An Artificial Intelligence Outlook for Colorectal Cancer Screening

Panagiotis Katrakazas
Zelus P.C.
Athens, Greece
ORCID ID: 0000-0001-7433-786X

Aristotelis Ballas
Dept. of Informatics and Telematics
Harokopio University of Athens
Athens, Greece
ORCID ID: 0000-0003-1683-8433

Marco Anisetti
Dipartimento di Informatica (DI)
Università degli Studi di Milano
Milan, Italy
ORCID ID: 0000-0002-5438-9467

Ilias Spais
Zelus P.C.
Athens, Greece
e-mail: ilias.spais@zelus.gr

***Abstract*—Colorectal cancer is the third most common tumor in men and the second in women, accounting for 10% of all tumors worldwide. It ranks second in cancer-related deaths with 9.4%, following lung cancer. The decrease in mortality rate documented over the last 20 years has shown signs of slowing down since 2017, necessitating concentrated actions on specific measures that have exhibited considerable potential. As such, the technical foundation and research evidence for blood-derived protein markers have been set, pending comparative validation, clinical implementation and integration into an artificial intelligence enabled decision support framework that also considers knowledge on risk factors. The current paper aspires to constitute the driving force for creating change in colorectal cancer screening by reviewing existing medical practices through accessible and non-invasive risk estimation, employing a straightforward artificial intelligence outlook.**

Keywords—colorectal cancer, screening, risk factor analysis, artificial intelligence approach

I. Introduction

Colorectal cancer is the third most common tumor in men and the second in women, accounting for 10% of all tumors worldwide. It ranks second in cancer-related deaths with 9.4%, following lung cancer. About 1.9 million new cases were diagnosed in 2020, translating into 0.9 million deaths, while incidence is projected to rise significantly over the next decade, with 3.2 million new diagnoses annually by 2040. In affected European Union (EU) individuals, 5-year survival ranges from 28.5% to 57% in men and from 30.9% to 60% in women, with pooled estimations in 23 countries of 46.8% and 48.4% respectively [1], [2]. Moreover, colorectal cancer is among the five most likely to metastasize cancers. Upon initial diagnosis, 22% of cases are metastatic, while about 70% of patients will eventually develop metastatic relapse [3].

In the colorectal cancer treatment domain, options include endoscopic and surgical excision, radiotherapy, immunotherapy, palliative chemotherapy, targeted therapy, extensive surgery and local ablative therapies for metastases [1]. Meanwhile, screening methods consisting of endoscopic tests (e.g., colonoscopy) and non-invasive alternatives such as the fecal occult blood test (FOBT) have been put into action. Applied pathways have successfully inhibited cancer progression [4] contributing to decreased mortality rates through 2017. Moreover, some EU countries have adopted population-based screening programs over the last 15 years, seeking to halt incidence and mortality rates. In this regard, studies have compared mortality rates for symptom-detected vs screening-detected colorectal cancer, stating the considerable impact of screening via quantified reduction estimates surpassing 30% for screening-based detections [5]. Particularly, 5-year survival rate can reach 90% for stage I diagnosis, being less than 15% for advanced stages [6]. Therefore, routine screening is key for reducing mortality and declining incidence rates, since colorectal cancer is now considered a highly preventable disease with a considerably wide temporal development window [7]. Namely, the transitional path from normal mucosa to pre-malignant growth and then to malignant lesion might spread over 15 to 20 years, with scientists seeking means for earlier, cost-effective and less taxing detection of pre-malignant states.

In determining the colorectal cancer risk status, factors such as age, body mass index, diet, smoking habits and family history have been pinpointed by researchers and clinicians alike [8]. Namely, age, sex and family history have been integrated into practice as flagships on risk stratification [2], although evidence on the complete risk factor set has not been analyzed in the context of a detailed assessment. Indicatively, Western registry data show an increased incidence in the age group of 40-44, considerably lower than the 50-year threshold [9]. This tendency is attributed to modern lifestyle alterations, although more assays are required as to the corresponding effects. Overall, despite the long-assumed colorectal cancer preventability based on modifiable risk factors, awareness and knowledge exploitation remain extremely low.

Moreover, despite the available arsenal of screening practices, citizen participation is hindered due to suboptimal performance and invasive or overall taxing nature besetting these methods [10]. EU reports indicate participation rates of 14%, a rather disappointing number compared to the >60% rate for breast cancer screening programs [11]. Poor screening outreach is augmented by limited penetration of Council recommendations into clinical practice. By 2019, only three member states had adopted population-based screening targeting all citizens at risk, albeit purely based on age thresholds of 50-74 years. Over the last three years, other states have also launched population-based screening or regional programs. Still, there is little progress on standardized programs that unite knowledge on colorectal cancer towards EU-wide regulations. Overall, taxing procedures, citizen reluctance, poor awareness and screening accessibility are hindering participation, forcing researchers into the survey of accessible, non-invasive biomarkers that bear the potential to render cancer screening less burdensome and more accessible to citizens. In this vein, liquid biopsy appears to be a promising new tool on non-invasive, quick and safe assessment [6]. However, lack of operating protocols, reproducibility issues and cost-effectiveness discrepancies are impeding clinical implementation. Among all liquid biopsy products, blood-derived proteins seem to constitute the most cost-effective solution regarding resources, sensitivity and

978-1-6654-5890-0/22/$31.00 ©2022 IEEE
DOI 10.1109/BigDataService55688.2022.00018

research maturity. On this premise, a vast protein pool has been tested, albeit evidence lacks comparative validation, perplexing standardization margins. The current research highlights (a) the existing practices in colorectal screening methodology along with a comparison matrix of their benefits and limitations, along with (b) a risk factor analysis that can be feasibly exploited for population-based screening and sustainably covered by health insurance bodies, via an artificial intelligence (AI)-enabled suggestion

II. Standard of Care on Colorectal Screening

Early cancer detection through effective screening constitutes the hallmark for providing patients and clinicians with the best possible "fighting chance", more so for colorectal cancer, a highly preventable cancer with long (>10 years) pre-malignant underlying processes. Specifically, the colorectal cancer field has been provided with a variety of screening tools [12], [13], which can be divided into two main groups (Table I). Each method bears specific benefits and limitations with regard to invasiveness, sensitivity, specificity, preparation, sampling process, risks and costs (Table 2).

TABLE I. Colorectal Cancer Screening Tests

Group	Description	Tests	Assessment
Stool-Based	They are simpler and less invasive than structural tests, however they require frequent reassessment	Fecal immunochemical test (FIT)	Every year
		guaiac-Based Fecal Occult Blood Test (gFOBT)	
		Multi-targeted stool DNA test (mt-sDNA)	Every three years
Structural	They scan colon and rectum structure for abnormal areas. They are invasive tests conducted either with a scope inside the rectum or via special imaging.	Flexible Sigmoidoscopy (FSIG)	Every five years
		Computed Tomography (CT) Colonography	
		Colonoscopy	Every ten years

TABLE II. Colorectal Cancer Screening Tests

Test	Benefits	Limitations
FIT	○ *No direct risk to the colon* ○ *No bowel preparation* ○ *No pre-test diet or medication changes* ○ *At-home sampling* ○ *Fairly inexpensive*	– *Can miss many polyps and some cancers* – *Potential false-positive results* – *Needs annual reassessment* – *Requires colonoscopy if abnormal*
gFOBT	○ *No direct risk to the colon* ○ *No bowel preparation* ○ *At-home sampling* ○ *Inexpensive*	– *Can miss many polyps and some cancers* – *Potential false-positive results* – *Pre-test diet changes and medication changes* – *Needs annual reassessment* – *Requires colonoscopy if abnormal*
mt-sDNA	○ *No direct risk to the colon* ○ *No bowel preparation* ○ *No pre-test diet or medication changes* ○ *At-home sampling*	– *Can miss many polyps and some cancers* – *Potential false-positive results* – *Needs reassessment every three years* – *Requires colonoscopy if abnormal*
FSIG	○ *Fairly quick and safe* ○ *Usually doesn't require full bowel prep* ○ *Sedation usually not used* ○ *Does not require a specialist* ○ *Performed every five years*	– *Not widely used as a screening test* – *Examines only about 1/3 of the colon* – *Can miss small polyps* – *May cause some discomfort* – *Small risk of bleeding, infection or bowel tear* – *Requires colonoscopy if abnormal*
CT Colonography	○ *Fairly quick and safe* ○ *Can usually examine the entire colon* ○ *Performed every five years* ○ *No sedation needed*	– *Can miss small polyps* – *Full bowel prep needed* – *Potential false-positive results* – *Exposure to radiation* – *Inability to remove polyps during testing* – *Requires colonoscopy if abnormal*
Colonoscopy	○ *Can usually examine the entire colon* ○ *Can include biopsy and polyp removal* ○ *Performed every ten years*	– *Can miss small polyps* – *Requires full bowel preparation & (usually) sedation* – *More expensive than other tests* – *Small risk of bleeding, bowel tears or infection*

Evidently, in the context of colorectal cancer screening, the medical community is faced with the paradox of a dedicated screening procedure pool that is heavily under-utilized by citizens. Although colonoscopy is recommended as the current gold standard, the long preparation and recovery procedures (~24 hours) as well as the potential adverse effects (colon tears, diverticulitis, abdominal pain, and risks in cardiovascular conditions) induce heavy reluctance for individuals to include colonoscopy in their regular check-up routine. Moreover, all existing Standard-of-Care (SoC) screening tests can miss small polyps, thus failing to detect potential pre-malignant lesions.

III. State-of-the-Art on Risk Factor Analysis

Numerous studies have investigated the association of colorectal cancer incidence with demographic, behavioral, and environmental risk factors including age, sex and lifestyle. For instance, a 25% higher incidence has been documented for males, varying among countries [2], while even race has been highlighted as a noteworthy parameter [14]. Overall, age comprises the main factor assessed by current guidelines, formulating at-risk groups for recommended screening. Guidelines suggest screening after 50 years, with healthy citizens advised to pursue regular testing through the age of 75. For people aged 76-85, guidelines are based on overall health and prior history, while beyond this range there are no strict recommendations. However, clinical practice has shown that these thresholds are gradually decreasing, a fact under investigation by the medical community. This status creates a pressing need for alternative practices beyond age-only recommendation onset. Currently, the only such criteria correspond to medical history, family history or symptom manifestation. Indicatively, a meta-analysis of observational studies found that having at least one affected first-degree relative (parents, siblings, or children) increased the risk of colorectal cancer by 2.2-fold and having at least two affected first-degree relatives increased the risk of colorectal cancer by almost 4-fold [15]. Moreover, patients with persistent Inflammatory Bowel Disease (IBD) are twice as likely to acquire colorectal cancer. Inflammation causes aberrant growth cytokines to be released, as well as increased blood flow, metabolic free radicals, and other variables that contribute to carcinogenesis [16]. Regarding symptoms, colorectal cancer may not cause any right away, thus constituting a particularly threatening cancer that necessitates early reliable screening. Alarming symptoms include rectal bleeding or changes in bowel habits such as diarrhea, constipation, or stool narrowing. In any other case, an individual is assigned with a “higher-than-average” risk status under the presence of one (or more) of the following:

- Personal history or family history of colorectal cancer or certain types of polyps
- Personal history of inflammatory bowel disease (ulcerative colitis or Crohn’s disease)
- Confirmed or suspected hereditary cancer syndrome (2%–5% of all colorectal cancers), such as familial adenomatous polyposis coli and its variants (1%), Lynch-associated syndromes (hereditary non-polyposis colon cancer) (2%–4%), Turcot, Peutz–Jeghers and MUTYH-associated polyposis syndrome
- Personal history of radiation treatment on the abdomen or the pelvic area for a prior cancer

The above elements comprise unalterable facts regarding early detection or even cancer prevention. On the other hand, several lifestyle-related factors have been identified, which are modifiable through suitable behavioral screening and personalized interventions. In fact, the links of diet, weight and exercise to colorectal cancer risk are some of the strongest among all cancer types. For example, being overweight raises both the incidence and the mortality risk for colorectal cancer in both men and women, but the link seems to be stronger in men. Body-mass index (BMI) and waist circumference (WC) are well-established risk factors for colorectal cancer, as evidenced by epidemiological research employing a variety of anthropometric measurements [17]. By extension, physical activity seems to constitute a key factor with evidence not favoring a sedentary lifestyle, as is the case for dietary habits. Namely, a diet that is high in red meats (such as beef, pork, lamb, or liver) and processed meats (like hot dogs and some luncheon meats) is assumed to raise colorectal cancer incidence risk. Even cooking-related processes seem to play a part, with very high temperatures (frying, boiling, or grilling) generating chemicals that might raise the associated risk. Similarly, lifestyle habits like smoking or alcohol consumption are linked to colorectal cancer incidence [18]. Although smoking is a well-known factor for lung cancer, research has displayed association with additional malignancies. Similarly, heavy alcohol use may cause a number of significant health-related outcomes, with colorectal cancer bearing a connection. By extension, all these modifiable factors have the potential to be addressed via lifestyle interventions promoting healthy behaviors including physical activity, BMI control, appropriate eating/cooking habits and refraining from smoking or excessive alcohol. On the whole, although a large pool of risk factors has been assumed to correlate with colorectal cancer (among other malignancies as well), the underlying regulatory processes remain largely unknown. The key to quantifying the corresponding transition mechanisms might lie into specific biomarkers that are extracted in a minimally invasive and cost-effective manner. Indicatively, it is a well-known fact that dietary habits are translated into alterations in routine blood biomarkers, such as the level of vitamin D in the blood, directly associated with eating patterns [19]. However, such phenotype manifestations have not been analyzed within a colorectal cancer-centered protocol. A handful of studies have assessed some indicative biomarkers, albeit without proposing an interpretation model for behavior effects on biomarker alteration and colorectal cancer incidence [20], [21].

IV. The Potential of AI in Colorectal Cancer Screening

Generally, outcome prediction has been the hallmark regarding AI utilization on cancer. Indicatively, AI models have attempted to predict 5-year colorectal cancer recurrence risk, outperforming grading and/or staging evaluation by expert pathologists. As far as screening is concerned, AI has been implemented for automated decision making in combination with various studied screening procedures [7]. Relying on the duration of the gradual transition path from normal mucosa to a premalignant growth and then to a malignant lesion spanning over 15-20 years, AI can detect suspect changes corresponding to abnormal tissue, which may be indicative of either a premalignant precursor lesion or an early-stage tumor. In particular, a high adenoma detection rate (ADR) has been validated as inversely correlated with adenoma miss rate and the risk of post-colonoscopy colorectal cancer), with every 1% increase in ADR corresponding to a 3% reduction in colorectal cancer development risk and 5% reduction in fatal colorectal cancer incidence [22]. However, ADRs may range from 7% to 53% among different endoscopists, creating the demand for objective and reproducible assessment towards attaining a robust ADR in clinical settings. In this regard, convolutional neural networks (CNNs) have been found to accurately detect and localize premalignant lesions on imaging data. In a prospective randomized trial under controlled conditions [23] with conventional vs computer-aided colonoscopy detection, ADR was significantly increased in favor of computer-aided detection (CAD). Similarly, virtual colonoscopy paired with machine learning (ML) modules was able to distinguish benign and precancerous colorectal polyps in an average-risk asymptomatic colorectal cancer screening samples with a sensitivity of 82% and specificity of 85%.

In the field of liquid biopsy, AI impact has been examined with the use of supervised learning methods such as support vector machines (SVMs). Hierarchical classification has shown feasible applicability, with an initial classification level filtering out non- colorectal cancer samples and a subsequent level differentiating between non-malignant lesions and a “no findings” class [24]. Additionally, blood biomarkers have been considered for use, alongside risk factors (e.g., demographic), within an AI-based framework for estimating cancer risks. These biomarkers were not limited in complex resource-demanding analytics, but included standard blood test markers as well, such as red cell distribution width (RDW) and anemia findings [25], [26] . Further knowledge can be drawn from electronic health records (EHR), thus incorporating screening in primary healthcare settings.

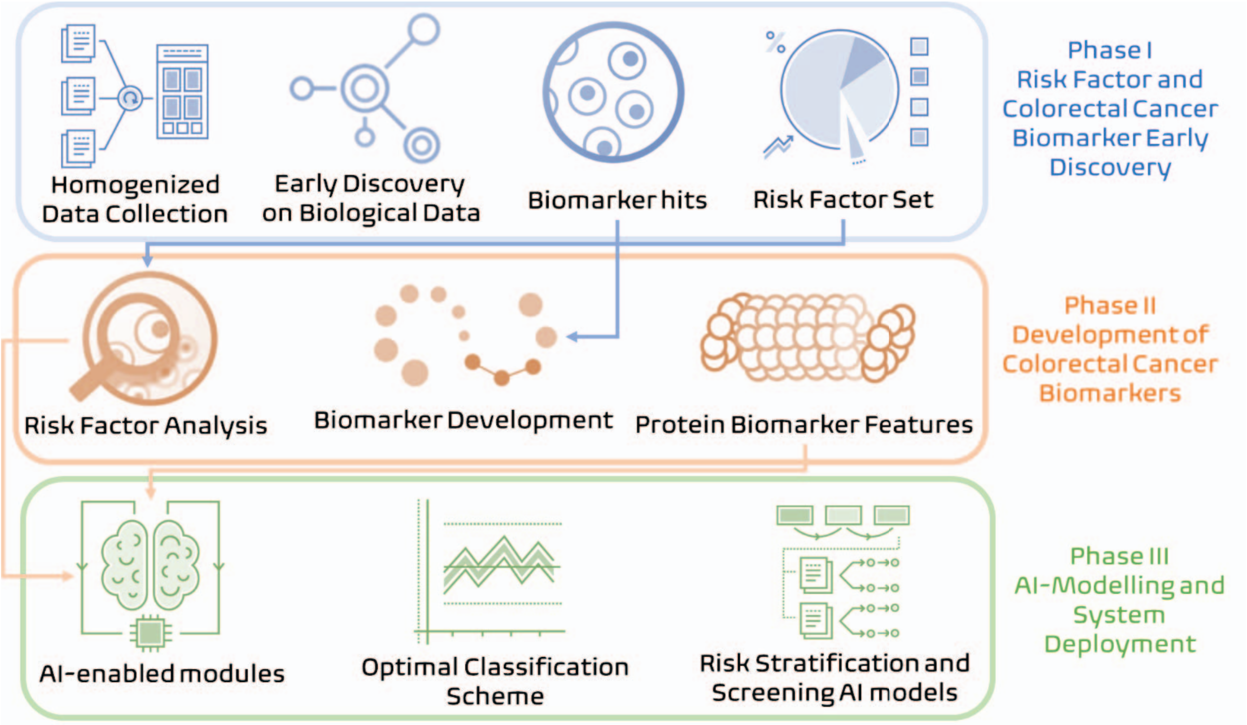
Homogenized Data Collection
Early Discovery on Biological Data
Biomarker hits
Risk Factor Set
Phase I Risk Factor and Colorectal Cancer Biomarker Early Discovery
Risk Factor Analysis
Biomarker Development
Protein Biomarker Features
Phase II Development of Colorectal Cancer Biomarkers
AI-enabled modules
Optimal Classification Scheme
Risk Stratification and Screening AI models
Phase III AI-Modelling and System Deployment

Figure 1 Suggested Methodological Architecture of the AI-based solution

Notably, combining such models with gFOBT seems to contribute to more than 2-fold increase in cancer detection capability for retrospective data cohorts. Based on this information pool, decision trees can be constructed for producing risk stratification scores and proposing clinical pathways. Finally, with regard to the Cancer of Unknown Primary (CUP) case, AI has been used with known metastatic colorectal cancer cases as training data, attempting subsequent testing and validation on both metastatic colorectal cancer and CUP data [27].

AI-powered virtual assistants can also provide personalized healthcare services and improve communication between patients and care providers [7]. By this principle, AI-based mobile applications such as the Colorectal Cancer Awareness Application (ColorApp®) [28] have sought to foster community education and participation in screening programs, achieving a score of 72 in the System Usability Scale Questionnaire for the Assessment of Mobile Apps. However, there is a lack of regional adaptations and performance surveys beyond usability scores.

In general, despite the above AI advances in colorectal cancer risk and progression assessment, the medical community is still skeptical and reluctant in trusting the outcomes of machine learning methods. This is mainly due to the depth of most neural network approaches, which are regarded as "black boxes", along with their confusing architecture. Furthermore, the majority of deep learning models fail to generalize adequately on previously 'unseen' data [29]. Since medical practices and methods are often altered and updated, AI models are expected to maintain their performance when evaluated on data that are sampled from a similar but different distribution than that of the model's training data (e.g., datasets across hospitals or signals captured by different devices). To this end, several Domain Generalization (DG) [30]–[32] methods have been proposed. DG methods attempt to push AI models to extract invariant features present in the data and therefore be less affected by the potential distributional shift in different datasets. In addition, explainable artificial intelligence (XAI) is gradually becoming a prerequisite by clinicians and policy makers, seeking to instill accountability and medical transparency into AI-assisted decisions for launching trustworthy clinical.

V. Suggestion of an AI-based Solution

Under these suggestions, AI can be exploited in the context of assessing an extended risk factor pool compared to current limited clinical consideration and by creating a clinical decision support for stratifying high-risk cases as an output from a front-line screening tool to the main evaluation phase. Figure 1 depicts the suggested approach.

Organization of knowledge into a standardized format, can generate a homogenized dataset. This will be utilized where risk factor investigation will take place, producing a holistic colorectal cancer risk factor set and a set of biomarker 'hits' to be further evaluated during a subsequent analytics step. This phase will in turn generate protein features for training and evaluating an AI model. In the latter phase, risk factor and protein features will undergo an AI-based analysis for generating a proposed subset of discriminative protein biomarkers as well as a full stratification model utilizing a hybrid approach based on the combination of biomarker and risk factor phenotypes.

A. Phase I: Risk Factor and Colorectal Cancer Biomarker Early Discovery

Demographic, behavioral, environmental, medical and history factors will be filtered from available data based on expert knowledge and existing research for determining the set that will be used during Phase III in the context of AI model training. In cases where retrospective data are deemed insufficient for AI training regarding individual parameters, those will be incorporated in the subsequent protocol, pending investigation and verification of their role based on prospective data. Regarding colorectal cancer biomarkers from blood samples, the necessary activities following the traditional biomarker discovery pipeline can be conducted by clinical entities, from early discovery and verification to clinical validation. Phase I corresponds to the early discovery phase, involving the identification of "hits" by searching thousands of proteins in limited number of samples that are present in blood serum/plasma using Proximity Extension Assay (PEA) by Olink, utilizing Next Generation Sequencing (NGS) methods to measure the concentration of thousands of human plasma proteins from small sample volumes [27]. PEA uses matched pairs of antibodies attached to unique DNA nucleotides for each protein to produce NGS readouts for the parallel quantification of up to 3000 proteins.

NGS readouts will be analyzed using statistical methods to identify a projected panel of ~30 protein markers that are representative of colorectal cancer incidence, showing a significant difference between healthy and diseased blood samples. However, due to the complex nature of cancer and the lack of mechanistic insight of proteome-wide or large-scale discovery screens, most of the identified biomarkers may still have poor diagnostic performance and clinical outcome. On that account, early discovery will be followed by main biomarker development in Phase II, exploiting an extended pool of retrospective as well as prospective data. Phase II will provide the complete list of colorectal risk factors to be included in the AI modelling process for risk stratification, behavioral monitoring, and interventions.

B. Phase II: Development of Colorectal Cancer Biomarkers

As stated above, due to the complex nature of cancer and the lack of mechanistic insight of proteome-wide or large-scale discovery screens, the majority of identified biomarkers can still have poor diagnostic performance and clinical outcome. This is evident by the limited translation of candidate biomarkers into clinical diagnostic assays, with less than two overall approvals per year [28]. Thus, a mechanism-based approach using multi-omics data is required to verify the role of the identified serum biomarkers in the complex colorectal cancer mechanism and select the most promising ones for clinical validation [33]. Recent advances in systems-based approaches have led to the development of interdisciplinary methods that utilize multi-omics data and network-based technologies to elucidate the complex disease mechanism [34]. More specifically, transcriptomic and proteomic data from patients' tissue samples are coupled with knowledge bases of bio-molecular interactions to create in-silico models that best explain experimental data and are representative of the disease mechanism(s) [35]. These models are often represented as signaling networks, with nodes being the signaling proteins and edges depicting the directed flow of information in the system.

Using pathway and network analysis methods, these models can be analyzed to identify the affected biological pathways and central protein nodes that are characteristic of the colorectal cancer mechanism. Being a representation of the colorectal cancer mechanism, these models can be applied to verify the connection and importance of the blood biomarkers identified during the early discovery phase. The result is a selection of verified biomarkers, present in blood, with mechanistic insight and improved performance for the early detection of colorectal cancer. Finally, the identified panel of biomarkers can be developed into highly specific and cost-effective multiplex proteomic assays to provide powerful features for an AI-based detection system.

A state-of-the-art method for multiplex diagnostic assay development involves the use of bead-based immunoassays. Multiplex assays utilize the xMAP technology (Luminex Corp) that relies on color-coded microspheres (bead regions) to allow for the simultaneous detection of responses against multiple protein targets from the same sample. Each bead region is coated with an antibody that recognizes and binds to a specific part of the protein. Mixtures of bead regions are used in a sandwich-type Enzyme-Linked Immunosorbent Assay (ELISA) assay to provide relative and absolute quantification of multiple proteins across the various conditions tested. These assays offer high multiplexability, sample throughput, quality of measurements and specificity for measurement of identified biomarkers in blood. Multiplex readouts from the developed assays can then be combined with logic-based and AI-based computational models to select the optimal combination of biomarkers that maximizes their detection and diagnostic performance [36]. In essence, this step constitutes the main proteinic feature selection process for AI modelling of Phase III towards the identification of the biomarkers associated with optimal screening capacity. Outputs of Phase II will equip with protein features the AI analysis aiming to pinpoint the most significant biomarkers bearing biological explainability through the demonstration of their role within colorectal cancer regulatory network paths.

C. Phase III: AI-Modelling and System Deployment

Features pinpointed during Phases I & II (risk factors & proteinic markers) will comprise the necessary feature vector for training and testing the AI-enabled modules. Initially, a pool of state-of-the-art AI methodologies, encompassing supervised learning, unsupervised learning and reinforcement learning for operating on retrospective datasets will be deployed. Parametric and non-parametric techniques will be incorporated with network-based techniques also included, maintaining interpretability by individually testing feature subsets. As such, application of multiple frameworks in pursue of the optimal classification scheme, testing both individual algorithms and ensemble schemes (e.g., majority-vote models) within various combinations and architectures can be pursued. Moreover, enhancing interpretability and transparency can be offered for avoiding class exclusion for cases close to the decision boundaries for colorectal cancer risk. This can be achieved by exploration of probabilistic and fuzzy classification, both for directly producing quantified estimates and for providing clinicians with comparative class predictions. This pursue will remain in the context of deep neural networks (DNNs), that bear high performance, although the multitude of layers and lack of understanding on prediction derivation render explainability a challenging task [37]. To address this, dedicated techniques for investigating the models' decisions such as SHapley Additive exPlanations (SHAP) values [38], saliency maps and Gradient-Based Localization [39] shall be deployed. Furthermore, Layer-wise Relevance Propagation (LRP) propagation-based approaches will take into account the internal structure of the neural network propagating the prediction from the output backwards to the input in order to assign relevance scores to the input variables and in compliance with some predefined rules [40]. Moreover, Local Interpretable Model-agnostic Explanations (LIME) will be applied to the DL architecture, explaining the model's response to local perturbations, occluding some of the inputs to investigate the importance of specific variables.

As part of AI modelling, validation will also constitute a major step with regard to the trustworthiness and reliability of the optimal AI model. In this respect, the large pool of retrospective data comprises the required foundation for extensive evaluation of the tested schemes. By utilizing cross-validation architectures and applying weighted multi-class splitting within the available datasets, it possible to extract unbiased estimations, while applying the traditional evaluation metrics (accuracy, specificity, sensitivity, precision etc.) on a class-specific scale. The above techniques will be applied on the extracted features consisting of serum sample protein features, demographic, behavioral and medical features. Classification will be attempted both individually for each feature class (i.e., separating protein data) and employing combinatorial schemes, attempting to link proteinic profiles with routine factors (modifiable or not), aiming for an enhanced model that valorizes routine data alongside protein markers. This will be done with the vision of generating two separate AI models appertaining to real-life use cases. Namely, the risk stratification AI model (RS-AI) will solely assess risk factors and potential risk adjustments based on suggested behavioral interventions. On the other hand, the screening AI model (S-AI) will assess liquid biopsy results – in combination with risk factors – to provide decision support on high-risk individuals that should undergo colonoscopy.

On the whole, the AI system should also by-principle and by-design fulfil the requirements comprising (i) accountability, (ii) privacy and data governance, (iii) societal and environmental wellbeing, (iv) technical robustness and safety, (v) human agency and oversight, (vi) diversity and fairness, and (vii) transparency. In such manner, the front-end interface will provide relevant knowledge to clinicians, medical professionals and citizens regarding risks and guidelines for colorectal cancer, raising awareness and contributing to enhanced prevention through population knowledge and evidence-based recommendations.

VI. Conclusion

Strengthening the collaborative environment and building AI-based bridges between different fields of practice- and theory-driven research in colorectal cancer screening is more than necessary given the cost of the existing healthcare pathway. By delivering an early screening and risk assessment solution via AI, solutions can associate significant advancements on many different levels, especially those related to patient acceptance and knowledge discovery. Embracing of new technologies in the colorectal screening pathway will improve public health, while generating knowledge to increase the quality of life. Thus, screening will become more accessible, offering risk assessment, health monitoring and better disease evaluation. Besides early diagnosis, identification of disease trajectories and relapse, as well as employment of cutting-edge technologies may shed light to the nature of the colorectal cancer biological initiation mechanisms, therefore improving the existing human condition.

While colorectal cancer screening tools have proven to be effective, the medical examinations usually employed present several obstacles. As such, a substantial colorectal cancer miss rate has been often reported, while the fact that certain screening methods (e.g., colonoscopy) usually require hospital/clinical settings puts excessive workload on the medical experts to program appointments and consultation. In this regard, information regarding colorectal cancer development could not only prioritize citizens clinic appointments (e.g., increasing the participation of high-risk colorectal cancer patients), but also allow healthcare professionals to be better positioned in terms of the required screening technique.

Allowing preliminary and evidence-based results, our suggested approach will provide medical experts with detailed information for an early and precise screening of colorectal cancer augmenting clinical decision reliability for preventing unneeded invasive examination and thus potential adverse outcomes. Furthermore, it will improve quality in the clinical actions and patient care, reducing the time spent on routine administrative tasks, while providing easy and faster access to colorectal cancer knowledge staging and intervention planning.

The knowledge generated by targeting subject-specific quantification of precise biomarkers determining divergencies and tolerance constraints on colorectal cancer determinants via Phases I & II of the methodology presented here, will be assessed based on cutting-edge AI tools (Phase III). This way, only the clinical validation will be left to be performed via medium or large-scale pilots. Thus, it will grant personalized staging and early protopathic/recurrence prediction, reducing human resources costs by unburdening the healthcare system of nonessential hospital/clinical visits.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation Programme under Grant Agreement No. 875351.

References

[1] Y. Xi and P. Xu, 'Global colorectal cancer burden in 2020 and projections to 2040', *Translational oncology*, vol. 14, no. 10, Oct. 2021, doi: 10.1016/j.tranon.2021.101174.

[2] G. Argilés *et al.*, 'Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up', *Annals of oncology : official journal of the European Society for Medical Oncology*, vol. 31, no. 10, Oct. 2020, doi: 10.1016/j.annonc.2020.06.022.

[3] M. Riihimäki, H. Thomsen, K. Sundquist, J. Sundquist, and K. Hemminki, 'Clinical landscape of cancer metastases', *Cancer Medicine*, vol. 7, no. 11, p. 5534, Nov. 2018, doi: 10.1002/cam4.1697.

[4] R. L. Siegel, K. D. Miller, and A. Jemal, 'Cancer statistics, 2020', *CA Cancer J Clin*, vol. 70, no. 1, pp. 7–30, Jan. 2020, doi: 10.3322/caac.21590.

[5] A. Gini *et al.*, 'Impact of colorectal cancer screening on cancer-specific mortality in Europe: A systematic review', *Eur J Cancer*, vol. 127, pp. 224–235, Mar. 2020, doi: 10.1016/j.ejca.2019.12.014.

[6] V. Vymetalkova, K. Cervena, L. Bartu, and P. Vodicka, 'Circulating Cell-Free DNA and Colorectal Cancer: A Systematic Review', *International journal of molecular sciences*, vol. 19, no. 11, Oct. 2018, doi: 10.3390/ijms19113356.

[7] A. Mitsala, C. Tsalikidis, M. Pitiakoudis, C. Simopoulos, and A. K. Tsaroucha, 'Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. A New Era', *Current Oncology*, vol. 28, no. 3, p. 1581, Jun. 2021, doi: 10.3390/curroncol28030149.

[8] D. E. O'Sullivan *et al.*, 'Combinations of modifiable lifestyle behaviours in relation to colorectal cancer risk in Alberta's Tomorrow Project', *Sci Rep*, vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-76294-w.

[9] K. Simon, 'Colorectal cancer development and advances in screening', *Clinical interventions in aging*, vol. 11, Jul. 2016, doi: 10.2147/CIA.S109285.

[10] G. Binefa, F. Rodríguez-Moranta, À. Teule, and M. Medina-Hayas, 'Colorectal cancer: From prevention to personalized medicine', *World J Gastroenterol*, vol. 20, no. 22, pp. 6786–6808, Jun. 2014, doi: 10.3748/wjg.v20.i22.6786.

[11] 'Colorectal Screening In Europe', *Digestive Cancers Europe*. https://digestivecancers.eu/publication/colorectal-screening-in-europe/ (accessed May 01, 2022).

[12] J. L. Petrick, L. E. Barber, S. Warren Andersen, A. A. Florio, J. R. Palmer, and L. Rosenberg, 'Racial Disparities and Sex Differences in Early- and Late-Onset Colorectal Cancer Incidence, 2001–2018', *Frontiers in Oncology*, vol. 11, 2021, Accessed: May 01, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fonc.2021.734998

[13] F. Bénard, A. N. Barkun, M. Martel, and D. von Renteln, 'Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations', *World J Gastroenterol*, vol. 24, no. 1, pp. 124–138, Jan. 2018, doi: 10.3748/wjg.v24.i1.124.

[14] A. S. Butterworth, J. P. T. Higgins, and P. Pharoah, 'Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis', *Eur J Cancer*, vol. 42, no. 2, pp. 216–227, Jan. 2006, doi: 10.1016/j.ejca.2005.09.023.

[15] T. Jess, C. Rungoe, and L. Peyrin-Biroulet, 'Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies', *Clin Gastroenterol Hepatol*, vol. 10, no. 6, pp. 639–645, Jun. 2012, doi: 10.1016/j.cgh.2012.01.010.

[16] Y. Dong *et al.*, 'Abdominal obesity and colorectal cancer risk: systematic review and meta-analysis of prospective studies', *Biosci Rep*, vol. 37, no. 6, p. BSR20170945, Dec. 2017, doi: 10.1042/BSR20170945.

[17] E. L. Amitay *et al.*, 'Smoking, alcohol consumption and colorectal cancer risk by molecular pathological subtypes and pathways', *Br J Cancer*, vol. 122, no. 11, pp. 1604–1610, May 2020, doi: 10.1038/s41416-020-0803-0.

[18] H. Kim *et al.*, 'Total Vitamin D Intake and Risks of Early-Onset Colorectal Cancer and Precursors', *Gastroenterology*, vol. 161, no. 4, pp. 1208-1217.e9, Oct. 2021, doi: 10.1053/j.gastro.2021.07.002.

[19] M. C. Hornbrook *et al.*, 'Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data', *Dig Dis Sci*, vol. 62, no. 10, pp. 2719–2727, Oct. 2017, doi: 10.1007/s10620-017-4722-8.

[20] G. Lech, R. Słotwiński, M. Słodkowski, and I. W. Krasnodębski, 'Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances', *World J Gastroenterol*, vol. 22, no. 5, pp. 1745–1755, Feb. 2016, doi: 10.3748/wjg.v22.i5.1745.

[21] E. Koncina, S. Haan, S. Rauh, and E. Letellier, 'Prognostic and Predictive Molecular Biomarkers for Colorectal Cancer: Updates and Challenges', *Cancers (Basel)*, vol. 12, no. 2, p. 319, Jan. 2020, doi: 10.3390/cancers12020319.

[22] S. Lei *et al.*, 'Adenoma detection rate is not influenced by the time of day in computer-aided detection colonoscopy', *Medicine (Baltimore)*, vol. 99, no. 51, p. e23685, Dec. 2020, doi: 10.1097/MD.0000000000023685.

[23] M. Y. Lu *et al.*, 'AI-based pathology predicts origins for cancers of unknown primary', *Nature*, vol. 594, no. 7861, Art. no. 7861, Jun. 2021, doi: 10.1038/s41586-021-03512-4.

[24] M. F. Mohamad Marzuki, N. A. Yaacob, N. M. Bin Yaacob, M. R. Abu Hassan, and S. B. Ahmad, 'Usable Mobile App for Community Education on Colorectal Cancer: Development Process and Usability Study', *JMIR Hum Factors*, vol. 6, no. 2, p. e12103, Apr. 2019, doi: 10.2196/12103.

[25] L. Xie *et al.*, 'Diagnostic Value of Methylated Septin9 for Colorectal Cancer Detection', *Front Oncol*, vol. 8, p. 247, Jul. 2018, doi: 10.3389/fonc.2018.00247.

[26] E. Tjoa and C. Guan, 'A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI', *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[27] L. Wik *et al.*, 'Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis', *Molecular & Cellular Proteomics*, vol. 20, Jan. 2021, doi: 10.1016/j.mcpro.2021.100168.

[28] Y. Zhao and A. R. Brasier, 'Qualification and Verification of Protein Biomarker Candidates', *Adv Exp Med Biol*, vol. 919, pp. 493–514, 2016, doi: 10.1007/978-3-319-41448-5_23.

[29] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study', *PLOS Medicine*, vol. 15, no. 11, p. e1002683, Nov. 2018, doi: 10.1371/journal.pmed.1002683.

[30] T. Dissanayake, T. Fernando, S. Denman, H. Ghaemmaghami, S. Sridharan, and C. Fookes, 'Domain Generalization in Biosignal Classification', *arXiv:2011.06207 [cs]*, Nov. 2020, Accessed: May 07, 2022. [Online]. Available: http://arxiv.org/abs/2011.06207

[31] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, 'Domain Generalization via Model-Agnostic Learning of Semantic Features', in *Advances in Neural Information Processing Systems*, 2019, vol. 32. Accessed: May 07, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/2974788b53f73e7950e8aa49f3a306db-Abstract.html

[32] L. L. Guo *et al.*, 'Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine', *Sci Rep*, vol. 12, no. 1, Art. no. 1, Feb. 2022, doi: 10.1038/s41598-022-06484-1.

[33] A. Antoranz, T. Sakellaropoulos, J. Saez-Rodriguez, and L. G. Alexopoulos, 'Mechanism-based biomarker discovery', *Drug Discov Today*, vol. 22, no. 8, pp. 1209–1215, Aug. 2017, doi: 10.1016/j.drudis.2017.04.013.

[34] C. Fotis, A. Antoranz, D. Hatziavramidis, T. Sakellaropoulos, and L. G. Alexopoulos, 'Network-based technologies for early drug discovery', *Drug Discov Today*, vol. 23, no. 3, pp. 626–635, Mar. 2018, doi: 10.1016/j.drudis.2017.12.001.

[35] I. N. Melas *et al.*, 'Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury', *Integr Biol (Camb)*, vol. 7, no. 8, pp. 904–920, Aug. 2015, doi: 10.1039/c4ib00294f.

[36] C. Fotis *et al.*, 'Accurate SARS-CoV-2 seroprevalence surveys require robust multi-antigen assays', *Sci Rep*, vol. 11, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41598-021-86035-2.

[37] L. Chen, X. Yang, G. Jeon, M. Anisetti, and K. Liu, 'A trusted medical image super-resolution method based on feedback adaptive weighted dense network', *Artif Intell Med*, vol. 106, p. 101857, Jun. 2020, doi: 10.1016/j.artmed.2020.101857.

[38] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions', in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2017, pp. 4768–4777.

[39] A. Das and P. Rad, 'Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey', *arXiv:2006.11371 [cs]*, Jun. 2020, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2006.11371

[40] W. Samek, T. Wiegand, and K.-R. Müller, 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models', *arXiv:1708.08296 [cs, stat]*, Aug. 2017, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1708.08296