Development of A Classification Framework for Colon Cancer

V.I. Ainoko, O.A. Odeniyi, O.Y. Ogunlola, I.A. Ainoko, B.K. Alese
Federal University of Technology Akure
Federal University of Technology Minna
Nigeria

Abstract

Colorectal cancer (CRC) is the third most commonly diagnosed cancer globally, causing significant public health concerns. Early screening and effective detection methods are crucial to improve survival rates and reduce mortality. However, accurately diagnosing colorectal cancer in a timely manner remains a challenge. Machine learning, a subset of artificial intelligence, emerging as a transformative tool in healthcare, particularly in disease classification and prediction tasks. The aim of this study is to design and implement A machine learning- based classification framework for aiding the prediction of colon cancer. The framework utilized decision trees, support vector machines, and random forests to accurately categorize patients into colorectal and healthy classes. Data from 362 patients were sourced from clinicaltrials.gov, identified as NCT030322874 and analyzed using an ensemble model after undergoing data preprocessing procedure. In addition dietary data were obtained from the American institute of cancer research as dietary information impacts the significantly in cancer risk and progression. The integration of data from clinicaltrials.gov and dietary information were combined into the data repository, standardized and populated to avoid overfitting. The results showed promising potential for machine learning to enhance early detection and facilitate targeted treatment strategies for Colorectal Cancer. This study highlights the significant role of machine learning in improving the diagnostic process for colorectal cancer. By employing advanced classification techniques and an ensemble modeling approach. Researchers can achieve more accurate predictions, ultimately aiding health practitioners in making informed decisions for patient care. The findings underscore the potential of integrating machine learning in clinical settings to enhance early detection, treatment efficacy, and patient outcomes.

1. Introduction

With the recent trends of sophisticated advancement sustainable research and durable computing solutions are essential. Thus, as we evolve from one century to another, computational advancement has proved to be a key factor for growth. These modern technologies and advancements have impacted the medical field, Education, government, entertainment, automated industries, banking, business, and so on. These trends can be traced to several sophisticated areas in Computer science; however, Machine Learning (ML) is a key factor in our research. Colorectal cancer disease is associated with the cells in the colon (large intestine), where the colon grows beyond control. This growth happens on the inner lining of the colon as little growths called polyps, which thereafter grow to become advanced malignant tumors that attack the human body. Colorectal cancer is recorded as the second death-related cancer associated with humans in the United States of America. The American Cancer Society estimated that one hundred fifty-three thousand and twenty (105020) will be diagnosed with colorectal cancer in America in 2023 [2]. Early detection impacts colorectal cancer metastasis; therefore two indications give likely tendencies to colon cancer; they include modifying risk or permanent risk, also called non-modifying risk. The modifying risk includes smoking, obesity, and eating habits. Excessive consumption of alcohol, animal fat, and red meat [15], [16], while the non-modifying risks are often genetic and hereditary-based. Machine learning has enabled the health sector to overcome healthcare challenges via the use of data science through machine learning models and data analysis. Providing ultimate improvement in healthcare delivery [11].

The focus of this study is to investigate and develop a model for colon cancer predictions using modifying and non- modifying risk datasets such as eating habits, bowel symptoms, stools, family history or records, and blood on the stools. The finding of the study journeys through the strength of the bagging ensemble technique and its application to the early detection of colon cancer reveals a narrative of innovation, collaboration, and a commitment to advancing the frontiers of knowledge.

2. Reviews of elated works

A comprehensive review of barriers and recommendations for colorectal cancer screening in

 | DOI: 10.20533/ijehis.2046.3332.2024.0039

Africa. The study objective was to review the current barriers and various recommendations used in the implementation of colorectal screening in African countries. The study deployed a keyword search on a PubMed publication; these keywords include colon cancer and screening, colorectal neoplasm, early detection, faecal immunochemical test (FIT), Africa, and colonoscopy [11]. The evaluation continued, and the full text was also passed to check its eligibility and relevance, so n equals 13. The final articles to be reviewed were about thirteen (13) articles whose authors were [1], [12] a combination of both interventional and non- interventional studies. In conclusion, the study highlighted the unavailability of endoscopic resources, limited capacity for colonoscopy, lack of awareness, lack of insurance or huge cost of testing, cultural issues in Africa, and lack of equipment facilities and infrastructure as barriers to colon cancer screening in Africa. In addition, the review recommended the ready availability of tests and educating more individuals who may not necessarily be health practitioners about colon care and screening would lower the cost of screening [11]. A machine learning-based colorectal cancer prediction using global dietary data was developed by [14]. The study's goal was to train datasets from a large sample of older and younger people. In addition, the study also provides an active, readily accessible health screening and prevention machine learning model to curb the excessive outbreak of colon cancer. The research collated dietary-related colorectal cancer data from 25 countries, such as Mexico, India, South Korea, the United States, Sweden, Argentina, Bangladesh, Bulgaria, Canada, China, Korea, Ecuador, Estonia, Ethiopia, Finland, Germany, Iran, Israel, Kenya, Malaysia, Mozambique, the Philippines, Portugal, Tanzania, Italy, and Japan. The original data contains several abnormalities, such as unbalanced values, missing values, and duplicates. The research methodology approach aims at developing a machine-learning-based model for predicting colon cancer [14]. The dataset was obtained from the original data from 25 countries, having about three million, five hundred twenty thousand, eight hundred sixty-six (3,520,866) samples of data. This data undergoes data processing using listwise detection. The listwise detection in turn gets random samples of 109,342 cases after analysis. The strength of the study lies in the large dataset extracted from several countries; however, the limitation of the study was recorded in the inability to predict the stage and type of colon cancer [14]. In addition, during the development phase, some features termed as not common were excluded, which could lead to data discrepancy. A colorectal cancer detection machine-learning model using conventional laboratory test data. The objective of the research work is to develop a model that would identify colorectal cancer using conventional test datasets extracted from patient records [10]. The methodology used captures electronic medical records (EMR) from patients who had been identified as healthy or colon cancerous patients from July 2017 to June 2018. A collection of one thousand one hundred and sixty- four (1164) medical reports was evaluated as a dataset with a label of 582 colorectal cancer patients and 582 healthy individuals. The machine learning algorithms deployed include logistic regression (LR), random forest (RF), K- nearest neighbor (KNN), support vector machine (SVM), and Naïve Bayes (NN). The extracted medical records from the electronic medical records (EMR) include demographics, tumor biomarkers, hospitalization information, vital signs, liver enzymes, complete blood counts, lipid profiles, diagnosis and treatments for each patient [10]. In conclusion, the study analyzed the comparison of the aforementioned models, and it was estimated that logistic regression had the best performance evaluation on the dataset collected, however, the limitation of the study lies in the selection criteria phase [10]. Patients were selected, and to this end, not all patients would participate in this cohort thereby limiting the generalization ability. Robust machine-learning model for stratifying and predicting colorectal cancer risk [13]. The study's goal was to find a useful screening technique for estimating an individual's risk of colorectal cancer based on their personal health information, utilizing seven algorithms. The methodology employed seven algorithms, which include logistic regression, decision trees, random forests, naive Bayes, support vector machines, and linear discriminant analysis [13]. Furthermore, the research work combines the learning algorithms in combination with six imputation methods for missing data, all trained and cross-tested with the NHIS and PLCO datasets. Among various machine learning algorithms using different imputation methods, the artificial neural network with Gaussian expectation- maximization imputation was found to be optimal, with a concordance of 0.70 ± 0.02, a sensitivity of 0.63 ± 0.06, and a specificity of 0.82 ± 0.04. In CRC risk stratification, this optimal model had a never-cancer misclassification rate of only 2% and a CRC misclassification rate of only 6%. The above results indicate that a trained artificial neural network can be used as an effective screening tool for early intervention and prevention of CRC in large populations [13]. In conclusion, the study is a TRIPOD level 3 where NHIS/PLCO datasets were used for the training sets while PLCO/ NHIS datasets were for the testing set however, the limitation of the study is seen in its dataset such that only personal health data were used. A prognosis random forest model for colorectal cancer patients. The motivation of the study is to determine the factors that affect colon cancer patients' survival in the city of Makassar and Indonesia [3], [4]. The objective of the research

 | DOI: 10.20533/ijehis.2046.3332.2024.0039

work is to collect data on the history of patients with colorectal cancer to determine the future survival rate of the patient through an ensemble technique. The methodology approach used the random forest, an algorithm used for data classification. The random forest technique merges two or more decision trees through training on sample data. This is also called an ensemble method because it is a combination of several decision trees as classifiers [3]. Data was collected first from colon cancer patients diagnosed in 2012 from four hospitals in Makassar City, and their survival was observed up until 2015 [3]. The predictor variables include comorbidity, age, stage of cancer, the location of cancer, treatment status, metastasis history of patients with colorectal cancer, and sex. The samples used in this research work were a combination of thirty-eight (38) colon cancer patients. The cancer stage is categorized into early and advanced stages; the comorbidity is either heavy or light; and the treatment status is either complete or incomplete. Complete treatment implies that either surgery, chemotherapy, or radiation has been carried out as recommended by the doctor with a medical record, while incomplete treatment implies that the patient has yet to adhere to the recommendation as prescribed by the doctor. The location of cancer in the research is categorized as the colon or rectum, and the metastatic history is either the development of primary tumor cells or no primary tumor cell development found in other organs. The random forest collects the data randomly through a bagging technique to generate a classification tree. By sampling the data, the variables for classification include a history of metastasis with colorectal cancer, the location of the cancer, gender, and comorbidity. The results of the study show that the analysis using a random forest model produced a classification accuracy of fifty per cent (50%) on the Area under Curve (AUC). The area under the curve is an essential metric used for identifying the accuracy of a random forest model. In conclusion, a random forest algorithm could predict the survival rate of a colorectal cancer patient, however, the study limitation lies in the accuracy of the analysis and its small data samples [3]. Machine-learning model using gender, complete blood count data, and age for the early detection of colon cancer. The objective of the study was to detect colon cancer in the US-based adult community based on their demographics and their results from laboratory tests [9]. The methodology extracted datasets from colonoscopy procedures from 2000 to 2003, while demographic data and a complete body count were extracted from 1998 to 2003 from the Kaiser Permanente Northwest Region Tumor Registry [9]. The model detection score requires at least gender, year of birth, and a combination of a complete blood count. These complete blood counts are either red blood count (RBC), Hemoglobin (HGB), hematocrit (HCT) or Red blood count (RBC), Hematocrit (HCT), mean corpuscular hemoglobin (MCH) or Red blood count (RBC), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC) or hemoglobin (HGB), Hematocrit (HCT), mean corpuscular hemoglobin (MCH) or hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC) or Hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC).The dataset consists of seventeen thousand and ninety-five (17095) patients. Nine hundred (900) patients had a label of colorectal cancer, of which 439 were female and 461 were male, while the healthy patients without colon cancer were about 9108 females and 7087 males [3].

The decision tree and statistical method were used to model the ColonFlag. The evaluation of the model was estimated using the area under curve (AUC). The AUC of KPNWN data for both combined genders is estimated at 0.81, while age alone is estimated at 0.73. The area under the receiving characteristic curve (AUROC) for the ColonFlag from the Maccabi Healthcare Services (MHS) Israel data performed better at 0.87 compared to the National Health Service (NHS) data [3]. In conclusion, the strength of the study includes using electronic medical record data that has a substantial number of colorectal cancer cases with a sizable control group and employing a machine learning detection algorithm while the limitation lies in the inability of the model to recognize patients without medical health records.

3. Methodology

The aims of the study are to develop and implement an ensemble framework for colon cancer predictions using modifying and non-modifying risk datasets such as eating habits, bowel symptoms, stool, family history or records, and blood on the stools. The study underwent the following for the development of the framework as shown in Figure 1. The experimental setup was carried out sequentially in distinct phases to achieve the goals of this study:

i. The first phase describes the collection of the dataset for both the training and testing model. The research project used a cross-sectional dataset acquired from clinicaltrials.gov as NCT030322874 containing 362 patients (a). These patients were assessed by colonoscopy and questionnaire. The study was conducted around January 2014 to July 2016 in Nigeria from three hospitals which are Obafemi Awolowo University Teaching Hospital (OAUTHC), Ile-Ife Osun State, University College Hospital (UCH), and the University of Ilorin Teaching Hospital, Kwara State (UITH) and dietary data from American Institutes of Cancer.

 | DOI: 10.20533/ijehis.2046.3332.2024.0039

ii. The second phase involves data preprocessing and feature extraction. The preprocessing stage involves tokenization, lowercasing, stopword removal, data augmentation, encoding, handling of missing values, feature extraction and splitting the dataset into training and testing sets using the ratio 75/25 resulting in 3750 symptomatic colon data in the training set and 1250 in the testing sets after the augmenting the 362 dataset to five thousand patients datasets.

iii. The third phase involves the model-building phase using the bagging ensemble technique. Three algorithms would be deployed, they are Support Vector Machine, decision tree and logistic regression base model for the bagging hybrid machine learning model.

iv. The last phase involves the evaluation of the model using standard evaluation metrics.

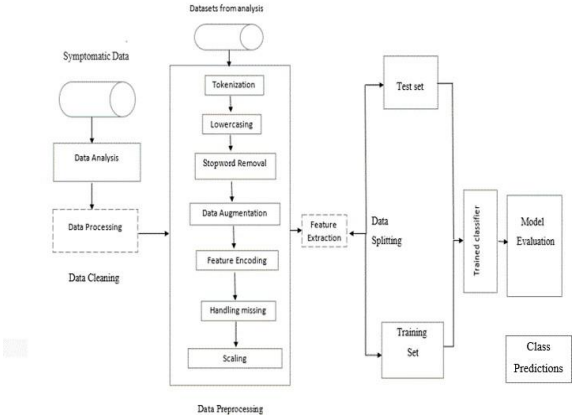

Figure 1. System Architectural Design for the Predictive Colon Cancer System

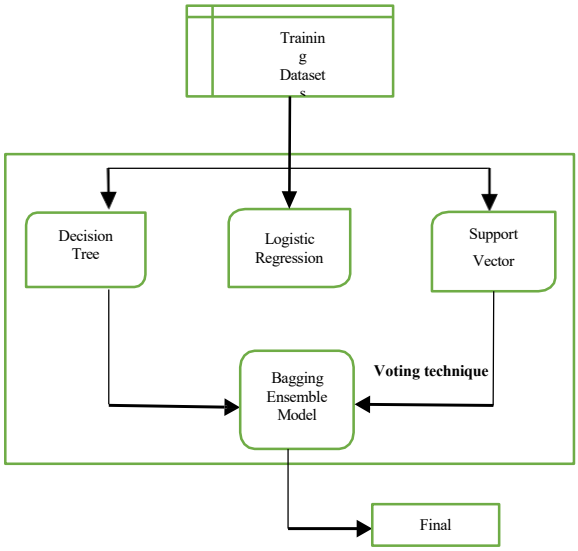

Figure 2. Description of the Bagging Ensemble for the Predictive Colon Cancer System

3.1. Bagging Ensemble Architecture

Bagging is an ensemble learning technique used for hybrid machine learning, the bagging ensemble method is deployed to improve accuracy and enable robustness for models. This model is a combination of two or more models usually called base model that aims to improve overall predictive performances. In addition, the bootstrap aggregate involves the random selection of a subset of a dataset that helps to prevent overfitting and the aggregation of their output results [5]. The aggregating technique for different tasks depends largely on the type of problem, usually, machine learning tasks could be either a regression or classification task [6], [7]. The regression tasks usually deploy the averaging method while the classification problems use the majority vote methods. The architecture of the Bagging model consists of two or more base classifier models that combine to make predictions.

a) Individual Machine Learning Classifier

The individual machine learning algorithms used to develop the hybrid machine learning classifier are known as the ensemble method. The following algorithms represent the base learners for the research work:

i. Support Vector Machine
ii. Logistic Regression
iii. Decision Tree

b) Hybrid Machine Learning Classifier Using Bagging

In essence, the bagging model merges two or more base learners with the ultimate objective of enhancing overall predictive capabilities, in order to reduce overfitting. The key component of bootstrap aggregation is the random selection of a dataset subset, which is followed by the aggregate of each output result. Two methods are widely used to estimate the output of the bagging method. They include aggregating averaging in regression and using majority voting in classification. The output of the bagging ensemble is frequently estimated in regression tasks by averaging the prediction given by each base model. This averaging can produce more precise results by lowering the variations of prediction while majority voting is usually utilized in classification problems where the prediction is made by choosing the class that receives the most base model votes, this method reduces the risk of overfitting while classification accuracy is improved. The base model used for the research includes the random forest, support vector machine and logistics regression. The idea involves that the predictions are aggregated using majority voting [5], [6].

 | DOI: 10.20533/ijehis.2046.3332.2024.0039

i. Logistic Regression Model

Logistic regression is used for binary classification problems; it uses the sigmoid function to model its probability. It is denoted as;

$$Sigmoid\ (z) = \frac{1}{(1+ (-z))} \quad 1.0$$

where:
$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
$(Y = 1|X)$ Equals the probability that the dependent variable Y is positive in class one (1) given that X is the predictor variables
e = The base logarithm
β_1, β_2 = The coefficient associated with the predictor variable
X_1, X^2 = The values of the predictor values
$(Y = 1|X) = 1 / (1 + e\text{^} - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k))$1.1

ii. Decision Tree Model

The decision tree algorithm is a supervised Machine Learning algorithm that builds a tree-like model based on data features for classification or prediction, with each decision splitting data into subsets and leaf nodes representing class labels.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

b) Augmented datasets

$$H(S) = -\sum C\ PxlogPx \quad 1.2$$

where,

S - Set of all instances,
N - Number of distinct class values,
Px - Event probability,
C – Number of classes

iii. Support Vector Machine

The SVM classifier model predicts the class of a new instance x by simply computing the decision function

$wT \cdot x + b = w1\ x1 + \cdots + wn\ xn + b$:
It can be denoted mathematically as: $(x) = w * x + b$1.3
where
w = weight vector
T = transpose
b = bias
x = symbolizes the training examples closest to the hyperplane

Finally, the Support vector machine deploys different kernel functions depending on their tasks

3.2. Dataset Description

This describes the dataset used for the model. It was obtained from clinicaltrials.gov and global dietary data. The dataset from clinicaltrials.gov contains 362 patients 66 patients had colorectal cancer while 296 patients didn't have colorectal cancer.

Table 1. Summary of the Dataset

a) Initial Dataset

Dataset (n=362)	Training set 75% (n =217)	Test set 25% (n=145)
Yes	38	28
No	179	117

b) Augmented datasets

Dataset (n=5000)	Training set 75% (n =3750)	Test set 25% (n=1250)
Yes	656	241
No	3094	1009

3.3. Analysis of the base learner(s) and the ensemble

Table 2: Accuracy of the models

Model	**Accuracy (%)**
SVM	94%
LR	85%
DT	84%
Bagged Ensemble	98%

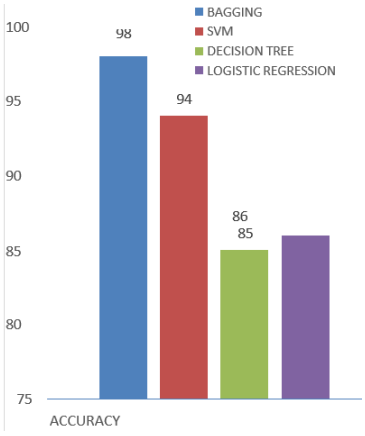

Figure 3. Model Performance Based On Accuracy

 | DOI: 10.20533/ijehis.2046.3332.2024.0039

The comparative analysis chart represents the result obtained from the base models and the ensemble. It was denoted that the bagging methods perform better.

4. Conclusion

The bagging ensemble technique for the hybrid machine learning approach is best for classification problems. In this study, the research attained a high sensitivity, specificity and accuracy.

The dataset originally derived was a combination of 362 patients alongside eating habits, these datasets were further augmented into five thousand (5,000) patients. This patient's data were further broken down into 3750 for the training sets (75%) and 1250 testing set (25%). The preprocessing phase played a pivotal role in refining the datasets. Employing space-based tokenization, the study utilized the word tokenize function from the Natural Language Toolkit (NLTK), complemented by lowercasing to standardize the textual data. Stop word removal, a key data preprocessing technique focused on eliminating punctuations and common words, further enhanced the quality of the dataset. Following the augmentation to five thousand samples, the dataset underwent encoding, with missing values completely addressed using Filna and linear interpolation. The subsequent feature extraction phase leveraged the use of principal component analysis (PCA) for dimensional reduction, facilitating a streamlined representation of the dataset's features while preserving their essential characteristics. When subjected to comparison with the work of Rahman et al. [14], the predictive model in this study exhibited a misclassification rate of 0.01, significantly outperforming the 3% misclassification rate reported for the colon cancer class. Notably, the dataset employed encompassed both non-modifying and modifying indicators, underscoring the versatility and comprehensiveness of the predictive learning model system. Beyond its statistical achievements, the model demonstrated practical utility by potentially assisting medical practitioners in the early detection of colon cancer, showcasing the transformative impact of modern computerized innovations when strategically applied to healthcare challenges

4.1. Recommendation

The research work developed a model built on a study cohort from low and middle-income countries as NCT030322874, whose datasets were extracted, and dietary data were collected from the American institutes of cancer. In future research, more options such as colon treatments and prognosis would be explored. In addition, Future research would endeavors to explore additional options, such as colon treatments and prognosis, thereby expanding the scope and applicability of the developed model. Furthermore, the study commits to deploy the model into a more

human interacting system, a crucial step in ensuring its practical usefulness in the medical field and promoting wider acceptance and adoption.

4.2. Contribution to Knowledge

The study established an ensemble Machine Learning Model for the early detection of colon cancer to assist medical health practitioners. In conclusion, the research work journeys through the strength of the bagging ensemble technique and its application to the early detection of colon cancer reveals a narrative of innovation, collaboration, and a commitment to advancing the frontiers of knowledge.

5. References

[1] Alatise, O.I., Ayandipo, O.O., Adeyeye, A., Seier, K., Komolafe, A.O., Bojuwoye, M.O., Afuwape, O.O., Zauber, A., Omisore, A., Olatoke, S., Akere, A., Famurewa, O., Gonen, M., Irabor, D.O., & Kingham, T.P. (2018). A symptom-based model to predict colorectal cancer in low-resource countries: Results from a prospective study of patients at high risk for colorectal cancer. Cancer, 124(13), 2766–2773. DOI: 10.1002/cncr.31399.

[2] American Cancer Society. Colorectal Cancer Facts & Figures 2020-2022. Atlanta, Ga: American Cancer Society; 2020.

[3] Anuraga, G., Fernanda, J.W., and Pebrianty. (2019, May 1). Random forest prognostic factor in colorectal cancer. Journal of Physics: Conference Series, 1217(1), 012098. DOI: 10.1088/1742-6596/1217/1/012098.

[4] Araghi M, Arnold M, Rutherford MJ, et al Colon and rectal cancer survival in seven high-income countries 2010–2014: variation by age and stage at diagnosis (the ICBP SURVMARK-2 project) Gut 2021;70:114-126.World Cancer Research Fund, American Institute for Cancer Research. Summary: food nutrition and the prevention of cancer: a global perspective. Washington, DC: AICR; 1997.

[5] Breiman, L. Bagging predictors. Machine Learning 24, 123–140(1996). DOI: 10.1007/BF00058655.

[6] Campbell, M., Hoane, A., & Hsu, F H. (2002). Deep Blue. Artificial Intelligence, 134(1–2), 57–83. Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. Multiple Classifier Systems,1–15. DOI: 10.1007/3-540-45014-91.

[7] Donmez, P. (2012). Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press 2010. ISBN: 978-0-262-01243-0.584. Natural Language Engineering, 19(2), 285– 288. DOI: 10.1017/s1351324912000290.

[8] Flood, A. (2003). Meat, Fat, and Their Subtypes as Risk Factors for Colorectal Cancer in A Prospective Cohort of Women. American Journal of Epidemiology, 158(1), 59–68

DOI: 10.1093/aje/kwg099.

[9] Hornbrook, M.C., Goshen, R., Choman, E.,O'Keeffe-Rosetti, M., Kinar, Y., Liles, E.G., & Rust, K.C. (2017). Correction to: Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. Digestive Diseases and Sciences, 63(1), 270–270. DOI: 10.1007/s10620-017-4859-5.

[10] Hu, F., Wang, J., Zhang, M., Wang, S., Zhao, L., Yang, H., Wu, J., & Cui, B. (2021). Comprehensive Analysis of Subtype-Specific Molecular Characteristics of Colon Cancer: Specific Genes, Driver Genes, Signaling Pathways, and Immunotherapy Responses. Frontiers in Cell and Developmental Biology, 9. DOI: 10.3389/fcell.2021.758776.

[11] R Holmes, D. (2023). Barriers and recommendations for colorectal cancer screening in Africa. Global Health Action,16(1). DOI: 0.1080/16549716.2023.2181920.

[12] Nartowt, B.J., Hart, G.R., Muhammad, W., Liang, Y., Stark, G F., & Deng, J. (2020a). Robust machine learning for colorectal cancer risk prediction and stratification. Frontiers in Big data DOI: 10.3389/fdata.2020.00006.

[13] Nartowt, B J., Hart, G R., Muhammad, W., Liang, Y., Stark, G.F., & Deng, J. (2020a). Robust machine learning for colorectal cancer risk prediction and stratification. Frontiers in Big Data. DOI: 10.3389/fdata.2020.00006.

[14] Rahman, H.A., Ottom, M.A.,and Dinov, I.D. (2022) .Machine Learning-based Colorectal Cancer Prediction using Global Dietary Data. Research Square (Research Square); Research Square (United States). DOI: 10.21203/rs.3.rs-2031672/v1.

[15] Rawla, P., Sunkara, T.,and Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Gastroenterology Review, 14(2), 89–103. DOI: 10.5114/pg.2018.81072.

[16] Robertson, R. (2006). Predicting colorectal cancer risk in patients with rectal bleeding. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1920716/ (Access Date: 4 April 2024).

Further Readings

Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians,69(1),7–34. DOI: 10.3322/caac.21551.

Witten, I.H. and Frank, E. (2005) Data mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann Publisher, Burlington.

Zhang, B. (2019). Colorectal cancer predictive test using germ-line DNA data and multiple machine learning methods. UC Irvine. ProQuesID:Zhang_uci_0030D_15657. MerrittID:ark:/1303 0/m5s80hqw. https://escholarship.org/uc/item/44f3f487 (Access Date: 8 May 2024).

