Predicting the Predisposition to Colorectal Cancer based on SNP Profiles of Immune Phenotypes using Supervised Learning Models[1]

*Ali Cakmak, Huzeyfe Ayaz, Soykan Arıkan, Ali R. Ibrahimzada, Seyda Demirkol, Dilara S nmez, Mehmet T. Hakan, Saime T. S rmen, Cem Horozoglu, Mehmet B. Dogan, Ozlem Kucukhuseyin, Canan Cacına, Bayram Kıran, Umit Zeybek, Mehmet Baysan, Ihan Yaylım

**Corresponding author*: Ali Cakmak, ali.cakmak@itu.edu.tr, Department of Computer Engineering, Istanbul Technical University, Ayazaga Campus, Reşitpaşa, 34467 Sarıyer, Istanbul, Turkey.

[1] *Manuscript is to appear in Medical and Biological Engineering & Computing*

Abstract: *This study explores the machine learning-based assessment of predisposition to colorectal cancer based on single nucleotide polymorphisms (SNP). Such a computational approach may be used as a risk indicator and an auxiliary diagnosis method that complements the traditional methods such as biopsy and CT scan. Moreover, it may be used to develop a low-cost screening test for the early detection of colorectal cancers to improve public health. We employ several supervised classification algorithms. Besides, we apply data imputation to fill in the missing genotype values. The employed dataset includes SNPs observed in particular colorectal-cancer-associated genomic loci that are located within DNA regions of 11 selected genes obtained from 115 individuals. We make the following observations: (i) Random Forest-based classifier using one-hot encoding and K-Nearest Neighbor (KNN)-based imputation performs the best among the studied classifiers with an F1 score of 89% and Area Under the Curve (AUC) score of 0.96. (ii) One-hot encoding together with K-Nearest-Neighbor-based data imputation increase the F1 scores by around 26% in comparison to the baseline approach which does not employ them. (iii) The proposed model outperforms a commonly employed state-of-the-art approach, ColonFlag, under all evaluated settings by up to 24% in terms of the AUC score. Based on the high accuracy of the constructed predictive models, the studied 11 genes may be considered as a gene panel candidate for colon cancer risk screening.*

Keywords: Colorectal Cancer, Machine Learning, Classification, Cancer Screening, Immune Checkpoints

I. INTRODUCTION

Colorectal cancer (CRC) is the most common type of cancer in the world after breast cancer in women, and lung and prostate cancer in men [1]. It is generally accepted that colon cancer is caused by multistage genetic changes and variations because of mutations in oncogenes, tumor suppressor genes, or DNA repair genes [2]. Unfortunately, there is still insufficient information about the molecular pathology of CRC.

Tumor cells exhibit six different biological characteristics during their multi-stage development: (i) maintaining proliferative signals, (ii) avoiding growth suppression, (iii) resistance to cell death, (iv) providing replicative immortality, (v) stimulating angiogenesis, (vi) invasion and metastasis activation [3]. The immune system has a central role in controlling and eliminating cancer. However, in the case of malignancy, there may be several immune suppression mechanisms that inhibit effective anti-tumor response [4]. Immune control points are cell surface molecules expressed on a variety of immune cells which are employed to prevent peripheral autoimmunity during the inflammatory response [5]. Some example control points are Programmed Cell Death-1 (PD-1), Cytotoxic T-Lymphocyte Associated Antigen-4 (CTLA-4), T-cell Immunoglobulin and Mucin Protein-3 (TIM-3), Lymphocyte Activation Gene-3 (LAG-3). These control points stimulate negative modulation of immune responses to prevent autoimmunity after binding to their ligands. This immune mechanism is driven by tumor cells, allowing a rapid proliferation of tumor cells.

In terms of cancer occurrence and progression, in addition to genetic changes and tumor, node, and metastasis (TNM) staging, tumor microenvironment and tumor-infiltrating immune evaluation of cells are also important. In this study, we investigate the effect of MDM2 (rs2279744), GAL3 (rs4644), TIM1 (rs9313422), TRAIL (rs 1131580), PD-1 (rs2227981), PD-L1 (rs2890658), p16 540 (rs11515), p16 580 (rs3088440), CD28 (rs3116496), CD27 (rs2267966), and CD40 (rs1883832) gene polymorphisms on the antitumor response of tumor microenvironment. This study aims to investigate whether specific genetic polymorphisms of the above immune checkpoint genes are associated with the development of colorectal cancer. In this study, we adopt a computational approach and explore the machine learning-based assessment of predisposition to colorectal cancer based on polymorphisms in the above set of genes. Such a computational approach may be used as a risk indicator and an auxiliary diagnosis method complementary to other diagnostic tools, such as biopsy, CT scan, and MRI. Moreover, it may be used to develop a low-cost screening test for the early detection of colorectal cancer to improve public health. Finally, individuals with elevated risk of developing colorectal cancer may be advised on appropriate lifestyles [6] to minimize the effect of the environmental factors.

Our dataset includes single nucleotide polymorphisms (SNP) observed in particular colorectal-cancer-associated genomic loci that are located within DNA regions of the above-discussed genes. The dataset is obtained from 115 individuals (50 healthy (control group) and 65 colorectal cancer patients). We employ several supervised classification algorithms, namely, Logistic Regression (LR), Random Forests (RF), and Support Vector Machine (SVM). Besides, we apply data imputation to fill in the missing genotype values. Based on our experimental results, we make the following observations:

i. Random Forest-based classifier using one-hot encoding for feature representation and K-Nearest Neighbor (KNN)-based imputation to complete the missing data with nearest integer conversion performs the best among the studied classifiers. On the other hand, a Logistic Regression-based classifier provides a more generalized model at a cost of slight decrease in prediction accuracy.

ii. The proposed model outperforms a commonly employed state-of-the-art approach, ColonFlag [7, 8], under all evaluated settings in terms of the AUC scores.

iii. Based on the high accuracy of the constructed classification models, the studied 11 genes may be considered a gene panel candidate for the risk screening of colon cancer.

The rest of the paper is organized as follows. The next section presents a review of the related works from the literature. Then, Section 3 describes our methods. Experimental evaluation results are presented in Section 4. In Section 5, we discuss the contributions and impact of this study along with potential future directions to extend this paper. Finally, Section 6 summarizes our main results in this study.

II. RELATED WORK

Several previous studies focus on colorectal cancer diagnosis and risk prediction. Yu and Helwig [9] present a comprehensive review of the recent machine learning and AI methods employed in CRC diagnosis and treatment. In this section, we comparatively review the most relevant ones to our work.

Colon cancer prediction based on gene expression profiling is a highly popular topic in the literature. As an example, Barrier et al. [10] employed stage 2 and 3 Tumor and non-neoplastic mucosa mRNA samples from 18 patients in total (10 men and 8 women). Patients were profiled using the Affymetrix HGU133A GeneChip. K-Nearest Neighbor (KNN) algorithm was used with 6-fold cross-validation to select the number of Neighbors and the number of informative genes to include in the predictors. After splitting data into 6 portions, 15 of them were used as training data to assign a prognosis (recurrence or no recurrence) and 3 of them as validation data. To evaluate their method, they studied only the false predictions out of 18 samples. A predictor model is constructed by using m informative genes as features that are obtained based on k-nearest neighbors. A separate model is trained for each unique combination of model parameters m and k. In particular, 50 different values of m (i.e., m = 5, 10, …, 250) and 3 different values of k (i.e., k = 1, 3, 5) were considered. Hence, in total, 150 different predictive models are trained and compared in terms of their accuracy. The lowest numbers of false predictions (2 out of 18) were obtained from the lowest numbers of informative genes in six-fold cross-validation. A 30-gene T-based predictor was built with 78% estimated accuracy, and a 70-gene NM-based predictor provided 83% accuracy.

Another study that employs gene expression profiles was carried out by Horaira et al. [11]. As a predictor model, they exploited a kernel-based Support Vector Machine (SVM). Their dataset consisted of gene expression profiles of individuals covering more than 6500 humanoid genes. They picked 2000 genes based on their confidence in the expression levels. Two datasets of gene expression namely Leukemia and Colon Tumor Datasets have been used for this research. In total, 62 samples have been collected from colon cancer patients. In particular, 40 biopsies of tumor and 22 normal cell biopsies were obtained and labeled as negative and positive, respectively. Feature selection was performed mainly using the t-statistic. That is, the top 10 and top 3 genes were identified among 2000 genes that have been considered originally, and used for classification (GenBank accession numbers of the top 3 genes were I.37937, Has.2097, and Has.2291). The classification accuracy of the SVM model was estimated at 85.8%.

Besides, Alladi et al. [12] employed the benchmark colon cancer dataset that contains the expression profiles of 40 tumors and 22 healthy colon tissue samples. Two thousand out of around 6500 genes were selected based on confidence in the measured expression levels. The dataset was split into training and test sets in the ratio 80:20, 70:30, 65:35, and 60:40. The best prediction accuracy was obtained by the 65:35 ratio. They employed 3 classifiers namely, SVM, Neural Nets, and Logistic Regression using the top 10 genes ranked by the t-statistic. The best accuracy scores were obtained with SVM (linear) at 84.19% and SVM (sigmoid) at 85.80%.

In addition, the prediction of colon cancer stages and survival period with a machine learning approach has been analyzed by Gupta et al. [13]. 40

21 patients were selected for the analysis from the colorectal cancer registry of the Chang Gung Memorial Hospital, Linkou, Taiwan. They mainly focused on the prediction of tumors in the TNM stage of colon cancer. The distribution of patients based on the tumor stage were T3: 2500, T4: 500+, T2: 500, and T1: 500. By considering the Tumor Aggression Score as a prognostic factor, performances of different ML algorithms were evaluated using five-fold cross-validation to predict the tumor stage of colon cancer. Six ML models were evaluated for this study, namely, RF, SVM, LR, Multilayer Perceptron, KNN, and Adaptive Boosting. RF achieved an F-measure of 0.89 when the Tumor Aggression Score was considered as an attribute along with the standard attributes normally used for the TNM stage prediction. Besides, the top-performing model (i.e., Random Forest) achieved an accuracy of 0.84 and an Area Under the Curve (AUC) of 0.82 + 0.1 for predicting the five-year disease-free survival of colon cancer patients.

On the other hand, some other researchers have developed their own models for the detection of colon cancer. As an example, Kinar et al. [7] present an algorithm, called MeScore (also known as ColonFlag). The underlying model consists of a combination of a thousand tree-based classifiers. Then, the sum of the weights of all the trees is used for the prediction. The main data were collected from 606,403 individuals among whom 3,135 were diagnosed with colorectal cancer. In addition, they employed a second dataset of 30,674 individuals, which contains 5,061 CRC cases and 25,613 control subjects. In addition to blood count records, the sex and birth years of all individuals were available in the dataset. Overall, they obtained an AUC score of 0.82 and 88% specificity in the Israeli dataset, while the AUC and specificity scores were 0.81 and 94%, respectively, in the UK dataset.

In addition to the above summarized colorectal cancer diagnosis prediction works, some studies aim to detect CRC early or predict the risk of developing CRC. A common approach among this group of studies is using Complete Blood Counts (CBC) of patients. As an example, Hornbrook et al. [14] worked with a dataset that was obtained from Kaiser Permanente Northwest Region's Tumor Registry. In total, 17,095 individuals were included in this analysis. They employed the ColonFlag algorithm [7, 8] to identify people at risk of colorectal cancer. To calculate CRC detection score, gender, year of birth, and at least one combination of the following blood count parameters {RBC, Hgb, Hct}, {RBC, Hct, MCH}, {RBC, MCH, MCHC}, {Hgb, Hct, MCH}, {Hgb, MCH, MCHC}, or {Hct, MCH, MCHC} are used as features of the classifier. Their model achieved an AUC score of 0.8 +- 0.01.

Another CRC risk prediction study was conducted by Nartowt et al. [15]. They worked with the National Interview Survey and the Prostate, Lung, Colorectal, Ovarian Cancer Screening datasets with 795,215 samples. They considered Linear Discriminant Analysis, SVM, Naïve Bayes, Decision Tree, Random Forest, and Artificial Neural Networks (ANN) as machine learning models. To deal with the missing data, they employed methods like Mean, Gaussian, Lorentzian, One-Hot encoding, Gaussian Expectation-Maximization, and Listwise deletion. 2-fold cross-validation was applied during model evaluation. ANN achieved the max AUC score of 0.75 with family history data, and an AUC score of 0.70 when family history data is not available. Their improved model achieved an AUC score of 0.82.

Kinar et al. [8] in another study focused on the performance analysis of a machine learning flagging system. As a pre-screening test, the fecal occult blood test was used and a positive test result was referred to colonoscopy. CBC was used in the prediction. Their dataset contained 112,584 samples. In total, 133 patients were diagnosed with CRC in 2008. They employ an algorithm called as MeScore (i.e., ColonFlag) which was developed in their previous study [7]. There is no accuracy or AUC score-based evaluation presented in this study.

In addition to gene expression and CBC, SNP profiles have also been exploited in several studies to detect colorectal cancer. Patidar and Bhojwani [16] analyzed the SNP patterns involved in CRC. They mostly focused on the 4 major alternative signaling pathways defined for colorectal cancer, namely, Wnt, P53, TGF-beta, and K-Ras. The employed dataset includes tissue biopsies of 45 colorectal cancer patients. They examined protein expression in situ. The work was done mainly on two software programs. First, QualitySNP [17] was used to detect reliable SNPs, intersections, and deletions both with and without quality files. Second, the FastSNP web server [18] was used to efficiently identify and prioritize high-risk SNPs according to their phenotypic risks and putative functional effects.

Lin et al. [19] proposed a two-stage machine learning approach for identifying SNP to SNP interactions. Their primary goal was to identify a subset of significant SNPs and detect interaction patterns by combining two machine learning models, namely, Random Forest (RF) and Multivariate Adaptive Regression Splines (MARS). 1,151 prostate cancer cases (492 non-aggressive and 659 aggressive) were gathered from the Cancer Genetic Markers of Susceptibility. 149 SNPs in the six ER-related genes which were CYP1B1, SRD5A2, ESR1, ESR2, CYP19A1, and CYP1A1 were examined. In the two-stage RF-MARS approach, they first applied RF to detect a subset of SNPs for prediction, then MARS was applied to identify the interaction patterns based on selected SNPs. They followed two alternative approaches for RF. In the first one, the "Out-of-Bag" method was used to estimate the classification accuracy for each RF tree. In the second approach, the importance spectrum of the original data was compared with that of the permutated data where class labels were randomly permuted. They studied the true positive rate over the increased penetrance contrast to compare the results.

Jenkins et al. [20] worked on the ability of known susceptibility SNPs to predict colorectal cancer risk for persons with and without a family history. They estimated the association of CRC with 45 SNPs. The data contains 1,181 cancer cases and 999 healthy individuals. They preferentially selected cancer patients younger than 50 years with a 10% sampling from all other ages in diagnosis and control groups that did not have a family history of CRC. They applied multiple logistic regression to estimate the association between SNPs and CRC risk. They assessed a risk gradient and discrimination in risk between the cancer and control group by estimating the change in odds ratio per adjusted standard deviation (OPERA). OPERA is used to compare the strengths of associations for risk factors measured on different scales, across diseases and populations. As a result, they predicted that the 20% of the population with the highest number of risk alleles would have a 1.8 times higher risk of CRC than people with the average number of risk alleles. Also, they predicted that the 45 CRC risk-associated SNPs identified in the literature could explain about 22% of the familial component of CRC risk.

Based on earlier results in the literature that suggest a possible link between microbes and physiological conditions (e.g., [21]), Qu et al. studied the prediction of CRC cases based on a particular selection of gut microbiota [22]. To this end, they employed a dataset of rRNA sequences from the gut microbiota. The features were selected via multiple dimension reduction methods that also considered the taxonomy of the microbe species. Once features are selected, the authors employ random forest, Naïve Bayes, and decision tree classification models for prediction, and show that taxonomy-aware multi-dimension reduction methods improve the earlier results in the literature. Even though the accuracy of the best-performing models is promising, the study requires an extensive metagenomics study as a pre-processing step.

Zhao et al. [23] extend the above study by considering independent risk factors such as BMI, age, tumor type, tumor grade, and gender along with gut microbiota. They demonstrate that combining genetic factors with traditional risk items increases classification performance. Nevertheless, the collection of employed features requires costly biopsy procedures to determine tumor type and grade.

Another promising non-invasive approach is analyzing DNA methylation patterns of tumor samples obtained through blood-liquid biopsy [24]. Authors show that DNA methylation analysis provides reliable indicators for both diagnosis and survival prediction. On the other hand, the procedure requires whole genome-wide analysis, whereas we propose analyzing an exceedingly small set of SNP locations that provides comparable accuracy.

Metadherin mRNA expression levels in serum have been proposed as another non-invasive biomarker of CRC [25]. Despite the high AUC scores reported in this study, a major drawback of the study design is that the majority of the patients had advanced-stage cancer, and high metadherin mRNA expression is particularly associated with late-stage CRC. Hence, it is not feasible to use it for screening and early diagnosis purposes.

Another focus of interest in the field was microRNAs' role in CRC etiology [26, 27]. The authors demonstrated that all cancer samples manifested changes in four pathways where 21 micro RNAs are detected as the most significantly changing ones. Such models have the disadvantage that microRNA expression analysis requires a costly, and to some extent, invasive procedure of tumor sample biopsy.

Birks et al. [28] have studied the records of 2.5 million patients to evaluate the effectiveness of a compelete blood count-based model on the UK population. The strength of the prediction model is that it only requires a simple blood test. On the other hand, the resulting specificity and AUC scores in this large cohort are on the lower end of the spectrum of available methods. A similar study [29] was done to evaluate 14 risk models on a cohort of 500,000 individuals. The selected risk models employed only patient answers to a general questionnaire regarding their eating habits, smoking status, age, exercise habits, etc. with no additional biochemical or genetic results. The best AUC score out of the evaluated models was around 0.67 on average which motivates the need for identifying additional preferably non-invasive factors to predict colon cancer.

Li et al. also considered SNP data over 116 genes to predict the risk of CRC [30]. However, their best model provides a 0.61 AUC score, which is significantly less than the discriminative power of the 11 immune checkpoint gene SNPs considered in this study.

In cancer prediction model development, feature selection and stability may be also important when the employed dataset includes substantial numbers of features some of which may contain noisy measurements. Cueto-López et al. [31] compared several machine learning models on a dataset of 47 SNPs and 53 environmental variables including family history, BMI, physical activity, etc. The best-performing model provides an AUC score of 0.69 at the low end. Nevertheless, the contribution of this study is that it demonstrates that feature selection improves overall model performance and that the most accurate model may not be the best in terms of stability.

Zhou et al. [32] show that their deep learning-based classification model trained on thousands of endoscopy images performs better (AUC = 0.88) than the endoscopists' average accuracy. However, like biopsy studies, endoscopy is a costly and invasive procedure.

In another direction, Molparia et al. [33] investigate the use of copy number variation (CNV) as a possible indicator of CRC. The advantage of this study is that their analysis is also based on circulating DNA in the blood, which is minimally invasive. On the other hand, despite high specificity scores, their sensitivity scores are much lower. In addition, accurate CNV detection requires whole genome (or at least whole exome) sequencing [34].

III. METHODS

We employ several supervised classification algorithms, namely, LR, RF, and SVM. We implemented our scripts in Python programming language using the Scikit-learn library. We initially perform an exploratory analysis of the underlying data to determine features for the classification models.

A. Exploratory Analysis and Data Engineering

Our dataset includes polymorphisms observed in specific colorectal-cancer-associated genomic loci that are located within DNA regions of 11 selected genes, namely, MDM2, GAL3, TIM1, TRAIL, PD-1, PD-L1, p16540, p16580, CD28, CD27, and CD40. The dataset was collected from 115 people in total with 50 healthy individuals (control group) and 65 colorectal cancer patients. The dataset also includes additional information for patients only, such as the age of the patient, the stage of cancer, the location of the tumor, perineural invasion of the tumor, and tumor differentiation. Ethical approval for this study was obtained from the Clinical Research Ethics Committee of Istanbul Education & Research Hospital (Date: 23.06.2017; Number:1015)

An initial exploratory analysis of the dataset revealed that the mean age values of colon cancer patients (mean age = 60) are considerably higher than the control group (mean age = 35). Hence, we remove the age information from consideration to avoid the bias which would cause the classification models to overfit the current dataset, and limit the generalization of the model that would work for unseen future patients.

We next explored SNP data for individual genes. The SNP data for some genes were not complete. What is more, the individuals with missing SNP data mostly belonged to a particular category. For instance, for gene TRAIL, 53 individuals had no SNP information, and the majority (i.e., 39) of the individuals with no SNP data belong to the control group. Likewise, five other genes, namely, CD28, PD1, PDL1, CD27, and CD40 have unbalanced missing data.

1) Handling Missing Data with Data Imputation (DI)

Most statistical learning models cannot handle missing values in the data. Hence, data columns with missing values have to be taken care of. One option is to discard such fields with missing values from the data. The advantage of this option is that it requires almost no effort. Nevertheless, the omitted features in most cases may contain significant semantics, and discarding such features may decrease the accuracy of the constructed statistical models. A more viable option is to employ data imputation methods to estimate the missing values in a column based on the existing data values in the same data column. In particular, we employ two commonly used imputation techniques, namely, KNN-based Imputation and Multivariate Imputation by Chained Equations (MICE) [35].

2) Dealing with Categorical Data

The SNP data is considered categorical data. However, machine learning models work with numeric data. To address this problem, we explore two transformation methods.

One-Hot Encoding (OHE): One-hot encoding transformation converts categorical SNP data into binary (0 or 1) numbers. More specifically, one-hot encoding introduces a separate column into the data for each distinct value under each categorical column. For instance, the CD40 gene contains the following distinct genotypes: C/C, C/T, and T/T. One-hot encoding would create three columns for the CD40 gene, one for each distinct genotype, i.e., CD40_CC, CD40_CT, and CD40_TT. Then, those individuals having a particular genotype will have a value of 1 for the corresponding column and 0 for the remaining other genotype columns for CD40.

Custom Decimal Transformation (CDT): In this approach, we transform our dataset to a decimal form. In particular, we represent each possible 4 letter of the DNA alphabet in two-bit binary numbers. That is, A: 00, T: 01, C: 10, G: 11. Since we have pairs of nucleotides in SNP genotypes, we will convert them by just replacing each character with its number form. For instance, TG: 0111 = 7 (in decimal form), CC: 1010 = 10, AC: 0010 = 2. Hence, for each column, we use the corresponding decimal value after this transformation.

B. Baseline Approach

The baseline approach does not employ any data imputation, and simply omits genes that have missing SNP genotypes for some individuals in the dataset. More specifically, 6 genes, namely, Trail, CD28, PD1, PDL1, CD27, and CD40 are omitted. Hence, the classification models are built on the following remaining 5 genes: MDM2, GAL3, TIM1, p16540, and p16580. Then, one-hot encoding is applied for categorical data.

C. Approach OHE_DI

In this approach, we first perform one-hot encoding on our original data set. Then, we employ KNN-based data imputation to estimate the values for missing fields based on the nearest value in the same data column.

D. Approach OHE_DI_FI

The imputation in the above approach provides float numbers for the missing values. In this approach, we add an extra step to OHE_DI to convert each float value to the nearest integer (FI). This is because, we observe that the imputation provides unique float values almost for every cell, which may introduce some bias.

E. Approach CDT_DI_OHE

This approach employs the above-discussed custom decimal transformation to handle the categorical data, and then data imputation is applied to complete the missing values. Finally, one-hot encoding is performed.

F. Approach CDT_DI_FI_OHE

This approach is a variation of Approach CDT_DI_OHE in that, after imputation, we converted floating-point values that are filled for missing values to the nearest integer to remove the possible bias as explained before. Finally, one-hot encoding is performed.

G. Approach CDT_DI

This approach differs from Approach CDT_DI_OHE mainly in the last step. That is, it does not use one-hot encoding. Instead, it employs the decimal values of the genotype data that are obtained as explained above.

H. Approach CDT_DI_FI

This approach is a slight variation of the above approach in that, after imputation, we convert floating-point values that are filled for missing values to the nearest integer.

I. Evaluation Methodology

We perform stratified 10-fold cross-validation to evaluate the performance of the employed classifiers in all experiments. More specifically, we randomly divide the data into 10 parts. Then, we employ 9 parts for training the models, the remaining 1 part to test the trained model. Then, we repeat these steps 10 times, where each time, a new classification model is trained on a different combination of 9 parts and tested on the remaining 1 part. Then, the *average* (i.e., mean) performance from these 10 iterations is reported as the overall performance of a model. The splitting is done in a stratified manner keeping the same proportions of control and cancer groups in both the training and test sets.

J. Evaluation Metrics

As the performance metric, we use the F1 score, which is a commonly employed measure in machine learning tasks. Moreover, we also generate Receiver Operating Characteristics (ROC) curves and report the Area Under the Curve (AUC) as an alternative performance metric for the classifiers.

IV. RESULTS

1) Classification with the Baseline Approach

Figure 1 reports the F1 scores of the three employed classifiers for the baseline approach. The best classification performance is obtained with Logistic Regression with an F1 score of around 63%. These performance figures are not favorable since we had to omit 6 genes out of 11.

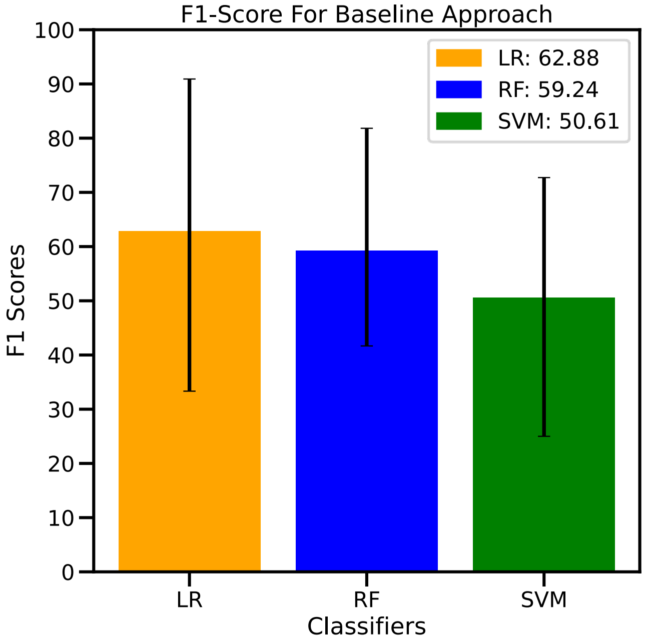

Figure 1: F1 scores of the employed classifiers with the baseline approach

Figure 2 reports the AUC scores of the same classifiers as an alternative performance metric. In this category, RF provides the best performance. We observe that LR has a higher true positive rate and a lower false negative rate than SVM, and this is what the AUC metric measures. Therefore, LR outperforms SVM under this metric.

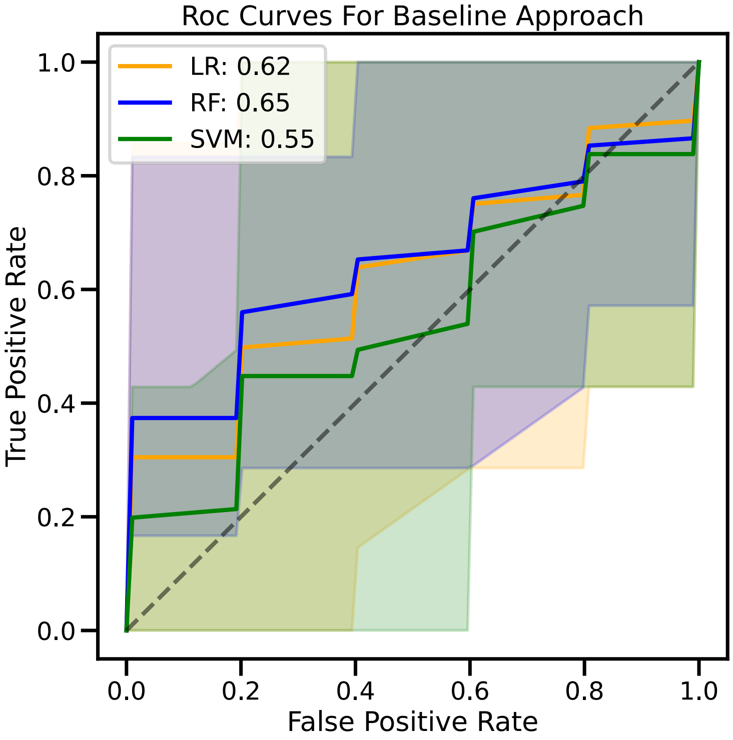

Figure 2: AUC scores of the employed classifiers for the baseline approach

2) Evaluating Different Data Imputation Methods

Next, we evaluate the approaches that employ data imputation and include all the genes in the analysis. We consider two different imputation approaches, namely, K-Nearest Neighbor (KNN)-based imputation and Multivariate Imputation by Chained Equations (MICE). First, we evaluate these two approaches to determine the imputation method to be used in the following approaches. Figure 3 compares the F1 scores of these two imputation methods within the setting of Approach CDT_DI_FI_OHE.

KNN-based imputation provides better estimates than the MICE algorithm in this case. Besides, we also compared the standard deviations of different classifiers under these two imputation methods between different folds in k-fold cross-validation. We observe that with the KNN-based imputation, the standard deviation is much smaller than that with the MICE algorithm. This shows that KNN-based imputation provides more stable results. For the rest of the experiments, we employ KNN-based imputation as our default imputation method. We next compare the performance of the proposed approaches.

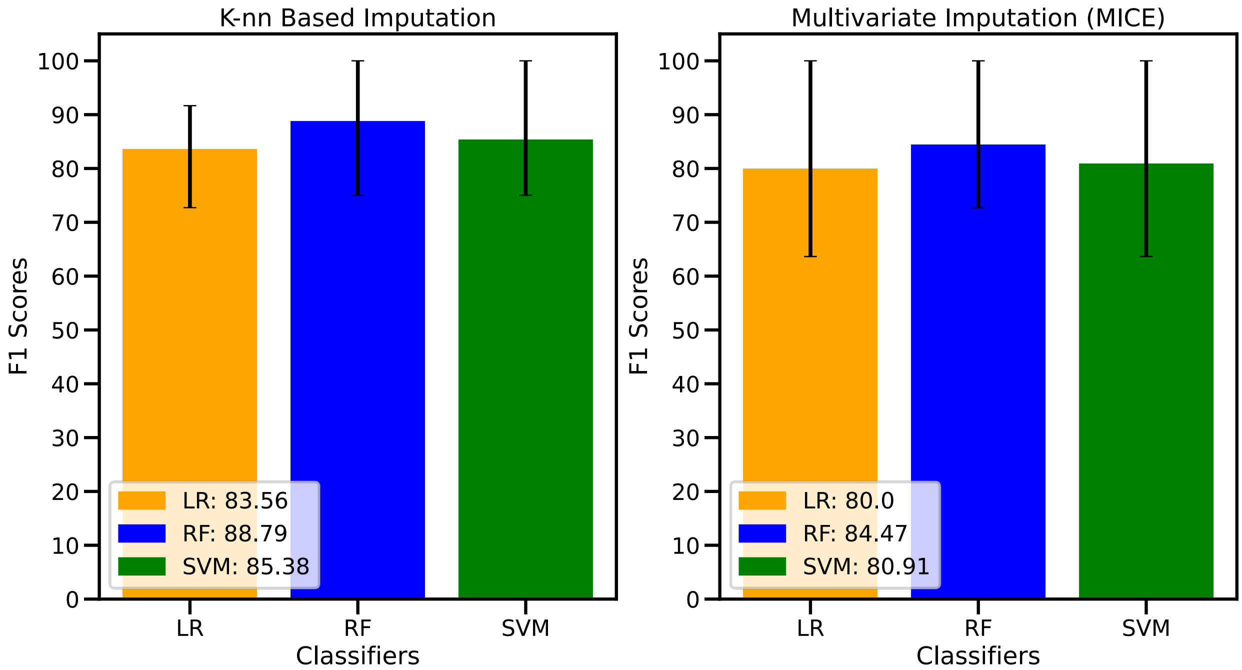

Figure 3: F1 scores of different imputation techniques on decimal forms of data

3) Evaluating Different Approaches

Figure 4 shows that the approach OHE_DI in general provides better performance than the approach OHE_DI_FI (except for LR). We further investigated why the approach OHE_DI performs better. We observe that data imputation in the last step of OHE_DI fills in distinct values for the missing values for different individuals. Thus, we hypothesize that such a way of data imputation may introduce some kind of bias in favor of the healthy (i.e., control) or cancer group depending on which group has the highest number of missing values. To test this hypothesis, we study the comparison between another pair of approaches, CDT_DI_OHE and CDT_DI_FI_OHE, as the latter approach removes the "uniquely-filled missing values" bias from the former one, and provides us with an opportunity to test the effect of this bias.

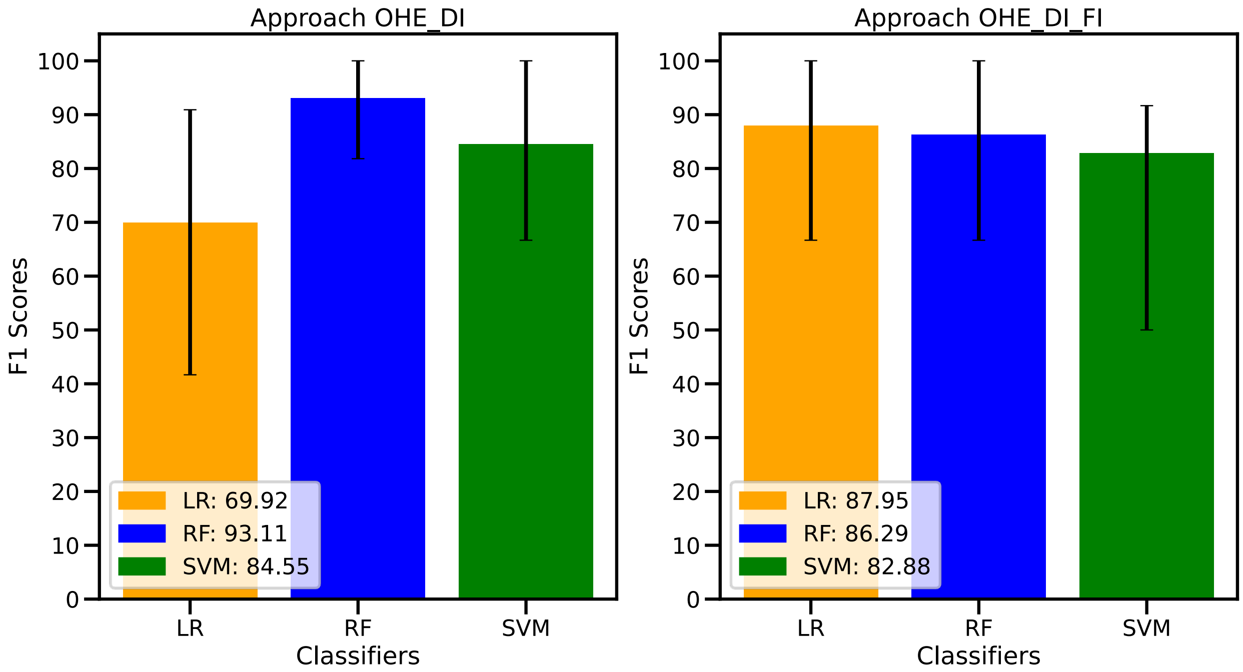

Figure 4: F1 scores of approaches OHE_DI (left) and OHE_DI_FI (right)

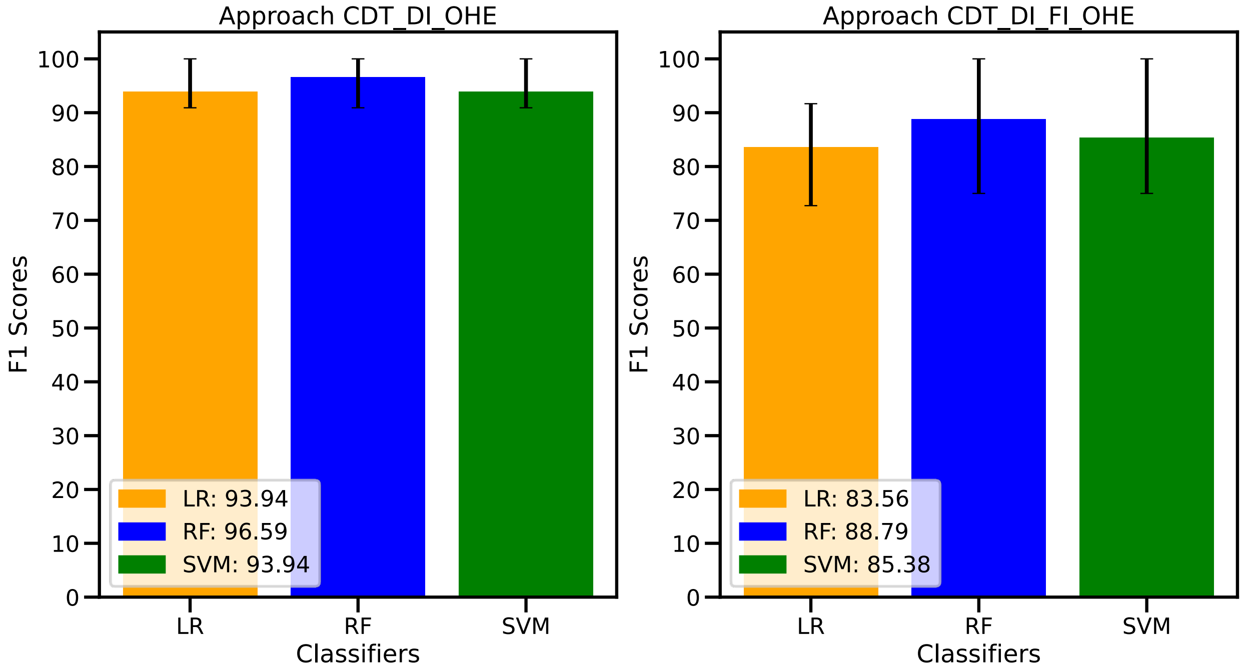

Figure 5: F1 scores for approaches CDT_DI_OHE (left) and CDT_DI_FI_OHE (right)

Figure 5 confirms that the approach without the “uniquely-filled missing values” bias (i.e., CDT_DI_FI_OHE) has lower F1 scores than those with this bias (i.e., CDT_DI_OHE). These results confirm the bias introduced by uniquely-filled values. Therefore, we ignore approaches CDT_DI_OHE and OHE_DI.

We next evaluate the approaches CDT_DI and CDT_DI_FI. Figure 6 shows that the approach CDT_DI outperforms the approach CDT_DI_FI in terms of all compared classifiers. However, based on our observation regarding the “uniquely-filled missing values” bias in the above approaches, we further evaluated the effect of this bias within the context of the approach CDT_DI. The approach CDT_DI_FI removes the “uniquely-filled missing values” bias by converting the floating-point missing values to the nearest integer. The right part of Fig. 6 shows that after removing the effect of the “uniquely-filled missing values” bias, the performance figures of the models decrease significantly. Hence, for comparison purposes, we also ignore the approach CDT_DI. To sum up, among all the evaluated approaches without the “uniquely-filled missing values” bias, CDT_DI_FI_OHE has the best F1 score. In particular, Random Forest performs the best (F1 score: 89%).

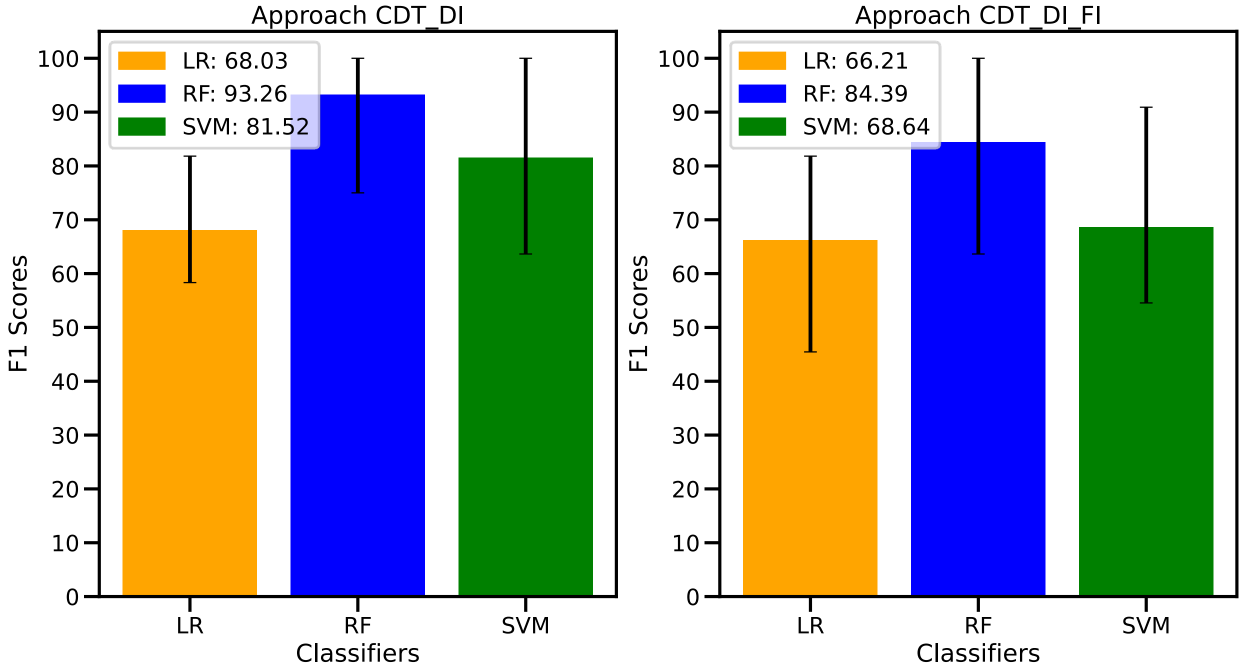

Figure 6: F1 scores for approaches CDT_DI (left) and CDT_DI_FI (right)

4) Comparison to the State of the Art

In this section, we compare our proposed approaches to a widely adopted approach, ColonFlag (i.e., MeScore) [7, 8]. In order to make a fair comparison, we employ ColonFlag under the same settings that we apply our algorithms. The improvement figures in terms of the AUC and F1 scores in each setting are presented in Table 1 and Table 2, respectively. We employ Z-test to check whether the difference between the results from different folds of each experiment is statistically significant. In brief, we make the following observations:

- In terms of the AUC score, our models outperform ColonFlag by up to ~24% in all evaluated settings.

- The improvement figures are statistically significant (p-val < 0.05) under all settings except for OHE_DI_FI.

- In terms of F1 score, our models are better than ColonFlag under all settings (except for OHE_DI_FI) with improvement figures up to 10%.
- The improvement (i.e., 7.5%) of the best performing approach, i.e., CDT_DI_FI_OHE, over ColonFlag is statistically significant (p-val = 0.047).

TABLE I: AUC-BASED COMPARISON OF COLONFLAG ALGORITHM WITH OUR BEST MODEL FOR EACH APPROACH

Approach	Our best algorithm	Our AUC score (%)	ColonFlag AUC score (%)	Improvement (%)	p-val
OHE_DI_FI	LR	**91.76**	88.64	3.52	0.420000
CDT_DI_OHE	RF	**99.43**	91.19	9.04	**0.001400**
OHE_DI	RF	**99.14**	92.50	7.18	**0.001100**
CDT_DI_FI_OHE	RF	**96.07**	81.98	17.19	**0.000001**
CDT_DI	RF	**98.71**	90.64	8.90	**0.002300**
CDT_DI_FI	RF	**94.43**	76.45	23.52	**0.000110**

TABLE II: F1 SCORE-BASED COMPARISON OF COLONFLAG ALGORITHM WITH OUR BEST MODEL FOR EACH APPROACH

Approach	Our best algorithm	Our F1 score (%)	ColonFlag F1 score (%)	Improvement (%)	p-val
OHE_DI_FI	LR	87.95	**88.86**	-1.03	0.820
CDT_DI_OHE	RF	**95.83**	91.44	4.80	0.190
OHE_DI	RF	**93.11**	**93.11**	0.00	1.000
CDT_DI_FI_OHE	RF	**88.79**	82.58	7.51	**0.047**
CDT_DI	RF	**93.26**	90.61	2.92	0.510
CDT_DI_FI	RF	**84.39**	76.52	10.28	0.130

In Figures 7, 8, and 9, we present ROC curves for all of our approaches along with the ColonFlag algorithm. Figure 7 shows that RF provides the best AUC score for both OHE_DI (0.99) and OHE_DI_FI (0.93). In comparison, the ColonFlag model achieves the AUC scores of 0.92 (under the OHE_DI configuration) and 0.88 (under the OHE_DI_FI configuration), which are lower than all of our classifiers except for LR. The results are similar for the approaches CDT_DI_OHE and CDT_DI_FI_OHE in Figure 8, and the approaches CDT_DI and CDT_DI_FI in Figure 9. Figure 10 presents the tabular information in Tables I and II graphically in a single chart.

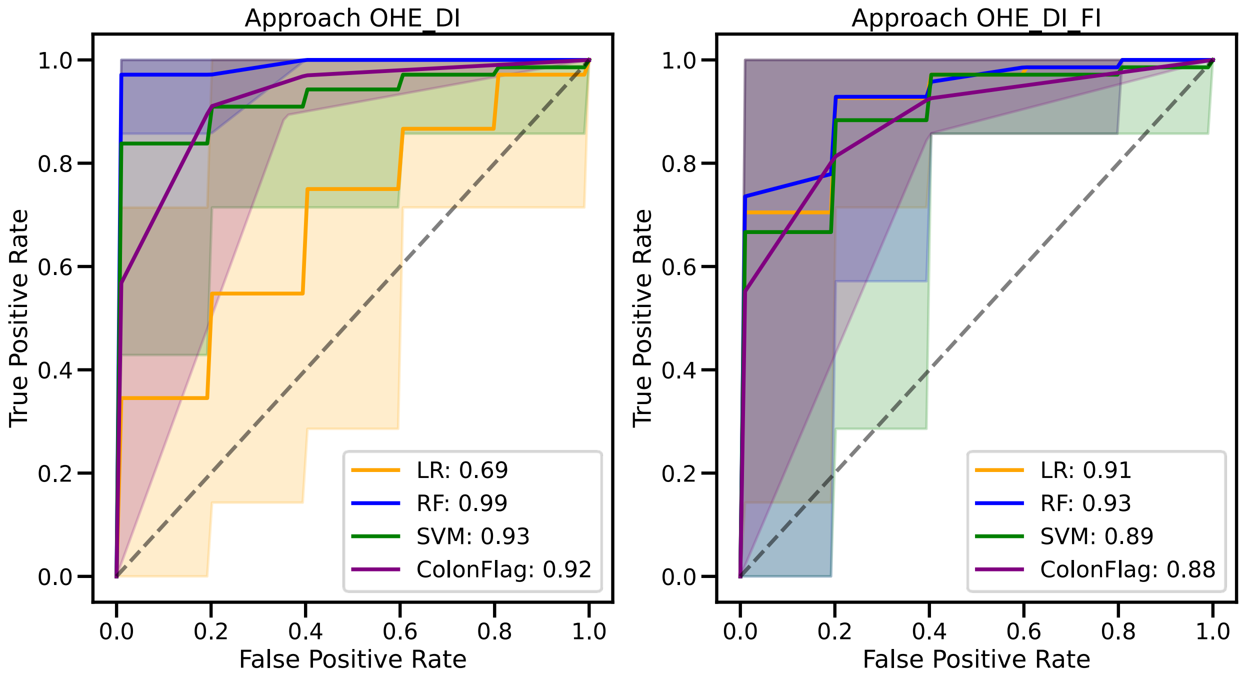

Figure 7: ROC Curves of OHE_DI and OHE_DI_FI approaches

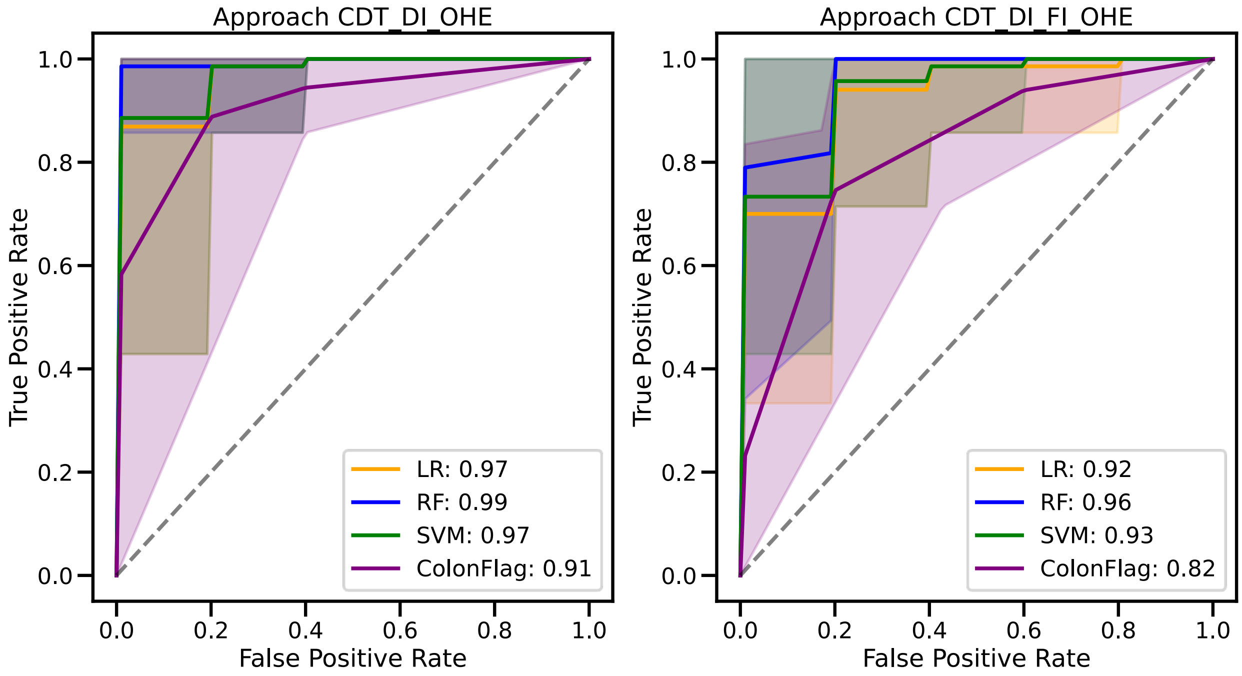

Figure 8: ROC Curves of CDT_DI_OHE and CDT_DI_FI_OHE approaches

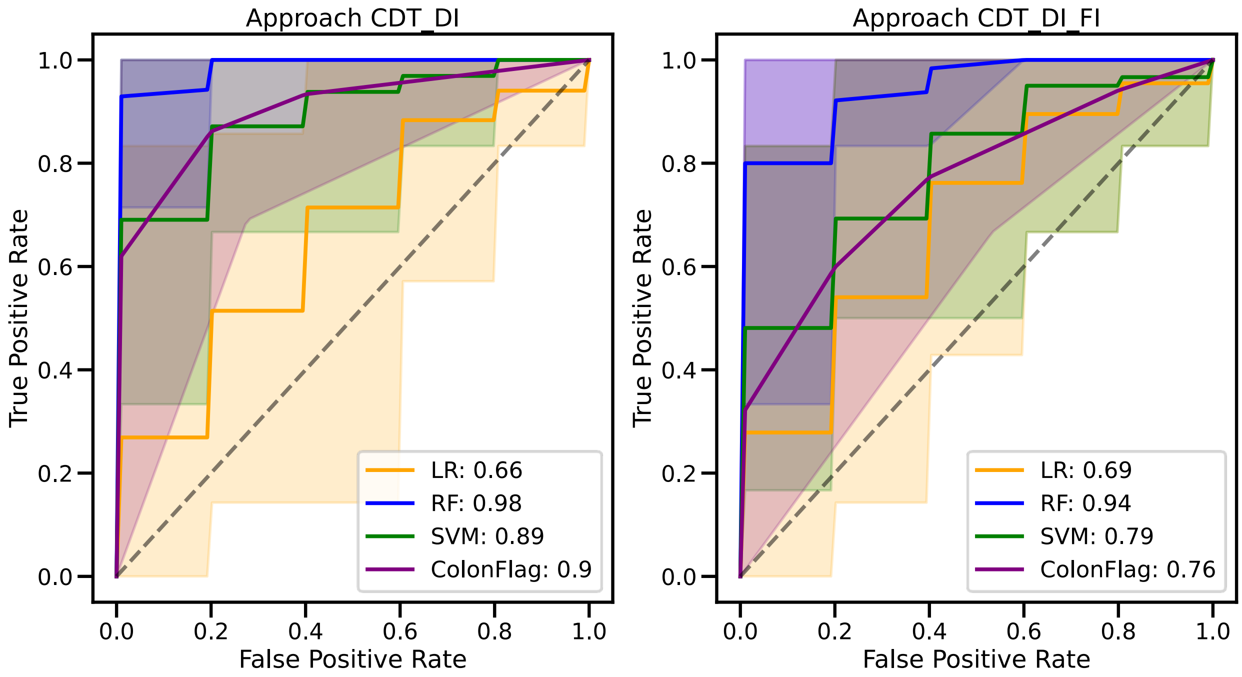

Figure 9: ROC Curves of CDT_DI and CDT_DI_FI approaches

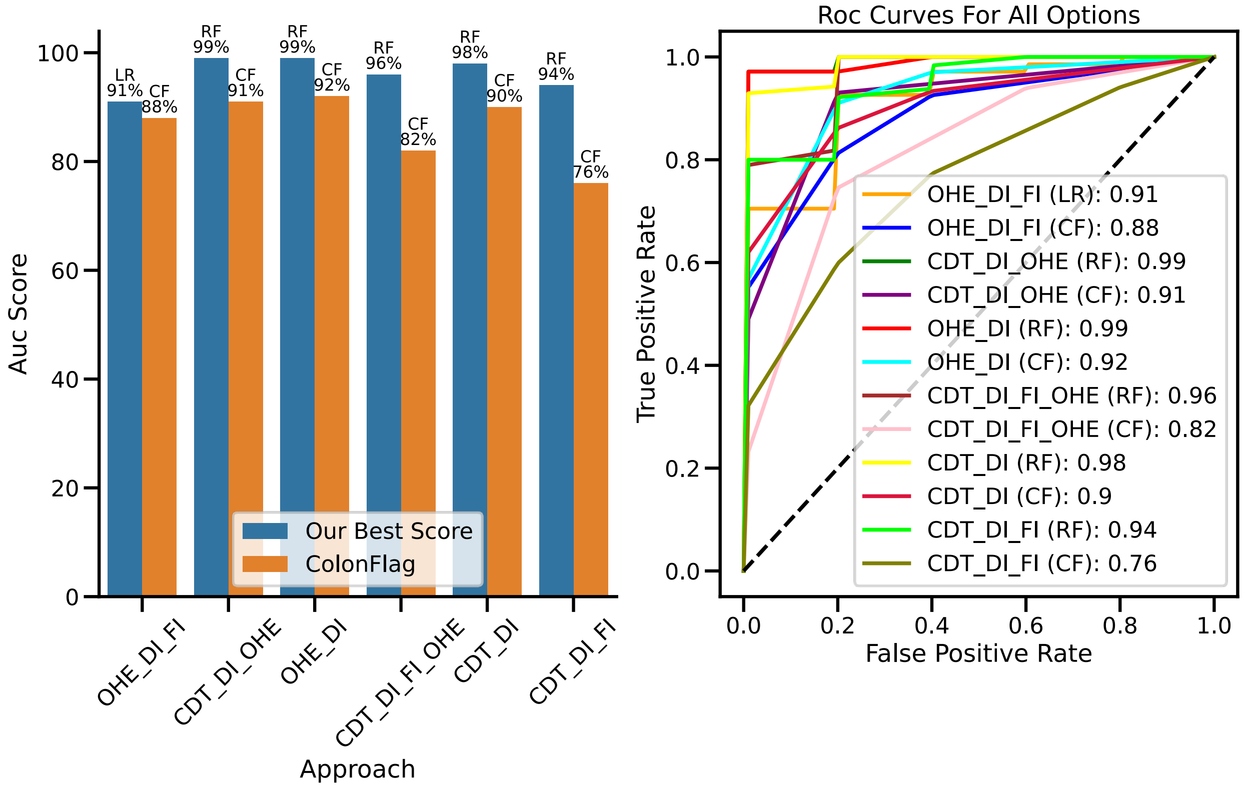

Figure 10: Comparison of the AUC scores (left) and ROC curves (right) of our best algorithm to the ColonFlag algorithm under all configurations

5) Exploratory analysis of Approach CDT_DI_FI_OHE

In this section, we further present an exploratory analysis of our best-performing approach, i.e., CDT_DI_FI_OHE (after filtering approaches with floating-point bias). First, we investigate the most prominent features that differentiate between healthy and colon cancer subjects. To this end, we compute Random Forest Classifier's feature importance values. Figure 11 lists the features in their decreasing order of importance. According to this list, CD40, CD27, and CD28 SNPs constitute the top-5 most important feature set.

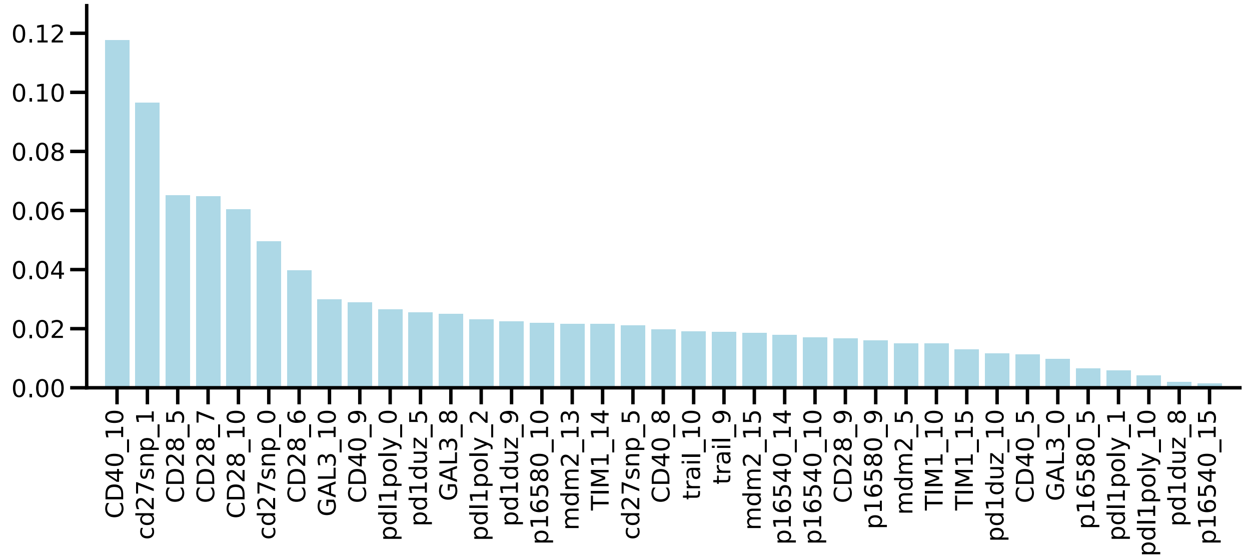

Figure 11: Feature importance scores for the Approach CDT_DI_FI_OHE

Second, we perform an ANOVA analysis and compute the statistical significance of features in differentiating healthy subjects and colon cancer patients. Table 3 provides F-values and p-values of features where the list is sorted by F-values. Genotypes that are represented with numbers (e.g., CD28_7) belong to the imputed data cells. Hence, they are not mapped to a particular genotype. Based on the ANOVA analysis, the following genotypes are the statistically significant (i.e., p-value < 0.01) discriminating features between the control group and colon cancer patients: C/C, C/T, and C/A in CD40, A/T and A/A in CD27, C/C and T/T in CD28. This shows that colon cancer is associated with multiple genes in a complex interaction.

Table III Anova Analysis of Features Sorted by F-values and P-values in Approach CDT_DI_FI_OHE

	F-Values	P-Values
CD40_C/C	**37.5083**	**0.00000001**
cd27snp_A/T	**26.6308**	**0.00000106**
CD28_7	**22.4677**	**0.00000627**
CD28_C/C	**18.6100**	**0.00003450**
CD28_T/T	**13.3250**	**0.00039810**
cd27snp_A/A	**12.7459**	**0.00052518**
CD28_6	**12.2826**	**0.00065641**
CD40_C/T	**7.4010**	**0.00754996**

CD40_C/A	**6.8955**	**0.00983858**
cd27snp_T/T	**6.1596**	**0.01454092**
pdl1poly_A/A	**5.6481**	**0.01915619**
GAL3_C/C	**4.8505**	**0.02966828**
trail_C/T	**4.3091**	**0.04017778**
trail_C/C	**4.3091**	**0.04017778**

Third, we study the decision boundaries of the employed classifiers (see Figure 12). We observe that the RF model demonstrates some level of overfitting, while the LR model is more generic with no sign of overfitting. Hence, LR may be preferred over RF for future datasets, as the performance figures of LR and RF are close.

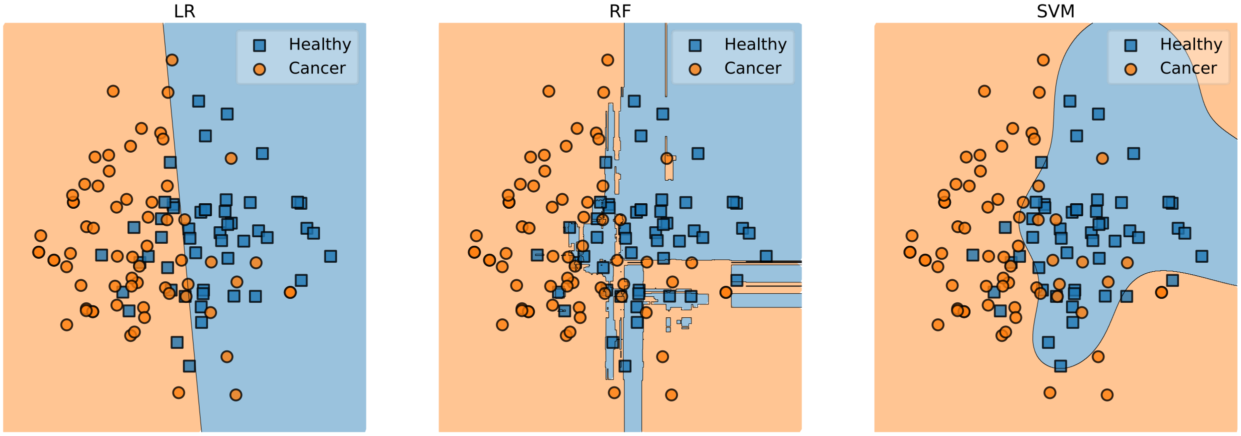

Figure 12: Decision boundaries of the LR, RF, and SVM Classifiers for CDT_DI_FI_OHE

V. DISCUSSION

In this work, we employ supervised learning methods to predict the disposition of individuals to colorectal cancer based on polymorphisms observed in genes p16, MDM2, GAL3, TIM1, TRAIL, PD-1, PD-L1, CD28, CD27, and CD40.

Despite the existence of many works [9-33] on colorectal cancer diagnosis and risk prediction, we note the following drawbacks: First, many of the studies [15, 31] assume that family history data is available, which limits their applicability in practice. Moreover, several other works [10-11] require genome-wide expression data. However, such data may not always be available, and/or it may be costly to produce them. Second, in terms of the performance metrics, several papers [11-12] employ “accuracy”. However, the accuracy score would be misleading in unbalanced datasets. For instance, suppose that a data set contains 40 colon cancer cases and 22 control groups. Then, for a dummy classification model that predicts all subjects as CRC patients, the accuracy would be roughly 65%. Many papers [8, 19] do not even employ well-known metrics like AUC score, F1 score, or even accuracy. Instead, they overly focused on false positive or true positive rates which may provide an incomplete view of the evaluated models’ performance. Last but not the least, none of the previous studies focused on CRC prediction based on major immune checkpoint genotypes. We show that SNPs in immune checkpoint genes are informative for predicting the predisposition to CRC. Besides, since SNP profiles implicitly contain information on family links, family history data is not required as an additional input. The main contributions of this study are as follows:

- Based on the high accuracy of the proposed statistical learning models, we propose the studied 11 SNPs as a novel screening panel for CRC predisposition prediction. The small size of the SNP set makes it practical in clinical settings.
- Since the specimens used in the analysis are obtained from blood, the proposed screening test is non-invasive and has a lower cost when compared to many other prediction approaches in the literature that, for instance, require tumor biopsy [24] or whole genome or exome sequencing [34] which are costly and/or invasive.
- The proposed models under all settings outperform the state-of-the-art approach, ColonFlag [7, 8], by a significant margin in terms of the AUC score.
- We employ custom as well as traditional data encoding methods to represent the non-numeric SNP dataset and demonstrate their impact on improving the prediction accuracy of the models.
- Taking advantage of each and every sample in machine learning studies is vital in particular when the dataset is limited. Missing fields are commonly observed in almost all datasets. Hence, we explore various data imputation methods and offer strategies to fully exploit the available dataset despite the missing fields.

Looking forward, applying transfer learning and fine-tuning pre-trained artificial neural network algorithms are highly popular in recent studies. As part of our future work, we may consider using one of the existing well-known artificial neural network algorithms to obtain better results from the SNP profiles (with more data). Besides, another future work item would be developing a new model from scratch that helps to extract more information from SNPs.

VI. CONCLUSION

In this work, we employ supervised learning methods to predict the disposition of individuals to colorectal cancer based on polymorphisms observed in genes p16, MDM2, GAL3, TIM1, TRAIL, PD-1, PD-L1, CD28, CD27, and CD40. We show that a Random Forest-based classifier performs best in distinguishing colon cancer patients from healthy individuals with an F1 score of 89% and an Area-Under-Curve of 0.96. On the other hand, a Logistic Regression-based classifier provides a more generalized model. We also demonstrate that our presented models outperform a widely-used state-of-the-art algorithm, ColonFlag, by up to 24% improvement in the AUC score.

ACKNOWLEDGEMENTS

This study was funded by the Scientific Research Projects Coordination Unit of Istanbul University (BYP-2018-28139 and TYO-2009-33675).

Author Biograhies

Ali Cakmak received his B.Sc. degree in 2003 from the Computer Engineering Department at Bilkent University (Ankara, Turkey), and his Ph.D. degree in 2008 from the Electrical

Engineering and Computer Science Department at Case Western Reserve University (Cleveland, OH). Then, he moved to the Silicon Valley, and worked as a senior software engineer as part of the Query Optimization Group at Oracle, Inc (Redwood Shores, CA). He is currently a faculty member in the Department of Computer Engineering at Istanbul Technical University. His research interests include bioinformatics, machine learning, data mining, databases.

Huzeyfe Ayaz received his B.Sc. degree in Computer Engineering from Marmara University in 2022. He is now a graduate student in the Department of Computer Engineering at Technical University of Munich. He is interested in Data Science, Machine Learning, and AI.

Soykan Arıkan received his MD degree from the Medical College of Gazi University (Ankara, Turkey) in 1986. He completed his residency in general surgery at Ankara Dışkapı Yıldırım Beyazıt Research and Training Hospital in 1991. He currently serves as a surgent at Başakşehir Çam ve Sakura City Hospital.

Ali R. Ibrahimzada received his B.Sc. degree in Computer Engineering from Marmara University (Istanbul, Turkey) in 2022. He is now a PhD student in the Department of Computer Science at the University Illinois, Urbana Champaign. Ali Reza is mainly interested in Machine Learning, Artificial Intelligence and Data Science.

Seyda Demirkol received her B.Sc. degree in Food Engineering from Gaziantep University in 2009, MSc degree in Medical Biology and Genetics from Istanbul Bilim University in 2015, and PhD degree in Molecular Medicine from Istanbul University. She is currently a faculty member in the Department of Molecular Biology and Genetics at Biruni University. She conducts research primarily in medical biochemistry.

Dilara Sönmez received her B.Sc. and M.Sc degrees in biology from Mustafa Kemal University. She is currently a PhD candidate in the Department of Molecular Medicine at Istanbul University’s Aziz Sancar Institute of Experimental Medicine. Her research interests include medical biochemistry.

Mehmet T. Hakan received his B.Sc. and M.Sc. degrees in biology from Suleyman Demirel University in 2009 and 2013. He is currently a PhD. candidate in the Department of Molecular Medicine at Istanbul University’s Aziz Sancar Institute of Experimental Medicine. His research interests include cancer metabolism and signaling.

Saime T. Sürmen received her PhD degree from the Department of Molecular Medicine at Istanbul University’s Aziz Sancar Institute of Experimental Medicine in 2020. Her research interests include the molecular mechanisms of cancer.

Cem Horozoğlu received his B.Sc. degree in biological sciences from Gaziantep University in 2009, MSc and PhD degrees from in the Department of Molecular Medicine at Istanbul University's Aziz Sancar Institute of Experimental Medicine in 2012 and 2016, respectively. He is currently a faculty member in the Medical College at Biruni University. His research interests include medical biology and genetics.

Mehmet B. Doğan received his MD degree from the Medical College of Istanbul University (Ankara, Turkey) in 1989. He completed his residency in general surgery at Istanbul Haseki Research and Training Hospital in 1998. He currently serves as a surgent at Samatya Research and Training Hospital.

Özlem Küçükhüseyin received her B.Sc. degree in biology from Istanbul University in 2005, MSc and PhD degrees in molecular medicine from Istanbul University in 2008 and 2012, respectively. She is currently a faculty member in the Department of Molecular Medicine at Istanbul University's Aziz Sancar Institute of Experimental Medicine. Her research interests include medical biology and medical genetics.

Canan Cacına received her B.Sc. degree in biology from Marmara University in 2006, MSc and PhD degrees in molecular medicine from Istanbul University in 2008 and 2012, respectively. She is currently a faculty member in the Department of Molecular Medicine at Istanbul University's Aziz Sancar Institute of Experimental Medicine. Her research interests include medical biology and medical genetics.

Bayram Kıran received his B.Sc. degree in chemical engineering from Galatasaray University in 1980, MSc and PhD degrees in immunology from Istanbul University in 1996 and 2003, respectively. He is currently a faculty member in the Department of Genetics and Bioengineering at Kastamonu University. His research interests include immunogy and cancer.

Umit Zeybek received his B.Sc. degree in biology from Istanbul University in 1993, MSc and PhD degrees in molecular medicine from Istanbul University in 1996 and 2003, respectively. He is currently a faculty member in the Department of Molecular Medicine at Istanbul University's Aziz Sancar Institute of Experimental Medicine. His research interests include molecular medicine, neurobiology, and neurochemistry.

Mehmet Baysan received his B.Sc. degree from Bilkent University in Computer Science in 2003, M.S. and Ph.D. degrees in Computer Science from the University of Texas at Dallas in 2005 and 2008, respectively. After his Ph.D., he worked for a year at the Management Department of University of Toronto as a post-doctoral fellow. In 2010, Dr. Baysan joined as a post-doctoral fellow at the Neuro-Oncology Branch of the National Cancer Institute (NCI) of the NIH, USA. After his fellowship, he moved to NYU Cancer Institute and later to Weill Cornell

Medical College as an Associate Research Scientist in 2013 and 2015, respectively. He is currently a faculty member at the Department of Computer Engineering at Istanbul Technical University. Dr. Baysan's research areas are Translational Cancer Research Using Computational Techniques, Algorithm Design and Development, Graph Theory, Combinatorial Optimization.

İlhan Yaylım received her B.Sc. degree in medical biology from Istanbul University in 1991, MSc and PhD degrees in molecular medicine from Istanbul University in 1995 and 2000, respectively. She is currently a faculty member in the Department of Molecular Medicine at Istanbul University’s Aziz Sancar Institute of Experimental Medicine. Her research interests include Medicine, Health Sciences, Fundamental Medical Sciences, and Biochemistry.

Bibliography

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and 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), pp.394-424.

2. Yamada, S., Ajioka, Y., Watanabe, H., Hashidate, H., Takaku, H., Kazama, S., Yokoyama, J., Nishikura, K., Fujiwara, T. and Asakura, H., 2001. Heterogeneity of p53 Mutational Status in Intramucosal Carcinoma of the Colorectum. *Japanese Journal of Cancer Research*, *92*(2), pp.161-166.

3. Hanahan, D. and Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. *Cell, 144*(5), pp.646-674.

4. Patel, M.A., Kim, J.E., Ruzevick, J. and Lim, M., 2015. Present and Future of Immune Checkpoint Blockade: Monotherapy to Adjuvant Approaches. *World J Immunol*, *5*(1).

5. Nirschl, C.J. and Drake, C.G., 2013. Molecular Pathways: Coexpression of Immune Checkpoint Molecules: Signaling Pathways and Implications for Cancer Immunotherapy. *Clinical Cancer Research*, *19*(18), pp.4917-4924.

6. Kirkegaard, H., Johnsen, N.F., Christensen, J., Frederiksen, K., Overvad, K. and Tjønneland, A., 2010. Association of Adherence to Lifestyle Recommendations and Risk of Colorectal Cancer: A Prospective Danish Cohort Study. *Bmj*, *341.*

7. Kinar, Y., Kalkstein, N., Akiva, P., Levin, B., Half, E.E., Goldshtein, I., Chodick, G. and Shalev, V., 2016. Development and Validation of a Predictive Model for Detection of

Colorectal Cancer in Primary Care by Analysis of Complete Blood Counts: A Binational
Retrospective Study. *Journal of the American Medical Informatics Association*, *23*(5),
pp.879-890.

Kinar, Y., Akiva, P., Choman, E., Kariv, R., Shalev, V., Levin, B., Narod, S.A. and
Goshen, R., 2017. Performance Analysis of a Machine Learning Flagging System used to
Identify a Group of Individuals at a High Risk for Colorectal Cancer. *PloS one*, *12*(2),
8. p.e0171759.

Yu, C. and Helwig, E.J., 2022. The role of AI technology in prediction, diagnosis and
9. treatment of colorectal cancer. *Artificial Intelligence Review*, *55*(1), pp.323-343.

Barrier, A., Lemoine, A., Boelle, P.Y., Tse, C., Brault, D., Chiappini, F., Breittschneider,
J., Lacaine, F., Houry, S., Huguier, M. and Van der Laan, M.J., 2005. Colon Cancer
10. Prognosis Prediction by Gene Expression Profiling. *Oncogene*, *24*(40), pp.6155-6164.

Horaira, M.A., Ahmed, M.S., Kabir, M.H., Mollah, M.N.H. and Shah, M.A.R., 2018,
February. Colon Cancer Prediction from Gene Expression Profiles using Kernel-based
Support Vector Machine. In *2018 International Conference on Computer, Communication,*
11. *Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

Alladi, S.M., Ravi, V. and Murthy, U.S., 2008. Colon Cancer Prediction with Genetic
12. Profiles using Intelligent Techniques. *Bioinformation*, *3*(3), p.130.

Gupta, P., Chiang, S.F., Sahoo, P.K., Mohapatra, S.K., You, J.F., Onthoni, D.D., Hung,
H.Y., Chiang, J.M., Huang, Y. and Tsai, W.S., 2019. Prediction of Colon Cancer Stages
13. and Survival Period with Machine Learning Approach. *Cancers*, *11*(12), p.2007.

Hornbrook, M.C., Goshen, R., Choman, E., O'Keeffe-Rosetti, M., Kinar, Y., Liles, E.G.
and Rust, K.C., 2017. Early Colorectal Cancer Detected by Machine Learning Model using
Gender, Age, and Complete Blood Count Data. *Digestive Diseases and Sciences*, *62*(10),
14. pp.2719-2727.

Nartowt, B.J., Hart, G.R., Muhammad, W., Liang, Y., Stark, G.F. and Deng, J., 2020.
Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification.
15. *Frontiers in Big Data*, *3*, p.6.

Patidar, P. and Bhojwani, J., 2013. Identification and Pattern Analysis of SNPs Involved in
16. Colorectal Cancer. *J. Stem Cell Res. Ther*, *3*(144.10), p.4172.

17. Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A., 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC bioinformatics*, *7*(1), pp.1-15.

18. Yuan, H.Y., Chiou, J.J., Tseng, W.H., Liu, C.H., Liu, C.K., Lin, Y.J., Wang, H.H., Yao, A., Chen, Y.T. and Hsu, C.N., 2006. FASTSNP: An Always Up-to-Date and Extendable Service for SNP Function Analysis and Prioritization. *Nucleic Acids Research, 34* (suppl_2), pp.W635-W641.

19. Lin, H.Y., Ann Chen, Y., Tsai, Y.Y., Qu, X., Tseng, T.S. and Park, J.Y., 2012. TRM: A Powerful Two-Stage Machine Learning Approach for Identifying SNP-SNP Interactions. *Annals of Human Genetics*, *76*(1), pp.53-62.

20. Jenkins, M.A., Win, A.K., Dowty, J.G., MacInnis, R.J., Makalic, E., Schmidt, D.F., Dite, G.S., Kapuscinski, M., Clendenning, M., Rosty, C. and Winship, I.M., 2019. Ability of Known Susceptibility SNPs to Predict Colorectal Cancer Risk for Persons With and Without a Family History. *Familial Cancer*, *18*(4), pp.389-397.

21. Fan, C., Lei, X., Guo, L. and Zhang, A., 2019. Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. Neurocomputing, 323, pp.76-85.

22. Qu, K., Gao, F., Guo, F. and Zou, Q., 2019. Taxonomy dimension reduction for colorectal cancer prediction. *Computational biology and chemistry*, *83*, p.107160.

23. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D. and Lyu, C., 2019. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Medical & biological engineering & computing*, *57*(4), pp.901-912.

24. Luo, H., Zhao, Q., Wei, W., Zheng, L., Yi, S., Li, G., Wang, W., Sheng, H., Pu, H., Mo, H. and Zuo, Z., 2020. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Science translational medicine, 12*(524), p.eaax7533.

25. Abdel Ghafar, M.T., Gharib, F., Abdel-Salam, S., Elkhouly, R.A., Elshora, A., Shalaby, K.H., El-Guindy, D., El-Rashidy, M.A., Soliman, N.A., Abu-Elenin, M.M. and Allam, A.A., 2020. Role of serum Metadherin mRNA expression in the diagnosis and prediction of survival in patients with colorectal cancer. *Molecular biology reports*, *47*(4), pp.2509-2519.

26. Xu, P., Zhu, Y., Sun, B. and Xiao, Z., 2016. Colorectal cancer characterization and therapeutic target prediction based on microRNA expression profile. *Scientific reports, 6*(1), pp.1-11.

27. Di, Z., Di, M., Fu, W., Tang, Q., Liu, Y., Lei, P., Gu, X., Liu, T. and Sun, M., 2020. Integrated analysis identifies a nine-microRNA signature biomarker for diagnosis and prognosis in colorectal cancer. *Frontiers in genetics*, *11*, p.192.

28. Birks, J., Bankhead, C., Holt, T.A., Fuller, A. and Patnick, J., 2017. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. *Cancer medicine*, *6*(10), pp.2453-2460.

29. Usher-Smith, J.A., Harshfield, A., Saunders, C.L., Sharp, S.J., Emery, J., Walter, F.M., Muir, K. and Griffin, S.J., 2018. External validation of risk prediction models for incident colorectal cancer using UK Biobank. *British journal of cancer*, *118*(5), pp.750-759.

30. Li, X., Timofeeva, M., Spiliopoulou, A., McKeigue, P., He, Y., Zhang, X., Svinti, V., Campbell, H., Houlston, R.S., Tomlinson, I.P. and Farrington, S.M., 2020. Prediction of colorectal cancer risk based on profiling with common genetic variants. *International Journal of Cancer*, *147*(12), pp.3431-3437.

31. Cueto-López, N., García-Ordás, M.T., Dávila-Batista, V., Moreno, V., Aragonés, N. and Alaiz-Rodríguez, R., 2019. A Comparative Study on Feature Selection for a Risk Prediction Model for Colorectal Cancer. *Computer Methods and Programs in Biomedicine, 177*, pp.219-229.

32. Zhou, D., Tian, F., Tian, X., Sun, L., Huang, X., Zhao, F., Zhou, N., Chen, Z., Zhang, Q., Yang, M. and Yang, Y., 2020. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. *Nature communications*, *11*(1), pp.1-9.

33. Molparia, B., Oliveira, G., Wagner, J.L., Spencer, E.G. and Torkamani, A., 2018. A feasibility study of colorectal cancer diagnosis via circulating tumor DNA derived CNV detection. *PloS one*, *13*(5), p.e0196826.

34. Zhao, L., Liu, H., Yuan, X., Gao, K. and Duan, J., 2020. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC bioinformatics, 21*(1), pp.1-10.

35. Van Buuren, S. and Oudshoorn, K., 1999. *Flexible multivariate imputation by MICE* (pp. 1-20). Leiden: TNO.

List of Figures