The Role of Digital Health Literacy and Socioeconomic Factors in Colorectal Cancer Screening: Machine Learning Analysis of HINTS Data

Sujin Kim[1][a], Madhav Dahal[1][b], Avinash Bhakta[2][c] and Jihye Bae[3][d]
[1] *Division of Biomedical Informatics, University of Kentucky, 725 Rose Street, Lexington, KY, U.S.A.*
[2] *Deptment of Electrical and Computer Engineering, University of Kentucky, 512 Administration Drive, Lexington, KY, U.S.A.*
[3] *Department of Surgery, University of Kentucky, 800 Rose Street, Lexington, KY, U.S.A.*

Keywords: Colorectal Cancer Screening, Digital Health Literacy, Socioeconomic Factors, Machine Learning, Health Information National Trend Surveys (HINTS).

Abstract: While colorectal cancer (CRC) screening rates are on the rise, significant disparities persist, particularly among underserved populations, highlighting ongoing challenges in achieving equitable access to preventive care. This study utilizes machine learning models to analyze multi-year data from the Health Information National Trends Survey (HINTS), identifying critical factors influencing CRC screening adherence across three distinct time periods (2003–2008, 2011–2013, 2018–2020). Using Random Forest and Logistic Regression models, interpreted through Shapley Additive exPlanations values, we examine the impact of sociodemographic characteristics, digital health engagement, and digital literacy on CRC screening behaviors. Findings reveal that age, prior screening behavior, and digital literacy are key predictors; individuals with higher digital literacy, for example, exhibited a 22% higher likelihood of adhering to CRC screening guidelines. Age emerged as a dominant factor, with screening rates peaking at 43% in the 50–64 age group. These results suggest that interventions targeting digital health literacy and enhancing provider communication may effectively improve CRC screening rates among underserved populations. This study underscores the value of data-driven approaches in informing public health strategies to increase CRC screening adherence and reduce health disparities.

1 INTRODUCTION

Colorectal cancer (CRC) remains one of the most prevalent cancers globally and is the second leading cause of cancer-related deaths, even as advancements in screening and treatment have contributed to declining incidence and mortality rates (Siegel, 2022; Bray, 2018). In the United States, CRC predominantly affects adults aged 65-74, with screening rates increasing over recent decades due to the adoption of colonoscopy and non-invasive methods such as multitarget stool DNA (FIT-DNA) testing (Keum, 2019). However, a concerning trend is the rising incidence of early-onset colorectal cancer (EOCRC) among adults under 50, a rate projected to double by 2030 and driven by complex, interacting risk factors not fully understood (Zhen, 2024; Sun, 2024). Disparities in CRC screening persist, influenced by sociodemographic factors such as income, education, and race, as well as health behaviors. These multifaceted barriers, often studied in isolation, underscore the need for a more integrated approach to understanding and addressing CRC screening uptake.

Digital health interventions, including telehealth, patient portals, and mobile health (mHealth) applications, offer promising avenues to address these complex challenges in CRC screening by making screening information, reminders, and test results more accessible to diverse populations (Miller, 2018;

[a] https://orcid.org/0000-0002-7878-4322
[b] https://orcid.org/0009-0002-1016-428X
[c] https://orcid.org/0000-0003-2471-3681
[d] https://orcid.org/0000-0002-6609-9782
Corresponding Author: sujinkim@uky.edu

Kim, S., Dahal, M., Bhakta, A. and Bae, J.
The Role of Digital Health Literacy and Socioeconomic Factors in Colorectal Cancer Screening: Machine Learning Analysis of HINTS Data.
DOI: 10.5220/0013310700003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 228-239
ISBN: 978-989-758-731-3; ISSN: 2184-4305

McIntosh, 2024). For instance, studies have shown that digital health interventions can improve screening adherence in vulnerable groups compared to standard care, and patient portal reminders paired with mailed test kits have increased adherence among average-risk populations (Miller, 2018; McIntosh, 2024). The Health Information National Trends Survey (HINTS) provides a valuable, nationally representative dataset on health behaviors, digital literacy, and sociodemographic factors that impact CRC screening. Leveraging machine learning (ML) such as Random Forest (RF) and Logistic Regression (LR) to analyze multi-year HINTS data enables the identification of nuanced relationships among digital health literacy, socioeconomic variables, and screening adherence, offering a more holistic, data-driven approach to improve CRC screening rates and reduce disparities in an increasingly digital healthcare environment.

2 BACKGROUNDS

CRC screening uptake is shaped by a wide range of demographic, psychosocial, and access-related factors. Insights from the HINTS dataset reveal crucial predictors that influence CRC screening decisions, providing a nuanced view of how different factors affect adherence. For example, Atarere (2024c) found that smokers are 30% less likely to adhere to CRC screening protocols than non-smokers, highlighting the potential of health information technology interventions to increase participation among high-risk groups (Atarere, 2024b). Further, Atarere et al. (2024) reported that patients engaged in telehealth primary care visits had a 20% higher likelihood of discussing CRC screening with their providers, suggesting that Health IT (HIT) tools like telehealth can significantly improve screening discussions and adherence in populations typically resistant to CRC screening (Atarere, 2024a; Atarere, 2024c).

Beyond access to HIT, cultural and social influences are critical in shaping CRC screening behaviors. Jun and Oh (2013) found that cancer fatalism among Asian and Hispanic Americans was associated with a 15% reduction in CRC screening likelihood, pointing to cultural perceptions as barriers to uptake. Similarly, Idowu et al. (2016) observed that U.S. adults born outside the United States were 18% less likely to be up-to-date with CRC screening, underscoring the informational and cultural barriers faced by immigrant populations. Additionally, Finney Rutten et al. (2009) found that only 56% of the general public understood CRC risk and prevention guidelines accurately, linking low public knowledge to reduced screening adherence. These findings highlight a complex interplay of sociocultural and informational barriers, which Nawaz et al. (2014) further supported by showing that CRC screening offered in hospital settings resulted in a 35% higher uptake rate, suggesting that more accessible, opportunistic screening efforts could be effective in improving national screening rates.

Early studies using HINTS 123 data (2003–2008) identified foundational barriers to CRC screening, including limited awareness, inadequate knowledge of guidelines, and socioeconomic and cultural disparities. For instance, Geiger et al. (2008), analyzing HINTS 1 data, pinpointed knowledge gaps and access issues as central barriers. Similarly, Hay et al. (2006) reported that perceived risk was a decisive factor in CRC screening uptake, with substantial demographic differences in risk perception. More recent HINTS data (HINTS 5, 2018-2020), however, indicate a shift toward increased digital literacy and greater use of HIT, both of which are positively correlated with screening adherence. The growth of telehealth and online health resources appears to support CRC screening behaviors, marking a significant change in how technology influences preventive health actions.

ML applications in CRC research have increasingly used electronic health record (EHR) data to enhance predictive models, especially in EOCRC studies. For example, studies by Sun et al. (2024) and Zhen et al. (2024) showed high predictive accuracy for EOCRC among individuals under the standard screening age, with area under the curve (AUC) scores reaching up to 0.888 when using RF models. These studies illustrate the clinical value of EHR data in identifying risk factors such as immune and digestive disorders, allowing for the development of models that can flag high-risk patients who may benefit from early diagnostic interventions.

While EHR-based ML models provide valuable insights into clinical risk factors, they often overlook behavioral elements influencing CRC screening adherence, such as health literacy, digital literacy, and risk perception—areas extensively covered in the HINTS dataset. Unlike EHR data, which primarily capture clinical encounters, HINTS data offer a comprehensive view of health behaviors, allowing for broader exploration of sociobehavioral factors like health literacy and digital engagement. Integrating these broader factors is essential for understanding screening behaviors, as digital literacy, socioeconomic status, and perceived cancer risk have become key determinants of screening adherence, especially with the evolution of HIT over time.

In this study, we examine these broader behavioral factors by analyzing HINTS data across three significant temporal milestones—HINTS 123 (2003-2008), HINTS 4 (2011-2013), and HINTS 5 (2018-2020). This approach allows us to track changes in CRC screening attitudes, HIT engagement, and health information access over nearly two decades. By combining HINTS data with ML, we aim to uncover complex, time-varying relationships between behavioral factors and screening outcomes, thereby informing tailored interventions designed to improve CRC screening rates across diverse populations.

3 METHODS

3.1 Data Source and Study Variables

The HINTS datasets offer a robust, nationally representative sample capturing U.S. adults' cancer knowledge, perceptions, information-seeking behaviors, and digital health tool adoption over time. Each HINTS cycle introduces unique variables to reflect emerging HIT trends, while maintaining core questions that allow for cross-cycle comparisons on foundational topics such as cancer knowledge, perception, and health information access. For this study, we grouped three survey cycles within each of three temporal groups, focusing on CRC screening-related variables while noting differences in HIT priorities across cycles:

- Group 1: Merged HINTS Cycles 1-3 (2003, 2005, 2008) – This grouping consolidates early cycles with a focus on CRC-related variables, including “Awareness of CRC screening,” “Frequency of CRC screening conversations with healthcare providers,” and “CRC screening completion history.” Predating widespread digital tool adoption, this group establishes a baseline of public knowledge and screening behaviors (Ford, 2006; Geiger, 2008; Hay, 2006).
- Group 2: Merged HINTS Cycle 4 (2011, 2012, 2013) – Reflecting the era’s shift toward digital health adoption, this group includes variables like “Access to EHRs,” “Use of telehealth for health information,” and “Frequency of patient portal usage,” though certain CRC-specific questions are omitted in favour of HIT-focused variables.
- Group 3: Merged HINTS Cycle 5 (2018, 2019, 2020) – Emphasizing advanced HIT usage and digital health literacy, this group includes variables such as “Confidence in locating reliable health information online,” “Frequency of telehealth usage,” and “Patient portal engagement for preventive care.” Although CRC-specific questions are less prominent, consistent HIT variables allow for comparison across cycles.

This grouping allows for a comprehensive examination of CRC screening behaviors over time, highlighting the growing influence of digital tools on CRC engagement. The primary outcome variable, *Ever_Tested_Colon*, is binary, coded as 1 for individuals who reported undergoing a colorectal cancer screening test and 0 for those who did not. Predictor variables include multiple categories representing demographic characteristics (e.g., age, gender, race/ethnicity), socioeconomic status (e.g., income, education level), health behaviors (e.g., smoking status, physical activity), access to healthcare (e.g., health insurance coverage, regular healthcare provider), and digital health engagement (e.g., frequency of internet use for health information, use of electronic health records, and telehealth usage). This comprehensive set of predictors allows for a nuanced analysis of factors influencing CRC screening uptake, considering both individual health characteristics and the broader context of healthcare access and technology use.

3.2 Data Processing and Feature Engineering

To ensure data consistency across cycles, we standardized variable names using the HINTS 5 Cycle 4 format and encoded all categorical variables numerically. Race was expanded into binary indicators (Hispanic, White, Black, Asian, and Other), and marital status was binarized, with “Married” coded as 1 and all others as 0. Missing values were replaced with -1 to retain cases, with verification that imputing -1 did not distort predictions. Group 1’s multiple-response categorical variables were converted into binary format for simplicity, and non-informative variables, duplicates, and weight variables were excluded. Feature selection focused on variables consistently present across cycles, including demographics, health information-seeking behaviors, digital health adoption, and CRC screening history. Final dataset sizes were 11,710 rows and 1,022 columns for Group 1, 10,534 rows and 475 columns for Group 2, and 12,391 rows and 535 columns for Group 3, where each row indicates each individual sample, and each column represents variables extracted from HINTS datasets.

3.3 Predictive Model Development

This study utilized Python (version 3.8) within the Google Colab environment, leveraging cloud resources for efficient data processing, model training, and interpretation. Key libraries included Pyreadstat for SAS file handling, Pandas and NumPy for data manipulation, and Scikit-Learn for machine learning model implementation, pre-processing, and validation. Model interpretability was achieved using SHapley Additive exPlanations (SHAP), enabling a detailed examination of feature contributions to specific predictions. All data were securely accessed and processed via Google Drive integration, ensuring data consistency and reproducibility. Regarding ethics and data privacy, this study exclusively analysed publicly available, de-identified data from the HINTS thereby maintaining strict confidentiality and compliance with data privacy standards. As no personally identifiable information (PII) was present, this secondary data analysis was exempt from Institutional Review Board (IRB) oversight.

We employed two predictive models, Random Forest (RF) and Logistic Regression (LR) with Elastic Net regularization, to predict colorectal cancer (CRC) screening outcomes (CRC Screened or Not). Following Min-Max normalization, 15% of the dataset was allocated for testing. Hyperparameter optimization was conducted through 5-fold cross-validation using a grid search approach. For RF, we searched across key hyperparameters, including the number of trees (n_estimators: 100–500, step 50), maximum tree depth (max_depth: 5–20, step 1), minimum samples required to split a node (min_samples_split: 2–10, step 1), and minimum samples per leaf node (min_samples_leaf: 1–5, step 1). For LR, optimization included searching over regularization strength (C: 0.1–1.0, step 0.1), the balance between L1 and L2 penalties (l1_ratio: 0.5–0.9, step 0.1), and iteration limits (max_iter: 100–1000, step 100). The configurations yielding the highest average cross-validation accuracy were applied to the testing set, with performance evaluated across key metrics: accuracy, precision, recall, F1 score, and area under the curve (AUC) (Table 3).

The optimal hyperparameters for RF varied across datasets. For HINTS123, the best RF configuration included 400 trees, a maximum depth of 12, a minimum of 7 samples per split, and 2 samples per leaf, achieving a cross-validation accuracy of 98.25%. In HINTS4, the optimal configuration used 400 trees, a depth of 14, a minimum of 8 samples per split, and 1 sample per leaf, achieving a cross-validation accuracy of 81.84%. For HINTS5, the best RF model used 400 trees, a depth of 15, 9 samples per split, and 2 samples per leaf, with a cross-validation accuracy of 80.61%.

For LR, the optimal configuration across datasets involved a regularization strength of 0.1, an L1/L2 ratio of 0.5, and 800 iterations, yielding a cross-validation accuracy of 80.86% in HINTS5. Test set performance metrics for both models across datasets were summarized in Table 3, demonstrating the models' strengths in predicting CRC screening outcomes under varying data conditions.

3.4 Model Interpretability with SHAP Analysis

To interpret model predictions and identify top predictors of CRC screening, we calculated SHAP values for the optimized RF model. SHAP values were chosen for their consistency and additive feature contributions, providing insights into how each variable influence on the CRC screening prediction. Focusing on the 15 most impactful variables in each group, we visualized these features to clarify the key drivers of CRC screening across different periods and demographic groups. This interpretability step enhances our understanding of the predictors behind screening behaviors and the evolving role of HIT.

This methodology—combining multiple HINTS cycles, consistent variable standardization, and advanced modelling with interpretability—offers a comprehensive, time-sensitive analysis of CRC screening behaviors. By integrating SHAP analysis, we achieved transparent, interpretable predictions, allowing us to pinpoint essential factors influencing CRC screening adherence across various demographic and HIT-related contexts.

4 RESULTS

4.1 Patient Characteristics of CRC Screened or not

Across the three groups, CRC screening uptake showed a clear upward trend over time. In Group 1 (2003–2008), the screening rates started with 40.12% tested in 2003, dipping to 33.33% in 2005, and rising to 37.52% in 2008, resulting in an overall group rate of 36.94% tested. Group 2 (2011–2013) saw a more substantial increase, with a 50.02% overall screening rate, increasing steadily from 49.32% in 2011 to 52.28% in 2013. Group 3 (2018–2020) had the highest screening rates, with a group average of 62.15%. Notably, 60.90% were tested in 2018, 60.02% in 2019, and this rose to 64.49% by 2020. This progression highlights a positive trend in CRC screening uptake, suggesting increased awareness and access to screening over the years.

SES Characteristics Among Three Groups

Table 1 presents a comparison of socioeconomic (SES) and digital factors among CRC screening-tested individuals across the three HINTS groups in abridged format highlighting underserved groups, while **Appendix A** provides comprehensive details on additional variables. The findings from **Table 1** reveal significant disparities in colorectal cancer (CRC) screening rates among underserved populations, particularly when compared to their counterparts—groups not explicitly represented in the table. Individuals aged 18–49 years displayed consistently lower screening participation, with rates rising from 5.36% in Group 1 to 10.90% in Group 2 but falling again to 7.56% in Group 3. In contrast, their counterparts aged 50 and above, particularly the 50–64 age group, dominated screening adherence, consistent with guideline recommendations targeting this demographic. This trend highlights a critical age disparity, emphasizing the need for enhanced outreach and interventions tailored to younger populations.

Table 1: Comparison of CRC Screening Tested Among Three HINTS group – Underserved.

Variable	Category	Group 1 Tested N (%)	Group 2 Tested N (%)	Group 3 Tested N (%)
Age	18-49	232 (5.36)	574 (10.90)	582 (7.56)
Gender	Female	2735 (63.22)	2969 (56.35)	4327 (56.19)
Education	< 12 years	409 (9.45)	484 (9.18)	519 (6.75)
Income	< $20K	661 (15.28)	971 (18.43)	1269 (16.48)
Employment	Unemployed	2732 (63.15)	3099 (58.82)	2729 (35.44)
Race	Non-White	826 (19.07)	2101 (39.88)	3146 (40.88)

Further disparities emerged in education and income levels. Individuals with less than a high school education (<12 years) showed persistently lower screening rates, declining from 9.45% in Group 1 to 6.75% in Group 3. Compared to individuals with at least a high school diploma or higher, this group remains significantly underserved, underscoring the influence of education and health literacy on screening adherence. Similarly, CRC screening among individuals earning below $20,000 annually increased from 15.28% in Group 1 to 18.43% in Group 2 but slightly dropped to 16.48% in Group 3. This income disparity reflects barriers such as affordability and access to preventive healthcare, which disproportionately affect lower-income populations.

Racial disparities were also notable, with non-White populations increasing their representation from 19.07% in Group 1 to 40.88% in Group 3. While this improvement is promising, non-White individuals remain under-screened compared to White populations, who consistently demonstrate higher screening rates. Gender and employment disparities also persisted; although females consistently represented a higher proportion of those screened, their participation rates declined from 63.22% in Group 1 to 56.19% in Group 3. Conversely, males, historically less represented in screening, may have experienced incremental improvements over time. Meanwhile, unemployed individuals saw significant reductions in screening rates—from 63.15% in Group 1 to 35.44% in Group 3—remaining significantly underserved compared to their employed counterparts, who likely benefit from better healthcare access.

In summary, underserved groups—such as younger individuals aged 18–49, those with lower educational attainment, lower incomes, unemployment, and non-White populations—consistently lag behind their counterparts (older individuals, those with higher education or income, the employed, and White populations) in CRC screening participation. These disparities highlight the need for targeted public health interventions, policy reforms, and culturally sensitive outreach strategies to promote CRC screening among these vulnerable groups. Addressing these gaps is essential to reducing health inequities and improving population-level outcomes.

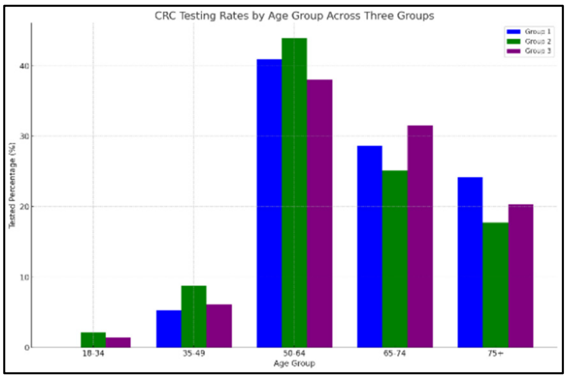

Figure 1: CRC Screening Uptake by Age Groups Among 3 HINTS Groups.

Our further findings on age-specific screening reflect the impact of recent screening uptake policy recommendations. Updates to CRC screening guidelines by the American Cancer Society (ACS) and the U.S. Preventive Services Task Force (USPSTF), which now recommend initiating CRC screening at age 45, may be contributing to a modest increase in screening rates among individuals aged 45-49, particularly in Groups 2 and 3 (Atarere, 2024b). **Figure 1** reveals a gradual increase in testing rates within the 30-49 age bracket, which could indicate early adoption of these guideline shifts by healthcare providers and patients. However, screening rates for those under 50 remain low compared to older age groups, highlighting an area with significant potential for improving adherence.

In the 50-64 age group, screening rates reach their peak, confirming this as the primary age range for CRC screening. As shown in **Figure 1**, Group 2 leads with a screening rate of 43.88%, followed closely by Group 1 at 40.92% and Group 3 at 38.06%. This pattern reflects the age group's alignment with recommended screening ages and highlights the effectiveness of CRC screening initiatives targeted toward this demographic. Recent studies indicate that high adherence in this group is often driven by regular healthcare provider recommendations and routine medical care access (Atarere, 2024b; Wu, 2022).

Among older adults aged 65-74, screening rates remain substantial, likely due to Medicare coverage, which facilitates preventive screenings. **Figure 1** illustrates that Group 3 has the highest adherence at 31.58%, followed by Group 1 at 28.62% and Group 2 at 25.09%. The sustained adherence in this age group is likely bolstered by Medicare's support for routine screenings, as well as a focus on preventive health among this population (Atarere, 2024a). Group 3's elevated rate could reflect improved preventive health practices among the cohort or more proactive Medicare utilization.

In the 75+ age group, CRC screening rates drop markedly, which aligns with current guidelines that discourage routine screening for older seniors due to increased procedural risks and lower expected benefit. In **Figure 1,** Group 1 has the highest rate at 24.23%, followed by Group 3 at 20.31% and Group 2 at 17.73%. This decline may reflect adherence to recommendations, though the relatively higher rate in Group 1 suggests that some seniors continue screening based on individual decisions or physician recommendations, despite standard guidelines (Atarere, 2024c).

Overall, **Figure 1**, **Table 1,** and **Appendix A** collectively underscore age as a critical factor in CRC screening uptake, with the highest adherence observed in the 50-64 and 65-74 age groups. The variations across groups reveal stronger adherence in the 50-64 range for Group 2 and the highest rates in the 65-74 range for Group 3, highlighting the role of Medicare in supporting screening behaviors. These findings suggest that CRC screening is highly age-dependent, following established guidelines, with **Figure 1** clearly depicting age-specific screening trends. This evidence points to a need for targeted strategies to improve screening rates among younger adults, who remain under-screened.

Digital Characteristics Among Three Groups

The findings in **Table 2** reveal evolving trends in digital connectivity, device ownership, and access to health information resources among CRC-screened individuals across three HINTS groups. Internet usage has significantly increased over time, with Group 3 showing the highest rate at 76.73%, compared to 69.39% in Group 2 and 54.65% in Group 1, indicating a growing reliance on the internet for health-related information and resources. Among connection types, broadband, Wi-Fi, and mobile internet use have also expanded in recent groups, with mobile access (“Cell” in **Table 2**) notably rising from 21.94% in Group 2 to 39.54% in Group 3, underscoring a shift toward mobile connectivity and more flexible access to health information.

Device ownership also demonstrates upward trends, particularly in smartphone ownership, which reached 71.5% in Group 3, and tablet ownership, rising from 9.47% in Group 2 to 53.6% in Group 3. This increased adoption of mobile and digital devices likely supports easier access to health resources, potentially influencing CRC screening behaviors. Access to electronic health information similarly improved across groups, with 65.03% of CRC-tested individuals in Group 3 accessing digital health resources, up from just 15% in Group 1, reflecting an increase in both digital engagement and the availability of electronic health information systems in recent years. Social media engagement saw a dramatic rise as well, from 13.44% in Group 2 to 57.17% in Group 3, suggesting an emerging role of social media in health information dissemination and community support for CRC-screened individuals.

Table 2: Digital Factors Comparison Across Three HINTS Groups - CRC Screening Tested.

Variable	Category	Group 1 Tested 4326 (100%)	Group 2 Tested 5269 (100%)	Group 3 Tested 7701 (100%)
Use Internet	Yes	2364 (54.65)	3656 (69.39)	5909 (76.73)
Internet Type	Dial-Up	n/s	272 (5.56)	166 (2.16)
	Cell	n/s	1156 (21.94)	3045 (39.54)
	Broadband	n/s	2551 (48.42)	1915 (24.87)
	Wi-Fi	n/s	1968 (37.35)	4221 (57.41)
Where Use Internet	Home	1630 (37.68)	n/s	3323 (43.15)
	Work	78 (1.8)	n/s	1535 (19.93)
	Public Place	n/s	n/s	33 (0.43)
	Mobile	n/s	n/s	3286 (42.67)
Electronic Health Info	Accessed	649 (15)	2860 (54.28)	5008 (65.03)
Health Device Owner	Tablet	n/s	499 (9.47)	4128 (53.6)
	Smartphone	n/s	633 (12.01)	5506 (71.5)
Social Media	Visited	n/s	708 (13.44)	4403 (57.17)
Electronic Medical Record	Maintained	1225 (28.32)	4684 (88.9)	6543 (84.96)
Provider Access Online	Offered	n/s	n/s	3374 (43.81)
Health Tracking Devices	Wearable	n/s	n/s	1222 (15.87)

Healthcare provider-driven digital health initiatives also saw substantial growth. By Group 3, 84.96% of respondents reported that their provider maintained EHRs, and 43.81% had online access to their health records, indicating expanding digital infrastructure that may be contributing to preventive health engagement, including CRC screening. Additionally, wearable health-tracking devices were more commonly used in Group 3, with 15.87% reporting usage, which suggests an increasing emphasis on self-monitoring and preventive actions that could positively impact CRC screening adherence.

However, an important limitation is the presence of several "n/s" (not surveyed) entries, indicating that specific questions were omitted in certain HINTS cycles. This inconsistency limits our ability to conduct longitudinal comparisons for some digital factors, such as types of internet connections, usage locations, and wearable health-tracking device adoption. While trends are apparent, these gaps suggest caution when interpreting results as fully representative across all cycles. This limitation emphasizes the need for more consistent data collection in future cycles to enable comprehensive trend analysis over time.

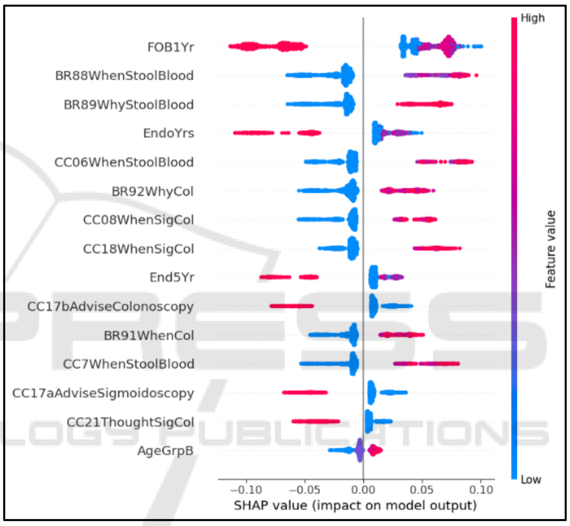

Figure 2a: SHAP graph of Group 1 (HINTS123).

4.2 Feature Selection – Critical Factors Influencing CRC Screening Uptake

Overall Features Among Three Groups

The SHAP analysis across Groups 1, 2, and 3 highlights the most influential factors impacting CRC screening adherence over time. In Group 1 (2003–2008), recent fecal occult blood test (FOBT) behavior (FOB1Yr) emerged as the most impactful feature, with individuals who had previously undergone FOBT screening being significantly more likely to adhere to CRC screening recommendations. This aligns with the importance of reinforcing preventive behaviors through past engagement. Related features, such as stool blood test timing (BR88WhenStoolBlood, BR89WhyStoolBlood), and endoscopic procedure timing (EndoYrs), also showed substantial contributions, emphasizing the role of recent screening experiences and provider recommendations in fostering adherence during this earlier period. Demographic factors like age (AgeGrpB) and socioeconomic elements, while included, played a less dominant role. **Figure 2a** shows 15 important features from Group 1. Extended Features for further analysis are included in **Appendix B**.

In Group 2 (2011–2013), age (AgeGrpB) became the most significant predictor, reflecting the growing adherence to age-specific screening guidelines. The importance of healthcare provider discussions about CRC screening (DrTalkColCaTest) was also prominent, highlighting the critical role of provider-patient communication in increasing screening uptake. Preventive health behaviors, such as mammogram participation (WhenMammogram) and PSA test history (EverHadPSATest), emerged as secondary influencers, suggesting that individuals engaged in other preventive health measures were more likely to comply with CRC screening. Notable contextual features included household composition (ChildrenInHH) and routine healthcare checkups (MostRecentCheckup), which indicated that family settings and regular medical care contributed to adherence during this period. **Figure 2b** shows 15 important features from Group 2.

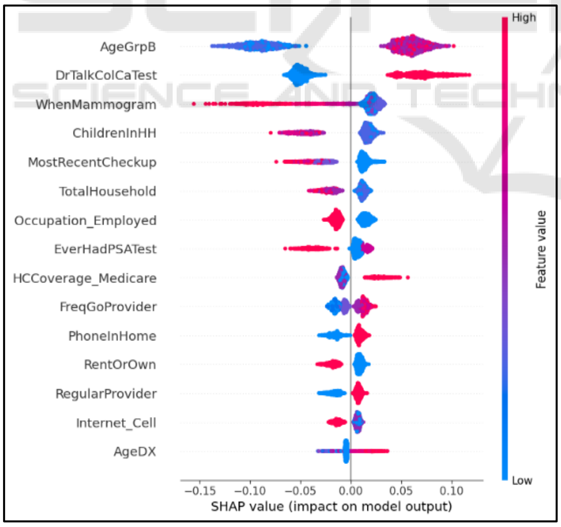

Figure 2b: SHAP graph of Group 2 (HINTS4).

In Group 3 (2018–2020), the relative importance of digital health literacy and access to healthcare resources became increasingly evident. Age (AgeGrpB) remained the most significant predictor, but health insurance coverage through Medicare (HealthIns_Medicare) and preventive health behaviors, such as mammograms (WhenMammogram), gained prominence. Features related to chronic health conditions (MedConditions_HighBP) and consistent provider relationships (RegularProvider) also demonstrated meaningful contributions. Moreover, digital engagement variables like internet use through mobile devices (WhereUseInternet_MobileDevice) and electronic medical record (EMR) maintenance by providers (ProviderMaintainEMR) highlighted the increasing role of digital tools in influencing screening behaviors. These shifts reflect the growing integration of digital health technologies and access disparities into screening decision-making. **Figure 2c** shows 15 important features from Group 3.

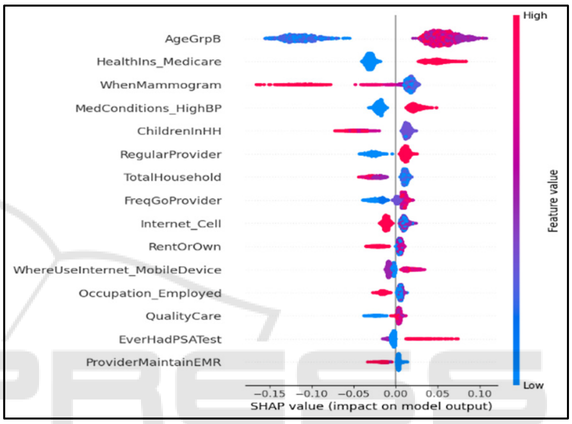

Figure 2c: SHAP graph of Group 3 (HINTS5).

The SHAP analysis underscores the evolving predictors of CRC screening adherence, with a shift from prior screening behaviors and demographic factors in earlier groups (Group 1) to greater emphasis on healthcare provider interactions, preventive health engagement, and digital health access in later groups (Groups 2 and 3). This evolution highlights the importance of adapting public health strategies to leverage digital tools and target demographic disparities while reinforcing the role of consistent healthcare provider engagement in promoting screening adherence.

Digital Features Among Three Groups

In order to focus on critical digital factors, **Figure 3** provides a comparative overview of key digital engagement trends across three HINTS groups, highlighting significant shifts in digital health utilization over time. Our findings include:

- Mobile Device and Internet Usage: " MobileDevice" and "Internet_Cell" consistently rank high in importance across all groups, reflecting a growing reliance on mobile and cellular internet access for health information and engagement across Group 1 (HINTS123), Group 2 (HINTS4), and Group 3 (HINTS5).

- Social Media Engagement: "SocMed_Visited" shows persistent significance across all groups, underscoring social media's role as a major platform for health-related digital interactions, particularly for sharing and seeking health information.
- Device Ownership: Variables like "HaveDevice_SmartPh" and "Tablet_AchieveGoal" score higher in Group 3 (HINTS5) and Group 1 (HINTS123), suggesting a substantial increase in smartphone and tablet use for health purposes, which supports broader digital access and convenience in recent years.
- Provider Digital Interactions: "ProviderMaintainEMR" and "HCPEncourageOnlineRec" reach their highest scores in Group 3 (HINTS5), indicating a strengthened focus on provider-supported digital health tools, particularly electronic medical records (EMR), demonstrating deeper integration of digital interactions within healthcare over time.
- Electronic Health Information Access: "Electronic_HealthInfo" and "Electronic_TalkDoctor" have significantly higher scores in recent groups, particularly in Group 3 (HINTS5) and Group 2 (HINTS4), reflecting an upward trend in patients' access to electronic health information and digital communication with healthcare providers.

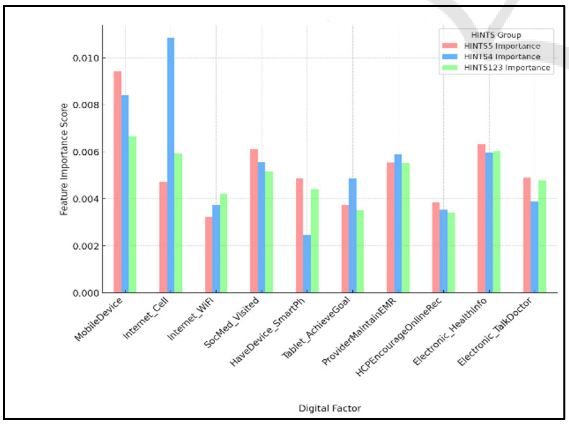

Figure 3: Comparison of Critical Digital Factors Across three HINTS groups.

Overall, **Figure 3** illustrates a progressive shift toward digital health resources across the HINTS groups, marked by increasing mobile internet usage, social media engagement, provider-supported digital tools, and enhanced electronic health information access. This trend underscores the expanding role of digital tools in facilitating health engagement and access over time, and subsequent analyses will focus on these key digital factors to evaluate their impact on health behaviors and outcomes.

4.3 Comparison of the Machine Learning Models' Performance for CRC Screening Uptake Prediction

This analysis evaluates the performance of LR and RF models across three HINTS datasets—HINTS123, HINTS4, and HINTS5—by examining metrics such as accuracy, precision, recall, F1 score, and AUC. Each dataset represents unique characteristics that challenge the models differently, providing insight into the models' suitability for various data complexities.

Table 3: Comparison of LR and RF Models on HINTS Datasets.

Group	Model	Pr	Re	F1	AUC
Group 1 HINTS123	**LR**	99.55	99.1	99.32	99.41
	RF	97.55	95.93	96.74	97.23
Group 2 HINTS4	**LR**	81.2	83.14	82.16	81.84
	RF	80.26	85.41	82.75	82.08
Group 3 HINTS5	**LR**	86.45	82.27	84.31	80.54
	RF	83.75	90.92	87.18	80.95

In Group 1, using the HINTS123 dataset, LR performed exceptionally well across all metrics, outstripping RF. LR achieved an accuracy of 99.49, a precision of 99.55, and a recall of 99.1, indicating that it could classify instances with remarkable accuracy and minimal error. Its F1 score of 99.32 and AUC of 99.41 further underscore its capability in distinguishing between classes effectively. In contrast, while RF also showed strong performance, its lower recall (95.93) and AUC (97.23) metrics indicate that it was slightly less effective than LR in managing this dataset's characteristics.

In Group 2, with the HINTS4 dataset, RF had a slight advantage over LR, particularly in metrics such as recall and AUC. While both models performed comparably in accuracy—RF at 82.1 and LR at 81.85—RF excelled in recall, achieving 85.41 compared to LR's 83.14. This advantage in recall suggests that RF was more sensitive in identifying positive instances within this dataset, a key benefit for applications prioritizing true positive detection. RF's F1 score (82.75) and AUC (82.08) also outpaced

Logistic Regression, highlighting its suitability for more complex data structures like those found in HINTS4.

In Group 3, which used the HINTS5 dataset, RF continued to outperform LR, particularly in recall and F1 score, demonstrating its strength in identifying positive cases in this dataset. RF achieved an accuracy of 83.28 and a recall of 90.92, compared to Logistic Regression's accuracy of 80.96 and recall of 82.27. Additionally, the F1 score for RF was significantly higher at 87.18 compared to Logistic Regression's 84.31, indicating that RF maintained a better balance between precision and recall. This improved performance highlights RF's ability to capture nuanced, non-linear patterns in the HINTS5 dataset, making it a more suitable model for datasets with complex relationships.

In summary, LR showed exceptional performance in Group 1, making it ideal for datasets with straightforward, linear relationships like HINTS123. In contrast, RF proved advantageous in Groups 2 and 3, excelling in datasets with greater complexity, as seen in HINTS4 and HINTS5. These findings suggest that LR is highly effective for simpler datasets, while RF is better suited for complex datasets requiring high sensitivity and the ability to handle intricate, non-linear patterns. This comparison serves as a guide for model selection based on dataset characteristics and the importance of specific metrics such as recall or precision in future applications.

5 DISCUSSIONS

This study explored the impact of digital health literacy and socioeconomic factors on CRC screening behaviors across different time periods using data from the HINTS datasets. By applying machine learning models, we identified critical determinants influencing CRC screening adherence and uncovered variations in the influence of socioeconomic and digital health literacy factors over time.

Our analysis confirms several important trends in CRC screening behavior, as observed in prior studies, while also highlighting new insights specific to digital health engagement. Age, socioeconomic stability, and digital literacy emerged as consistent predictors of CRC screening uptake across the HINTS cycles. This indicates that while digital health interventions, such as patient portals and telehealth, have gained prominence, traditional demographic factors continue to play a substantial role in CRC screening adherence. The finding aligns with studies like Atarere et al. (2024a, 2024b, 2024c), which emphasize that while digital health tools may enhance adherence, addressing fundamental socioeconomic and demographic barriers is essential for achieving equity in CRC screening rates.

The SHAP analysis demonstrated that prior CRC screening behaviors, age, and patient-provider interactions were among the strongest predictors of screening adherence, especially in the earlier HINTS groups. This reinforces the idea that positive prior experiences with CRC screening and effective communication with healthcare providers are crucial in fostering long-term adherence to screening recommendations (Wu, et al.2023). Specifically, our results underscore the role of healthcare providers in reinforcing CRC screening messages, particularly for at-risk populations who may be less engaged with digital health tools.

Chronic conditions, including hypertension and diabetes, also play a role in shaping CRC screening behaviors across the HINTS groups. These conditions, often requiring routine medical attention, may increase patients' engagement with healthcare providers, creating additional opportunities for providers to recommend CRC screening as part of comprehensive preventive care. The consistent significance of variables such as "MedConditions_HighBP" and "MedConditions_Diabetes" across multiple groups indicates that individuals managing chronic illnesses may be more attuned to preventive health measures. Additionally, health coverage factors, particularly Medicare (represented by "HCCoverage_Medicare"), are associated with higher screening rates among those with chronic conditions, likely reflecting the expanded access to preventive services Medicare offers to older adults with chronic health needs.

Collectively, these insights suggest that psychological readiness, chronic condition management, and continuous healthcare engagement are influential in CRC screening decisions. Addressing psychological barriers, reinforcing supportive patient-provider communication, and leveraging routine chronic care visits for screening recommendations could help enhance screening adherence, especially among high-risk or less-engaged populations.

The influence of digital health literacy on CRC screening behaviors appeared most prominent in the later HINTS cycles, particularly HINTS 5 (2018-2020). This suggests that as digital health technologies become more integrated into routine healthcare, the ability to navigate these tools may increasingly shape preventive health behaviors. For instance, confidence in locating reliable health information online and regular telehealth usage were associated with higher screening adherence, indicating that digital health literacy is becoming an important factor in promoting preventive behaviors like CRC screening. These findings suggest a need for targeted interventions to enhance digital literacy among populations with historically low screening rates, such as rural communities and lower-income groups.

Our comparison of LR and RF models across different HINTS datasets highlighted that model performance varies with data complexity. LR outperformed RF on the HINTS123 dataset (2003-2008), likely due to the dataset's simpler, more linear structure, while RF excelled in HINTS4 and HINTS5 datasets, which introduced more complex, non-linear relationships as HIT variables became more prevalent. This performance variation indicates that non-linear models like RF may be more effective for analyzing recent datasets where digital health literacy factors play a larger role. Future studies aiming to incorporate digital health engagement factors should consider leveraging non-linear models to capture the nuanced behaviors associated with these variables.

This study has several limitations. First, HINTS data are based on self-reported responses, which may introduce reporting bias. Additionally, the focus on the U.S. population limits the generalizability of our findings to other regions where digital health adoption and socioeconomic structures differ significantly. The machine learning models, while effective in identifying predictive factors, may not fully capture the dynamic and complex interactions that influence health behaviors over time. Further studies could benefit from incorporating longitudinal data or using more advanced modelling techniques, such as neural networks, to capture temporal patterns in digital health engagement.

Future research should explore integrating EHRs with survey data to enhance the predictive accuracy of CRC screening models. Additionally, examining the role of social determinants of health, such as social support and community engagement, could provide a more holistic view of factors influencing CRC screening adherence.

6 CONCLUSIONS

CRC screening rates in the United States have shown improvements over time, yet significant disparities persist, particularly among underserved populations such as younger individuals, non-White racial groups, and those with lower socioeconomic status or limited digital health literacy. Leveraging machine learning models and SHAP analysis on HINTS data across three temporal groups revealed evolving predictors of CRC screening adherence. Key findings emphasize the persistent influence of traditional sociodemographic factors like age, income, and education, alongside the growing importance of digital health literacy and access to health technologies. Targeted interventions focusing on enhancing digital health engagement, improving access to preventive care, and addressing socioeconomic barriers are critical to bridging these disparities. These findings highlight the importance of integrating digital tools with equitable public health strategies to improve CRC screening uptake and reduce health inequities across diverse populations.

ACKNOWLEDGEMENTS

Funding: This publication was supported by the University of Kentucky College of Medicine Artificial Intelligence in Medicine Research Alliance and the Center for Clinical and Translational Science, which is funded by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

Atarere, J., Chido-Amajuoyi, O., Mensah, B., Onyeaka, H., Adewunmi, C., Umoren, M., & Kanth, P. (2024a). Primary care telehealth visits and its association with provider discussion on colorectal cancer screening in the United States. *Telemedicine and e-Health*, 30(5), 1325-1329.

Atarere, J., Haas, C., Akhiwu, T., Delungahawatta, T., Pokharel, A., Adewunmi, C., & Barrow, J. (2024b). Prevalence and predictors of colorectal cancer screening in the United States: Evidence from the HINTS database 2018 to 2020. *Cancer Causes & Control,* 35(2), 335-345.

Atarere, J., Haas, C., Onyeaka, H., Adewunmi, C., Delungahawatta, T., Orhurhu, V., & Barrow, J. (2024c). The role of health information technology on colorectal cancer screening participation among smokers in the United States. *Telemedicine and e-Health*, 30(2), 448-456.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.

CA: A Cancer Journal for Clinicians, 68(6), 394–424.

De La Garza, Á. G., Blanco, C., Olfson, M., & Wall, M. M. (2021). Identification of suicide attempt risk factors in a national US survey using machine learning. *JAMA Psychiatry*, 78(4), 398-406.

Finney Rutten, L. J., Hesse, B. W., Moser, R. P., McCaul, K., & Rothman, A. J. (2009). Public understanding of cancer prevention, detection, and survival/cure: Comparison with state-of-science evidence for colon, skin, and lung cancer. *Journal of Cancer Education*, 24(1), 40-48.

Ford, J. S., Coups, E. J., & Hay, J. L. (2006). Knowledge of colon cancer screening in a national probability sample in the United States. *Journal of Health Communication*, 11(Suppl 1), 19-35.

Geiger, T. M., Miedema, B. W., Geana, M. V., Thaler, K., Rangnekar, N. J., & Cameron, G. T. (2008). Improving rates for screening colonoscopy: Analysis of the health information national trends survey (HINTS I) data. *Surgical Endoscopy*, 22(2), 527-533.

Hay, J., Coups, E., & Ford, J. (2006). Predictors of perceived risk for colon cancer in a national probability sample in the United States. *Journal of Health Communication,* 11(Suppl 1), 71-92.

Idowu, K. A., Adenuga, B., Otubu, O., Narasimhan, K., Kamara, F., Hunter-Richardson, F., Larbi, D., Sherif, Z. A., & Laiyemo, A. O. (2016). Place of birth, cancer beliefs, and being current with colon cancer screening among US adults. *Annals of Gastroenterology*, 29(3), 336-340.

Jun, J., & Oh, K. (2013). Asian and Hispanic Americans' cancer fatalism and colon cancer screening. American Journal of Health Behavior, 37(2), 145-154.

Keum, N., & Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature reviews Gastroenterology & Hepatology*, 16(12), 713-732.

McIntosh, J. G., Jenkins, M., Wood, A., Chondros, P., Campbell, T., Wenkart, E., ... & Emery, J. D. (2024). Increasing bowel cancer screening using SMS in general practice: the SMARTscreen cluster randomised trial. *British Journal of General Practice*, 74(741), e275-e282.

Miller Jr, D. P., Denizard-Thompson, N., Weaver, K. E., Case, L. D., Troyer, J. L., Spangler, J. G., ... & Pignone, M. P. (2018). Effect of a digital health intervention on receipt of colorectal cancer screening in vulnerable patients: a randomized controlled trial. *Annals of Internal Medicine,* 168(8), 550-557.

Mirzaei, A., Aslani, P., & Schneider, C. R. (2022). Healthcare data integration using machine learning: A case study evaluation with health information-seeking behavior databases. *Research in Social and Administrative Pharmacy*, 18(12), 4144-4149.

Nawaz, H., Via, C., Shahrokni, A., Ramdass, P., Raoof, A., Sunkara, S., & Petraro, P. (2014). Can the inpatient hospital setting be a golden opportunity to improve colon cancer screening rates in the United States? *Health Promotion Practice*, 15(4), 506-511.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1), 7–33.

Sun, C., Mobley, E., Quillen, M., Parker, M., Daly, M., Wang, R., & Xu, J. (2024). Predicting early-onset colorectal cancer in individuals below screening age using machine learning and real-world data. medRxiv, 2024-07.

Wu, S., Zhang, X., Chen, P., Lai, H., Wu, Y., Shia, B. C., & Qin, L. (2022). Identifying the predictors of patient-centered communication by machine learning methods. *Processes*, 10(12), 2484.

Zhen, J., Li, J., Liao, F., Zhang, J., Liu, C., Xie, H., & Dong, W. (2024). Development and validation of machine learning models for young-onset colorectal cancer risk stratification. *NPJ Precision Oncology*, 8(1), 239.