Machines and Algorithms
http://www.knovell.org/mna

Research Article

Predicting Colorectal Cancer Using Machine Learning and Worldwide Dietary Data

Muhammad Sanaullah[1, *], Muhammad Kashif[1]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan, 60800, Pakistan

[*]Corresponding Author: Muhammad Sanaullah. Email: muhammad.sanaullah@aumc.edu.pk

Received: 29 November 2024; Revised: 08 January 2025; Accepted: 10 February 2025; Published: 20 March 2025

AID: 004-01-000048

Abstract: Colorectal Cancer (CRC) is considered to be a substantial catastrophic disease and the third most commonly reported type of cancer worldwide. By performing proactive screening of patients for CRC detection, it has been found that its is most prominently diagnosed in younger adults. However, most of the recently published papers have primarily focused upon the implication of statistical machine learning algorithms for CRC diagnosis in older adults with the aid of small-scale datasets, which are unable to depict acceptable performance in practice for large populations. So, it is crucial to assess machine learning algorithms on big datasets from varied areas and socio demographics, including both younger and older persons. The Centre for Disease Control and Prevention acquired a dataset of 109,343 individuals from colorectal cancer research in South Korea, India, Canada Mexico, Italy, Sweden, and the US. This worldwide dietary database was supplemented using publicly available information from several sources. In this study, we have evaluated performance of nine supervised and unsupervised machine learning methods on the aggregated dataset. Both type of tested models (i.e., supervised and unsupervised) models accurately predicted CRC and non-CRC traits. Among the nine tested models, artificial neural network (ANN) has achieved best performance, while attaining a misclassification rate of 1% and 3% for CRC and non-CRC respectively. ANN model has depicted extraordinary performance over diverse datasets, which make it a suitable choice for CRC diagnosis in both young and elderly persons. Using optimum algorithms and ensuring high screening compliance can significantly enhance early cancer detection and increase the success rate of prompt treatments.

Keywords: Colonel Cancer; Machine Learning; Cancer Screening; Early Cancer Detection;

1. Introduction

The recent revolution of artificial intelligence (AI) in 21st century has opened up new opportunities to enhance healthcare services by introducing advanced healthcare data analytics solutions, while overcoming conventional statistical and research constraints [1, 2]. Colorectal cancer (CRC) is one of the problems facing healthcare today. After lung and breast cancers, colorectal cancer (CRC) is the second most prevalent cause of cancer-related death globally and the third most often diagnosed disease [3, 4]. An estimated 1.93 million new cases of colorectal cancer were detected in 2020, representing 10% of all cancer cases worldwide [5]. Effective population-wide screening and monitoring initiatives that have been rapidly and proactively implemented may be responsible for the rising number of CRC cases worldwide [6, 7].

Nonetheless, CRC death rate remains high, with as estimated 0.94 million mortalities documented for this disease in 2020, accounting for 9.4% of all cancer deaths worldwide [5]. These statistics highlight the need of active health screening for the prevention of CRC in younger generation (under 50 years) specifically due to the high reporting rate of early-onset cases in technologically advanced countries, while an elevated observed rate in CRC incidence detection in developing and emerging economies [8, 9]. The medical advancements to enhance treatment options for CRC, such as by performing surgical and endoscopy-based interventions, targeted chemotherapy, immunotherapy and radiotherapy have led to improved survival rates and quality of life [9, 10]. Towards this end, Saudi Ministry of Health (MoH) has also taken a prominent measure i.e., by recommending early and periodic screening for CRC primarily on the basis of patient's history and symptoms. The two primary groups who are primarily focused for the early diagnosis of this catastrophic disease are individual having low risk i.e., between age 45 to 75, and secondly individuals with high risk of CRC i.e., who may have a family history of cancer or gets exposed to radiation therapy in their childhood. The colorectal examination is divided into two types: 1) the fecal occult blood test (FOBT) also known as fecal immunological test (FIT) and secondly 2) the whole colonoscopy [11]. Early detection of CRC leads to a better prognosis for treatment, but it still poses significant public health and financial issues (9). In 2015, the economic impact of CRC in Europe was projected to be 19 billion euros, including hospital bills, lost productivity, premature mortality, and informal care costs [11]. Early-onset CRC pathological characteristics are sporadic and require further investigation to properly understand the underlying processes and risk factors [9]. With the advancement of digital technologies, health information systems can efficiently collect high-quality CRC data from a larger patient population. This has allowed data science to provide a new route for increasing understanding about CRC through research and development.

Machine-learning methods have successfully predicted CRC based on genomic data, indicating inherited propensity in some situations [12, 13]. However, genetic disorders are persistent and unchangeable risk factors. Dietary restriction is a highly effective way to prevent CRC, as it is linked to a lifestyle associated with globalization [4, 14, 15-16]. As the food business and supply chain become more globalized, it's crucial to do data science study on how global diets impact CRC prediction.

This research aims to develop ML models to identify key dietary components influencing CRC risk. By utilizing publicly available global dietary data, we seek to improve understanding of how dietary habits contribute to CRC and enhance predictive analytics for early detection and prevention strategies. In this research, we used exploratory unsupervised and supervised machine learning-based models to examine the key dietary components in predicting CRC labelling.

2. Problem Statement

CRC is the second leading cause of cancer-related deaths worldwide. Despite improvements in early detection methods, such as fecal tests and colonoscopy, and improvements in treatment options, CRC death count is high. Genetic predisposition is a major risk factor, but lifestyle and dietary habits, influenced by globalization, also play a crucial role in CRC development. The existing studies primarily focus on genomic data, overlooking the impact of dietary patterns on CRC risk.

3. Related Work

The authors of [17] employed feature selection methods and machine learning algorithms to identify colon cancer. For feature selection, the maximum degree greedy (MDG) and malondialdehyde (MDA) algorithms were employed. AdaBoost, logistic regression (LR), KNN, SVM, and RF Algorithms have been used on a public dataset consisting of 2000 genes and 62 instances. Among them are 22 normal patients and 40 abnormal patients. The outcome demonstrated that, the random forest classifier together with a features selection approach has achieved the best accuracy of 95.2%. Only genes are used as features in the model.

The study in [18] uses an ensemble classifier approach to divide tissues into normal and pathological categories. Filtering and wrapping are the feature selection techniques that were applied. At Pablo de

Olavide University's Bioinformatics Research Group, 62 patients and 1200 gene expressions were subjected to the machine learning algorithms KSVM, RF, eXtreme ensemble, KNN and Gradient Boosting (XGB). Among them are 22 normal patients and 40 abnormal patients. This model's results showed that the proposed ensemble learning based approach has achieved the best accuracy of 91.7%.

The authors of [19] examined the challenge of identifying colorectal cancer. The primary model that has been employed in this study is modified Harmony Search Algorithm (Z-FS-KM-MHS). The Princeton University Gene Expression Project provided 2000 genes in all. Z-FS-KM-MHS obtained an accuracy of up to 94.4%, according to the results. Many genes were employed in the model. In contrast to other research, this approach can be used to study genetically based disorders like breast cancer.

Hamida et al. has presented a deep Convolutional Neural Network (CNN) model for dividing colon pictures into normal and nonnormal categories. In Germany, the UNET and SEGNET models were the primary models used for 100,000 histopathology scans [46–50]. SEGNET achieved 99.5% high-performance accuracy. The scientists came to the conclusion that DL performs better at picture classification than ML when dealing with large-scale photos. In contrast to, colon cancer was classified using pictures [20].

The authors of [21] enhanced the colon cancer diagnosis. To do so, selected machine learning models have been trained over two publicly available datasets, which primarily incorporate 98 samples and 9457 genes. The ML models that have been selected and trained in this study are decision trees (DT), naive Bayes (NB), Support Vector Machine (SVM) and KNN. The results of the study reveal that the KNN and DT models have achieved best classification performance over the first selected dataset, while NB model has attained best results on second dataset.

The issue of colonoscopy failure to detect polyps is examined by the authors in [22]. The primary technique used was DL on 27,113 colonoscopy pictures and 1290 patients which belongs to Sichuan Provincial People's Hospital's Endoscopy Center. The employed DL classifier has achieved a Per-image detection rate of 91.6%. But the algorithm just finds polyps. In order to identify colon tumors from biopsy data, another prominent methodology has been presented by author of [23], who basically employed the Density-Based Spatial Clustering of Applications with Noise (DB-SCAN) algorithm, which classifies healthy cells from dangerous ones. One hundred photos gathered from Zendo repositories were used to test the algorithm. According to the findings, the model detected colon cancers with 99% accuracy score, 85.4% sensitivity score, and 87.6% specificity score [23].

Jørgensen et al. extracted information from cell nuclei to determine whether the tissue was malignant or benign. To do so, author has exploited a multi-classifier based approach, which include RF, k-means clustering, color deconvolution, local adaptive thresholding, and separation of cell within ROI on 87 distinct colon tissue slides. Consequently, the proposed algorithm's sensitivity, specificity, accuracy, and area under the curve (AUC) were 0.96, 0.88, 0.92, and 0.91, respectively [24].

Using ANNs and a feature selection approach, authors in [25] has presented a model for the classification of lung and colon cancer. The creators of this model used a publicly available dataset with 62 cases and 2000 genes. For the two classifications of cancer and normal, the classification accuracy was 98.4%. Furthermore, the authors discovered that the feature selection approach might improve the model's classification accuracy.

In his study, Choi et al has presented a deep learning-based computer-aided diagnosis (CAD) system for the multi-classification of pathologic histology of colorectal adenoma in four different classes. The model developed a diagnostic method that primarily forecast tissue adenoma of the colon and rectum by using CNN's algorithm, while employing 3400 computed tomography (CT) images gained from a Korea based Hospital (KUMC) and a CAD. The authors then contrasted the system's output with the experts' findings. According to the findings, the classification had a sensitivity of 77.25% and a specificity of 92.42%. Furthermore, it closely matched the experts' findings. Lack of sufficient samples to evaluate the model's validity is one of the authors' challenges [26].

Using a self-speed transmission network, Yao et al. suggested automatically classifying and segmenting colorectal images into three groups: cancers, polyps, and normal tissue. To enhance the outcome, a pre-

trained ImageNet network was first used, and then 3061 photos were used. The trained STVGG network was employed for further analysis for colon rectal classification after the Unet network architecture was utilized for segmentation. The model has achieved good segmentation and classification accuracy. The combination of two goals—classification and segmentation—sets this paper apart from the other research. Additionally, it employed self-learning to learn the challenging sample, address the imbalance issue, and improve performance [27].

Table 1: Literature Review Analysis

Ref.	Dataset Instances	ML Model	Achieved Accuracy
[17]	62 patients and 1200 gene expression	AdaBoost, KNN, logistic regression (LR), SVM, and RF	95.161%
[18]	62 patients and 1200 gene expression	RF, KSVM, eXtreme Gradient Boosting (XGB), KNN,	91.67%.
[19]	2000 genes	Z-FS-KM-MHS	94.36%
[20]	100,000 histopathology scans	CNN, SEGNET	99.5%
[21]	9457 genes and 98 samples	SVM, Naïve Bayes, Decision Tree, KNN	-
[23]	27,113 colonoscopy pictures and 1290 patients	DB-SCAN	99%
[24]	87 colon tissue	RF, K mean Clustering	92%
[25]	62	ANN	98.4%.
[26]	3400 computed tomography (CT) images	CNN	sensitivity = 77.25%, specificity = 92.42%.
[27]	3061 photos	Self-paced Transfer VGG (STVGG)	-

The comparison analysis of literature is given in table 1 where 1st column represents the references of selected paper for analysis, 2nd column represents the dataset used in study, 3rd column represents the ML model used and last column represents the highest achieved accuracy. The datasets used in the studies range from small-scale gene expression data with a few dozen patients [17, 18], [25] to large-scale histopathology and colonoscopy image datasets containing thousands of samples [20], [23], [26]. The nature of the datasets significantly influences the choice of ML models. Gene expression-based studies [17-19], [21], [24] used traditional ML models such as 1) RF, 2) KNN, 3) SVM and 4) eXtreme Gradient Boosting (XGB), which are effective for structured data analysis. On the other hand, image-based datasets in studies [20], [23], [26] based on deep learning models like Convolutional Neural Networks (CNN) and SEGNET, which are well-suited for feature extraction from complex medical images.

Performance comparisons indicate that deep learning models perform exceptionally well on medical imaging tasks. CNN and SEGNET achieved the highest accuracy of 99.5% in histopathology image classification [20], while DB-SCAN [23] achieved 99% accuracy in colonoscopy image analysis. In traditional ML approaches, AdaBoost, KNN, Logistic Regression (LR), SVM, and RF [17] reached 95.161% accuracy, demonstrating the strength of ensemble methods in gene expression data analysis. Additionally, Artificial Neural Networks (ANN) [25] achieved an impressive 98.4% accuracy, indicating that neural networks remain competitive in structured data classification. Moderate performance was observed in RF and clustering-based approaches [24], which achieved 92% accuracy, and Z-FS-KM-MHS [19] with 94.36% accuracy. Notably, study [26] reported sensitivity (77.25%) and specificity (92.42%) instead of overall accuracy, offering a different perspective on model effectiveness. However, some studies [21], [27] did not report accuracy, limiting direct comparisons.

Overall, the findings indicate that deep learning models, particularly CNN-based architectures, dominate medical image classification, while traditional ML methods such as SVM, RF, and XGB continue to excel in gene expression data analysis. Hybrid and clustering-based approaches, such as DB-SCAN [23] and self-paced transfer learning [27], show potential in specific applications. Future research should focus on integrating multimodal data (gene expression and imaging) to improve predictive accuracy while enhancing the interpretability of AI-driven healthcare solutions.

4. Methodology

The proposed methodology composed of Dataset selection, Data Preprocessing, Feature Selection, Data Normalization and Classification as shown in Fig. 1.

4.1. Dataset

The Centers for Disease Control and Prevention also renowned as the Global Dietary database, and publicly available institutional websites provided the dietary-related colorectal cancer data [28-34]. Canada, Korea, Argentina, Ecuador, Bangladesh, Estonia, Bulgaria, Finland, China, India, Ethiopia, Israel, Germany, Malaysia, Iran, Kenya, Mexico, Portugal, United States, the Philippines, Japan, Mozambique, Italy, Sweden and Tanzania were among the 25 nations that made up the original combined data. These data sets were collected using comparable techniques, such as cross-sectional surveys and food questionnaires. Following that, the various data sets were combined and extrapolated using the same dietary features. Excluded were characteristics that did not appear in all of the data sets.

4.2. Data Preprocessing

Only English-language data sets are included in this analysis. For the purpose of standardization, features with disparate measurement units were transformed. A process of cleaning was used, such as listwise elimination of features having greater than 50% missing values, duplicate characteristics, and ineligible situations. There were still 3,520,586 valid data points at this point.

During preprocessing a specific sample of dataset containing 109,342 cases have been extracted to tackle the computational cost issue. This data sampling has been done through a multi-stage, proportionate random sample method. Of these, 7,326 (6.7%) cases had positive colorectal cancer labels, which are derived for seven distinct countries worldwide.

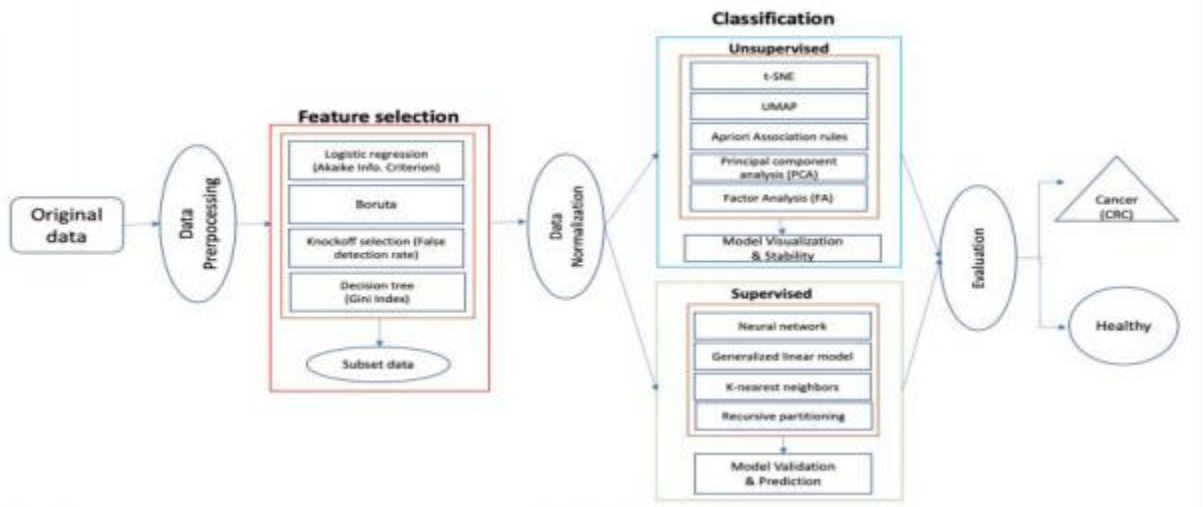

Figure 1: Proposed Methodology

4.3. Feature Selection

A two-phase feature selection process was used. Three distinct processes are involved in step one: Boruta, Knock of selection, and logistic regression (LR). Each index was filtered out using LR in order to decrease redundant features by utilizing the step AIC function in the MOSS package to calculate a stepwise iterative process of forward addition and backward removal. The primary aim of forward addition is to add significant features to a null set of features, while the aim of backward removal is to remove the worst-performing features from the list of full features [35]. Based on statistical analysis, the most statistically significant features are considered to be most economical model with the lowest Akaike Information Criterion (AIC) were used to choose variables. Then, a randomized wrapper technique called Boruta was used, which gradually eliminates characteristics that are more important and statistically insignificant than those of random probes [36].

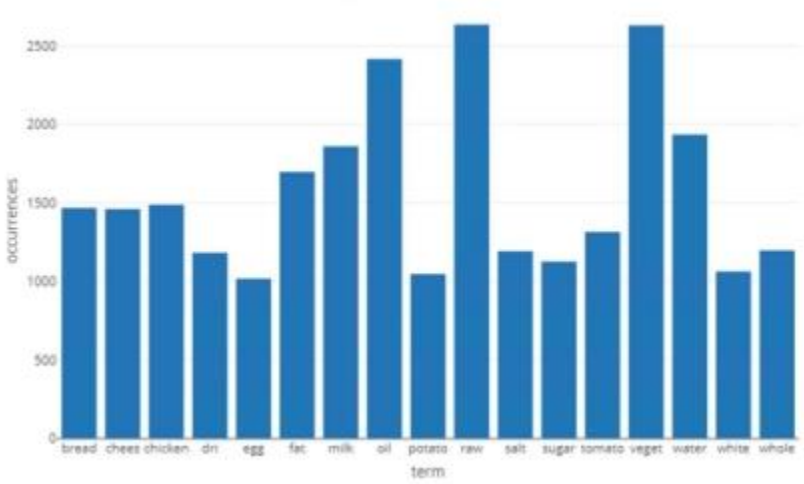

Figure 2: Frequent Text Chunks in Data (1,000 occurrences)

$$False\ Discovery\ Rate\ (FDR) = E\left(\frac{\#\ False\ Positives}{total\ number\ of\ selected\ features}\right) \quad (1)$$

Where E is the expectation and the given ratio is False Discovery Proportion.

Formulae for Gini Index has been mentioned in equation below, which primarily assist in final attributes selection on the basis of variable importance:

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

where k represents how many classes are there.

4.4. Data Normalization

Data normalization was then used to further process the data set with the finalized features in order to assist achieve improved classification accuracy and prevent the effects of extreme numerical ranges [37, 38]. The following is how the features were scaled:

$$V' = \frac{V-Min}{Max-Min'} \quad (3)$$

where Min and Max values depict the upper and lower data boundaries, and V′ is the scale value that corresponds to the initial value V. Ultimately, the most noteworthy characteristics to be applied to both supervised and unsupervised classifications were those that intersected throughout the two-step variable selection processes.

4.5. Classification

The dimensions of the data were investigated using four different forms of unsupervised machine learning for nonlinear relationships: uniform manifold approximation, factor analysis (FA), t-distributed stochastic Neighbor embedding (t-SNE), Apriori association rules, principal component analysis (PCA), and projection (UMAP) [38]. High-dimensional data may be embedded into lower-dimensional spaces with the help of the t-SNE technique, which is a renowned ML approach for nonlinear dimensionality reduction. For each pair (xi, xj), t-SNE calculates the probabilities pi,j that are proportionate to their corresponding similarities, pj|i, if the high dimensional data (N D) is x_1, x_2, … x_N.

$$p_{j|i} = \frac{exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \tag{4}$$

The kNN classifier does computations in two phases. specified a certain similarity metric d, a new testing case x, and a specified k.

- Computes d(x, y) after running through the entire training dataset (y). Let A stand for the k points in the training data y that are closest to x.
- Calculates the proportion of points in A that have a certain class label, or the conditional probability for each class. If an indicator function, is I(z).

$$P(y = j|X = x) = \frac{1}{k}\sum_{i \epsilon A} I(y^{(i)} = j) \tag{5}$$

Based on the model visualization, unsupervised procedures were assessed as the most effective means of assessing the models' appropriateness. In contrast, the ML-classifiers employed certain parameters. Automatic parameter tweaking has been employed with the assistance of repeated technique set at the caret package. Ten rounds of a 15-folded cross-validation resampling were conducted [39]. Based upon the results obtained from k-folds validation, confusion matrix calculation is done, which further assist in gauging metrics like 1) specificity, 2) accuracy, 3) sensitivity and 4) kappa. The performance of the ML-model classifiers was assessed using these metrics. The following formula was used to determine these measures:

$$sensitivity = \frac{TP}{TP+FN} \tag{6}$$

$$specificity = \frac{TN}{TN+FP} \tag{7}$$

$$kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

TP in above formulae depicts the number of instances that are correctly classified as positive/Yes class, while ***TN*** depicts the number of instances that are correctly categorized as negative/No class. On the other hand, the instances that are mis-categorized as positive/yes class are referred as ***FP*** and the count instances that are misclassified as negative class are termed as ***FN***.

5. Results and Discussion

Ten noteworthy factors (Fig. 3) that are significant contributors to CRC were identified from the common characteristics obtained from the variable selection techniques. These variables are, in order important factors, including age, total fat, cholesterol, fibre, vitamin E, monounsaturated and saturated fats, carbs, and vitamin B12. The following stage of machine learning modelling made use of these attributes.

The neural network seems to function better than the others. We mapped out the network schematic, and concluded that the employed neural network classification model with a single layer having three hidden nodes was performed best when weight decay was taken into consideration. Additionally, sensitivity analysis identified seven characteristics in the neural network model that should be considered going forward.

In this work, we demonstrate that supervised and unsupervised machine learning techniques may be used to predict colorectal cancer based on a list of significant dietary data. The current study's high prediction accuracy is consistent with other research showing that misclassification rates only varied between 1% and 2% [40, 41]. These machine learning algorithms can be used to forecast the clinical outcomes of colorectal cancer as well as to identify people at risk early on [42, 43]. One of the most effective preventative and changeable strategies for cancer that the public may use is dietary management. Dietary characteristics, such as those of the distal colon and rectum, might provide indicators of the risk of developing certain types of colon-rectal cancer early on [44]. Indeed, a comprehensive analysis of research conducted over 17 years found that there is substantial evidence connecting dietary variables to the incidence of colorectal cancer (CRC). There were few elements in this connection [45].

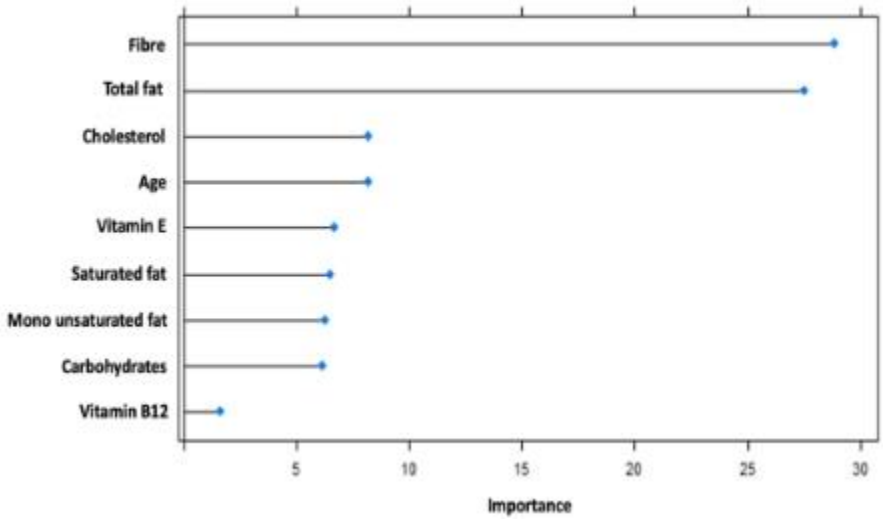

Figure 3: A variable significance plot that illustrates how factors contribute to the prediction of colorectal cancer

Table 2: Two-factor dimensionality reduction model results

Factor based Analysis	Two Factor Model	
	Factor 1	**Factor 2**
Age	0.8	
Carbohydrates	-0.1	0.97
Energy	0.4	0.5
Total fat	0.99	0.1
Fiber	-0.1	0.7
Omega-6	0.5	
Mono unstructured fats	0.9	0.1
Vitamin B12	0.2	0.2
Cholesterol	0.5	
Colorectal cancer	0.65	
Linoleic acid	0.6	

The current investigation found that dietary characteristics that were moderately to highly connected with positive colorectal cancer included total fat, monounsaturated fats, linoleic acid, cholesterol, and omega-6. On the other hand, there is a negative association between colorectal cancer incidences and fiber and carbs. These characteristics are consistent with precision nutrition research showing that dietary parameters, especially those included in the healthy eating index (1) saturated fats, 2) whole fruit, and 3) grains), are more accurate than those found in a single dietary index (like the glycaemic index) when it comes to modifiable behavior for cancer prevention [43, 46]. Furthermore, our apriori algorithm and text mining revealed that 1) vegetables, 2) margarine, 3) eggs, and 4) cheese had significant effects on colorectal cancer [47].

This study's strength is its extensive datasets, which include instances from seven major nations. Owing to processing limitations, we generated estimates, model fts, and classification predictions by randomly sampling observations. Confounding effects might arise because some of the less prevalent elements have to be left out of the model's development.

The CRC outcome label is based on instances that have been discovered and may not reflect risk stratification in different stages and kinds of the disease or early, new, or delayed start of CRC [48]. However, this study has identified key elements that future researchers may take into consideration in a more comprehensive manner, especially multi-dimensional approaches that concurrently take genetics, lifestyle, nutrition, and other relevant aspects into account for the prediction of colorectal cancer [49].

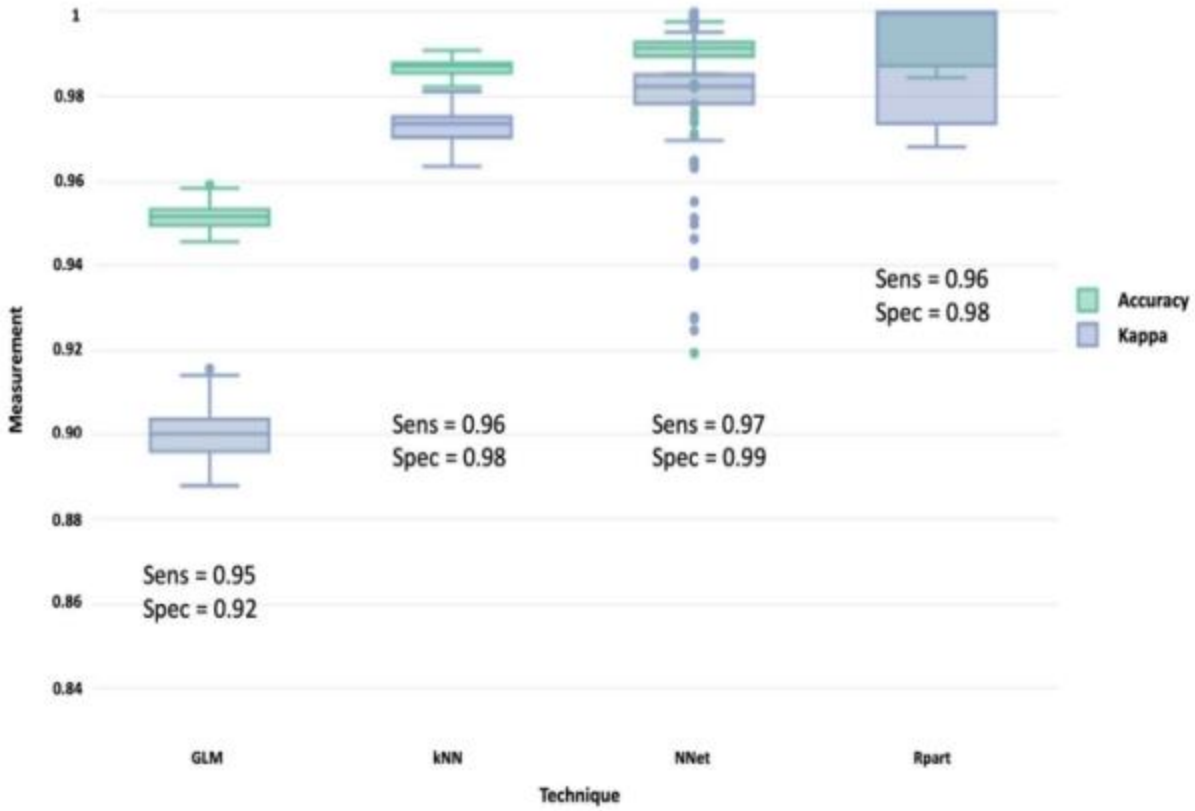

Figure 4: Using dietary data, box plots for the evaluation of performance metrics for colorectal cancer classification models

6. Conclusion

We concluded in this work that the critical dietary factors for colorectal cancer prediction may be explored using a combination of supervised and unsupervised machine learning techniques. With a 1% misclassification of colorectal cancer and a 3% misclassification of non-CRC, the artificial neural network was determined to be the best algorithm for more practical and feasible cancer screening methods. Additionally, using dietary data for screening is a non-invasive method that may be employed on a broad population. Therefore, the success rate of cancer prevention will be significantly increased by using optimum algorithms in conjunction with high cancer screening compliance. Future research should focus on integrating multimodal data, combining gene expression profiles with medical imaging to enhance colorectal cancer prediction accuracy. The use of advanced deep learning architectures, such as transformer-based models, can further improve feature extraction and interpretation in medical diagnostics. Additionally, explainable AI techniques should be explored to increase the transparency and trustworthiness of ML models in healthcare. Developing cost-effective and scalable ML frameworks for early disease detection, particularly in resource-limited settings, will be crucial. Lastly, large-scale, diverse datasets and federated learning approaches can help address data privacy concerns while improving model generalizability across different populations.

Funding Statement: The authors declare that this work was carried out without financial support from any funding agency.

Conflicts of Interest: The authors of this paper have no potential conflicts of interest.

Data Availability: The dietary-related colorectal cancer data are publicly available from the CDC, the Global Dietary Database, and institutional websites

References

[1] Hassibi, K. "Machine learning vs. traditional statistics: Different philosophies, different approaches-DataScienceCentral. com." *Data Science Central* (2016).

[2] Stewart, Matthew. "The Actual Difference Between Statistics and Machine Learning." *Medium: TDS Archive.* https://medium.com/data-science/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3

[3] Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 68, no. 6 (2018): 394-424.

[4] Xi, Yue, and Pengfei Xu. "Global colorectal cancer burden in 2020 and projections to 2040." *Translational oncology* 14, no. 10 (2021): 101174.

[5] World Health Organization, Cancer, (2022). Retrieved 20 April 2022 from https://www.who.int/news-room/fact-sheets/detail/cancer.

[6] Bénard, Florence, Alan N. Barkun, Myriam Martel, and Daniel von Renteln. "Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations." *World journal of gastroenterology* 24, no. 1 (2018): 124.

[7] Schreuders, Eline H., Arlinda Ruco, Linda Rabeneck, Robert E. Schoen, Joseph JY Sung, Graeme P. Young, and Ernst J. Kuipers. "Colorectal cancer screening: a global overview of existing programmes." *Gut* 64, no. 10 (2015): 1637-1649.

[8] Araghi, Marzieh, Isabelle Soerjomataram, Aude Bardot, Jacques Ferlay, Citadel J. Cabasag, David S. Morrison, Prithwish De et al. "Changes in colorectal cancer incidence in seven high-income countries: a population-based study." *The lancet Gastroenterology & hepatology* 4, no. 7 (2019): 511-518.

[9] Guren, Marianne Grønlie. "The global challenge of colorectal cancer." *The Lancet Gastroenterology & Hepatology* 4, no. 12 (2019): 894-895.

[10] Dekker, Evelien, Pieter J. Tanis, J. L. Vleugels, Pashtoon M. Kasi, and Michael Wallace. "Pure-amc." *Lancet* 394, no. 10207 (2019): 1467-1480.

[11] Alboaneen, Dabiah, Razan Alqarni, Sheikah Alqahtani, Maha Alrashidi, Rawan Alhuda, Eyman Alyahyan, and Turki Alshammari. "Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities." *Big Data and Cognitive Computing* 7, no. 2 (2023): 74.

[12] Hossain, Md Jakir, Utpala Nanda Chowdhury, M. Babul Islam, Shahadat Uddin, Mohammad Boshir Ahmed, Julian MW Quinn, and Mohammad Ali Moni. "Machine learning and network-based models to identify genetic risk factors to the progression and survival of colorectal cancer." *Computers in Biology and Medicine* 135 (2021): 104539.

[13] Zhao, Dandan, Hong Liu, Yuanjie Zheng, Yanlin He, Dianjie Lu, and Chen Lyu. "A reliable method for colorectal cancer prediction based on feature selection and support vector machine." *Medical & biological engineering & computing* 57, no. 4 (2019): 901-912.

[14] Bingham, Sheila A., Nicholas E. Day, Robert Luben, Pietro Ferrari, Nadia Slimani, Teresa Norat, Françoise Clavel-Chapelon et al. "Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study." *The lancet* 361, no. 9368 (2003): 1496-1501.

[15] Keum, NaNa, and Edward Giovannucci. "Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies." *Nature reviews Gastroenterology & hepatology* 16, no. 12 (2019): 713-732.

[16] Murphy, Neil, Victor Moreno, David J. Hughes, Ludmila Vodicka, Pavel Vodicka, Elom K. Aglago, Marc J. Gunter, and Mazda Jenab. "Lifestyle and dietary environmental factors in colorectal cancer susceptibility." *Molecular aspects of medicine* 69 (2019): 2-9.

[17] Shafi, A. S. M., MM Imran Molla, Julakha Jahan Jui, and Mohammad Motiur Rahman. "Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques." *SN Applied Sciences* 2, no. 7 (2020): 1243.

[18] Islam, Ashraful, Mohammad Masudur Rahman, Eshtiak Ahmed, Faisal Arafat, and Md Fazle Rabby. "Adaptive feature selection and classification of colon cancer from gene expression data: an ensemble learning approach." In *Proceedings of the international conference on computing advancements*, pp. 1-7. 2020.

[19] Bae, Jin Hee, Minwoo Kim, J. S. Lim, and Zong Woo Geem. "Feature selection for colon cancer detection using k-means clustering and modified harmony search algorithm." *Mathematics* 9, no. 5 (2021): 570.

[20] Hamida, A. Ben, Maxime Devanne, Jonathan Weber, Caroline Truntzer, Valentin Derangère, François Ghiringhelli, Germain Forestier, and Cédric Wemmert. "Deep learning for colon cancer histopathological images analysis." *Computers in Biology and Medicine* 136 (2021): 104730.
[21] Al-Rajab, Murad, Joan Lu, and Qiang Xu. "A framework model using multifilter feature selection to enhance colon cancer classification." *Plos one* 16, no. 4 (2021): e0249094.
[22] Wang, Pu, Xiao Xiao, Jeremy R. Glissen Brown, Tyler M. Berzin, Mengtian Tu, Fei Xiong, Xiao Hu et al. "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy." *Nature biomedical engineering* 2, no. 10 (2018): 741-748.
[23] Rajesh, Gundlapalle, Boda Saroja, M. Dhivya, and A. B. Gurulakshmi. "DB-scan algorithm based colon cancer detection and stratification analysis." In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 644-648. IEEE, 2020.
[24] Jørgensen, Alex Skovsbo, Anders Munk Rasmussen, Niels Kristian Mäkinen Andersen, Simon Kragh Andersen, Jonas Emborg, Rasmus Røge, and Lasse Riis Østergaard. "Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides." *Cytometry Part A* 91, no. 8 (2017): 785-793.
[25] Rahman, Md Akizur, and Ravie Chandren Muniyandi. "Feature selection from colon cancer dataset for cancer classification using artificial neural network." *Int. J. Adv. Sci. Eng. Inf. Technol* 8, no. 4-2 (2018): 1387-1393.
[26] Choi, Seong Ji, Eun Sun Kim, and Kihwan Choi. "Prediction of the histology of colorectal neoplasm in white light colonoscopic images using deep learning algorithms." *Scientific Reports* 11, no. 1 (2021): 5311.
[27] Yao, Yao, Shuiping Gou, Ru Tian, Xiangrong Zhang, and Shuixiang He. "Automated Classification and Segmentation in Colorectal Images Based on Self-Paced Transfer Network." *BioMed Research International* 2021, no. 1 (2021): 6683931.
[28] 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.
[29] Global Dietary Database, Microdata Surveys, (2018). Retrieved March 2022 from https://www.globaldietarydatabase.org/management/micro data-surveys.
[30] 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.
[31] Inter-university Consortium for Political and Social Research, Find Data, (2022). Retrieved March 2022 from https://www.icpsr.umich.edu/web/ pages/.
[32] China Health and Nutrition Survey, China Health and Nutrition Survey, (2015). Retrieved March 2022 from https://www.cpc.unc.edu/projects/ china.
[33] 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-nutritionsurveys/canadian-community-health-survey-cchs.html.
[34] Data.world, Data.world, (2022). Retrieved March 2022 from https://ourwo rldindata.org.
[35] Ripley, Brian, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. "Package 'mass'." *Cran r* 538, no. 113-120 (2013): 822.
[36] Kursa, Miron B., and Witold R. Rudnicki. "Feature selection with the Boruta package." *Journal of statistical software* 36 (2010): 1-13.
[37] Zhao, Mingyuan, Chong Fu, Luping Ji, Ke Tang, and Mingtian Zhou. "Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes." *Expert Systems with Applications* 38, no. 5 (2011): 5197-5204.
[38] Dinov, Ivo D. "Data science and predictive analytics." *Cham, Switzerland: Springer* (2018).
[39] Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, and R. Core Team. "Package 'caret'." *The R Journal* 223, no. 7 (2020): 48.
[40] Nartowt, Bradley J., Gregory R. Hart, Wazir Muhammad, Ying Liang, Gigi F. Stark, and Jun Deng. "Robust machine learning for colorectal cancer risk prediction and stratification." *Frontiers in big Data* 3 (2020): 6.
[41] Hornbrook, Mark C., Ran Goshen, Eran Choman, Maureen O'Keeffe-Rosetti, Yaron Kinar, Elizabeth G. Liles, and Kristal C. Rust. "Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data." *Digestive diseases and sciences* 62, no. 10 (2017): 2719-2727.
[42] Gründner, Julian, Hans-Ulrich Prokosch, Michael Stürzl, Roland Croner, Jan Christoph, and Dennis Toddenroth. "Predicting clinical outcomes in colorectal cancer using machine learning." In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, pp. 101-105. IOS Press, 2018.

[43] Shiao, S. Pamela K., James Grayson, Amanda Lie, and Chong Ho Yu. "Personalized nutrition—genes, diet, and related interactive parameters as predictors of cancer in multiethnic colorectal cancer families." *Nutrients* 10, no. 6 (2018): 795.
[44] Hofseth, Lorne J., James R. Hebert, Anindya Chanda, Hexin Chen, Bryan L. Love, Maria M. Pena, E. Angela Murphy et al. "Early-onset colorectal cancer: initial clues and current views." *Nature reviews Gastroenterology & hepatology* 17, no. 6 (2020): 352-364.
[45] Tabung, Fred K., Lisa S. Brown, and Teresa T. Fung. "Dietary patterns and colorectal cancer risk: a review of 17 years of evidence (2000–2016)." *Current colorectal cancer reports* 13, no. 6 (2017): 440-454.
[46] Li, Tian, Chunqiu Zheng, Le Zhang, Ziyuan Zhou, and Rong Li. "Exploring the risk dietary factors for the colorectal cancer." In *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 570-573. IEEE, 2015.
[47] Zuhri, Mohammad AZ Abu, Mohammed Awad, Shahnaz Najjar, Nuha El Sharif, and Issa Ghrouz. "Colorectal cancer risk factor assessment in Palestine using machine learning models." *International Journal of Medical Engineering and Informatics* 16, no. 2 (2024): 126-138.
[48] Zheng, Ling, Elijah Eniola, and Jiacun Wang. "Machine learning for colorectal cancer risk prediction." In *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, pp. 1-6. IEEE, 2021.
[49] Jørgensen, Alex Skovsbo, Anders Munk Rasmussen, Niels Kristian Mäkinen Andersen, Simon Kragh Andersen, Jonas Emborg, Rasmus Røge, and Lasse Riis Østergaard. "Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides." *Cytometry Part A* 91, no. 8 (2017): 785-793.