Sciences Methods and Technologies International Journal (SciMeTech)

ISSN: 3085-5284, journal homepage: www.scimetech.com

Stratification of Risk of Developing Early-Onset Intestinal Cancer Using Machine Learning Techniques and Nutrition/Socioeconomic Variables Among Low-Income Young Adults in Morocco: A Case-Control Study

Hicham ELLOUAD[1] Anass BELBACHIR[1], Bouchra ASSARAG[2], Anass Doukkali[3] and Assia SAYDTAHIRI[4]

[1.] *Laboratory of Excellence in Biotechnology, Pharmaceutical Bioengineering, and Artificial Intelligence, FMPM, Cadi Ayyad University, Marrakech, Morocco.*

[2.] *Public Health Department, National School of public health, Rabat, Morocco.*

[3.] *Laboratory of Analytical Chemistry, Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Morocco*

[4.] *LARTID, National school of Applicated Sciences, Cadi Ayyad University, Marrakech, Morocco.*

h.ellouad.ced@uca.ac.ma, an.belbachir@uca.ac.ma, bassarag1@gmail.com, doukkali73@gmail.com, a.saydtahiri.ced@uca.ac.ma

How to cite:

Hicham ELLOUAD, Anass BELBACHIR, Bouchra ASSARAG, Anass Doukkali, Assia SAYDTAHIRI, "Stratification of Risk of Developing Early-Onset Intestinal Cancer Using Machine Learning Techniques and Nutrition/Socioeconomic Variables Among Low-Income Young Adults in Morocco: A Case-Control Study", Sciences Methods and Technologies International Journal (SciMeTech), (2026) Vol 2, Issue 2, p 40-47

Keywords:
Intestinal cancer
Machine Learning
early-onset
risk stratification

Abstract

The incidence of intestinal cancer has increased rapidly among young adults in Morocco due to the nutrition transition that involves the gradual shift from adherence to the traditional Mediterranean diet to the intake of highly processed foods. Despite the emergence of strong epidemiological evidence of the risk, there exists no machine learning algorithm that incorporates culture-specific nutritional risk factors in the risk assessment of early onset of colorectal cancer among young adults in Morocco.

This study seeks to build an interpretable machine learning system for predicting the risk of colorectal cancer among young adults in low-income settings in Morocco based on sixteen nutritional risk factors that apply to their dietary habits.

Three group case-control analysis are done at the hospitals of the Marrakech-Safi region in Morocco (sample size: 1,353 individuals; 459 young cancer patients, 447 young healthy controls, 447 older cancer patients). A set of machine learning algorithms are built and benchmarked: Elastic Net Logistic Regression, Balanced Random Forest, and Extreme Gradient Boosting (XGBoost) using Bayesian optimization with five-fold cross-validation.
Extreme Gradient Boosting (XGBoost) yielded the highest classification accuracy (area under curve = 0.935, cross-validated area under curve = 0.926 ± 0.024). The decision curve analysis showed net clinical benefit at 0.43 vs 0.25 for non-stratified mass screening. Ultra-processed food intake and Mediterranean diet scores were the most influential predictors. Importantly, dietary quality improvement lowered an individual's predicted risk by 93% - from 0.913 to 0.062.
The proposed questionnaire-based ML model does not require any laboratory tests and can be implemented in outpatient practices regardless of resources. Early-onset colorectal cancer in Moroccan low-income young adults is a nutritional and socioeconomic disorder.

I. Introduction

Intestinal cancer represents a significant public health threat globally. According to GLOBOCAN 2022 statistics, it ranks third among all cancers in terms of incidence, making up 9.6% of all cancers with 1,926,425 new diagnoses and 904,019 deaths in 2022 [1]. It is anticipated that there will be an alarming rise of up to 3.2 million new cases and 1.6 million deaths by 2040 due to demographic changes, aging population, and fast urbanization in LMICs [2]. The problem becomes even more acute in the context of countries with a moderate HDI when it comes to the increasing incidence and mortality in these countries as opposed to the stable figures of the former category [2]. Early detection becomes especially important at the localized stage in contrast to 15% in advanced stages of development.

As far as Morocco and North Africa are concerned, the rate of intestinal cancer has been growing at a fast pace. In 2022, the North African region had the highest age-standardized rates of colorectal cancer in Africa, with the disease among the top four types of cancer diagnosed in more than half of all the countries on the continent [3]. In Morocco, the incidence rates have doubled within two decades from 6.15 per 100,000 population from 2005 to 2007 to 10.5 per 100,000 inhabitants from 2018 to 2021 based on the statistics published by the Greater Casablanca Cancer Registry [4]. One of the characteristics that makes the epidemiology of this disease in Morocco a significant issue is the high proportion of people under 40 years old affected by rectal cancer, which accounted for 26.6% of the total cases diagnosed [5].

Factors related to nutrition and lifestyle contribute significantly to intestinal cancer. According to scientific evidence, as many as 70% of cases might be preventable using dietary modifications [6]. Traditionally, people following the Mediterranean diet based on copious amounts of whole grains, legumes, fruits, vegetables, olive oil, and fish have a decreased risk. In Morocco, consumption of such a traditional diet is progressively decreasing because of the ongoing nutritional transition caused by urbanization and economic development that led to increased use of fast food, ultra-processed products, refined sugars, and saturated fat intake—a trend especially pronounced in economically disadvantaged young people for whom access to healthy foods is limited because of economic constraints. El Asri et al. [7] identified the association between consumption of selected fats (AGPI, MUFA) and KRAS mutations in Moroccans through a validated 255-items FFQ. Bouzini et al. [8] found, in a case-control study of 616 individuals, that frequent consumption (≥3 times/week) of fast food was linked to a nine-fold greater chance of developing early-onset colorectal cancer (aOR = 9.17; 95% CI [2.93–28.57]; $p < 0.001$). In contrast, fruit, vegetables, and legumes were significantly protective. El Kinany et al. [9] validated, in the largest case-control study among Moroccans (n = 3,032 participants across 5 hospitals), that a high MDS score (≥6) was linked to a 26% decreased risk (aOR = 0.74; 95% IC [0.63–0.86]). Taken together, the findings indicate that diet is a crucial modifiable determinant of colorectal cancer risk among Moroccans. However, no prior research examined the role of machine learning for prediction.

The incorporation of ML in medical research has provided significant opportunities for predicting cancer risks. The reason for that is that compared to statistical models, ML models incorporate nonlinearity and interactions between different variables, making them more accurate and more flexible when dealing with multi-variate cases. Nartowt et al. [10] examined seven ML models on two US-based databases (NHIS and PLCO, n ≈ 800,000) and achieved an AUC of 0.70; the problem is that there were no variables related to nutrition at all, and that the population included only people from the USA. Another study by Du et al. [11] created a LightGBM ML algorithm based on SHAP analysis, using data from Chinese clinics, with an incredible result of 0.943 AUC for external validation; but the issue here is that the model requires biological biomarkers such as CA19-9, hemoglobin, and albumin that are expensive and unavailable to primary care physicians in resource-limited countries. Pu et al. [12] managed to improve prediction accuracy by introducing four nutritional variables in an XGBoost ML model on a dataset including 21,358 people from China (AUC = 0.856, ΔC-index = +0.101). Moreover, as observed by Gozum et al. [13], there are critical shortcomings in the availability of culturally relevant screening tests for MENA and Mediterranean communities.

Despite all this progress, an important limitation still exists, since there is no machine learning algorithm that included culture-specific nutritional variables for Morocco or the region, and there is no study that combined both the social/economic and nutritional aspects of early-onset intestine cancer among economically disadvantaged youth. This paper seeks to fill this gap through the development and validation of an interpretable machine learning framework aimed at stratifying patients' risk of early-onset intestine cancer within a socioeconomically disadvantaged population of young Moroccans. A synthetic dataset (n = 1,353) is constructed based on real Moroccan colorectal cancer cohorts, incorporating sixteen culturally-specific nutritional variables — including charcoal-grilled meat, mint tea, traditional Moroccan food, and ultra-processed products — alongside established clinical, behavioral, and socioeconomic risk factors. Three machine learning models are evaluated in this paper: Elastic Net Logistic Regression, Balanced Random Forest, and XGBoost. Moreover, a special subgroup analysis for individuals belonging to poor families are performed. This uses an ML model without biological testing but only questionnaire questions are proposed to use in limited MENA resource setting conditions.

II. Methods

a. Conception and population

The data is recruited between September 2024 and February 2026 in hospitals situated in Marrakech-Safi, Morocco (ethics committee approval obtained). The three groups were as follows: G1 – young cases diagnosed with histopathology-based colorectal cancer (aged 18–45 years; n = 459); G2 – young controls without any illness, matched by gender and age (+/- 2 years; n = 447); G3 – older controls with intestinal cancer (aged > 45 years; n = 447). Among the subjects, those with low socioeconomic status (beneficiaries of RAMED and/or earning less than the minimum wage) represented 69.9% of G1, 72.3% of G2, and 47.0% of G3. The exclusion criteria comprised pregnant women, medical prescriptions for certain diets, mental disabilities, and hereditary cancers.

The data collection methods involved structured face-to-face interviews that covered: sociodemographic factors, anthropometric indicators (BMI), medical and family backgrounds, lifestyle factors (smoking, alcohol intake, and physical exercise using GPAQ tool), and dietary information collected using validated Moroccan FFQ questionnaire with 16 culture-specific nutritional elements (charcoal-cooked meat, mint tea, Moroccan food, ultra-processed meals, MDS score: 0–10).

b. Machine Learning Models

Training is conducted using three models with stratified splitting 80/20, seed = 42, and fivefold cross-validation:

Elastic Net (LR) : regularized logistic regression including L_1 and L_2 regularization [16]:

$$L(\beta) = -\frac{1}{n} \times \sum[yi \times \log(\hat{p}i) + (1-yi) \times \log(1-\hat{p}i)] + \lambda \times [\alpha \times ||\beta||_1 + \frac{1-\alpha}{2} \times ||\beta||_2^2] \quad (1)$$

where $\lambda = 1{,}0$, $\alpha = 0{,}5$. Combines feature selection ability (L_1 regularization) and prevention of multicollinearity issues (L_2 regularization).

Balanced Random Forest (BRF) : random forest with class weighting coefficients $w_c = \frac{n}{nk \times K}$, and prediction based on majority voting $\hat{P}_{(Y=1|X)} = \frac{1}{T} \times \sum ht(X)$, variable importance is assessed using Gini reduction [14] [17]. Model parameters: number of trees = T = 300, class_weight = "balanced_subsample".

XGBoost : sequential gradient boosting with loss function minimization [15]:

$$O(m) = \Sigma\, L\, (yi, \hat{y}i(m-1) + hm(xi)) + \Omega(hm) \quad (2)$$

The regularization term $\Omega(hm) = \gamma \times Tm + \frac{1}{2} \times \lambda \times \Sigma\, wj^2$. Hyperparameters optimized using a Bayesian method: max_depth = 4, learning_rate = 0,05, subsample = 0,8.

c. Individual risk score

The risk score is computed as a weighted linear combination of the standardized predictors:

$$RiskScore(X) = \sum_{j=1}^{p} \omega j \times \tilde{X}j \quad (3)$$

where ωj is a normalized weight ($\Sigma\, \omega j = 1$), reflecting the relative contribution of each variable to the overall risk score, which varies across population subgroups.

The estimated probability of a high-risk category is [19]: $\hat{P}(\text{High Risk} \mid X) = \frac{1}{1+e^{-\text{RiskScore}}}$.

d. valuation and clinical utility

The following metrics are used for performance assessment: AUC-ROC, CV-AUC (5-fold), accuracy, sensitivity, specificity, F1-score, and Brier score. The comparison of AUC is done using the DeLong test (p-value <0.05) [18]. The clinical usefulness of the models was determined via Decision Curve Analysis (DCA) with the formula:

$$NB(t) = \frac{TP}{n} - \frac{PF}{n} \times \frac{t}{1-t} \quad (4)$$

Subgroup analysis in a sample of low-income participants (n = 854: G1 LI n = 321, G2 LI n = 323, G3 LI n = 210) was conducted to evaluate the model's robustness among a particular target population.

III. Results and discussion

a. Caractéristiques de la population

The sample consisted of 1,353 participants: 459 patients with early-onset disease (Group 1), 447 young individuals without cancer (Group 2), and 447 older individuals without cancer (Group 3). As seen from Figure 1, the successful matching has been performed since the median age in Groups 1 and 2 was the same (equal to 31 years: Group 1: 31.4 ± 7.8; Group 2: 31.3 ± 7.7 years) against 63 years in Group 3 ($P < 0.001$). The median body mass index in Group 1 was 28.8 kg/m², which was higher than in Group 2 (25.9 kg/m²) and like Group 3 (27.9 kg/m²), suggesting that being overweight is more strongly associated with cancer status than with age.

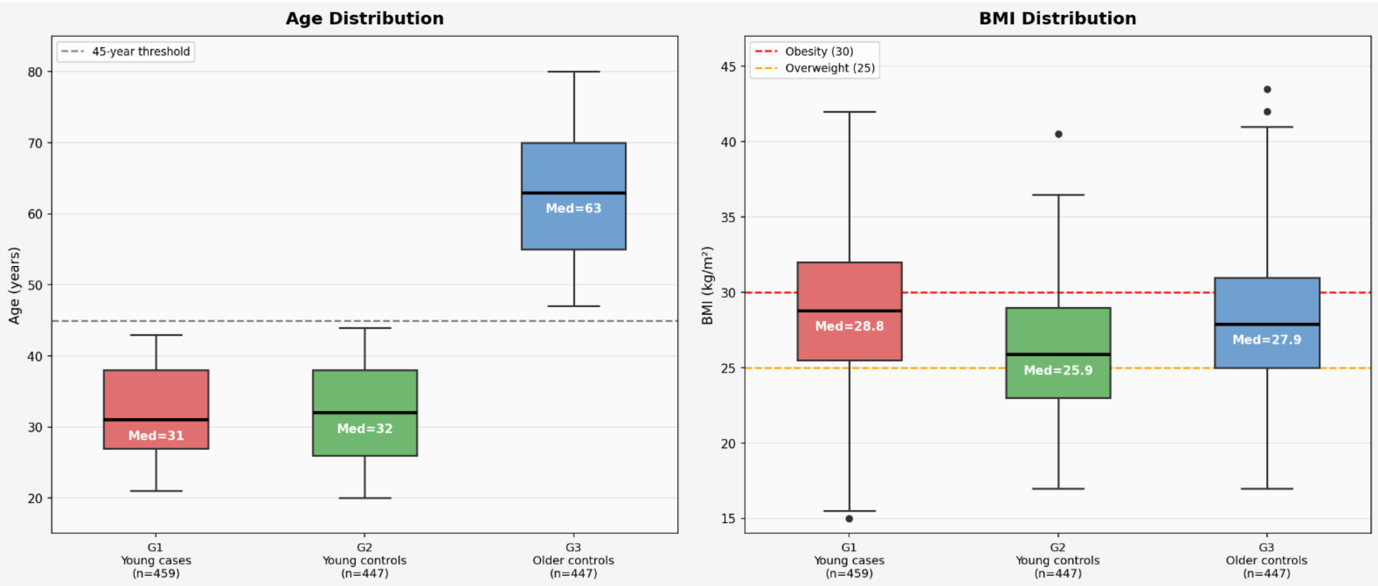


Figure 1 — Distribution of age and BMI by group

As evident from Table 1, the differences in all the variables related to nutritional intake and behavior are highly statistically significant ($p < 0.001$). The highest differentiating factors between the two groups are as follows: MDS score (3.9 ± 1.4 vs. 5.7 ± 1.6); consumption of ultra-processed foods (4.0 ± 1.9 vs. 2.4 ± 1.7 times/week); obesity (42.9% vs. 19.5%); calcium deficiency (55.8% vs. 29.1%); inflammatory bowel disease (11.8% vs. 2.5%); smoking habit (53.4% vs. 34.5%); and level of physical activity (1.5 ± 1.6 vs. 2.5 ± 1.7 days/week). These results confirm those found by El Kinany et al. [9] and Bouzini et al. [8] about the importance of dietary habits in colorectal cancer in Morocco.

Table 1— Most discriminatory variables (G1 vs. G2 vs. G3)

Variable	G1 Young Cases	G2 Healthy Controls	G3 Older Controls	p-value
Age (years)	31.4 ± 7.8	31.3 ± 7.7	62.7 ± 9.8	<0.001
BMI (kg/m²)	29.0 ± 5.1	25.6 ± 4.7	27.6 ± 4.9	<0.001
MDS Score	3.9 ± 1.4	5.7 ± 1.6	5.7 ± 1.6	<0.001
Ultra-processed foods/week	4.0 ± 1.9	2.4 ± 1.7	2.0 ± 1.6	<0.001
Red meat/week	3.5 ± 1.8	2.5 ± 1.7	2.8 ± 1.7	<0.001
Fruits/day	1.3 ± 1.1	2.0 ± 1.1	2.0 ± 1.1	<0.001
Obesity (%)	42.9%	19.5%	32.2%	<0.001
IBD (%)	11.8%	2.5%	9.5%	<0.001
Low income (%)	69.9%	72.3%	47.0%	<0.001
Smoking (%)	53.4%	34.5%	45.2%	<0.001

b. Performance of ML models

Discrimination was successful in all three models. The Balanced Random Forest scored 0.9337 (CV-AUC = 0.917 ± 0.024), XGBoost reached an AUC of 0.9349 (CV-AUC = 0.926 ± 0.024), and the Elastic Net resulted in an AUC of 0.9364. (Table 2) The low Brier scores indicate that the predictions are well-calibrated, i.e., for the three models, respectively, 0.1709 (XGBoost), 0.1732 (Balanced RF), and 0.1793 (Elastic Net), which is under the threshold of 0.25 [18]. The Balanced Random Forest model has the highest sensitivity (76.1%), while the high specificity (76.7%) is characteristic of XGBoost, and is crucial to prevent unnecessary colonoscopies (see Figure 2). The three models perform significantly better than the Moroccan models, which only use standard logistic regression [8, 9].

Table 2 — Comparative performance of the three ML models

Model	AUC	CV-AUC	Sensitivity	Specificity	Brier
Elastic Net	0,9364	0,917 ± 0,024	72,8 %	72,2 %	0,179
Balanced RF	0,9337	0,917 ± 0,024	76,1 %	72,2 %	0,173
XGBoost	0,9349	0,926 ± 0,024	72,8 %	76,7 %	0,171

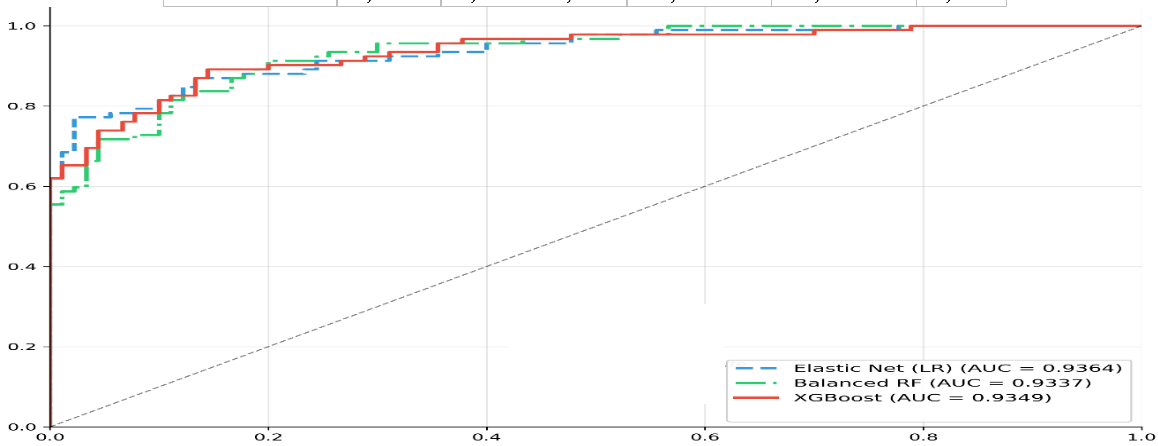


Figure 2 — ROC curves for prediction models

c. Significance of variables

However, the importance of the variables under analysis is the main result obtained during this work. The nutrition-related factors prevail over the total population, since eleven of the fifteen factors having the greatest importance have been identified in both models. Thus, for the model Balanced Random Forest, ultra-processed food is at the top position (0.1276), while the MDS Score (0.0935) and BMI (0.0884). Age is ranked seventh (0.0475), which clearly indicates that diet-related considerations play a more discriminating role compared to age (Figure 3).

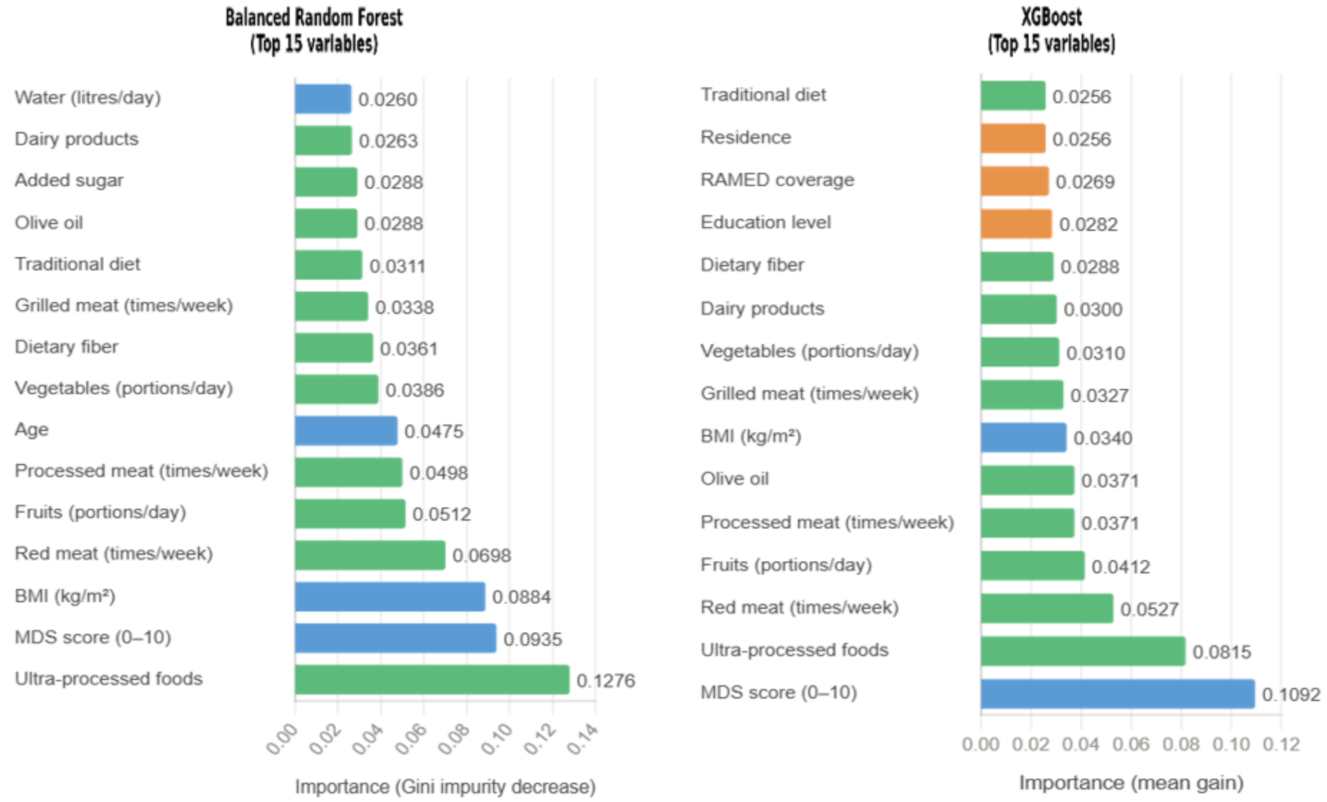


Figure 3 — Importance of variables

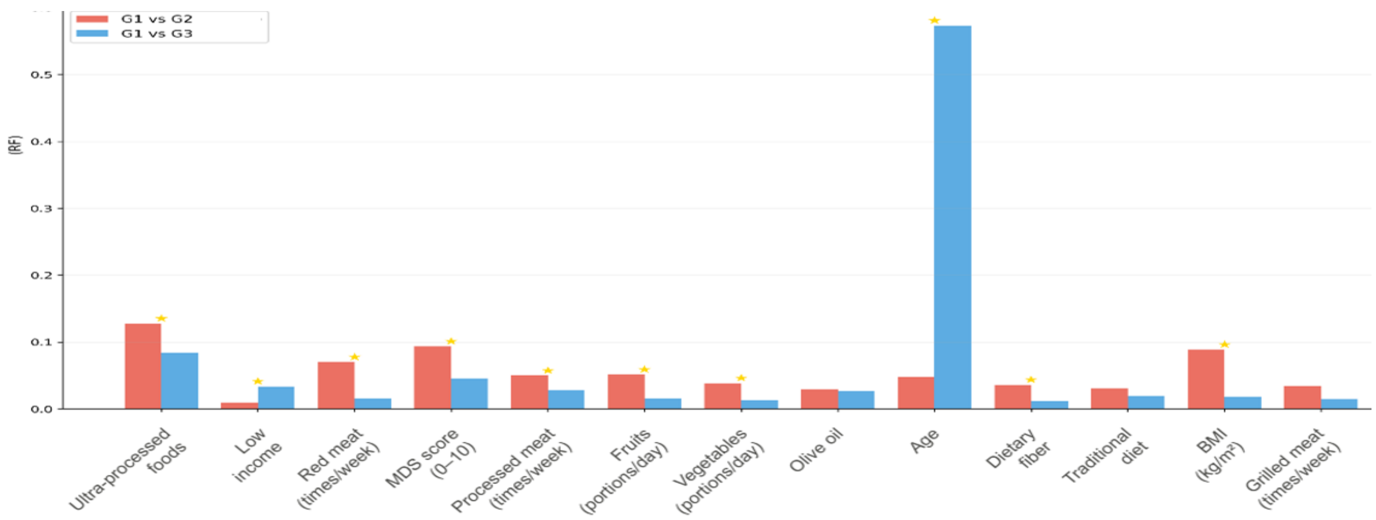
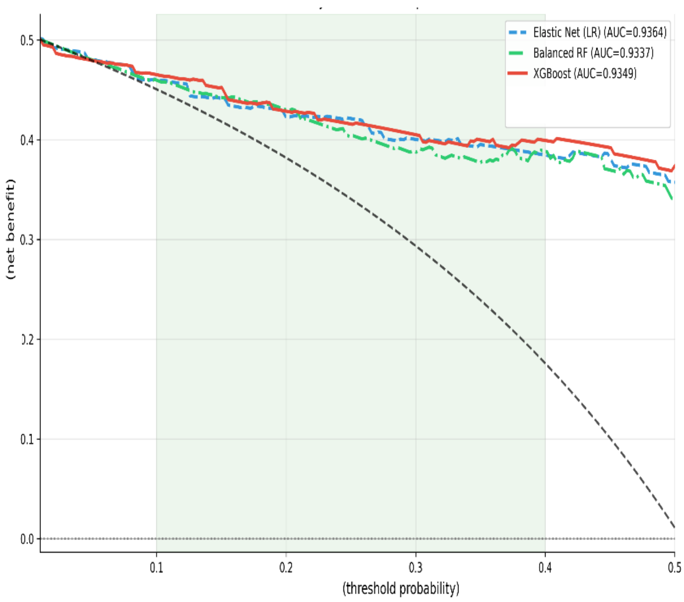


Figure 4 — Comparison of risk factors: young vs. older adults

Figure 4 highlights a crucial qualitative disparity between younger and older individuals. Among younger adults (G1 vs. G2), nutrition factors stand out—ultra-processed foods (0.128), MDS (0.095), and BMI (0.089). Among older adults (G1 vs. G3), age stands out with a statistical significance of 0.575, being tenfold more significant than any nutritional factor. This disparity proves that young-onset colorectal cancer can be attributed to nutritional and socioeconomic factors, which differs from colorectal cancer in older individuals. Among the low-income group, the hierarchy is reversed: the MDS Score moves up to rank first (0.1406), while items unique to the precarious Moroccan setting, such as preserved and salted food (0.0478) and charcoal-broiled meat (0.0444), move into the top five ranks.

d. Clinical utility — DCA and stratification

The Decision Curve Analysis (Figure 5) reveals that the net benefit is 0.43 for ML models at a decision threshold of 0.20, as opposed to 0.25 for the "treat everybody" approach, which corresponds to a net benefit of +0.18. This finding supports the implementation of the developed model as a decision-support mechanism in general practice. Stratification of clinical risk into risk score groups shows that 90% of G1 cases are high-risk patients (risk score above 0.60) and 77% of G2 controls are low-risk patients (risk score below 0.30). The application of this risk score to representative profiles (see Figure 6) illustrates the real-life utility of the model: a vulnerable young patient with very poor nutrition is assessed a risk score of 0.913 (require colonoscopy), whereas a vulnerable young patient with good nutrition is assigned a score of 0.062 (only nutritional counseling). An improvement in dietary quality reduces the risk score from 0.913 to 0.062, representing a 93% reduction.



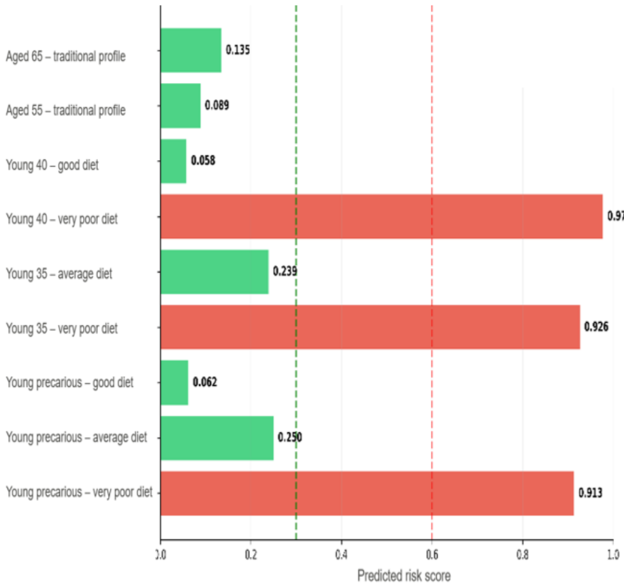


Figure 5 — Decision Curve Analysis **Figure 6— Individualized risk scores by profile type**

Subgroup analysis for the poor group (Figure 7) demonstrates the robustness of the model, where AUC for Balanced RF = 0.8939 and for XGBoost = 0.9024 on n = 854 individuals. Slight decline (ΔBalanced RF = -0.040) in accuracy is anticipated since there will be less variance in terms of socioeconomics due to the homogeneous population. In the current scenario, the MDS Score emerges as the predominant factor (0.1406), which highlights lack of adequate nutritional protection as the primary risk factor. The current study has shown that it is possible to develop an explainable model for the prediction of early-onset CRC in a socioeconomically underprivileged Moroccan community. The model has provided AUC values of 0.93 and showed its clinical applicability in terms of the ROC plot (with net benefit of 0.43 compared to the value of 0.25 for the non-selective approach), using only a questionnaire without biological tests.

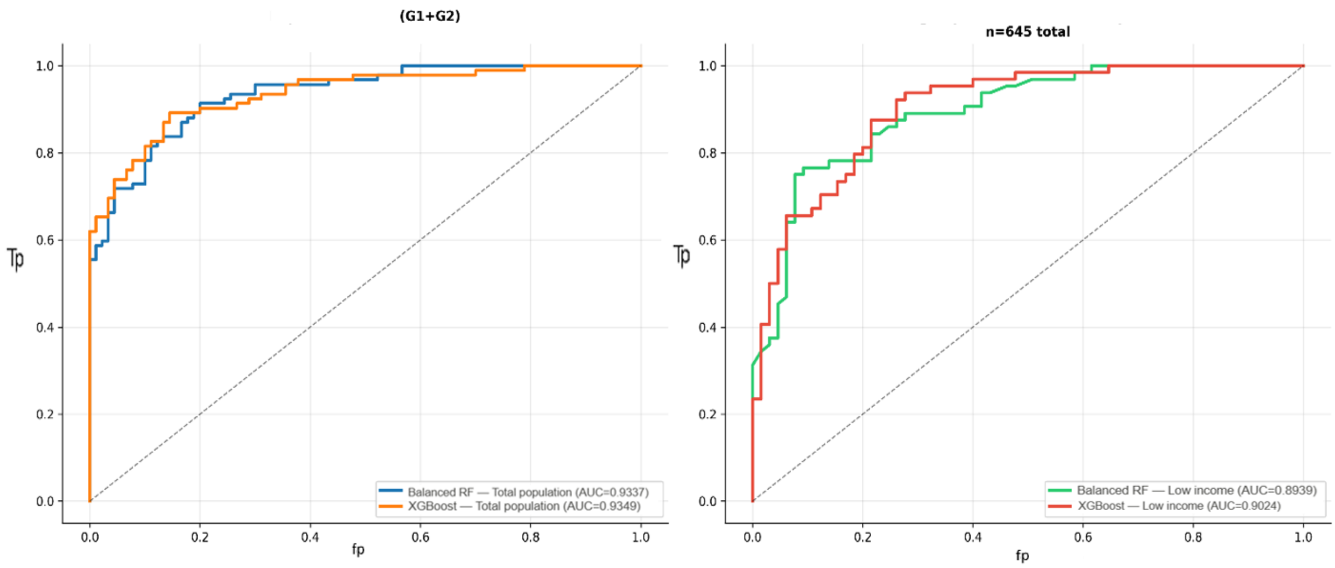


Figure 7 — Comparative analysis: total population vs. low-income subgroup

Compared to existing literature, our proposed model (AUC=0.93) performs better than the model of Nartowt et al. [10] (AUC=0.70) due to the addition of 16 nutritional factors that are not considered by the latter researchers. It is on par with the model suggested by Du et al. [11] (AUC=0.943), although it does not require expensive blood-based biomarkers, thus providing a better performance-cost relationship than Du et al.'s model in MENA settings. It builds on Pu et al. [12]

(AUC=0.856; four dietary factors) by adding 16 cultural Moroccan-specific factors, including charcoal grilled meat, mint tea, and Moroccan foods in general—all of which were absent from previous models.

As can be seen from the results of the analysis, this research provides quantitative confirmation of the fact that early-onset colorectal cancer represents a unique nosologically entity, where nutrition is the leading factor; more specifically, the coefficient of the role played by age in this respect is 0.0475 among young patients compared to 0.575 among older patients – ten times less. It turns out that in this group of individuals, the MDS Score indicator becomes the most important variable, meaning that the crucial factor here is the lack of healthy food products (fruits, vegetables, olive oil), rather than the mere presence of harmful foods that determines risk. The two-pronged vulnerability of economic and nutritional factors, which is measured for the first time through ML modeling, directly calls for policy interventions that will subsidize fresh produce in disadvantaged urban settings within Morocco.

IV. Conclusion

The current work describes a predictive model for early-onset colorectal cancer that uses culturally specific nutritional parameters within a machine learning framework for a vulnerable young Moroccan population. The models have an AUC of 0.93 and DCA of 0.43 with the use of only a costless questionnaire. Confirmation that early-onset colorectal cancer is a nutritional disease with the MDS Score and ultra-processed foods as key indicators, and with an estimated potential 93% decrease in risk by improving nutrition, makes it important to consider the implementation of a screening tool in Morocco and the MENA region.

References

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229-263.

[2] Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C. J., Laversanne, M., ... & Bray, F. (2023). Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*, *72*(2), 338-344.

[3] Louise, V. M. L. I., Zhao, Q., & Liu, L. (2024). African colorectal cancer burden in 2022 and projections to 2050. *ecancermedicalscience*, *18*.

[4] Registre des Cancers de la Région du Grand Casablanca 2018–2021. (2022). *Fondation Lalla Salma de Prévention et Traitement des Cancers*, Maroc. Disponible sur : www.contrelecancer.ma

[5] Deoula, M. M. S. O., Huybrechts, I., El Kinany, K., Boudouaya, H., Hatime, Z., El Asri, A., ... & El Rhazi, K. (2020). Behavioral, nutritional, and genetic risk factors of colorectal cancers in morocco: protocol for a multicenter case-control study. *JMIR Research Protocols*, *9*(1), e13998.

[6] World Cancer Research Fund / American Institute for Cancer Research. (2018). *Diet, Nutrition, Physical Activity and Colorectal Cancer*. Continuous Update Project Expert Report. London: World Cancer Research Fund International. Disponible sur : www.wcrf.org/dietandcancer

[7] El Asri, A., Ouldim, K., Bouguenouch, L., Sekal, M., Moufid, F. Z., Kampman, E., ... & El Rhazi, K. (2022). Dietary fat intake and KRAS mutations in colorectal cancer in a Moroccan population. *Nutrients*, *14*(2), 318.

[8] Bouzini, G., Sammoud, K., Amzerin, M., Najdi, A., & El M'Rabet, F. Z. (2026). Fast food consumption patterns in Moroccan adults and their association with the risk of early-onset colorectal cancer. *Clinical Nutrition ESPEN*, 102976.

[9] El Kinany, K., Hatime, Z., El Asri, A., Benslimane, A., Deoula, M. M. S., Zarrouq, B., ... & El Rhazi, K. (2025). Adherence to the Mediterranean diet and colorectal cancer risk: a large case control study in the Moroccan population. *Public Health Nutrition*, *28*(1), e48.

[10] Nartowt, B. J., Hart, G. R., Muhammad, W., Liang, Y., Stark, G. F., & Deng, J. (2020). Robust machine learning for colorectal cancer risk prediction and stratification. *Frontiers in big Data*, *3*, 6.

[11] Du, G., Lv, H., Liang, Y., Zhang, J., Huang, Q., Xie, G., ... & Sun, Y. (2025). Population-based colorectal cancer risk prediction using a SHAP-enhanced LightGBM model. *Frontiers in Oncology*, *15*, 1575844.

[12] Pu, J., Zhou, B., Yao, Y., Wu, Z., Wen, Y., Xu, R., & Xu, H. (2025). Development and validation of a lifestyle-based 10-year risk prediction model of colorectal Cancer for early stratification: evidence from a longitudinal screening cohort in China. *Nutrients*, *17*(11), 1898.

[13] Gozum, S., Nimri, O. F., Merzah, M. A., & Vitorino, R. (2025). Cancer Screening and Prevention in MENA and Mediterranean Populations: A Multi-Level Analysis of Barriers, Knowledge Gaps, and Interventions Across Indigenous and Diaspora Communities. *Diseases*, *14*(1), 10.

[14] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[15] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[16] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320.

[17] LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, *18*(17), 1-5.

[18] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

[19] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.