A Machine Learning Framework for Personalized Lifestyle Recommendations in Colorectal Cancer Prevention

Christos Androutsos[1], Traianos Tsiokris[1], Zheshen Jiang[2], Nicolas Gillain[2], Ioannis S. Papanikolaou[3], Eleni Koukoulioti[3], Constantina Cloconi[4], Antria Savva[4], Sisse H. Njor[5], Susanne F. Jørgensen[5], Maja Ravnik[6], Sergej Černčič[6], María González Oter[7,8], Raquel Alcaraz Ortega[8], Vasilis Giannakopoulos[9], Dimitrios Kypreos[9], Dimitrios Dimitroulopoulos[9], George K. Matsopoulos[10], Dimitrios I. Fotiadis[1,11]

[1] Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
[2] Department of Information System Management, Hospital Center of University of Liège, Liège, Belgium
[3] Hepatogastroenterology Unit, Second Department of Internal Medicine, National and Kapodistrian University, Attikon University General Hospital, Athens, Greece
[4] Radiation Oncology Center, German Oncology Center
[5] Lillebaelt Hospital, Research Unit for Screening and Epidemiology, Department of Biochemistry and Immunology, Denmark University of Southern Denmark, Department of Regional Health Research, Denmark.
[6] Department of Oncology, University Medical Centre Maribor, Maribor, Slovenia
[7] Cancer Genetics Group, Unit of Excellence Institute of Biomedicine and Molecular Genetics, University of Valladolid Spanish National Research Council (IBGM; UVa-CSIC), 47003 Valladolid, Spain
[8] Unidad de Investigación, Hospital Universitario de Burgos, Burgos, España
[9] Agios Savvas, Cancer Hospital, Athens, Greece
[10] Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technological University of Athens, Athens, Greece
[11] Biomedical Research Institute, FORTH, Ioannina, Greece

***Abstract*—Colorectal cancer (CRC) is a largely preventable disease influenced by modifiable behavioral risk factors such as diet, smoking, alcohol consumption, physical inactivity, and chronic stress. This study proposes a machine learning–based framework that generates personalized lifestyle recommendations aimed at CRC prevention. The system consists of two components: the Behavioral Recommendation Mapping Engine, which maps behavioral questionnaire responses to expert-validated recommendations, and the Risk Assessment Module, which classifies participants into specific recommendations using supervised learning models. Eight domain-specific classifiers were developed, each targeting a key behavioral risk factor. Random Forests consistently outperformed Decision Trees, achieving high macro-averaged F1 scores even in imbalanced categories such as smoking (F1 = 0.88) and stress (F1 = 0.80). The system also identifies the most influential behavioral variable per domain to highlight actionable risk factors. This framework will be integrated into the DIOPTRA mobile application to support real-time, personalized prevention.**

Funded by the European Union (DIOPTRA, 101096649). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI). Funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10056682].

***Index Terms*—Colorectal cancer prevention, personalized recommendations, machine learning, behavioral profiling, risk assessment**

I. Introduction

Colorectal cancer (CRC) remains a leading cause of cancer-related morbidity and mortality worldwide [1]. Although largely preventable through early screening and behavioral changes, CRC incidence continues to rise, particularly among individuals under the age of 50 [2]. Modifiable behavioral factors such as smoking, physical activity, alcohol consumption, diet, and chronic stress significantly contribute to CRC behavioral risk [3]. Addressing these factors with personalized, evidence-based interventions is essential in prevention. Recent advancements in Artificial Intelligence (AI) have enabled the development of systems that analyze behavioral data to generate personalized health recommendations. These systems offer scalable solutions to support preventive care by tailoring recommendations to a person's unique risk profile. In the context of CRC, while ML has been applied extensively to cancer risk prediction and clinical decision support, few studies have addressed the generation of explicit, domain-specific behavioral recommendations. Most existing approaches either focus

on estimating disease risk or use simulations to model potential outcomes of lifestyle changes, without directly classifying or validating recommendation outputs. The present study aims to bridge this gap by developing and evaluating a modular ML framework that generates personalized CRC-related lifestyle recommendations based on behavioral profiles.

Herrera *et al.* [4] developed and internally validated an interpretable ML model for CRC risk prediction based on easily obtainable lifestyle and clinical factors. Utilizing data from 154,887 older adults who participated in the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial (including behavioral variables like smoking and weight, and medical history), a LightGradient Boosting Machine (LightGBM) classifier was developed to estimate an individual's probability of developing CRC. The final model incorporated 12 predictors and age, body weight, and smoking history emerged as the strongest risk factors, while use of heart medication showed a slight protective effect. The LightGBM model achieved a moderate discriminative performance with an area under the receiver operating characteristic curve of approximately 0.73 in internal validation. Beyond risk scoring, the model's output was structured to stratify individuals into average, increased, or high-risk categories, accompanied by measures that highlighted modifiable lifestyle contributors to each individual's risk profile. This feature enables the model to support targeted clinician–patient discussions and facilitates personalized behavioral recommendations.

Dogan *et al.* [5] proposed a novel AI-driven framework for generating personalized lifestyle recommendations aimed at cardiovascular disease (CVD) prevention. The system was designed to optimize behavioral interventions such as diet or exercise changes by simulating their projected impact on individual risk profiles. Using a publicly available dataset of clinical and lifestyle risk factors, the authors implemented a three-component pipeline. First, a supervised classification model predicted an individual's baseline CVD risk. Second, a generative adversarial network (GAN) was trained to simulate how hypothetical modifications to one or more risk factors such as reducing sodium intake or increasing physical activity would affect the individual's overall risk profile and third, a personalized utility function evaluated the trade-off between the expected risk reduction and the perceived effort or cost associated with each lifestyle change. This enabled the system to identify and recommend the optimal intervention for each person. Validation results demonstrated that the proposed system could successfully generate individualized lifestyle modifications that meaningfully reduced predicted CVD risk. Although no external labels or ground-truth recommendations were available for direct comparison, the effectiveness of the generated lifestyle changes was validated through simulation, demonstrating meaningful reductions in predicted CVD risk across individuals.

This paper presents a framework for generating and validating personalized CRC prevention recommendations through two interconnected components: a Behavioral Recommendation Mapping Engine (BRME) and a Risk Assessment Module (RAM). The BRME maps behavioral questionnaire responses to evidence-based recommendations sourced from authoritative guidelines utilized as annotations for ML models. The RAM then predicts these recommendations using classifiers trained on the annotated dataset. The objective of this study is to compare the performance of various classifiers across behavioral categories and to demonstrate the feasibility of a modular, ML-driven system for delivering CRC-specific lifestyle recommendations. The outcome of this work will inform the integration of these modules into the DIOPTRA mobile application, contributing to the broader goal of personalized CRC prevention.

II. Materials and Methods

A. Dataset

The dataset utilized in this study comprises exclusively behavioral data collected through a structured questionnaire developed within the framework of the DIOPTRA prospective study. Participants completed the questionnaire during their colonoscopy screening across multiple clinical sites. The questionnaire was designed to capture a wide range of lifestyle factors associated with CRC prevention, with a particular focus on modifiable behaviors. Each participant was categorized into one of four predefined health status groups: Healthy, Non-Advanced Adenomas (NAA), Advanced Adenomas (AA), and CRC patients. These groupings were not employed as input features or target variables in the development of the ML models. The emphasis of this study remains solely on behavioral patterns and the generation of personalized lifestyle recommendations.

The behavioral questionnaire covered multiple thematic categories, such as smoking status, alcohol consumption, physical activity, diet, supplement usage, stress levels and sociodemographic background. It gathers information on smoking habits, and exposure to secondhand smoke, alcohol intake frequency and volume, physical activity levels and sedentary behavior, as well as dietary habits, including the consumption of fruits, vegetables, whole grains, processed meats, sugary products, and fast food. Additionally, it records the frequency of dietary supplement use such as multivitamins, probiotics, omega-3, calcium, and vitamin D and includes stress-related questions adapted from the perceived stress scale. Socioeconomic indicators, including income perception, education level, and employment status, are also captured.

After excluding incomplete records, the final dataset comprised 756 fully completed records, each containing 45 structured features. The annotations corresponding to each participant's profile are described in detail in Section II.B.

B. Behavioral Recommendation Mapping Engine

The BRME is the intermediate layer between raw behavioral data and ML model training. Its primary function is to map questionnaire responses into structured, personalized lifestyle recommendations based on internationally accepted CRC prevention guidelines. These recommendations are the annotation labels that guide the supervised learning process of the RAM.

The mapping process was developed by integrating public health guidelines from organizations such as the World Cancer Research Fund, the National Comprehensive Cancer Network, the National Cancer Institute, and the American Institute for Cancer Research. Each of these institutions provides evidence-based recommendations related to diet, physical activity, alcohol consumption, smoking cessation, stress management, and supplement use, categories strongly associated with CRC behavioral risk modulation.

To implement the annotation process, a rule-based logic was developed that translated specific patterns of responses from the behavioral questionnaire into discrete, guideline-aligned recommendations. Each rule corresponded to a specific combination of responses and was defined in close collaboration with clinical experts to ensure medical validity. Recommendations were generated independently for each behavioral category captured in the questionnaire. For each participant one recommendation per category was assigned, resulting in a structured, multi-label annotation format. For example, in the physical activity category, participants who reported exercising daily or several times a week were mapped to a recommendation such as "Congratulations on staying physically active! Your commitment to maintaining an active lifestyle is commendable. However, to further enhance your health, consider reducing your sedentary time. Physical Activity convincingly protects against CRC and balancing movement with less sitting time can maximize its benefits." This approach allowed each behavioral category to be treated as an individual classification task.

C. Risk Assessment Module

The RAM constitutes the ML component of the system and is designed to predict personalized behavioral recommendations based on participants' questionnaire responses. Its primary objective is to enable automated recommendation delivery by learning the mapping between individual behavioral profiles and the expert-validated annotations generated by the BRME. By modeling this relationship, the RAM is the core of the AI-based decision support system, which will be integrated into the DIOPTRA mobile application.

Unlike conventional classification systems that aim to detect disease or estimate risk scores, the RAM is focused on recommendation generation. It classifies a specific behavioral recommendation for each lifestyle category, based on participants' responses. To achieve this, the RAM comprises eight independent classifiers, each dedicated to the categories of the behavioral questionnaire. Although a multi-label or multi-output classification strategy could theoretically address all eight behavioral categories simultaneously, this approach would result in extreme label sparsity and reduced performance due to the high number of possible label combinations. To avoid this, the classification task was decomposed into eight independent problems. This modular design allows each model's prediction to be directly attributed to category-specific features.

Tree-based models such as Decision Trees (DTs) and Random Forests (RF) were employed for classification due to their robustness, and effectiveness on structured questionnaire data. The model development process followed a consistent pipeline across all categories. Categorical variables were encoded utilizing either ordinal or one-hot encoding and continuous variables were normalized when necessary. The dataset was partitioned utilizing a stratified 70%-30% train-test split to preserve class distribution. To evaluate performance robustness, five-fold cross-validation was conducted. Hyperparameter tuning was performed separately for each classifier. For tree-based models, optimization included parameters such as maximum tree depth, number of estimators and minimum sample thresholds for splitting. Final model selection was based on a combined assessment of classification accuracy and F1-score.

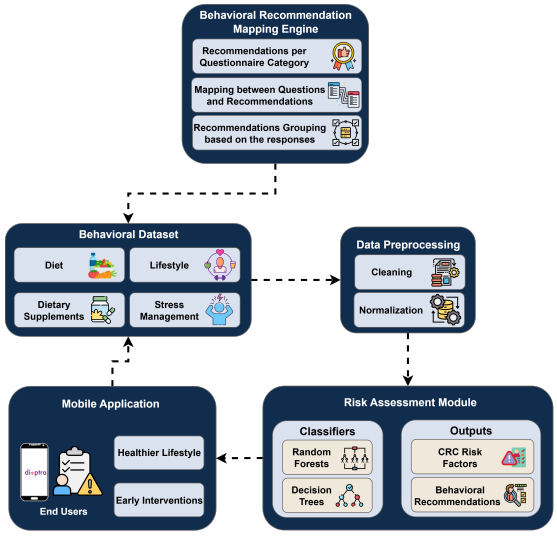

Fig. 1. Overview of the proposed ML–based framework.

Beyond classification, the RAM also supports the identification of key behavioral risk factors within each behavioral category. For every trained model, feature importance scores were utilized to determine which variables had the greatest influence on the predicted recommendation. The most influential feature for each category was identified to highlight the primary modifiable behavior contributing to the recommended action.

III. Results

The performance of the RF and DTs classifiers was evaluated across eight behavioral categories using five-fold cross-validation. RF was selected as the primary model due to its overall strong predictive accuracy, and capacity to handle multi-class classification tasks effectively. Although DTs models also demonstrated acceptable performance, RF consistently outperformed DTs and produced more stable predictions, particularly under conditions of class imbalance. Class imbalance was expected in several categories due to the limited dataset size and the he inherent variability in real-world behavioral

patterns. To mitigate this, macro-averaged F1 score was reported alongside classification accuracy. Unlike accuracy, macro F1 score assigns equal weight to each class, and provides a more balanced measure, particularly important when the accurate prediction of minority class recommendations carries significant health implications. The testing accuracy and macro-averaged F1 scores for both classifiers across each category are presented in Table I, along with the number of classes per category. Within the diet-related categories, RF achieved high performance in sugar intake (Accuracy = 0.87, F1 = 0.82) and meat consumption (F1 = 0.66), both of which exhibited relatively balanced label distributions. However, performance in the fruits and vegetables category was slightly lower (F1 = 0.67), likely reflecting class imbalance.

TABLE I
COMPARISON OF RF AND DTS CLASSIFIERS BY BEHAVIORAL CATEGORY.

Category	# Classes	RF Acc. / F1	DTs Acc. / F1
Diet			
Sugar Intake	6	**0.87** / 0.82	0.87 / **0.84**
Fruits & Vegetables	9	**0.75** / **0.67**	0.73 / 0.66
Meat Consumption	6	**0.78** / **0.66**	0.78 / 0.66
Lifestyle			
Physical Activity	7	**0.69** / **0.74**	0.68 / 0.72
Smoking Status	6	**0.97** / **0.88**	0.97 / 0.77
Alcohol Consumption	9	**0.73** / **0.68**	0.73 / 0.66
Dietary Supplements			
Supplements Use	6	0.83 / **0.64**	**0.86** / 0.63
Stress Management			
Stress Level	3	**0.76** / **0.80**	0.74 / 0.78

In the lifestyle-related categories, smoking status achieved particularly strong results (F1 = 0.88) demonstrating the model's ability to generalize well even in imbalanced conditions. Physical activity and alcohol consumption achieved moderate performance levels (F1 = 0.74 and 0.68, respectively), with variability driven by differences in label balance. Similarly, RF performed well in the dietary supplements (F1 = 0.64) and stress management categories (F1 = 0.80), both of which were affected by skewed class distributions.

DISCUSSION

The development of data-driven systems capable of delivering tailored health recommendations based on individual behavior profiles represents an important step toward personalized prevention in CRC. In this study, this challenge was addressed by implementing a classification framework that predicts domain-specific lifestyle recommendations utilizing structured behavioral data. The combination of the BRME and the RAM enabled the transformation of questionnaire responses into targeted, guideline-aligned advice. RF classifier achieved robust performance across eight behavioral categories, maintaining predictive strength even in the presence of class imbalance.

Compared to existing literature, this work advances beyond traditional CRC risk estimation models, such as those by Herrera *et al* [4], which focus on predicting disease probability. While such models are valuable for screening prioritization, they typically do not provide concrete, individualized behavioral recommendations. Similarly, simulation-based frameworks like Dogan *et al* [5] demonstrate promising methods for modeling behavioral change impact, but do not evaluate recommendation outputs directly as predictive targets. In contrast, the proposed RAM explicitly learns to predict domain-specific recommendations, allowing for direct evaluation through classification metrics.

Despite these promising results, the study has several limitations. First, the dataset used, while diverse in behavioral domains, remains limited both in terms of class distribution and in number of participants. Several categories exhibited class imbalance, which likely impacted F1 scores for underrepresented recommendations. Although techniques such as stratified sampling and macro-averaged F1 scoring were applied to address this, performance in rare classes remains an area for improvement. Second, while recommendations were derived from expert-reviewed guidelines, the current annotation logic does not cover the full spectrum of possible behavioral variations, meaning that some recommendation classes are underrepresented or absent. Future work will include the collection of more prospective data to represent more behavioral patterns and the inclusion of longitudinal data to assess behavioral change over time. Finally, to scale the framework to larger and more imbalanced datasets, imbalance handling methods will be explored.

CONCLUSIONS

This study presented a ML–based system for generating personalized lifestyle recommendations to support CRC prevention. By combining expert-guided annotation with domain-specific classifiers, the proposed framework offers a targeted approach to behavioral risk assessment. The modules developed here will be integrated into the DIOPTRA mobile application, enabling real-time, personalized guidance.

REFERENCES

[1] J. Li, J. Pan, L. Wang, G. Ji, and Y. Dang, "Colorectal Cancer: Pathogenesis and Targeted Therapy," MedComm, vol. 6, pp. e70127, March 2025.

[2] F. Karam, Y. E. Deghel, R. Iratni, A. H. Dakroub, and A. H. Eid, "The Gut Microbiome and Colorectal Cancer: An Integrative Review of the Underlying Mechanisms," Cell Biochemistry and Biophysics, vol 5, pp. 1–14, February 2025.

[3] V. V. Tsukanov, A. V. Vasyutin, and J. L. Tonkikh, "Risk factors, prevention and screening of colorectal cancer: A rising problem," World Journal of Gastroenterology, vol 31, pp. 98629, February 2025.

[4] D. J. Herrera, D. M. Seibert, K. Feyen, M. van Loo, G. van Hal and W. van de Veerdonk "Development and Internal Validation of a Machine Learning-Based Colorectal Cancer Risk Prediction Model," Gastrointestinal Disorders, vol 7, pp. 26, January 2025.

[5] A. Dogan, Y. Li, C. P. Odo, K. Sonawane, Y. Lin and C. Liu "A utility-based machine learning-driven personalized lifestyle recommendation for cardiovascular disease prevention," Journal of Biomedical Informatics, vol 141, pp. 104342, May 2023.