the Center for Disease Control and Prevention. This global dietary database was augmented with other publicly accessible information from multiple sources. Nine supervised and unsupervised machine learning algorithms were evaluated on the aggregated dataset.

Results Both supervised and unsupervised models performed well in predicting CRC and non-CRC phenotypes. A prediction model based on an artificial neural network (ANN) was found to be the optimal algorithm with CRC misclassification of 1% and non-CRC misclassification of 3%.

Conclusions ANN models trained on large heterogeneous datasets may be applicable for both younger and older adults. Such models provide a solid foundation for building effective clinical decision support systems assisting healthcare providers in dietary-related, non-invasive screening that can be applied in large studies. Using optimal algorithms coupled with high compliance to cancer screening is expected to significantly improve early diagnoses and boost the success rate of timely and appropriate cancer interventions.

Keywords Colorectal cancer, Machine learning, Dietary information

Introduction

In the current twenty-first century, the re-emergence of machine learning (ML) and advancement in artificial intelligence (AI) through data science provide unique opportunities to go beyond traditional statistical and research limitations, and advance health data analytics in solving healthcare challenges and ultimately improve the delivery of health services [1, 2].

One of the contemporary healthcare challenges is colorectal cancer (CRC). CRC is the third most commonly diagnosed malignancy after breast and lung cancers, and is also the second leading cause of cancer-related mortality worldwide [3, 4]. In 2020, an estimated 1.93 million new CRC cases were diagnosed, which accounts for 10% of the global cancer incidence [5]. The increasing number of global CRC cases could be attributed to successful population-based screening and surveillance programs that have been rapidly and actively implemented [6, 7]. Nonetheless, the number of CRC mortality is still high

*Correspondence:
Hanif Abdul Rahman
hanifr@umich.edu; hanif.rahman@ubd.edu.bn
[1] University of Michigan, Ann Arbor, USA
[2] PAPRSB Institute of Health Sciences, Universiti Brunei Darussalam, , Bandar Seri Begawan, Brunei
[3] Yarmouk University, Irbid, Jordan

where 0.94 million deaths were recorded in 2020 that accounts for 9.4% of cancer deaths globally [5]. Active health screening and prevention of CRC activities have yielded an increasingly younger generation (below 50 years) of early-onset CRC in developed countries and overall increase in CRC incidence detection in developing and emerging economic nations [8, 9]. Increased pathophysiological understanding of CRC progression and the advancement of treatment options, including endoscopic and surgical interventions, radiotherapy, immunotherapy, and targeted chemotherapy, have effectively prolonged survival years and improved quality of life of CRC patients [9, 10]. The prognosis after CRC therapy is generally good when CRC is detected at a younger age, however, there is still huge public health challenges and financial burden associated with CRC [9]. In 2015, the economic cost of CRC in Europe due to hospital-care costs, loss of productivity, premature death, and costs of informal care was estimated at 19 billion euros [11]. Furthermore, the underlying mechanisms and risk factors of early-onset CRC pathological features are sporadic and not fully understood and require more research [9].

In this era of digital technology, the vast amount of high-quality CRC data (owing to an increase in the number of patients) can be rigorously collected through health information systems. This has enabled data science to offer a new avenue of enhancing knowledge of CRC through research and development. Currently, the extant evidence using machine-learning models have made great strides in predicting CRC based on available genetic-based data, which have shown that some CRC cases have a component of hereditary predisposition [12, 13]. However, genetic disorder is a permanent and non-modifiable risk factor. In contrast, dietary control is one of the most effective protective measures against CRC that the population can modify [4, 14] especially because CRC susceptibility is mainly resulting from adopting dietary lifestyle associated with globalization [15, 16]. With the globalization of the food industry and supply chain, it is thus important data science research to look into global diet features in relation to CRC prediction. In this study, we obtained global dietary-based data from publicly accessible databases and investigate the important dietary factors of predicting CRC labels using exploratory unsupervised and supervised ML-based models.

Methods

Dataset and data preprocessing

Several end-to-end procedures were systematically performed, as illustrated in Fig. 1. Dietary-related colorectal cancer data was obtained from the Center for Disease Control and Prevention, Global Dietary database, and publicly accessible institutional sites [17, 18, 19, 20, 21, 22, 23]. The initial combined data contained 25 countries consisting of Argentina, Bangladesh, Bulgaria, Canada, China, Korea, Ecuador, Estonia, Ethiopia, Finland, Germany, India, Iran, Israel, Kenya, Malaysia, Mexico, Mozambique, Philippines, Portugal, Sweden, Tanzania,

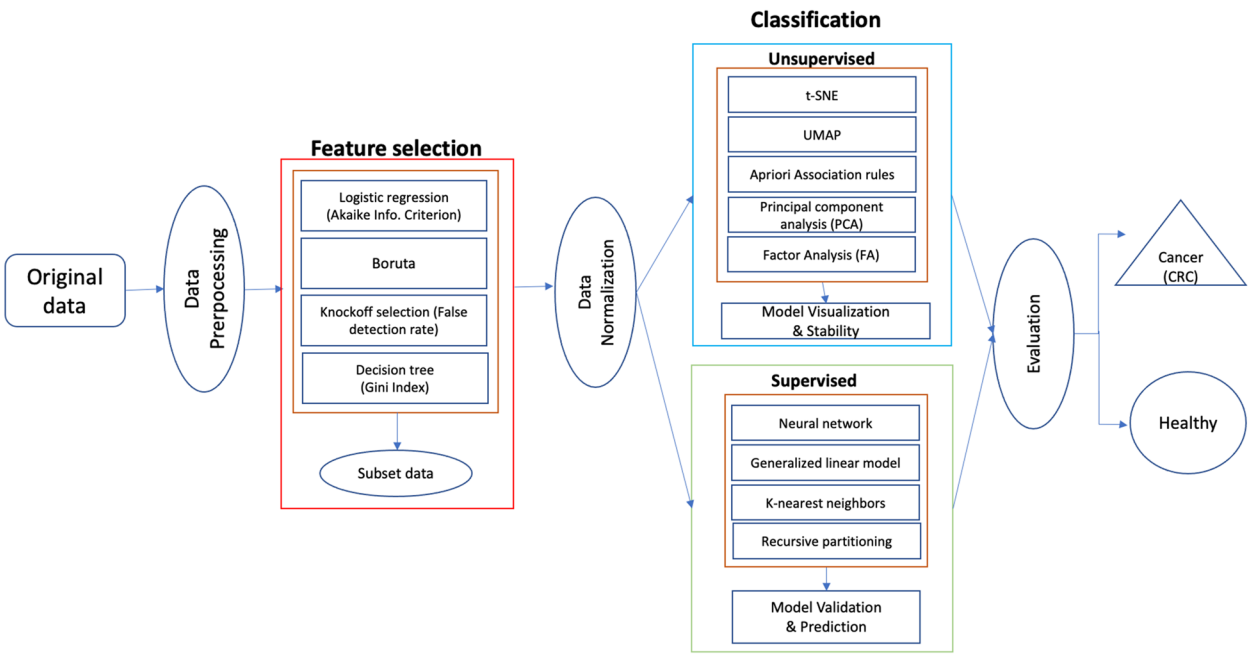

Fig. 1 A schematic of the procedures undertaken in this study to classify CRC labels

Italy, Japan, and the United States. The data collection methodology of these data sets were similar, i.e., cross-sectional and employed dietary questionnaires. The different sets of data were then merged and extrapolated based on the same dietary characteristics. Features that were not common across the data sets were excluded. This study only includes data sets that are of the English language. Features with different units of measurements were converted for standardization. A cleaning procedure was employed including removal of ineligible cases, duplicate characteristics, and features with more than 50% missing values (listwise deletion). At this stage, a total of 3,520,586 valid data remained. Due to computational limitations, a multi-stage, proportionate random sample of 109,342 were extracted for analysis, that maintains the percentage by country and CRC distribution, of which 7,326 (6.7%) cases were positive colorectal cancer labels that are derived for seven countries that comprised of Canada, India, Italy, South Korea, Mexico, Sweden, and United States. A sample size of 5,000 cases was sufficient to achieve a power of 0.8 [24]. Considering the computation ability of our machine could handle up to 110,000 data points, we randomly selected the maximum data load for this study. Table 1 presents the characteristics of the data.

Missing data in these valid cases was handled using multiple imputation techniques—MICE (Multivariate Imputation via Chained Equations) set at 10 multiple imputations to replace missing with predicted values, using R package *mice* [25]. The data set also consists of textual elements that describe the ingredients used such as milk, salt, chicken, and so on. Texts were converted into corpus objects and processed for standardization such as using English stop words, lower case, and removal of punctuation. The corpus item was then converted to a document term matrix to enable counting of most frequent terms occurring (Fig. 2), which are illustrated as a Wordcloud (Fig. 3). The important terms are converted into a data frame that is subsequently merged with the full data set. The dataset also has unbalanced binary CRC outcome, which was then re-balanced using the Synthetic Minority Oversampling Technique (SMOTE) [26].

Feature selection

Two-step feature selection method was employed. Step one involves three separate procedures including Logistic regression (LR), Boruta, and Knockoff selection. LR was used to screen each single index out to reduce redundant features by computing a stepwise iterative process of forward addition (adding important features to a null set of features) and backward elimination (removing worst-performing features from the list of complete features) using the stepAIC function in the MASS package [27]. Variable selection was determined by the most significant features ($p < 0.05$) in the most parsimonious model with the lowest Akaike Information Criterion (AIC). Next, a randomized wrapper method, Boruta, which iteratively removes features that are statistically not significant and relevant than that of random probes, was employed [28]. Finally, the Knockoff selection based on the Benjamini–Hochberg False Discovery Rate method was implemented, that controls for expected proportion of false rejection of features in multiple significance testing [29], which could be expressed as follows:

Table 1 Data characteristics and sample statistics

	Positive		Negative		Total	
	n	%	n	%	n	%
Overall	7326	6.7	102,016	93.3	109,342	100
Country						
Canada	6014	5.5	103,328	94.5	31,381	28.7
India	4702	4.3	104,640	95.7	18,807	17.2
Italy	14,652	13.4	94,690	86.6	8966	8.2
South Korea	2406	2.2	106,936	97.8	16,292	14.9
Mexico	2406	2.2	106,936	97.8	10,387	9.5
Sweden	17,604	16.1	91,738	83.9	10,497	9.6
United States	11,153	10.2	98,189	89.8	12,902	11.8
Gender						
Male	7763	7.1	101,579	92.9	51,172	46.8
Female	6998	6.4	102,344	93.6	58,170	53.2
Age (years) [Mean (SD)]	48.9	(16.7)	36.4	(23.3)	41.6	(21.7)

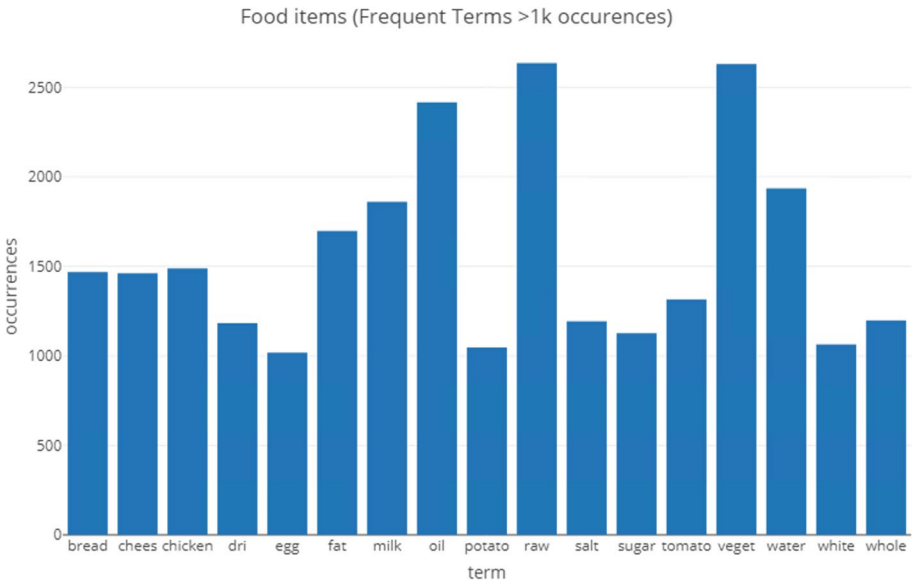

Fig. 2 Frequent text items (1,000 occurrences) of the data

Fig. 3 A word-cloud on the most frequent text items in the data set

$$\underbrace{FDR}_{\text{False Discovery Rate}} = \underbrace{E}_{\text{expectation}} \underbrace{\left(\frac{\#\text{FalsePositive}}{\text{total number of selected Features}}\right)}_{\text{False DIscovery Proportion}}$$

which determines the final selection based on variable importance using the Gini Index that is expressed as follows:

$$GI = \sum_k p_k(1-p_k) = 1 - \sum_k p_k^2,$$

where k is the number of classes.

The data set with the finalized features then was further processed using data normalization to avoid effects of extreme numeric ranges and to help obtain higher classification accuracy [30, 31]. The features were scaled as follows:

$$V^{'} = \frac{V - Min}{Max - Min},$$

where $V^{'}$ is the scale value corresponding to the original value V, and *Min* and *Max* are the upper and lower range limits.

Finally, the features that intersect among the two-step variable selection procedures were selected as the most salient features to be used for unsupervised and supervised classifications.

Unsupervised techniques

Four types of unsupervised machine learning for nonlinear relationship were used to explore the dimensions of the data including t-distributed stochastic Neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), Apriori association rules, principal component analysis (PCA), and factor analysis (FA) [31, 32].

t-SNE technique is a machine learning strategy for nonlinear dimensionality reduction that is useful for embedding high-dimensional data into lower-dimensional spaces. If the high dimensional data (N D) is $x_1, x_2, ..., x_N$ then, for each pair (x_i, x_j), t-SNE estimates

the probabilities $p_{i,j}$ that are proportional to their corresponding similarities, $p_{j|i}$:

$$p_{j|i} = \frac{\exp\left(\frac{-||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-||x_i - x_k||^2}{2\sigma_i^2}\right)}.$$

t-SNE performs a binary search for the value σ_i that produces a predefined value *perp*. The perplexity (*perp*) of a discrete probability distribution, p, is defined as an exponential function of the entropy, $H(p)$, over all discrete events: $perp(x) = 2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)}$.

UMAP relies on local approximations of patches on the manifold to construct local fuzzy simplicial complex (topological) representations of the high dimensional data. For example, if $S1$ the set of all possible 1-simplexes, let's denote by $\omega(e)$ and $\omega'(e)$ the weight functions of the 1-simplex e in the high dimensional space and the corresponding lower dimensional counterpart. Then, the cross-entropy measure for the 1-simplexes is:

$$\sum_{e \in E} \left[\underbrace{\omega(e)\log\left(\frac{\omega(e)}{\omega'(e)}\right)}_{\text{attractive force}} + \underbrace{(1-\omega(e))\log\left(\frac{1-\omega(e)}{1-\omega'(e)}\right)}_{\text{repulsive force}} \right].$$

The iterative optimization process would minimize the objective function composed of all cross entropies for all simplicial complexes using a strategy like stochastic gradient descent.

The optimization process balances the push–pull between the *attractive forces* between the points favoring larger values of $\omega'(e)$ (that correspond to small distances between the points), and the *repulsive forces* between the ends of e when $\omega(e)$ is small (that correspond to small values of $\omega'(e)$).

The Apriori algorithm is based on a simple apriori belief that *all subsets of a frequent item-set must also be frequent*. We can measure a rule's importance by computing its support and confidence metrics. The support and confidence represent two criteria useful in deciding whether a pattern is "valuable." By setting thresholds for these two criteria, we can easily limit the number of interesting rules or item-sets reported.

For item-sets X and Y, the support of an item-set measures how (relatively) frequently it appears in the data:

$$support(X) = \frac{count(X)}{N},$$

where N is the total number of transactions in the database and *count(X)* is the number of observations (transactions) containing the item-set X.

In a set-theoretic sense, the union of item-sets is an item-set itself. In other words, if $Z = X, Y = X \cup Y$, then

$$support(Z) = support(X, Y).$$

For a given rule $X \rightarrow Y$, the rule's confidence measures the relative accuracy of the rule:

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}.$$

The confidence measures the joint occurrence of X and Y over the X domain. If whenever X appears Y tends to also be present, then we will have a high $confidence(X \rightarrow Y)$.

Note that the ranges of the support and the confidence are $0 \leq support, confidence \leq 1$.

PCA (principal component analysis) is a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables through a process known as orthogonal transformation. In general, the formula for the first PC is $pc_1 = a_1^T X = \sum_{i=1}^{N} a_{i,1} X_i$ where X_i is a $n \times 1$ vector representing a column of the matrix X (representing a total of n observations and N features). The weights $a_1 = \{a_{1,1}, a_{2,1}, ..., a_{N,1}\}$ are chosen to maximize the variance of pc_1. According to this rule, the k^{th} PC is $pc_k = a_k^T X = \sum_{i=1}^{N} a_{i,k} X_i$. $a_k = \{a_{1,k}, a_{2,k}, ..., a_{N,k}\}$ has to be constrained by more conditions:

Variance of pc_k is maximized
$Cov(pc_k, pc_l) = 0, \forall 1 \leq l < k$
$a_k^T a_k = 1$ (the weights vectors are unitary)

FA optimization relies on iterative perturbations with full-dimensional Gaussian noise and maximum-likelihood estimation where every observation in the data represents a sample point in a higher dimensional space. Whereas PCA assumes the noise is spherical, Factor Analysis allows the noise to have an arbitrary diagonal covariance matrix and estimates the subspace as well as the noise covariance matrix.

Under FA, the centered data can be expressed in the following from:

$$x_i - \mu_i = l_{i,1} F_1 + ... + l_{i,k} F_k + \epsilon_i = LF + \epsilon_i,$$

where $i \in 1, ..., p$, $j \in 1, ..., k$, $k < p$ and ϵ_i are independently distributed error terms with zero mean and finite variance.

Supervised classifiers

The data was split into 80% for training and 20% for testing. The data was trained using machine learning (ML) algorithms including neural network (Neuralnet),

k-nearest neighbors (kNN), generalized linear model (GLM), and recursive partitioning (Rpart).

Neuralnet model mimics the biological brain response to multisource stimuli (inputs). When we have three signals (or inputs) x_1, x_2 and x_3, the first step is weighting the features (w's) according to their importance. Then, the weighted signals are summed by the "neuron cell" and this sum is passed on according to an activation function denoted by **f**. The last step is generating an output **y** at the end of the process. A typical output will have the following mathematical relationship to the inputs.

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right).$$

kNN classifier performs two steps calculations. For a given k, a specific similarity metric d, and a new testing case x,

- Runs through the whole training dataset (y) computing $d(x, y)$. Let A represent the k closest points to x in the training data y.
- Estimates the conditional probability for each class, which corresponds to the fraction of points in A with that given class label. If $I(z)$ is an indicator function

$I(z) = \begin{cases} 1 & z = true \\ 0 & otherwise \end{cases}$, then the testing data input x gets assigned to the class with the largest probability, $P(y = j|X = x)$:

$$P(y = j|X = x) = \frac{1}{k}\sum_{i \in A} I(y^{(i)} = j).$$

Generalized linear model, specifically, logistic regression, is a linear probabilistic classifier. It takes in the probability values for binary classification, in this case, positive (0) and negative (0) mental well-being and estimates class probabilities directly using the logit transform function [33].

Recursive partitioning (Rpart) is a decision tree classification technique that works well with variables with definite ordering and unequal distances. The tree is built similarly as a random forest with a resultant complex model, however, Rpart procedure also consists of a cross-validation stage to trim back the full tree into nested terminals. The final model of the sub-tree provides the decision with the 'best' or lowest estimated error [34].

Model validation and performance assessment

Unsupervised techniques were evaluated based on the model visualization, as the best way to determine suitability of the models. Whereas, the ML-classifiers used specific parameters. The caret package was used for automated parameter tuning with *repeatedcv* method set at 15-folded cross-validation re-sampling that was repeated with 10 iterations [35]. The k-fold validation results and values were then used to calculate the confusion matrix that determines the measures of sensitivity, specificity, kappa, and accuracy. These measures were used to evaluate the performance of the ML-model classifiers. These measures were calculated as follows:

$$sensitivity = \frac{TP}{TP + FN}.$$

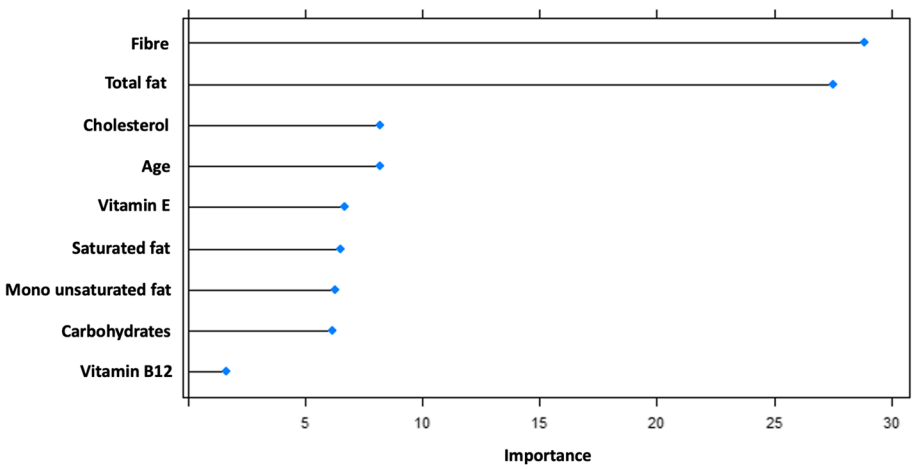

Fig. 4 Variable importance plot showing contribution of features to predicting colorectal cancer

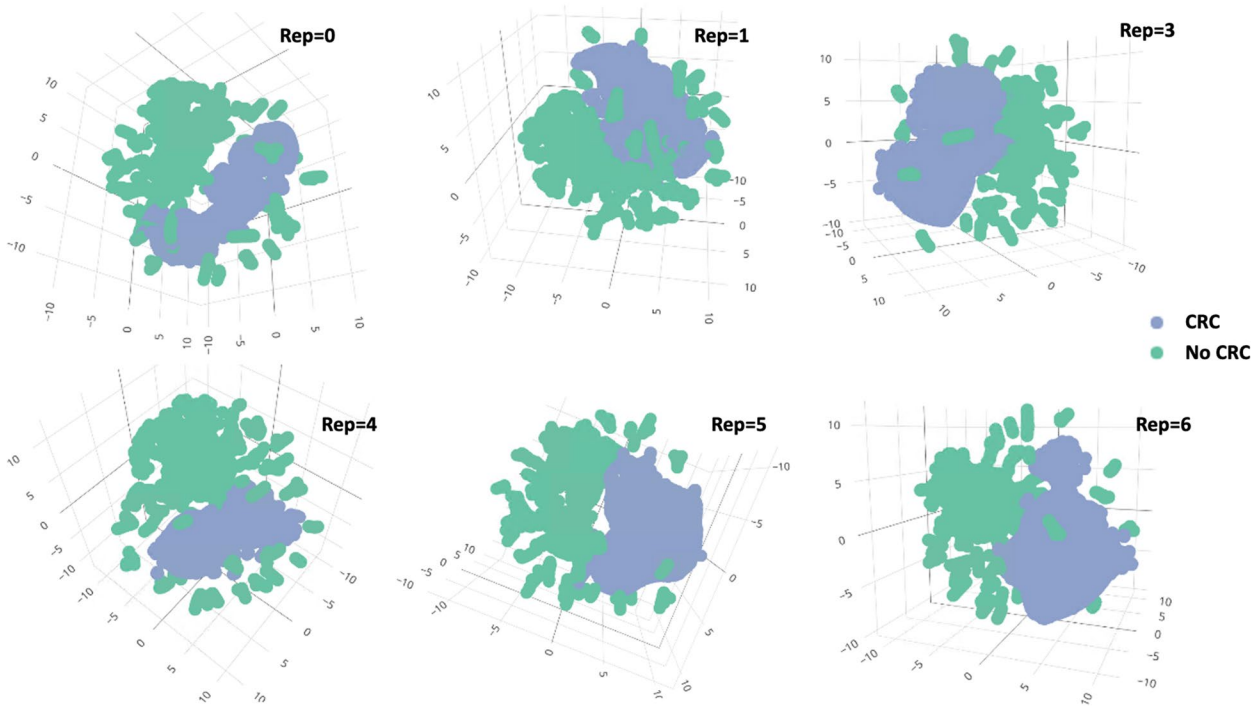

Fig. 5 Stability of t-SNE 3D embedding (Perplexity = 50) with six repeated (Rep) computations of the classification of no- and yes- colorectal cancer (CRC) labels

$$specificity = \frac{TN}{TN + FP}.$$

$$kappa = \frac{P(a) - P(e)}{1 - P(e)}.$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{\text{Total number of observations}}$$

where, True Positive(TP) is the number of observations that are correctly classified as "yes" or "success." True Negative(TN) is the number of observations that are correctly classified as "no" or "failure." False Positive(FP) is the number of observations that are incorrectly classified as "yes" or "success." False Negative(FN) is the number of observations that are incorrectly classified as "no" or "failure" (Dinov 2018).

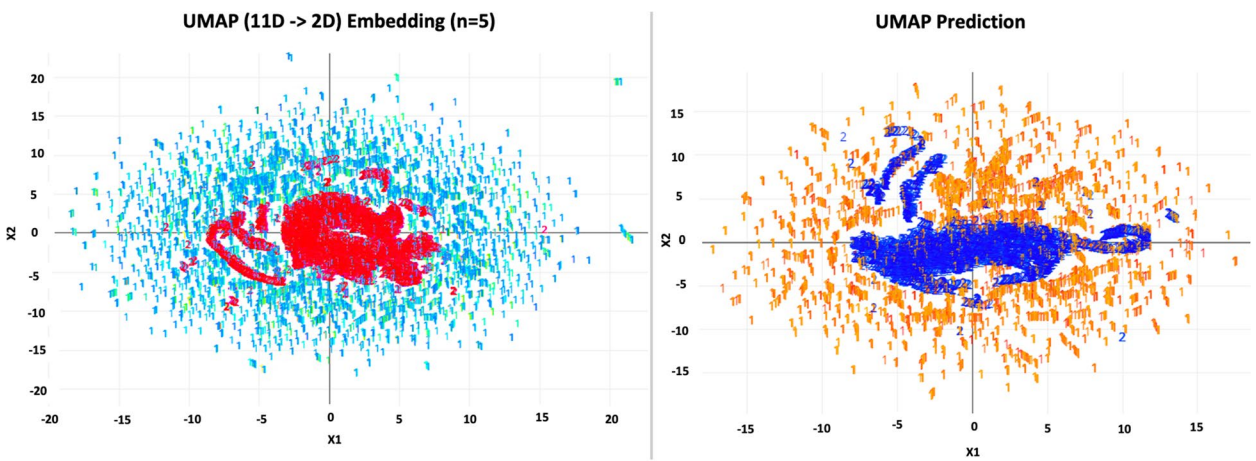

Fig. 6 Uniform Manifold Approximation (UMAP) 2D embedding model (n-neighbor = 5) (L) and UMAP prediction on testing data (R) in the classification of no- and yes- colorectal cancer labels

Results

Feature importance

The common features derived from the procedures of variable selection yielded ten salient variables (Fig. 4) that are important contributors of CRC including, by order of importance, fiber, total fat, cholesterol, age, vitamin E, saturated fats, monounsaturated fats, carbohydrates, and vitamin B12. These features were used in the next step of machine learning modeling.

Unsupervised learning

Among the unsupervised classifiers, t-SNE (Fig. 5) was the best performer. By visual inspection, t-SNE has maintained good stability of classifying positive CRC labels over several repeated computations. UMAP (Fig. 6) prediction also appears to be able to distinguish positive and negative CRC labels. Apriori association rules (Fig. 7) were able to map the textual features correlated to positive CRC labels, and the text items, by order of count, are listed in Table 2. PCA (Fig. 8) and FA (Table 3) showed that the data could be reduced to two dimensions where CRC is negatively correlated with fiber and carbohydrates, and positively correlated with the rest of the features.

Supervised learning

Model evaluation

In supervised classifiers, all techniques performed very well where accuracy, kappa, sensitivity, and specificity

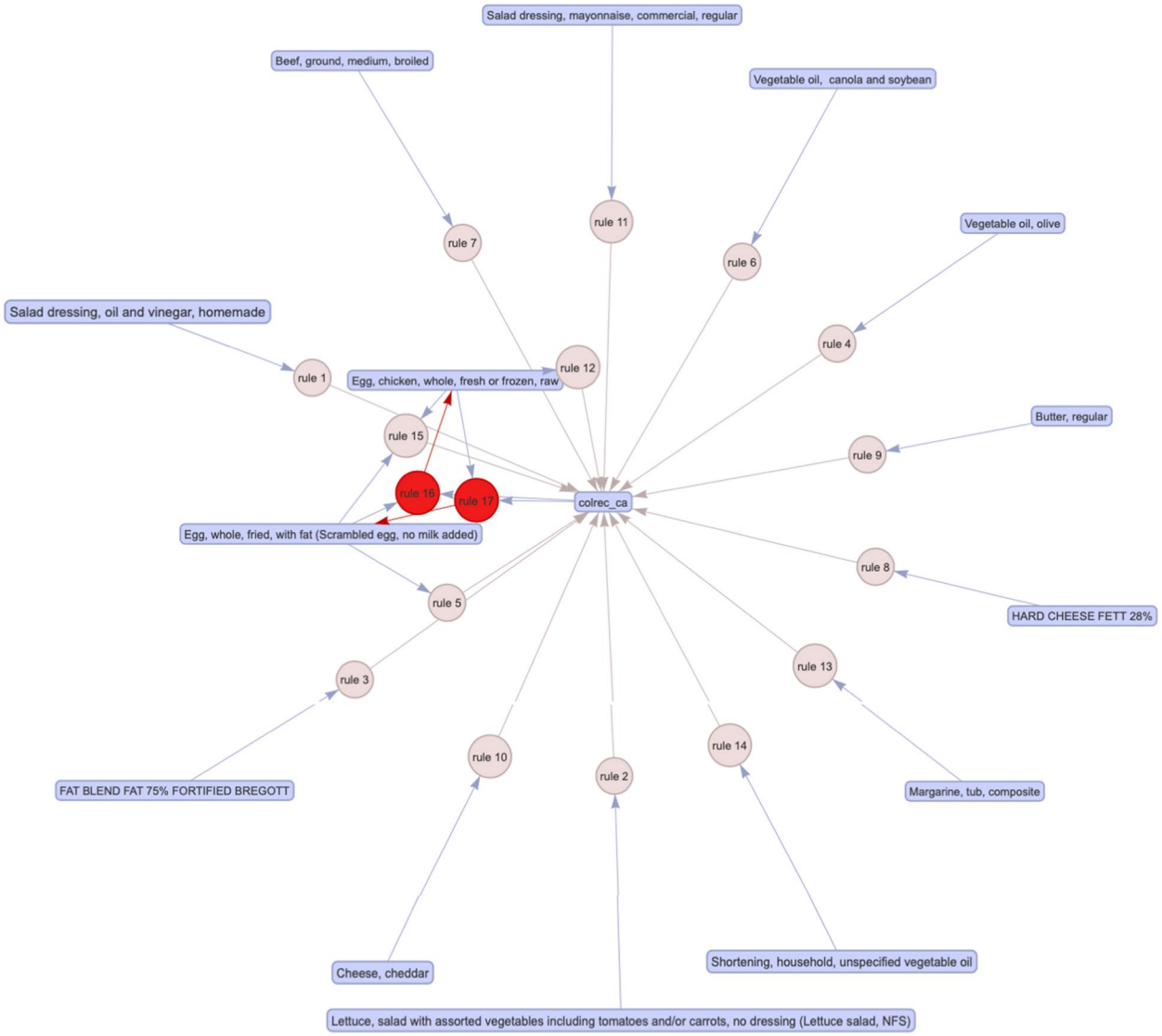

Fig. 7 Apriori association rules of text features that are associated with the labelled yes colorectal cancer ("colrec_ca") (See Supplementary 1 for this interactive html widget)

Table 2 Summary description of apriori association rules of no colorectal cancer ("No_colrec_ca") label

lhs	rhs	Support	Count
{Shortening, household, unspecified vegetable oil}	{colrec_ca}	0.062	3870
{Margarine, tub, composite}	{colrec_ca}	0.045	2818
{Egg, chicken, whole, fresh or frozen, raw}	{colrec_ca}	0.036	2277
{Cheese, cheddar}	{colrec_ca}	0.030	1903
{Salad dressing, mayonnaise, commercial, regular}	{colrec_ca}	0.029	1821
{HARD CHEESE FETT 28%}	{colrec_ca}	0.029	1804
{Butter, regular}	{colrec_ca}	0.028	1777
{FAT BLEND FAT 75% FORTIFIED BREGOTT}	{colrec_ca}	0.022	1360
{Beef, ground, medium, broiled}	{colrec_ca}	0.018	1150
{Vegetable oil, canola and soybean}	{colrec_ca}	0.018	1126
{Egg, whole, fried, with fat (Scrambled egg, no milk added)}	{colrec_ca}	0.016	998
{Vegetable oil, olive}	{colrec_ca}	0.015	971
{Lettuce, salad with assorted vegetables including tomatoes and/or carrots, no dressing (Lettuce salad, NFS)}	{colrec_ca}	0.014	905
{Egg, whole, fried, with fat (Scrambled egg, no milk added)} = > {Egg, chicken, whole, fresh or frozen, raw}	{colrec_ca}	0.012	751
{Egg, chicken, whole, fresh or frozen, raw} = > {Egg, whole, fried, with fat (Scrambled egg, no milk added)}	{colrec_ca}	0.012	751
{colrec_ca,Egg, whole, fried, with fat (Scrambled egg, no milk added)} = > {Egg, chicken, whole, fresh or frozen, raw}	{colrec_ca}	0.012	751
{colrec_ca,Egg, chicken, whole, fresh or frozen, raw} = > {Egg, whole, fried, with fat (Scrambled egg, no milk added)}	{colrec_ca}	0.012	751
{Egg, chicken, whole, fresh or frozen, raw, Egg, whole, fried, with fat (Scrambled egg, no milk added)}	{colrec_ca}	0.012	751
{Salad dressing, oil and vinegar, homemade}	{colrec_ca}	0.011	674

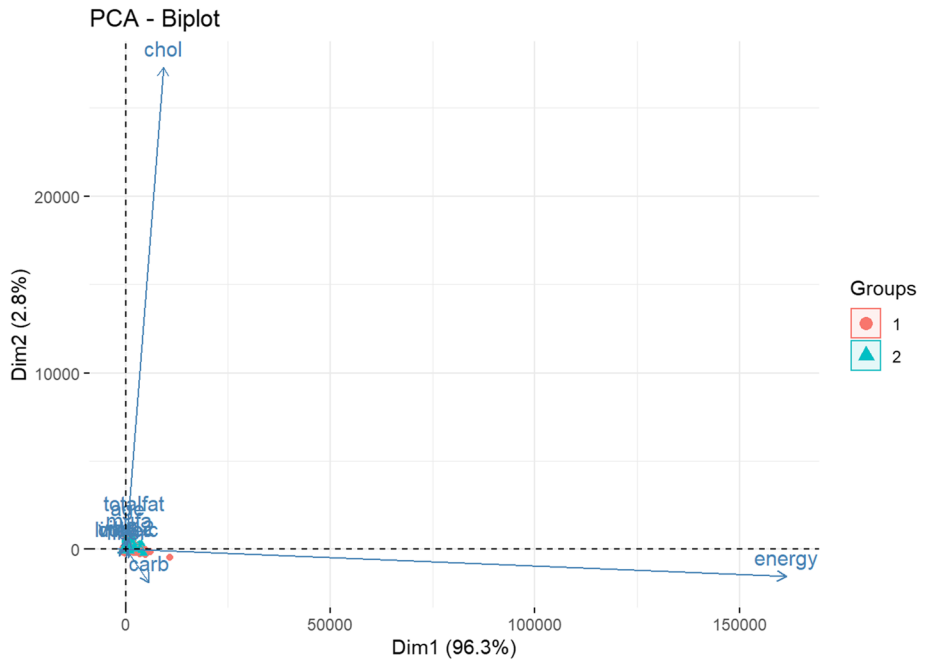

Fig. 8 A bi-plot of Principal component analysis on the most optimal number of dimensions in the data where Group 1 is no cancer label and Group 2 is the cancer label

Table 3 Two-factor model in the dimensionality reduction procedure of the colorectal cancer data

Factor Analysis	Two-factor model	
	Factor1	Factor2
Age	0.178	
Energy	0.433	0.525
carbohydrates	-0.121	0.972
fiber	-0.118	0.703
Total fat	0.990	0.123
Mono unsaturated fats	0.946	0.103
Omega-6	0.512	
cholesterol	0.483	
Vitamin B12	0.164	0.204
Linoleic acid	0.566	
Colorectal cancer	0.655	

were above 0.90 (Fig. 9). It appeared that the neural network performed better than the rest. By accounting the weight decay, the neural network model was optimal with a single layer of three hidden nodes, and we mapped out the schematic of the network, illustrated in Fig. 10. Sensitivity analysis also revealed seven features in the neural network model in future consideration (Fig. 11).

Discussion

Key findings

In this study, we show that colorectal cancer can be predicted based on a list of important dietary data using supervised and unsupervised machine learning approaches. The excellent level of prediction in the present study is congruent with previous findings where mis-classification only ranged from 1 to 2% [36, 37]. These machine learning models can be used both as an early tool to identify individuals at risk as well as predicting the clinical outcomes of colorectal cancer [38, 39].

Dietary control is one of the most effective protective and modifiable measures that the population can adopt for cancer prevention. Dietary features can signal clues of the likelihood of early-onset of specific type of colorectal cancer such as distal colon and rectum [40]. In fact, a systematic review of studies over a period of 17 years concluded that strong evidence linking dietary factors with CRC risk, however, specific food group components on this relationship, were limited [41]. The present study identified total fat, mono-unsaturated fats, linoleic acid, cholesterol, omega-6 as moderate to high correlated dietary features to positive colorectal cancer. In contrast, fiber and carbohydrates have negative correlation with colorectal cancer cases. These features reflects the evidence from precision nutrition that a combination of dietary parameters, particularly those in the healthy eating index (such as whole fruit, saturated fats, grains) are more accurate than single dietary index (such as glycemic index) is important in the modifiable behavior for cancer prevention [39, 42]. In addition, our text mining and apriori algorithm also indicated that vegetables, eggs, margarine, and cheese have great impacts on colorectal cancer.

Although all classifiers were very good predictors of CRC labels, artificial neural networks had the best accuracy and true positives and true negatives. The advantage of using neural networks over, for example, general linear models in cancer prediction, is having much lower uncertainty and better generalizability of the model [36, 43]. This is an important consideration since machine learning algorithms have increasingly been used in many medicine domains with varied success rates [44]. In addition, most or all data sets will have a clear imbalance between CRC and non-CRC labels. We used a smote technique

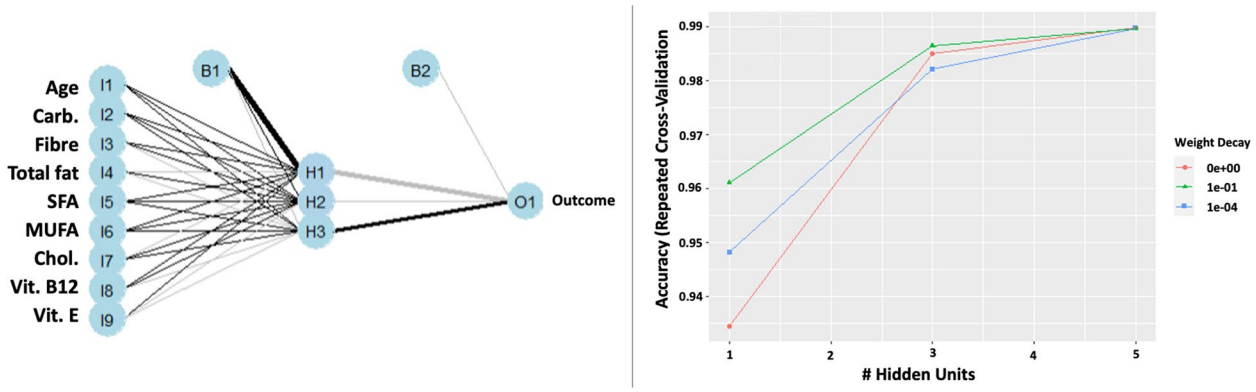

Fig. 9 A schematic of a neural network with a single hidden layer with three hidden nodes (L) and weight decay of optimal hidden node parameter using repeated cross-validation (R)

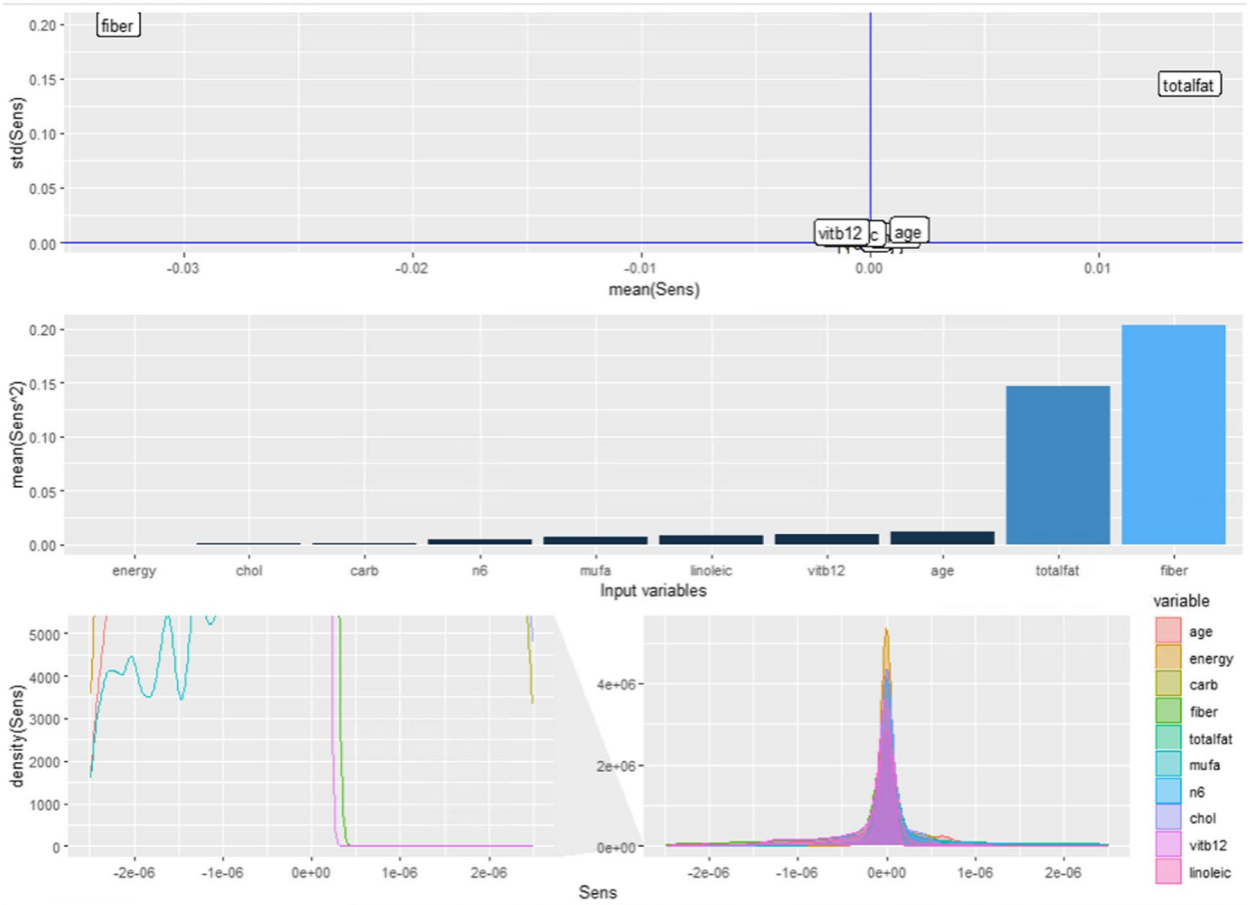

Fig. 10 Sensitivity analysis of the three hidden node neural network model in relation to the mean and standard deviation (top), mean square difference among the input variables (middle), and density plots (bottom)

to balance the data set, which otherwise, the machine learning models will predict all cases as non-CRC. Future work may need to consider controlling the sampling process to allow similar distribution of the two categories to minimize effects of down- or up- sampling. Another consideration is the age group of which this model is applicable. Unlike previous studies that account only for older people, this study includes younger adults in model training as well, therefore the models developed in this study may work well from young to older adults' CRC prediction. With early and regular screening assisted by an optimal machine learning algorithm, the incidence of CRC can be reduced even further.

Limitations

The strength of this study lies in the large datasets consisting of cases from seven major countries. Due to computational constraints, we randomly sampled observations to induce almost real-time estimates, model fits, and classification predictions. Some of the features that were not common had to be excluded from model development, which may result in confounding effects. The outcome label of CRC is based on detected cases and may not reflect early onset, new onset, or delayed onset of CRC as well as stratification of risk in different stages and types of CRC. Nevertheless, this study has narrowed down salient features that future researchers could consider in a more holistic approach, particularly, multi-dimensional that simultaneously accounts for diet, lifestyle, genetics, and related factors for CRC prediction.

Conclusion

In this study, we concluded that a combination of unsupervised and supervised machine learning approaches can be used to explore the key dietary features for colorectal cancer prediction. To help with feasibility and practicality, the artificial neural network was found to

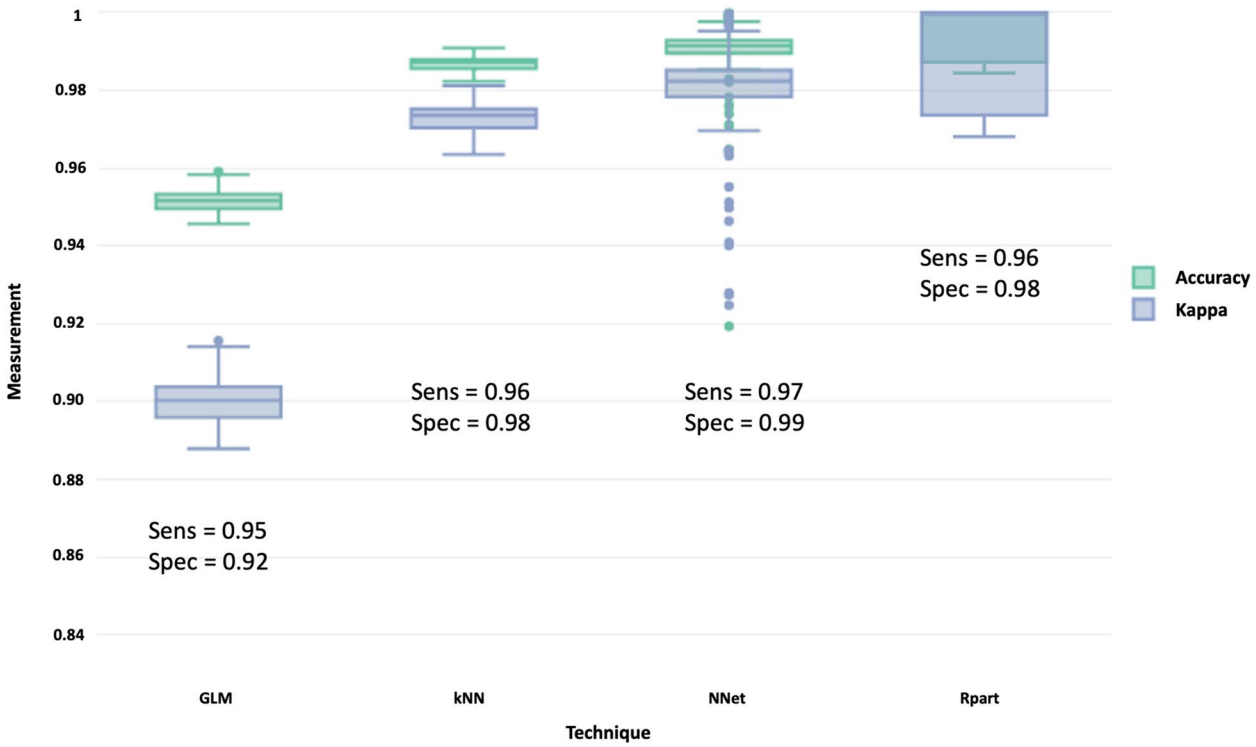

Fig. 11 Box plot evaluates the performance metrics of different classifiers in the prediction of colorectal cancer based on dietary data

be the optimal algorithm with misclassification of CRC of 1% and misclassification of non-CRC of 3%, for more effective cancer screening procedures. Furthermore, screening through dietary information can be used as a non-invasive procedure that can be applied in large populations. Using optimal algorithms coupled with high compliance to cancer screening will therefore significantly boost the success rate of cancer prevention.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-023-10587-x.

Additional file 1.

Acknowledgements
The authors expressed their sincere thanks to all databases used in this study.

Ethical guidelines
Not applicable.

Authors' contributions
HAR, MO, ID contributed to the conception or design of the paper. HAR conducted the data analysis. All authors contributed to data interpretation and drafting/editing the manuscript. All authors were involved in revising the manuscript, providing critical comments, and agreed to be accountable for all aspects of the work and any issues related to the accuracy or integrity of any part of the work.

Funding
This study was partially supported by grants from NSF (1916425, 1734853, 1636840, 1416953)and NIH (UL1TR002240, R01CA233487, R01MH121079, R01MH126137, T32GM141746).

Availability of data and materials
The datasets generated and/or analyzed during the current study are available upon reasonable request from the lead author, Dr Hanif Abdul Rahman.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 September 2022 Accepted: 27 January 2023
Published online: 10 February 2023

References

1. K. Hassibi, Machine learning vs. traditional statistics: different philosophies, different approaches, (2016). Data Science Central.
2. Stewart M. The actual difference between statistics and machine learning. Towar Data Sci. 2019;24:19.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A, Global cancer statistics,. GLOBOCAN estimates of incidence and

mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(2018):394–424.
4. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol. 2021;14:101174.
5. World Health Organization, Cancer, (2022). Retrieved 20 April 2022 from https://www.who.int/news-room/fact-sheets/detail/cancer.
6. Bénard F, Barkun AN, Martel M, von Renteln D. Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. World J Gastroenterol. 2018;24:124.
7. Schreuders EH, Ruco A, Rabeneck L, Schoen RE, Sung JJY, Young GP, Kuipers EJ. Colorectal cancer screening: a global overview of existing programmes. Gut. 2015;64:1637–49.
8. Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, De P, Tervonen H, Walsh PM, Bucher O. Changes in colorectal cancer incidence in seven high-income countries: a population-based study, Lancet. Gastroenterol Hepatol. 2019;4:511–8.
9. Guren MG. The global challenge of colorectal cancer, Lancet. Gastroenterol Hepatol. 2019;4:894–5.
10. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019;394:1467–80.
11. Henderson RH, French D, Maughan T, Adams R, Allemani C, Minicozzi P, Coleman MP, McFerran E, Sullivan R, Lawler M. The economic burden of colorectal cancer across Europe: a population-based cost-of-illness study, Lancet. Gastroenterol Hepatol. 2021;6:709–22.
12. Hossain MJ, Chowdhury UN, Islam MB, Uddin S, Ahmed MB, Quinn JMW, Moni MA. Machine learning and network-based models to identify genetic risk factors to the progression and survival of colorectal cancer. Comput Biol Med. 2021;135:104539.
13. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. Med Biol Eng Comput. 2019;57:901–12.
14. Bingham SA, Day NE, Luben R, Ferrari P, Slimani N, Norat T, Clavel-Chapelon F, Kesse E, Nieters A, Boeing H. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and nutrition (EPIC): an observational study. Lancet. 2003;361:1496–501.
15. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16:713–32.
16. Murphy N, Moreno V, Hughes DJ, Vodicka L, Vodicka P, Aglago EK, Gunter MJ, Jenab M. Lifestyle and dietary environmental factors in colorectal cancer susceptibility. Mol Aspects Med. 2019;69:2–9.
17. Centers for Disease Control and Prevention, National Health and Nutrition Examination Survey, (2022). Retrieved 20 April 2022 from https://www.cdc.gov/nchs/nhanes/index.htm.
18. Global Dietary Database, Microdata Surveys, (2018). Retrieved March 2022 from https://www.globaldietarydatabase.org/management/microdata-surveys.
19. U.S. National Library of Medicine, National Center for Biotechnology Information: dbGAP data, (2022). Retrieved March 2022 from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/collection.cgi?study_id=phs001991.v1.p1.
20. Inter-university Consortium for Political and Social Research, Find Data, (2022). Retrieved March 2022 from https://www.icpsr.umich.edu/web/pages/.
21. China Health and Nutrition Survey, China Health and Nutrition Survey, (2015). Retrieved March 2022 from https://www.cpc.unc.edu/projects/china.
22. Government of Canada, Canadian Community Health Survey, (2018). Retrieved March 2022 from https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs.html.
23. Data.world, Data.world, (2022). Retrieved March 2022 from https://ourworldindata.org.
24. Naing L, Bin Nordin R, Abdul Rahman H, Naing YT. Sample size calculation for prevalence studies using scalex and scalar calculators. BMC Med Res Methodol. 2022;22:209. https://doi.org/10.1186/s12874-022-01694-7.
25. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016;4:30.
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
27. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB. Package 'mass.' Cran R. 2013;538:113–20.
28. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36:1–13.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
30. Zhao M, Fu C, Ji L, Tang K, Zhou M. Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. Expert Syst Appl. 2011;38:5197–204.
31. Dinov ID, Data science and predictive analytics: Biomedical and health applications using R, Springer, 2018.
32. Dinov ID. Data Science and Predictive Analytics: Biomedical and Health Applications using R, 2nd edition, Springer Series in Applied Machine Learning, ISBN 978-3-031-17482-7. Cham, Switzerland: Springer; 2023.
33. Myers RH, Montgomery DC. A tutorial on generalized linear models. J Qual Technol. 1997;29:274–91.
34. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. Technical report Mayo Foundation. 1997;61:452.
35. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Core Team R. 2020 Package 'caret'. The R Journal 223, no. 7
36. Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF, Deng J. Robust machine learning for colorectal cancer risk prediction and stratification. Front Big Data. 2020;3:6.
37. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, Rust KC. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci. 2017;62:2719–27.
38. Gründner J, Prokosch H-U, Stürzl M, Croner R, Christoph J, Toddenroth D. Predicting Clinical Outcomes in Colorectal Cancer Using Machine Learning., in: MIE, 2018: pp. 101–105.
39. Shiao SPK, Grayson J, Lie A, Yu CH. Personalized nutrition—genes, diet, and related interactive parameters as predictors of cancer in multiethnic colorectal cancer families. Nutrients. 2018;10:795.
40. Hofseth LJ, Hebert JR, Chanda A, Chen H, Love BL, Pena MM, Murphy EA, Sajish M, Sheth A, Buckhaults PJ. Early-onset colorectal cancer: initial clues and current views. Nat Rev Gastroenterol Hepatol. 2020;17:352–64.
41. Tabung FK, Brown LS, Fung TT. Dietary patterns and colorectal cancer risk: a review of 17 years of evidence (2000–2016). Curr Colorectal Cancer Rep. 2017;13:440–54. https://doi.org/10.1007/s11888-017-0390-5.
42. T Li C Zheng L Zhang Z Zhou R Li 2015 Exploring the risk dietary factors for the colorectal cancer, in, IEEE Int. Conf. Prog. Informatics Comput IEEE 2015 570 573.
43. Abu Zuhri MAZ, Awad M, Najjar S, El Sharif N, Ghrouz I. Colorectal cancer risk factor assessment in Palestine using machine learning models, (2022).
44. L Zheng E Eniola J Wang M Learning for Colorectal Cancer Risk Prediction, in, 2021 Int. Conf. Cyber-Physical Soc. Intell IEEE 2021 1 6.

Publisher's Note

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com