Computers in Biology and Medicine 182 (2024) 109098

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

CCPred: Global and population-specific colorectal cancer prediction and metagenomic biomarker identification at different molecular levels using machine learning techniques

Burcu Bakir-Gungor [a], Mustafa Temiz [b,*], Yasin Inal [a], Emre Cicekyurt [a], Malik Yousef [c,d]

[a] Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, 38080, Turkey
[b] Department of Electrical and Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, 38080, Turkey
[c] Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel
[d] Galilee Digital Health Research Center (GDH), Zefat Academic College, Israel

ARTICLE INFO

Keywords:
Colorectal cancer
Machine learning
Biomarkers
Enzyme
Pathway
Species
Microbiome
Metagenomic

ABSTRACT

Colorectal cancer (CRC) ranks as the third most common cancer globally and the second leading cause of cancer-related deaths. Recent research highlights the pivotal role of the gut microbiota in CRC development and progression. Understanding the complex interplay between disease development and metagenomic data is essential for CRC diagnosis and treatment. Current computational models employ machine learning to identify metagenomic biomarkers associated with CRC, yet there is a need to improve their accuracy through a holistic biological knowledge perspective. This study aims to evaluate CRC-associated metagenomic data at species, enzymes, and pathway levels via conducting global and population-specific analyses. These analyses utilize relative abundance values from human gut microbiome sequencing data and robust classification models are built for disease prediction and biomarker identification. For global CRC prediction and biomarker identification, the features that are identified by SelectKBest (SKB), Information Gain (IG), and Extreme Gradient Boosting (XGBoost) methods are combined. Population-based analysis includes within-population, leave-one-dataset-out (LODO) and cross-population approaches. Four classification algorithms are employed for CRC classification. Random Forest achieved an AUC of 0.83 for species data, 0.78 for enzyme data and 0.76 for pathway data globally. On the global scale, potential taxonomic biomarkers include ruthenibacterium lactatiformanas; enzyme biomarkers include RNA 2′ 3′ cyclic 3′ phosphodiesterase; and pathway biomarkers include pyruvate fermentation to acetone pathway. This study underscores the potential of machine learning models trained on metagenomic data for improved disease prediction and biomarker discovery. The proposed model and associated files are available at https://github.com/TemizMus/CCPRED.

1. Introduction

Cancer remains as an important global health problem and one of the leading causes of preventable death worldwide. Its prevalence is characterized by high mortality rate that continues to rise exponentially, largely due to the factors such as population aging and environmental factors [1]. In particular, colorectal cancer (CRC) is the third most common cancer worldwide [2]. The complicated interplay between epigenetic, genetic, and environmental factors plays an important role in the development of CRC [3]. Remarkably, CRC is reported to have the fastest growth among all cancers worldwide, with 4.7 million cases expected by 2070 [4]. Furthermore, this malignancy accounts for approximately 10 % of all newly diagnosed cancers. Early detection of CRC has been shown to increase the 5-year relative survival rate by nearly 90 %, providing the impetus for increased research efforts to understand and combat this disease [5].

Given the close relationship between the human gut microbiome and the incidence of CRC, it is increasingly proposed to use the data resulting from the investigation of the gut microbiome as a diagnostic tool for colorectal cancer. Using multi-omics profiling and integrated approaches, it has been shown that CRC-related microbial taxa, metabolites, and changes in DNA methylation-related gene expression can serve as detectable biomarkers in multiple dimensions (Y. [6,7]). This multi-faceted approach exploits the dynamic interplay between the

* Corresponding author.
E-mail address: mustafa.temiz@agu.edu.tr (M. Temiz).

https://doi.org/10.1016/j.compbiomed.2024.109098
Received 20 April 2024; Received in revised form 29 August 2024; Accepted 31 August 2024
Available online 17 September 2024

human gut microbiome–comprising the collective genomes of the microbial community in our gastrointestinal tract–and various human "-omes." Therefore, the study of the human gut microbiome in people suffering from CRC is an extremely important area where intensive research is conducted.

Recently, a growing number of studies [8–10] (H. [11,12]) (F. [13–16]) have eagerly attempted to establish a link between the gut microbiota and colorectal cancer (CRC). These studies have provided compelling evidence of significant and notable changes in the microbiota of individuals suffering from colorectal cancer [1,4]. While these studies have underscored the central role of the gut microbiota in the pathophysiology of CRC, it is worth noting that the field is still in its early stages before it comes a fully established science. Different classification criteria and methods used in previous research studies have made the identification of key species involved in the development and progression of CRC somewhat challenging [17,18]. Given the complexity of metagenomic studies, the application of machine learning techniques has become increasingly important in this field, offering the possibility of answering a wide range of questions. In this context, the idea of finding taxonomic biomarkers for diseases by correlating the microbiome with disease states through taxonomically informed feature selection has been developed [19]. In a study conducted by Bose *et* al. [20], the researchers used data from different populations spanning four different regions—Argentina, Chile, Vietnam, and India. They used taxonomic microbiome data analyzed at the genus level and focused primarily on examining the microbiomes that exert significant influence in the Indian population. Their findings highlighted the prominent role of the *Prevotella* in colorectal cancer in the Indian population. Their machine learning model obtained a high Area Under the Curve (AUC) value of 86 % using the Random Forest classifier. This high AUC value indicates the success of their model in discriminating between positive and negative cases and demonstrates the effectiveness of the developed model in predicting CRC based on the analyzed microbiome data. A high AUC value also indicates that the model has a good balance between sensitivity and specificity, which is critical for accurate disease prediction. In addition, Bose *et* al. underscored the existence of a global CRC microbiome and pointed out that certain observations from their study are consistent with the findings from other studies in various regions. To this end, they emphasize that there may be common characteristics and microbial influences in colorectal cancer development and progression that transcend geographic boundaries and are important on a global scale. Zhen *et* al. conducted an in-depth study involving CRC patients and controls from diverse populations [21]. In their comprehensive review article, they meticulously examined the results from an extensive pool of 700 different studies. Through this exhaustive analysis, they brought to light the importance of *Fusobacterium nucleatum* in the context of CRC. This highlights the importance of this particular microorganism in the landscape of CRC research and its potential as a biomarker or target for further investigation and therapeutic intervention. Yu *et* al. conducted a thorough population survey on CRC that included diverse populations from various countries [22]. In their comprehensive review of 5696 different studies, they reported important insights into the microbiome associated with CRC and they identified numerous influential factors that contribute to the development and progression of this disease. Their comprehensive research sheds light on the complex interplay of factors and microbial communities involved in CRC and contributes significantly to the understanding of this disease.

Several studies have attempted to explore the composition and functionality of the gut microbiome in the context of CRC, but a comprehensive study of the gut microbiome in CRC patients has yet to be conducted. The present study attempts to close this gap by developing a robust classification model that facilitates the diagnosis of colorectal cancer. This can be accomplished by carefully analyzing a variety of CRC-related metagenomics datasets using a spectrum of feature selection techniques and machine learning methods. The objectives of the present study include the identification of biomarkers associated with CRC at the species level, as well as at the enzyme and pathway levels that influence host metabolism. The aim of the present study is to identify the most informative features for optimal CRC classification with a reduced feature set containing data on species, enzymes and pathways, leading to better classification results. Essentially, another goal is to develop a classification model that can perform best even with fewer features. To accomplish this, utilization of a dataset comprising metagenomic information from 9 different datasets, encompassing case and control groups from 8 diverse populations, is employed. Relative abundance values of the species, enzymes, and pathways are obtained from the same samples and presented as three different datasets. These three datasets are created both on the global scale and in a population-specific manner. In the present study, emphasis is placed on the utilization of feature selection algorithms to optimize feature sets, thereby achieving more accurate classification using fewer features. The performances of the models that use the union and the intersection of the features that are selected by different feature selection algorithms are investigated. Based on the superior performance of the union features that are selected by different feature selection methods, the union features are reranked by rescaling their importance value for each population and calculating the median of these values to produce a final ranking. Based on this ranking, top 20features are highlighted as potential biomarkers, and the top 5 features are biologically validated via conducting a literature search. Furthermore, the present study aims to identify population-specific CRC metagenomic biomarkers by developing population-specific models. In order to identify CRC associated taxonomic biomarkers, enzymes, and pathways that are specific to different populations, population-specific metagenomic datasets are utilized. In order to evaluate the performance of these models, within-population, leave-one-dataset-out (LODO) and cross-population analyses are conducted.

The rest of this manuscript is organized as follows. The Materials and Methods sections present the details of the CRC-associated metagenomic dataset and our methodology. The Results section presents our findings. In the Discussion section, the effective species, enzymes, and pathways for CRC identified by the proposed method are biologically validated by literature studies. The Conclusions section evaluates the findings, and the implications of the research are summarized. In addition, possible directions for future research are discussed in light of the current findings. This section highlights both the general contributions of the study and provides suggestions for future studies in this field.

2. Materials and methods

2.1. Dataset

Beghini *et* al. (2021) compiled a total of 1262 metagenomic samples (662 controls and 600 CRC patients) at different molecular levels (species, enzymes and pathways) from nine different BioProject datasets (PRJEB7774, PRJNA531273, PRJNA447983, PRJDB4176, PRJEB12449, PRJEB27928, PRJDB4176, PRJEB10878 and PRJEB6070) [23]In the original study, the raw microbiome DNA of each sample was downloaded from the respective project site and MetaPhlAn [24] and HUMAnN [23] were used to calculate the relative abundance values of all subgroups of each dataset. To ensure data quality, they applied quality filtering to meet the standards outlined in the Human Microbiome Project Consortium SOP, as referenced in (Thomas et al., 2019). The CRC-associated metagenomics dataset used in the present study includes the relative abundance values of 917 different species, 2895 different enzymes, and 551 different pathways calculated for 1262 samples from nine different datasets. The distribution of data by population, the number of patients (CRC) and healthy samples in each population, the number of male and female samples, the minimum age, maximum age and mean age, and the minimum and maximum body mass index (BMI) values are shown in Table 1.

Table 1
Distribution of data by populations and the numbers of samples in each population.

Name of population	# of Samples	# of CRC Samples	# of Healthy samples	Male/ Female	min age/ max age/ mean age	min BMI/ max BMI
Austria (AUT)	107	46	61	64/43	43/ 86/ 67,01	17.99/ 34.14
China (CHN)	128	75	53	81/47	34/ 89/ 64,23	17.10/ 35.10
Germany (DEU)	125	60	65	73/52	28/ 90/ 59,56	13.30/ 35.80
France (FRA)	114	53	61	57/57	25/ 87/ 63,47	15.00/ 40.00
Indian (IND)	60	30	30	30/30	22/ 75/ 50,65	17.69/ 36.40
Italy (ITA)	106	57	49	72/34	57/ 84/ 63,07	19.71/ 38.53
Japan (JP)	80	40	40	45/35	28/ 78/ 61,13	17.34/ 28.38
Japan (JPN)	438	187	251	253/ 185	21/ 79/ 61,33	16.14/ 39.32
United State of America (USA)	104	52	52	74/30	31/ 85/ 61,53	17.43/ 35.18

Note: The population of Japan is listed twice in this table. These are data for different regions of the same population (Japan), which are represented by different abbreviations (JP and JPN).

2.2. Proposed method

The proposed method involves a combination of computational analyses and advanced machine learning techniques with the following main objectives: (i) by applying well-known feature selection algorithms to metagenomic data globally, developing a robust CRC prediction model is aimed; (ii) analyzing the effect of the identified features on each population is aimed; (iii) Another objective is attaining a rigorous evaluation of the classification model developed at the global/ population-specific scale, having the potential to improve the accuracy and effectiveness of CRC diagnosis globally and in population specific manner; (iv) The final aim is to identify global and population-specific metagenomic biomarkers across different molecular levels, including species, enzymes, and pathways. The performance of these biomarkers is systematically assessed using machine learning techniques, providing insight into their diagnostic potential.

To this end, firstly, all features for each molecular level (species, enzyme, and pathway) are analyzed using a global perspective for CRC prediction. Then, feature selection algorithms are applied to reduce the number of features. Finally, CRC prediction is performed, using the union and intersection of the features that are selected by different feature selection algorithms. For population-based experiments, the following three approaches are utilized, i.e., within-population, leave-one-dataset-out (LODO) and cross-population. In these approaches, the intersection and union features that are obtained in the global analyses are also used, and the CRC prediction performance is evaluated. Furthermore, comparative performance evaluations are conducted against a grouping based feature-selection method that utilizes biological domain knowledge (i.e., microBiomeGSM [25]. microBiomeGSM is a novel approach developed for the identification of taxonomic biomarkers from metagenomic data, based on a methodology based on grouping, scoring, and modelling (G-S-M) [26]. This approach aims to detect disease-associated taxonomic biomarkers by developing an efficient machine learning model that analyses taxonomically transformed microbiome sequencing datasets. The microBiomeGSM tool utilizes species-level information and groups of taxonomic features at different levels such as genus, family, and order. Since microBiomeGSM performs best at genus taxon level, the results at this level are included within the comparison. G-S-M approach [26] is extensively usedto develop different tools like maTE [27], PriPath [61], GediNET [29], miRcorrNet [30], 3Mint [31], GeNetOntology [32], TextNetTopics [33], TextNetTopics Pro [34] microBiomeGSM [25], miRGediNET [35], miRdisNET [36], miRModuleNet [28], CogNet [62] and AMP-GSM [37], which integrate biological networks and prior knowledge to provide a comprehensive understanding of genetic interactions. For an extensive review of feature selection approaches based on the grouping of features, the reader is referred to Ref. [[60],38].

The following subsections present further details of the methodology.

2.3. Identification of important features (species, enzymes, and pathways), using different feature selection algorithms on the global scale

Using CRC-associated metagenomic datasets, a set of machine learning models is built to discriminate CRC from control samples using different feature selection algorithms and classification models. As illustrated in Fig. 1, there are two main parts in the workflow: (i) feature selection to identify the most relevant species, enzymes, and pathways for the development of CRC diagnostic model; (ii) model building and classification. As shown in Fig. 1, four different machine learning algorithms (Random Forest, LogitBoost, AdaBoost, and Decision Tree) are used for the classification task. For feature selection, Information gain (IG), SKB [39], and extreme gradient boosting (XGBoost) (T [40]) are utilized. In addition to using traditional feature selection algorithms individually, a hybrid approach is employed, combining SKB, IG, and XGBoost methods as follows.

1. The importance scores of each feature are calculated using the feature selection algorithms mentioned above. Of the importance scores generated by tree-based classification algorithms.
2. To ensure consistency, min-max scaling is applied to these values.
3. Features with importance scores below a certain threshold (0.5) are discarded.
4. If a feature that passes this filtering process is identified by all three feature selection algorithms, it is assigned to the intersection set.
5. If a feature that passes this filtering process is identified by at least one of the three feature selection algorithms, it is assigned to the union set.

Classification is performed using the following three sets of features.

i) ***Using all features without applying any feature selection algorithm:*** For each dataset (species, enzyme, and pathway), classification is performed using the above-mentioned machine learning algorithms and all features without applying any feature selection method. There are 917, 2895 and 551 features in species, enzyme, and pathway datasets, respectively.

ii) ***Using intersection of the features that are commonly identified by all three feature selection algorithms:*** In this method, the performance of machine learning methods is evaluated based on the intersection of the features that are selected by different feature selection algorithms. Each feature in the intersection set is among the top 100 features and has feature importance score higher than 0.5 for each feature selection algorithm. In the intersection set, different numbers of features were obtained for the species, enzyme, and pathway datasets. The number of

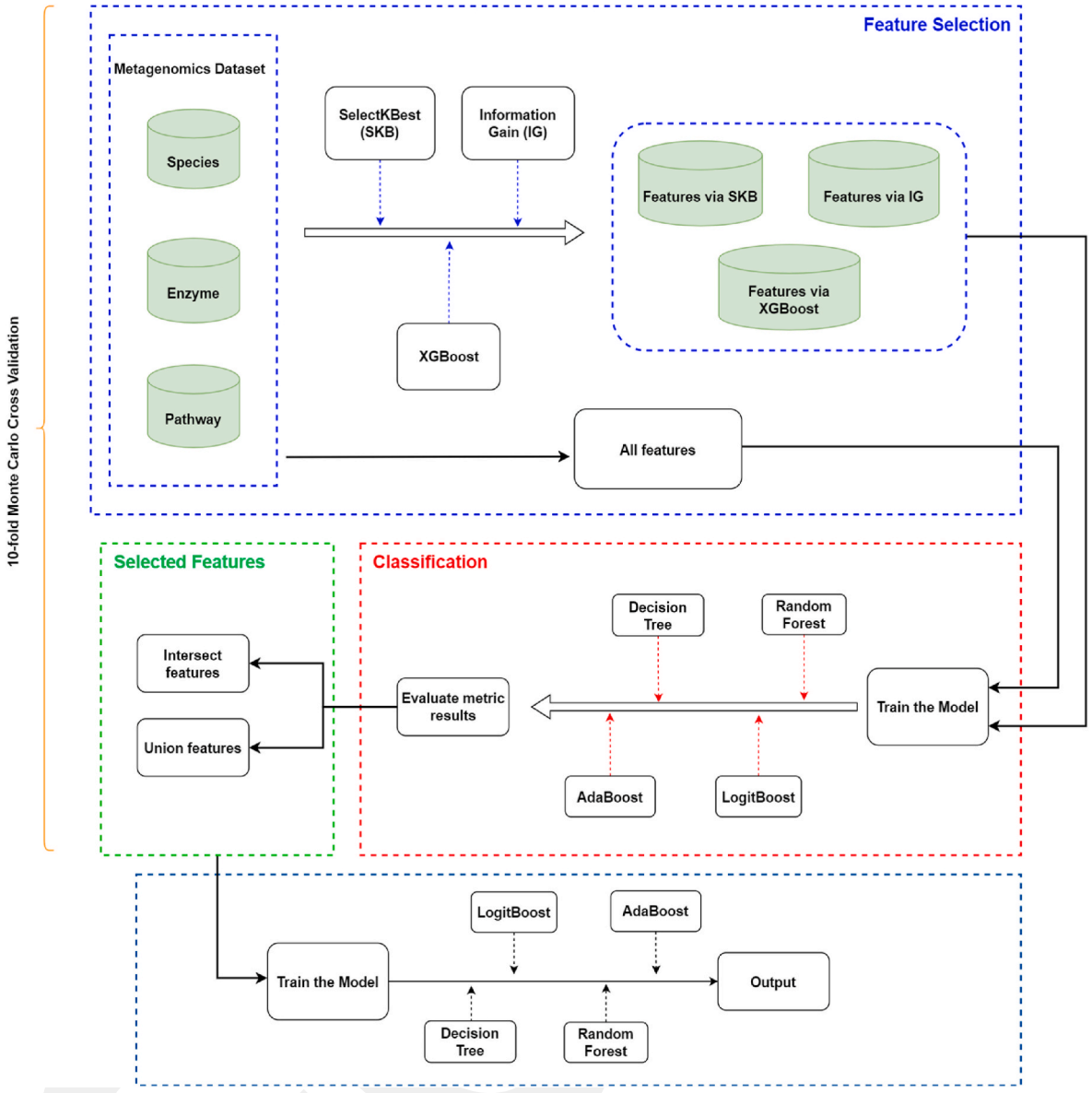

Fig. 1. Workflow of the methodology.

features in the intersection set is 9, 25, and 6 for species, enzyme, and pathway datasets, respectively.

iii) ***Using union of features that are selected by at least one of the three different feature selection algorithms:*** This method evaluates the performance of different machine learning methods by utilizing the union of the features that are selected by at least one of the feature selection algorithms. Since the focus is on the top 100 features selected by different feature selection algorithms and the features with importance scores higher than 0.5, different numbers of features are extracted for the species, enzyme, and pathway datasets. The number of features in the union set is 21, 295, and 38 for species, enzyme, and pathway datasets, respectively.

In this study, CRC-associated biomarkers are identified globally and in a population-based manner. Thus, the present study goes beyond biomarker identification by aiming to discover population-specific metagenomic biomarkers in three different datasets (species, enzymes, and pathways).

2.3.1. Population based CRC classification using metagenomic features (species, enzyme, and pathway)

To identify population-specific taxonomic biomarkers, enzymes, and pathways, the methods described in Section 2.2.1 are applied to population-specific datasets related to CRC. To this end, three different meta-analyses are performed: within-population, LODO and cross-population. This multi-faceted approach provides a deeper understanding of how metagenomic biomarkers differ across populations, and it also offers valuable insights into the potential applicability of these biomarkers in different contexts. In these experiments, the focus is on the RF algorithm because it outperforms the other classification algorithms in the preliminary analysis (global perspective) and RF is the most commonly used algorithm in human microbiome studies as reported by Ref. [19]. Fig. 2 demonstrates our workflow for the population-specific evaluation.

In order to calculate the performance metrics separately for each population dataset, within-population analysis was applied. A 10-fold Monte Carlo Cross Validation (MCCV) is performed for each population dataset. Data from each population are selected as 80 % for training and 20 % for testing. The average AUC values and standard deviations

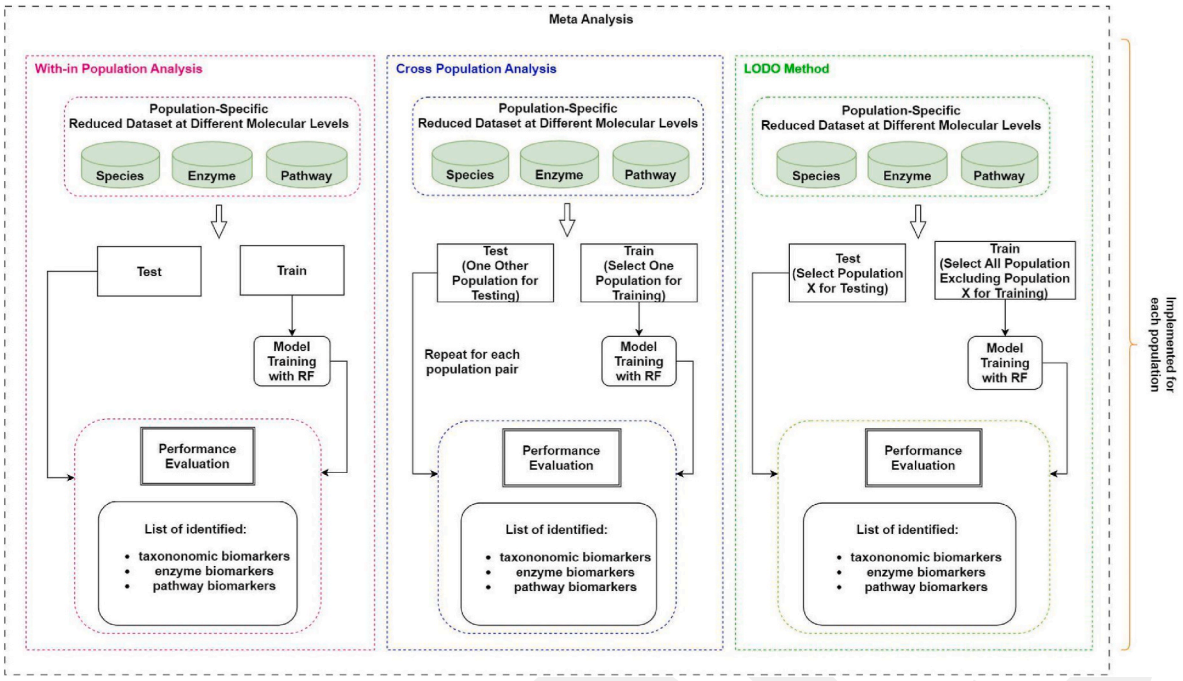

Fig. 2. Population-specific evaluation of the models that are developed with the intersection/union features of the CRC-associated metagenomics dataset.

are calculated.

In the cross-population analysis, the model is trained with the data from a specific population and the developed model is tested with the data from another population that was not used for training. The testing part of this experiment is repeated separately for each population that was not used in the training part. Each dataset is used to train the model, and each time the remaining datasets are used separately as test data.

In Leave One Dataset Out (LODO) analysis, the data from a specific population is kept as the test set, while the data from all other populations that were not selected for testing are combined and used for training the model. This experiment is repeated for each population.

2.3.2. Identification of CRC-associated species, enzymes, and pathways as potential biomarkers across different populations

In order to calculate the final importance score of each feature across different populations, the population-specific contribution of each feature to the Random Forest classifier, which outperforms other classification algorithms, is evaluated. In these evaluations, depending on the best mean scores obtained using the within-population analysis, possible biomarkers obtained by union features are used for species data, intersection features for enzyme data, and union features for pathway data. The intersection and union features determined by the feature selection algorithms for species, enzyme (EC Number) and pathway data are shared in Supplementary Tables. Then, using the min-max scaling method, the median values of these scaled feature importance scores are obtained and a ranking list is generated. Using this ranking list, the top 20 features for species, enzyme, and pathway datasets are identified, and the top 5 features are explored in depth via referring to the biological literature.

2.3.3. Implementation

The models are developed using the KNIME platform [41], and the H2O and scikit-learn libraries are utilized. The predictive performance of the models is evaluated using the metrics of accuracy, F1 score, and AUC (Area Under the Receiver Operating Characteristic Curve). As shown in Figs. 1 and 2, the generated models are tested on three separate datasets including the relative abundance values of microorganisms (species), enzymes, pathways that are calculated for CRC patients and healthy samples.

3. Results

The main objective of this study is i) to predict CRC with high machine learning performance using few features, and ii) to identify the species, enzymes, and pathways associated with this disease. Another important goal of this study is to provide more accurate and specific information for CRC diagnosis and treatment through global and population-specific experiments. Using different feature selection algorithms, dominant features will be highlighted and the effect of these features on CRC prediction performance will be investigated in detail using the proposed approaches. In this context, the results are evaluated from two different perspectives. Firstly, using a global dataset, the performance of the classification algorithms is comparatively evaluated using (i) all features, (ii) the intersection of the features that are selected by all feature selection algorithms (intersection features), and (iii) the union of the features that are selected by at least one feature selection algorithm (union features). Secondly, using population-specific datasets, within-population, LODO and cross-population experiments are performed; and the performance of the generated models using different feature sets, as explained for the global analysis, are comparatively evaluated. This performance evaluation is repeated for three different datasets containing species, enzymes, and pathways as features for the same samples.

3.1. Performance evaluation of the global models

The experiments conducted on the global CRC-associated metagenomics dataset, which includes samples from different populations, offer general insights into the CRC development.

3.1.1. Performance evaluation of the global models using all features

Using 10-fold cross-validation, 4 different classifiers (Random Forest, AdaBoost, LogitBoost, and Decision Tree) are run on CRC-associated metagenomic data without applying any feature selection method; and the average performance metrics and standard deviations are shown in Table 2. For species, enzyme, pathway and combined enzyme and

Table 2
Performance evaluation of the classification algorithms using all features within the CRC-associated species, enzyme, pathway and combined enzyme and pathway datasets.

CRC-Associated Species Dataset						
Model	Accuracy	Sensitivity	Specificity	F1-Measure	AUC	Precision
AdaBoost	0.53 ± 0.008	0.66 ± 0.06	0.76 ± 0.06	0.69 ± 0.002	0.79 ± 0.02	0.53 ± 0.004
DT	0.52 ± 9.93E-9	0.64 ± 0.05	0.70 ± 0.05	0.69 ± 0.001	0.68 ± 0.04	0.53 ± 9.93E-9
LogitBoost	0.53 ± 0.003	0.65 ± 0.07	0.76 ± 0.05	0.69 ± 0.001	0.81 ± 0.02	0.53 ± 0.001
RF	0.52 ± 9.93E-9	0.66 ± 0.02	0.84 ± 0.06	0.69 ± 0.01	0.83 ± 0.03	0.53 ± 9.93E-9
CRC-Associated Enzyme Dataset						
Model	Accuracy	Sensitivity	Specificity	F1-Measure	AUC	Precision
AdaBoost	0.69 ± 0.03	0.63 ± 0.1	0.73 ± 0.03	0.72 ± 0.01	0.76 ± 0.01	0.65 ± 0.04
DT	0.49 ± 0.02	0.60 ± 0.1	0.60 ± 0.05	0.66 ± 0.02	0.63 ± 0.01	0.49 ± 0.01
LogitBoost	0.68 ± 0.02	0.63 ± 0.06	0.73 ± 0.04	0.72 ± 0.009	0.76 ± 0.01	0.64 ± 0.03
RF	0.67 ± 0.02	0.61 ± 0.07	0.75 ± 0.07	0.73 ± 0.01	0.78 ± 0.009	0.61 ± 0.02
CRC- Associated Pathway Dataset						
Model	Accuracy	Sensitivity	Specificity	F1-Measure	AUC	Precision
AdaBoost	0.62 ± 0.05	0.62 ± 0.05	0.69 ± 0.03	0.70 ± 0.01	0.71 ± 0.02	0.58 ± 0.04
DT	0.49 ± 9.93E-9	0.59 ± 0.04	0.65 ± 0.07	0.66 ± 9.93E-9	0.59 ± 0.03	0.49 ± 9.93E-9
LogitBoost	0.61 ± 0.03	0.57 ± 0.08	0.69 ± 0.04	0.70 ± 0.01	0.70 ± 0.006	0.57 ± 0.03
RF	0.69 ± 0.02	0.59 ± 0.07	0.76 ± 0.08	0.73 ± 0.008	0.76 ± 0.01	0.65 ± 0.02
CRC-Associated Combined Enzyme and Pathway Dataset						
Model	Accuracy	Sensitivity	Specificity	F1-Measure	AUC	Precision
AdaBoost	0.55 ± 0.05	0.66 ± 0.04	0.72 ± 0.05	0.69 ± 0.01	0.75 ± 0.05	0.54 ± 0.03
DT	0.54 ± 0.04	0.62 ± 0.06	0.63 ± 0.06	0.68 ± 0.007	0.62 ± 0.04	0.54 ± 0.04
LogitBoost	0.54 ± 0.05	0.62 ± 0.06	0.74 ± 0.04	0.69 ± 0.01	0.75 ± 0.03	0.53 ± 0.02
RF	0.56 ± 0.07	0.59 ± 0.08	0.76 ± 0.05	0.70 ± 0.03	0.76 ± 0.04	0.55 ± 0.06

pathway datasets, the highest performance is obtained with the Random Forest classification algorithm. Accuracy, specificity, sensitivity, F1-measure, precision, and AUC are used as evaluation criteria. As shown in Table 2, the best results were obtained with the Random Forest classifier with an AUC of 0.83 for the species dataset. For the enzyme dataset, the best results were obtained with the Random Forest classifier with an AUC of 0.78. For the pathway dataset, the best results were obtained with the Random Forest classifier with an AUC of 0.76. For the combined enzyme and pathway dataset, the best results were obtained with the Random Forest classifier with an AUC of 0.76.

3.1.2. Performance evaluation of the global models using feature selection

In this subsection, the results obtained using feature selection algorithms are presented. 4 different feature selection algorithms including traditional feature selection algorithms (SelectKBest (SKB), Information Gain (IG), (XGBoost)) and microBiomeGSM [25] which is a biological domain-knowledge based feature selection technique; and 4 different classification algorithms (AdaBoost, LogitBoost, Decision Tree and Random Forest) are used for CRC prediction. The grouping of the features is performed at the genus level while running the microBiomeGSM tool. Fig. 3 provides a comparison of the AUC values obtained with different classification algorithms using different feature selection techniques on the global scale. Fig. 3 (A) shows a comparative assessment of the AUC values obtained using the CRC-associated species dataset. Fig. 3 (B) and 3 (C) show the comparative AUC values that are obtained for the enzyme and pathway datasets, respectively. Since microBiomeGSM uses Random Forest as the classification algorithm, its performance is compared with other feature selection algorithms only using RF classifier. In Fig. 3, the first 2 bars in each graph (colored in red and orange) represent the intersection and union features, respectively. The third bar in each graph (colored in yellow) shows the performance of the model using all features (when no feature selection method is used). The fourth bar in each graph (colored in light blue) shows the results obtained using XGBoost feature selection algorithm. The fifth bar in each graph (colored in green) shows the results obtained using Information Gain (IG) feature selection algorithm. The sixth bar in each graph (colored in dark blue) shows the results obtained using Select K Best feature selection algorithm. The last bar in the Random Forest classifier and in the mean values (colored in steel blue) show the results obtained with the microBiomeGSM tool, where the features are grouped on genus level.

Fig. 3 (A) illustrates the mean AUC values for different feature selection algorithms, calculated by averaging the AUC values obtained from various classification algorithms. The highest AUC averaged over different classifiers (an AUC of 0.78) is obtained by using all features. Among different classifiers, for each feature selection method except for the SKB feature selection method, the highest AUC values are obtained using the RF algorithm. This underlines the strength of the RF algorithm. Fig. 3 (A) shows that for the species dataset, the RF classifier achieves an AUC of 0.83 when all features (917 features) are used. For the same dataset, the RF classifier using the intersection of selected features by SKB, IG, and XGBoost algorithms (9 features) results in 0.78 AUC. For the same dataset, the RF classifier using the union of selected features by SKB, IG, and XGBoost algorithms (21 features) yields an AUC of 0.79. The union feature set including relative abundance values of only 21 species shows very close performance with the model that uses the relative abundance values of 917 species.

For the enzyme dataset, when the mean AUC values shown in Fig. 3 (B) are analyzed, it is observed that the highest AUC is achieved by using the union of the features selected by different feature selection methods. In the experiments performed on the enzyme dataset, the highest AUC values are obtained with the Random Forest classification algorithm. Examining Fig. 3 (B), an AUC of 0.76 is achieved with the RF classification algorithm using all features (2895 features); an AUC of 0.76 is obtained using the intersection of the features (25 features) that are identified by SKB, IG, and XGBoost feature selection algorithms; and an AUC of 0.79 is achieved using the union of the features (295 features) that are detected by different feature selection algorithms. These results imply that by examining only the abundance values of 25 enzymes in the intersection set, one can perform classification as successfully as analyzing the 2895 enzymes (all features in the enzyme dataset). Hence, it can be deduced that the method of the present study resulted in higher

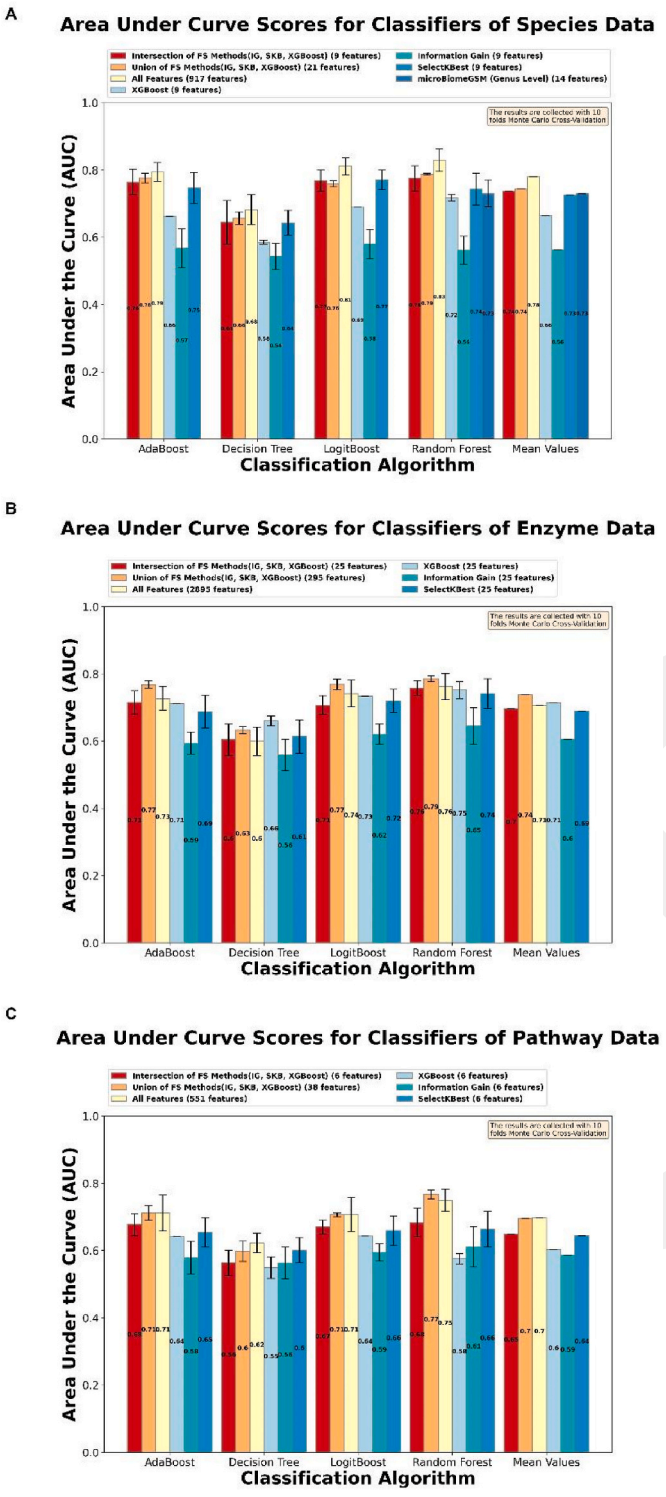

Fig. 3. Performance evaluation of different feature selection techniques using different classifiers on CRC-associated A) species, B) enzyme and C) pathway datasets.

AUC values using fewer features for the CRC-associated enzyme dataset.

When the mean values among different classifiers are analyzed on the pathway dataset, it is seen in Fig. 3 (C) that the AUC values that are obtained by using all features and by using the union features are similar. When the AUC values on the CRC-associated pathway dataset is analyzed, one can notice that the highest AUC values are obtained using the Random Forest classification algorithm among different classifiers. Examining Fig. 3 (C), an AUC of 0.75 is achieved with the RF classification algorithm using all features (551 features); an AUC of 0.68 is achieved using 6 features that are commonly identified by SKB, IG and XGBoost feature selection algorithms; and an AUC of 0.77 is obtained by using 38 features that are identified by at least one of the feature selection algorithms (union features). Obtaining higher AUC values using only 38 features that are identified by at least one of the feature selection algorithms emphasizes that these pathways are more informative compared to using all 551 pathways.

When the average performance metrics presented in Fig. 3 are analyzed, one can conclude that for the species and pathway datasets, the SKB algorithm as a feature selection method and the Random Forest algorithm as a classifier outperform other competitors. For the enzyme dataset, the XGBoost algorithm as a feature selection method and the Random Forest algorithm as a classifier outperform other tested algorithms. In the majority of the analyses performed for species, enzyme and pathway data, the SKB feature selection method and the Random Forest classification algorithm performed better than other tested feature selection algorithms. When the averages for the species, enzyme, and pathway dataset are analyzed using the all features (Fig. 3), the SKB feature selection method outperforms the other feature selection algorithms in terms of the AUC metric for the species and pathway datasets. For the enzyme dataset, the XGBoost feature selection method outperforms the other methods. When analyzing the performance of the classification algorithms, the best classification results for the dataset of species, enzyme and pathway were obtained with the Random Forest classification algorithm.

3.2. *Performance evaluation of population-specific models*

Table 2 and Fig. 3 show that the Random Forest classifier performs better than the Decision Tree, LogitBoost and AdaBoost algorithms in classifying CRC on the global-scale experiments. Therefore, the Random Forest classifier is deliberately utilized in the population-specific experiments. In particular, the potential of the Random Forest classification algorithm in population-based analysis of species, enzyme, and pathway datasets is investigated using the intersection/union of features identified by different feature selection algorithms. In order to evaluate the performance of the models, the following three methods are applied: i) within-population, ii) leave-one-dataset-out (LODO) and iii) cross-population. In this way, population-specific biomarkers associated with CRC are highlighted as a function of populations. Fig. 4 shows the experimental results of the within-population and the LODO approach for population-specific analyses. In Fig. 4, the mean AUC values among different populations are also plotted. Fig. 4 shows a comparison of the AUC values obtained for different population datasets using i) all features, and ii) the intersection and union of the features that are detected by the feature selection algorithms. In each figure, the first 2 bars (colored in red and orange) indicate the AUC values of the models, using intersection and union features, respectively. The third bar (colored in yellow) shows the performance of the model when all features are used without applying feature selection methods. For the species dataset, the last bar (colored in steel blue) in Fig. 4 indicates the findings of microBiomeGSM tool where the features are grouped on the genus level. Fig. 4 (A) shows the within-population and Fig. 4 (B) shows the LODO experimental results, respectively. Fig. 4 (A) and (B) illustrate the AUC values of the models for the species, enzyme, and pathway datasets, respectively. The experimental results for species data are shown in Fig. 4 (A.1), for enzyme data in Fig. 4 (A.2) and for pathway data in Fig. 4 (A.3).

3.2.1. *Findings for within-population analysis*

In the within-population analysis, each metagenomics dataset that is specific to a population is analyzed separately. In these experiments, the RF algorithm is used as the classification algorithm and a 10-fold Monte Carlo cross-validation method is applied. Through our within-

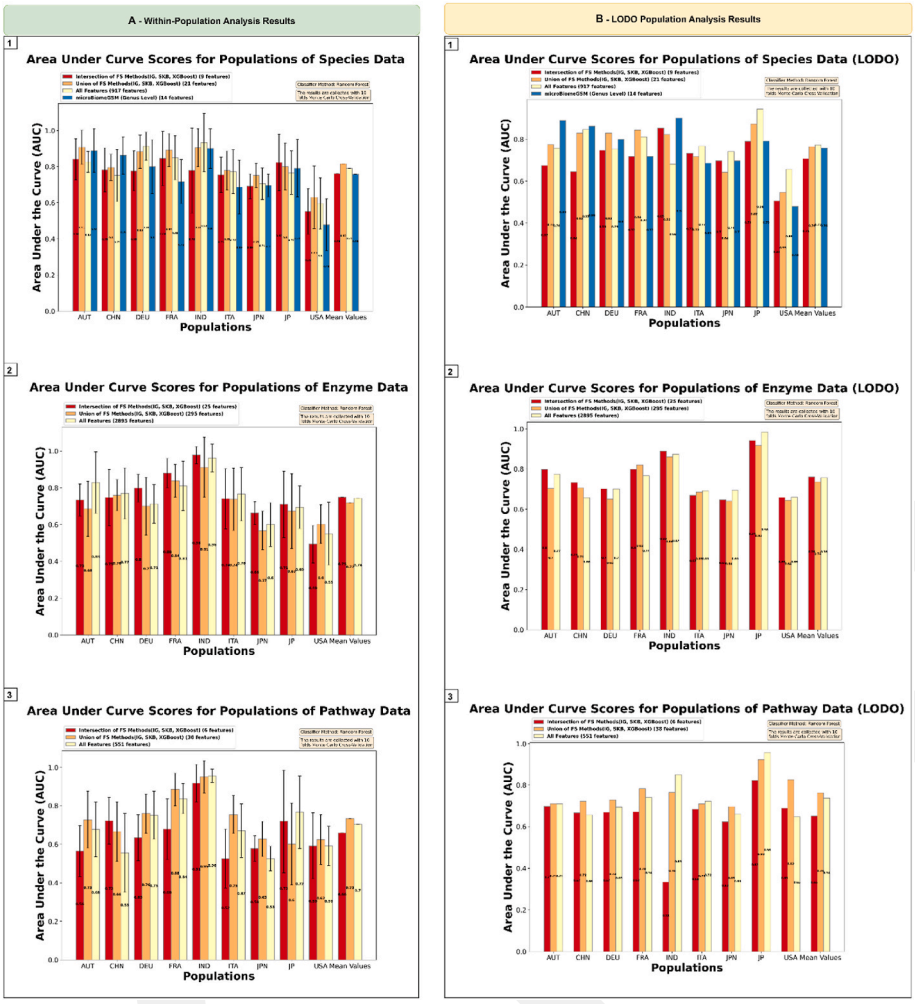

Fig. 4. AUC values that are obtained as part of the within-population analysis and LODO analysis using 1) species, 2) enzyme and 3) pathway features of the CRC-associated population-specific metagenomic datasets. AUC values of different feature selection methods are represented by different colors and comparatively evaluated.

population experiments using the species, enzyme, and pathway datasets, it was observed that for most of the cases (17 out of 27 comparisons), the AUC values that are obtained using the union features were higher than the AUC values obtained using all features (as shown in Fig. 4 (A)). More specifically, Fig. 4 (A.1) shows the AUC values obtained for the CRC-associated species dataset. As can be seen in Fig. 4 (A.1), for five out of nine populations ("AUT", "FRA", "ITA", "JPN" and "USA"), the AUC values obtained using the union features (21 features) are higher than the AUC values reported with other approaches (intersection of features, microBiomeGSM, and all features). Among the AUC values that are obtained for different CRC-associated species datasets belonging to different populations, the highest AUC value (an AUC of 0.91) is reported for "AUT" population using the 21 features included in the union set. In only one of the nine different datasets ("JP" population), the AUC value (0.82) that is obtained using the intersection of the features (9 features) is higher than the AUC values reported by other approaches. Using 9 common features from the species datasets, the highest AUC values (0.84) are noted for "AUT" and "FRA" populations. For two of the nine different datasets ("DEU", "IND"), the AUC metrics obtained using all features (917 features) are better than the AUC values reported by other approaches. The highest AUC value among these populations was obtained for "IND" with an AUC of 0.93 using all features. When mean AUC values are analyzed, one can observe from Fig. 4 (A.1) that the highest AUC value of 0.82 is obtained with the union features (relative abundance values of 21 species).

Fig. 4 (A.2) illustrates the AUC values derived from the CRC-associated enzyme dataset, showcasing outcomes obtained through the utilization of union features, intersection features, and the inclusion of all features. As shown in Fig. 4 (A.2), for five of the nine different datasets ("DEU", "FRA", "IND", "JPN", and "JP") the models that are developed using intersection features outperform other models in terms of the AUC metrics. The "IND" population dataset attains the highest AUC value of 0.98. In three out of the nine distinct datasets ("AUT", "CHN", and "ITA"), AUC metrics derived from utilizing all features exhibit a slight improvement over the AUC values obtained from alternative models. Among the AUC values observed across various CRC-associated enzyme datasets from diverse populations, the "AUT" population achieved the highest AUC value of 0.83 when utilizing all features (2895 features). Fig. 4 (A.2), the analysis exclusively focusing on all features (depicted by the yellow bar) reveals that the "IND" population yields the highest AUC value of 0.96 among all AUC values derived from the utilization of all features. In solely one of the nine distinct datasets (specifically, the "USA" population), the AUC value of 0.60 achieved through the utilization of feature union (comprising 295 features) surpasses the AUC values obtained through alternative methodologies. In Fig. 4 (A.2), exclusive analysis of union features (represented by the orange bar) comprising 295 features reveals that the "IND" population exhibits the highest AUC value of 0.91 among all AUC values derived from the utilization of union features. Upon examination of mean AUC values depicted in Fig. 4 (A.2), it becomes evident that the intersection

features (comprising relative abundance values of 25 enzymes) yield the highest AUC value of 0.75. Fig. 4 (A.2) indicates that the frequency of superior performances achieved through the intersection of features identified by feature selection methods exceeds that of other methodologies.

Fig. 4 (A.3) presents the AUC outcomes for pathway data, comparing the utilization of union features identified by various feature selection methods, intersection features identified by different feature selection methods, and all features without any feature selection method. As depicted in Fig. 4 (A.3), among six out of nine populations ("AUT", "DEU", "FRA", "ITA", "JPN", and "USA"), the utilization of union features (comprising 38 features) yields higher AUC values compared to other methodologies, including intersection of features, microBiomeGSM, and utilization of all features. Among the AUC values derived from distinct colorectal cancer (CRC)-associated species datasets across various populations, the most elevated AUC value, reaching 0.88, is documented for the "FRA" population. This result is achieved through the utilization of 38 features encompassed within the union set. In Fig. 4 (A.3), exclusive examination of union features (represented by colored orange bars) reveals the attainment of the most prominent outcome for the "IND" population concerning the AUC. Specifically, the utilization of union features yields an AUC value of 0.95 for this population subgroup. Among the nine distinct datasets examined, namely the "CHN" population, it is noteworthy that the utilization of the intersection of features results in a comparatively higher AUC value of 0.72, surpassing the AUC values obtained through alternative approaches in this specific population cohort. Fig. 4 (A.3) demonstrates that exclusive analysis of intersection features (depicted by colored red bars) yields the most superior outcome for the "IND" population concerning the area under the curve (AUC) values obtained using union features. Specifically, an AUC value of 0.92 is observed in this population subgroup. As depicted in Fig. 4 (A.3), it is notable that among the nine distinct datasets analyzed, specifically the "IND" and "JP" populations, the models constructed utilizing all features exhibit superior performance compared to alternative models, as evidenced by (AUC) metrics. Among the calculated AUC values derived from various colorectal cancer (CRC)-associated pathway datasets across diverse populations, the most elevated AUC value of 0.96 is documented for the "IND" population. This notable outcome is achieved through the utilization of all features, encompassing a total of 2895 features. Upon examination of the mean (AUC) values, it becomes apparent from Fig. 4 (A.3) that the most notable AUC value, reaching 0.73, is attained through the utilization of union features. These features comprise relative abundance values associated with 38 pathways. Fig. 4 (A.3) illustrates a greater frequency of superior performances achieved through the integration of features identified by feature selection methods, compared to alternative approaches.

3.2.2. *Findings for leave one dataset out (LODO) analysis*

In the Leave One Dataset Out (LODO) analysis, performance is evaluated by excluding one population dataset for repeated testing for species, enzyme, and pathway datasets, separately. In this experiment, one population is selected for testing and the remaining datasets are combined and used as training data. This experiment is repeated for each population. In these experiments, the RF algorithm is used as the classification algorithm and the 10-fold Monte Carlo cross-validation method is applied. A comparison of the AUC values obtained during the LODO analysis by using different feature selection methods (intersection, union, and all features) is shown for each population in Fig. 4 (B.1), Fig. 4 (B.2), and Fig. 4 (B.3) for the species, enzyme, and pathway datasets, respectively. The first 2 bars in each graph (colored red and orange) represent the intersection of features identified commonly by all three feature selection algorithms, and the union of features selected by at least one of the three different feature selection algorithms, respectively. The third bar shows the performance of the model when all features were used without applying a feature selection method. The fourth bar for the species data shows the resulting AUC scores determined using the microBiomeGSM tool. Since the microBiomeGSM is a tool that uses species data, only comparisons with species data are included in these experiments.

In Fig. 4 (B.1), it is evident that within the species dataset, the AUC outcomes derived from the union features, selected via feature selection methods, surpass those obtained through alternative approaches in two out of the nine distinct datasets ("DEU" and "FRA"). The highest AUC value among the various CRC-associated species datasets from diverse populations, standing at 0.84, is documented for the "FRA" population, employing the 21 features encompassed within the union set. In Fig. 4 (B.1), exclusive examination of the union features (depicted by the orange bar) reveals the most favorable outcome for the "JP" population, exhibiting an AUC of 0.87 among the AUC values derived from these features. As depicted in Fig. 4 (B.1), in the case of four out of the nine distinct datasets ("ITA", "JPN", "JP", and "USA"), the models constructed utilizing all features exhibit superior performance compared to alternative models, as evidenced by AUC metrics.

In LODO analysis, the "JP" population dataset yields the highest AUC value of 0.94. This result also represents the optimal one achieved for the species data. As depicted in Fig. 4 (B.1), the models constructed using intersection features do not demonstrate superior performance compared to other models in terms of AUC metrics across any of the nine distinct datasets. The highest achievement observed with the intersection of features identified by the feature selection methods is an AUC of 0.85 for the "IND" dataset. As illustrated in Fig. 4 (B.1), among the nine diverse datasets, the models employing microBiomeGSM exhibit superior performance in AUC metrics for three datasets, namely "AUT", "CHN", and "IND". In the array of AUC values derived from various CRC-associated species datasets across diverse populations, the apex AUC value of 0.90 is documented for the "IND" population, utilizing the 14 features amalgamated in the union dataset. Upon scrutiny of the mean AUC values, Fig. 4 (B.1) reveals that the utmost AUC value of 0.77 is acquired when employing all features, encompassing the relative abundance values of 917 species.

Fig. 4 (B.2) shows the AUC values obtained for the CRC-associated enzyme dataset using union features, intersection features, and all features. As can be seen in Fig. 4 (B.2) for the enzyme dataset, in three of the nine different datasets ("AUT", "CHN", and "IND"), the AUC results obtained with the intersection features obtained by the feature selection methods are higher than the AUC values obtained with the other approaches. Among the AUC values that are obtained for the different CRC-associated enzyme datasets belonging to different populations, the highest AUC value (an AUC of 0.89) is reported for the "IND" population using the 25 features included in the intersection set. In Fig. 4 (B.2), when only the intersection features are analyzed (colored in red), the highest result is obtained for the "JP" population among the AUC values obtained using the intersection features (AUC of 0.94). As shown in Fig. 4 (B.2), for three of the nine different datasets ("ITA", "JPN" and "JP"), the models that are developed using all features outperform the other models in terms of AUC metrics. Among the AUC values that are obtained for the different CRC-associated enzyme datasets belonging to different populations, the highest AUC value (an AUC of 0.98) is reported for the "JP" population using the 2895 features included in the all-feature set. This result is also the best result obtained for the LODO approach for enzyme data. When analyzing the mean AUC values, Fig. 4 (B.2) shows that the highest AUC value of 0.76 is obtained for the union features and all features.

As can be seen in Fig. 4 (B.3) for the pathway dataset, in five of the nine different datasets ("CHN", "DEU", "FRA", "JPN" and "USA"), the AUC results obtained with the union features achieved by the feature selection methods are higher than the AUC values obtained with the other approaches. Among the AUC values that are obtained for the different CRC-associated pathway datasets belonging to different populations, the highest AUC value (an AUC of 0.83) is reported for the "USA" population using the 38 features included in the union set. Fig. 4

(B.3), when only the union features are analyzed (colored in orange), the highest result is obtained for the "JP" population among the AUC values achieved using the union features (AUC of 0.92). As shown in Fig. 4 (B.3), for three of the nine different datasets ("IND", "ITA" and "JP"), the models that are developed using all features outperform the other models in terms of AUC metrics. The highest AUC value of 0.96 is obtained for the "JP" population dataset. This result is also the best result obtained for the pathway data. As shown in Fig. 4 (B.3), the models that are developed using intersection features do not outperform the other models in terms of AUC metrics for any of the nine different datasets. The highest AUC that is obtained with the intersection of the identified features is 0.82 for "JP" population. When analyzing the mean AUC values, it can be seen in Fig. 4 (B.3) that the highest AUC value of 0.76 is obtained with the union features (relative abundance values of 38 pathways).

3.2.3. Findings for cross-population analysis

In this evaluation method, the model is trained using data from a specific population and the developed model is tested separately using data from another population that was not used for training. This experiment is repeated for each population. One by one, every dataset is used for training the model and each time the remaining datasets are used separately as the test data. Fig. 5(A) and. 5 (B) and Fig. 5 (C) show the AUC values of the machine learning models that are developed using the cross-population technique and applied on species, enzyme, and pathway datasets, respectively.

By examining the average AUC values obtained throughout the within-population experiments, the feature selection approach yielding the highest mean AUC value is selected for the cross-population analysis. It can be observed from Fig. 4 (A.1) that for the species dataset, the mean AUC value generated by the union features (21 species) is higher than other tested methods. Hence, we used the reduced dataset including only these 21 features through cross-population experiments of the species dataset. For the enzyme dataset, as shown in Fig. 4 (A.2), the models using the intersection features (25 enzymes) resulted in the highest mean AUC value. Hence, the reduced dataset containing solely these 25 features was utilized in cross-population experiments of the enzyme dataset. For the pathway dataset, as shown in Fig. 4 (A.3), the models using the union features (38 pathways) generated the highest AUC values when averaged over different populations. Hence, the reduced dataset comprising solely these 38 features was employed in cross-population experiments of the pathway dataset. The Random Forest algorithm was used as the classifier in cross-population analyses since the Random Forest classifier resulted in the highest AUC in our previous experiments with other classifiers tested on the global perspective (as shown in Fig. 3).

As depicted in Fig. 5 (A), cross-population experiments revealed that an AUC of 0.80 or higher was achieved in 7 out of 72 instances when employing the union features, consisting of 21 features, within the species dataset. Fig. 5 (A) shows that using the relative abundance values of the 21 species within the species dataset, the highest AUC of 0.85 is obtained when "JPN" population is used as the training set and "JP" is used for the test set. In Fig. 5 (A), it is demonstrated that employing the relative abundance values of the 21 species within the species dataset yields the highest AUC of 0.85. This result is attained when utilizing the "JPN" population as the training set and the "JP" population as the test set.

This high-performance metric between two different datasets collected from the same country but different regions emphasize the success of the proposed approach.

The notable performance metric observed between two distinct datasets originating from the same population, but disparate regions underscore the efficacy of the proposed methodology.

The second highest AUC score of 0.83 is achieved when employing "DEU" as the training set and "JP" as the test set, while a similar AUC value of 0.83 is observed with "AUT" as the training set and "JP" as the test set.

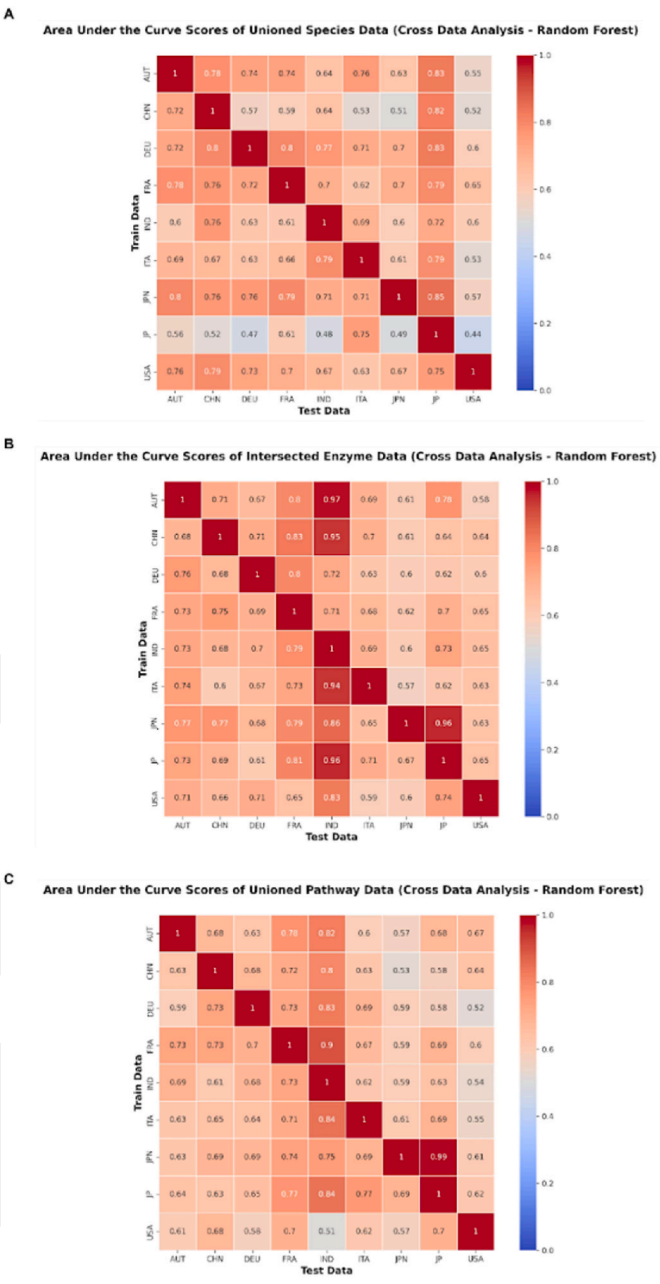

Fig. 5. Cross-population analysis using union features for species, enzyme, and pathway data.

As depicted in Fig. 5 (B), an AUC of 0.80 and above was obtained for 11/72 cases using the intersection of features for the enzyme dataset. Fig. 5 (B) shows that the best result is obtained with an AUC of 0.97 when using the intersection features selected by the feature selection methods for the enzyme data (25 enzymes/feature), using "AUT" as the training set and "IND" as the test set. The second-best result is obtained with 0.96 AUC when "JPN" is used as the training set and "JP" as the test set.

In Fig. 5 (C), AUC of 0.80 and above is obtained for 7/72 cases by using the union of features determined by feature selection methods in cross-population analyses for pathway data. Fig. 5 (C) shows that when "JPN" is used as the training set and "JP" as the test set, the best result of 0.99 AUC is obtained for the pathway data using the union features (38 features) selected by the feature selection algorithms. This high result between two data from the same population but from different regions

underlines the success of the proposed approach. The second-best result is obtained with an AUC of 0.90 using "FRA" as the training set and "IND" as the test data for the pathway data.

3.3. Potential biomarkers within the union/intersection features, as identified by different feature selection algorithms

In the present study, as shown in Fig. 3 (A), in the species dataset, union features selected by at least one feature selection algorithm are emphasized as potential biomarkers. Since the performance of the union features is superior to other models that are developed using individual feature selection algorithms and also superior to other models that are generated using the intersection of the features that are selected by all feature selection algorithms; for the species dataset, further analyses are conducted with the union set comprising 21 features. Intersection features selected by feature selection algorithms are highlighted as potential biomarkers in the enzyme dataset (see Fig. 3 (B)). Since the performance of the intersection features is superior to other models built with individual feature selection algorithms, and superior to other models built with a union of features selected by all feature selection algorithms, for the enzyme dataset, further analyses are performed with the intersection set containing 25 features. A similar trend can be seen in Fig. 3 (C), where the union feature set generates higher AUC than other tested feature selection algorithms. Hence, for pathway datasets, further analyses is performed utilizing the union set comprising 38 features.

For the features within the union features list, the importance scores of the features obtained after the classification process using the Random Forest classification algorithm are scaled using the min-max scaling method. The importance scores for each population are scaled between 0 and 100, where 0 denotes the lowest value and 100 represents the most important value. Then the median value of these scores is calculated for each feature and the final ranking is made based on this value. The feature with the highest median value is presented as the most valuable feature.

Recently, CRC prediction studies using metagenomic data, especially at the taxonomic level, have considerably increased. In this context, some species might influence colorectal cancer development and progression. The proposed method revealed a number of species associated with colorectal cancer. The top 20 species obtained with the proposed approach are shown in Supplementary Fig. 1. The relevance of those identified species has been assessed against the literature. In the "Discussion" section, the relevance of the top 5 species among these 20 species are analyzed in detail.

There are not many studies in the literature on the relationship between enzymes and CRC, but some enzymes may influence risk factors for colorectal cancer or play a role in CRC. The number of these studies is not large enough and the relationship between enzymes and colorectal cancer is still under investigation. Therefore, our discussion for the identified enzyme biomarkers is based on a limited number of studies. Although the number of studies is limited, the enzyme biomarkers identified by the proposed approach highlight a number of enzymes associated with colorectal cancer, some of which have been validated by the experiments in the literature. The top 20 enzymes obtained with the proposed approach are shown in Supplementary Fig. 2. In the "Discussion" section, the top 5 enzymes among these 20 species are analyzed in detail.

There are several studies suggesting that pathways may influence risk factors associated with colorectal cancer or may play a role in the development of cancer. The number of these studies is insufficient and the investigation of the relationship between pathways and colorectal cancer is still ongoing. In the present study, the union features selected by the feature selection algorithms determine the usage importance determined by the population-specific classification algorithms, take the median of these rankings, and produce a final ranking. The top 20 pathways identified using the proposed approach are shown in Supplementary Fig. 3. In the "Discussion" section, the top 5 pathways among these 20 pathways are analyzed in detail.

4. Discussion

In disease prediction using metagenomic data, feature selection algorithms play a crucial role by helping to identify biomarkers at different molecular levels, which not only contributes to a more comprehensive understanding of disease mechanisms at the molecular level, but also has implications for diagnosis and treatment [42]. Our proposed method, using robust feature selection algorithms such as Select K Best (SKB), Information Gain (IG), and Extreme Gradient Boost (XGBoost), and utilizing intersection and union of the features that are selected in multiple feature selection methods, has the potential to streamline microbiota analysis, reduce costs, and improve the effectiveness of CRC diagnosis. The impact of these features on colorectal cancer classification is thoroughly evaluated using state-of-the-art machine learning algorithms, namely, Decision Tree (DT), Random Forest (RF), AdaBoost, and LogitBoost for various metagenomic datasets. Systematic evaluation of the developed models includes a set of performance metrics that ensures a comprehensive assessment of their effectiveness.

4.1. Biological interpretation of the findings

In recent years, an increasing number of studies have used metagenomics data to identify biomarkers for CRC. Although there is increasing evidence that the gut microbiota is associated with CRC, the collective role of the gut microbiota is still under investigation. In the present study, feature selection algorithms are applied to metagenomic data on species, enzymes, and pathways levels, investigating the effects of the intersection and union features selected by these algorithms on the predictive performance of colorectal cancer. By developing global and population-specific models, the metagenomic data (species, enzymes, and pathways) associated with CRC is comprehensively investigated. Since the union features, which are identified by at least one of the feature selection algorithms exhibit high performance for the species and pathway datasets, these selected features are investigated from a biological perspective. For the enzyme dataset, the impact of the intersection features is investigated from a biological perspective, given their high performance across all feature selection algorithms. The potentially colorectal cancer-associated species, enzymes, and pathways identified by the proposed method are compared with the previous studies in the related literature. The promising candidates are indicated as possible biomarkers for CRC.

The scores of the top 20 species among different populations and their final ranks are visualized in Supplementary Fig. 1. The top 5 important species in this final ranking list are searched in literature for their possible roles in CRC development. Among the top 5 species, all species identified by the proposed method are previously reported by the following studies as associated with colorectal cancer: Ruthenibacterium Lactatiformans [43], Parvimonas Micra [44,45], Odoribacter Splanchnicus [46], Streptococcus Salivarius [47], and Gemella Morbillorum [48,49]. Among other taxonomic biomarkers identified in the top 20 species list of the present study, several species (e.g., Alistipes Finegoldii [50], Peptostreptococcus stomatis [44], Lactobacillus fermentum (J.-E [51])) have already been associated with colorectal cancer in the literature. This suggests that the proposed method is an effective method for species level metagenomics data and can be used in future studies to investigate the potential taxonomic biomarkers that are associated with other types of cancer.

The scores of the top 20 enzymes among different populations and their final ranks are visualized in Supplementary Fig. 2. According to this ranking, the top 5 enzymes are RNA 2′,3′-cyclic 3′-phosphodiesterase, methylaspartate mutase [8,52], glycine dehydrogenase, aminomethyltransferase [53], and peptide-methionine (R)-S- oxide reductase. Among these top 5 enzymes, two enzymes (methylaspartate mutase and

aminomethyltransferase) have been previously reported as directly linked with colorectal cancer. Studies in literature suggest that other enzymes are indirectly associated with colorectal cancer. For example, in the following studies, dihydrolipoyl dehydrogenase (P. J [59]), methylmalonyl CoA mutase [54] and tryptophanase [55] have not been directly associated with CRC, but some analyses have been conducted on other structures with which these enzymes interact, and an association with colorectal cancer has been reported. The enzymes found by the proposed method other than those mentioned above may be inspiring for future studies to reveal the undiscovered relationships between colorectal cancer and candidate enzyme biomarkers.

The scores of the top 20 pathways among different populations and their final ranks are visualized in Supplementary Fig. 3. According to this ranking, the top 5 pathway features are PWY-6588: pyruvate fermentation to acetone, HISDEG-PWY: L-histidine degradation I [56], PWY0-162: superpathway of de novo biosynthesis of pyrimidine ribonucleotides, P108-PWY: pyruvate fermentation to propanoate I, and HISTSYN-PWY: L-histidine biosynthesis. Among these top 5 pathways, only one pathway (HISDEG-PWY: L-histidine degradation I) is reported in literature as directly associated with colorectal cancer [56]. In addition, among the top 20 pathways, three pathways, e.g. L-lysine fermentation to acetate and butanoate [57], thiamine salvage II [56] and the 1,3-propanediol-glycerol degradation superpathway [58], are known as directly associated with colorectal cancer. The pathways found by the proposed method, other than those mentioned above, may be useful for future studies to reveal the relationship between colorectal cancer and the candidate pathway biomarkers.

The proposed method is used to predict colorectal cancer using machine learning methods and it identifies potential biomarkers associated with this disease. Research was conducted on nine different datasets including data from eight different populations. Using CRC-associated metagenomic data on species, enzyme, and pathway levels; global and population-specific analyses were performed. Potential biomarkers identified by the proposed method are validated with different studies in the existing literature. These identified species, enzymes and pathways were suggested as potential biomarkers since the performance of the models developed using these features were superior to other models. As illustrated in Supplementary Figs. 1, 2, 3, the scores of the top 20 species, enzymes and pathways are calculated separately for each population, and median scores are computed for species, enzyme, and pathway datasets. The top 5 species, enzymes and pathways in the final ranking were studied in detail in comparison to biomedical literature. For most of the species identified by our proposed approach, their possible roles in CRC development have been validated by the existing studies in literature, which highlights the effectiveness of our methodology. In the related literature, the number of studies on the association between enzymes and colorectal cancer is limited. One of the first 5 enzyme biomarkers have been previously reported in literature as associated with colorectal cancer. There are also relatively few studies on the association between pathway data and colorectal cancer. Research to identify pathways associated with colorectal cancer is ongoing. The fact that at least one biomarker from the top 5 biomarkers has been validated for each approach highlights the robustness of the proposed method, despite the limited number of studies available. This study not only provides valuable insights into colorectal cancer-associated species, enzymes, and pathways, but also serves as a guidepost for future research efforts aimed at uncovering previously unknown associations in this field.

5. Conclusion

Using various machine learning algorithms, the present study identifies the species, enzyme groups and pathways that influence the microbiota of the colorectal cancer patients; and hence contributes to the diagnosis and treatment of colorectal cancer. In addition, extensive experiments are performed to evaluate the effect of the identified species, enzyme, and pathway biomarkers on predicting the CRC at the population level. The population-based analyses of the present study include within-population, between population analysis (where one population is used as test data and all the remaining populations are used as training data (leave-one-dataset-out, LODO)) and cross-population (where one population is used as training data and another population is used as test data). This comprehensive set of experiments reveals molecular groups (species, enzymes, and pathways) that are effective in CRC diagnosis, customized for populations. For CRC-related species, enzyme and pathway datasets, models using i) all features, ii) the intersection of features identified by all feature selection algorithms and iii) the union of features identified by at least one feature selection algorithm are comparatively evaluated. The models that are generated by using union features show higher success rates compared with the success rates of the models that utilize all features and intersection features. When analyzing the performance metrics obtained with various feature selection methods and classifiers, one can notice that different feature selection methods and classifiers perform well on different datasets (i.e., species, enzyme and pathway data). Still, for all three datasets, the random forest classification algorithm shows superior performance compared to other classifiers. When analyzing the performance measures of different feature selection methods in Fig. 3, one can realize that XGBoost shows superior performance on the enzyme metagenomic data; and the SKB feature selection methods show superior performance on the species and pathway metagenomic data. On the other hand, the use of the XGBoost feature selection algorithm for enzyme data and the SKB algorithm for species and pathway data improves the performance in CRC prediction from metagenomic data. The species, enzymes, and pathways that we have identified as potential biomarkers are confirmed by studies in the related literature, enlightening their association with CRC. It is believed that the method developed by the present study will make an important contribution to the future studies of colorectal cancer.

CRediT authorship contribution statement

Burcu Bakir-Gungor: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis. **Mustafa Temiz:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Yasin Inal:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Emre Cicekyurt:** Visualization, Software, Methodology, Investigation. **Malik Yousef:** Writing – review & editing, Software, Project administration, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109098.

References

[1] A. Dokht Khosravi, S. Seyed-Mohammadi, A. Teimoori, A. Asarehzadegan Dezfuli, The role of microbiota in colorectal cancer, Folia Microbiol. 67 (5) (2022) 683–691, https://doi.org/10.1007/s12223-022-00978-1.

[2] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R.E.M. Pirozzi, F. Corcione, Worldwide burden of colorectal cancer: a review, Updates in Surgery 68 (1) (2016) 7–11, https://doi.org/10.1007/s13304-016-0359-y.

[3] F. Shi, G. Liu, Y. Lin, C. liutao Guo, J. Han, E.S.H. Chu, C. Shi, Y. Li, H. Zhang, C. Hu, R. Liu, S. He, G. Guo, Y. Chen, X. Zhang, O.O. Coker, S.H. Wong, J. Yu, J. She, Altered gut microbiome composition by appendectomy contributes to

colorectal cancer, Oncogene 42 (7) (2023), https://doi.org/10.1038/s41388-022-02569-3. Article 7.
[4] L. Zhou, Z. Jiang, Z. Zhang, J. Xing, D. Wang, D. Tang, Progress of gut microbiome and its metabolomics in early screening of colorectal cancer, Clin. Transl. Oncol. (2023), https://doi.org/10.1007/s12094-023-03097-6.
[5] Y. Wu, N. Jiao, R. Zhu, Y. Zhang, D. Wu, A.-J. Wang, S. Fang, L. Tao, Y. Li, S. Cheng, X. He, P. Lan, C. Tian, N.-N. Liu, L. Zhu, Identification of microbial markers across populations in early detection of colorectal cancer, Nat. Commun. 12 (1) (2021), https://doi.org/10.1038/s41467-021-23265-y. Article 1.
[6] Y. Zhang, A. Bhosle, S. Bae, L.J. McIver, G. Pishchany, E.K. Accorsi, K. N. Thompson, C. Arze, Y. Wang, A. Subramanian, S.M. Kearney, A. Pawluk, D. R. Plichta, A. Rahnavard, A. Shafquat, R.J. Xavier, H. Vlamakis, W.S. Garrett, A. Krueger, E.A. Franzosa, Discovery of bioactive microbial gene products in inflammatory bowel disease, Nature 606 (7915) (2022), https://doi.org/10.1038/s41586-022-04648-7. Article 7915.
[7] Q. Wang, J. Ye, D. Fang, L. Lv, W. Wu, D. Shi, Y. Li, L. Yang, X. Bian, J. Wu, X. Jiang, K. Wang, Q. Wang, M.P. Hodson, L.M. Thibaut, J.W.K. Ho, E. Giannoulatou, L. Li, Multi-omic profiling reveals associations between the gut mucosal microbiome, the metabolome, and host DNA methylation associated gene expression in patients with colorectal cancer, BMC Microbiol. 20 (1) (2020) 83, https://doi.org/10.1186/s12866-020-01762-2.
[8] C.S. Casimiro-Soriguer, C. Loucera, M. Peña-Chilet, J. Dopazo, Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer, Sci. Rep. 12 (1) (2022) 450, https://doi.org/10.1038/s41598-021-04182-y.
[9] P. Li, H. Luo, B. Ji, J. Nielsen, Machine learning for data integration in human gut microbiome, Microb. Cell Factories 21 (1) (2022) 241, https://doi.org/10.1186/s12934-022-01973-4.
[10] S. Yachida, S. Mizutani, H. Shiroma, S. Shiba, T. Nakajima, T. Sakamoto, H. Watanabe, K. Masuda, Y. Nishimoto, M. Kubo, F. Hosoda, H. Rokutan, M. Matsumoto, H. Takamaru, M. Yamada, T. Matsuda, M. Iwasaki, T. Yamaji, T. Yachida, T. Yamada, Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer, Nat. Med. 25 (6) (2019) 968–976, https://doi.org/10.1038/s41591-019-0458-7.
[11] H. Zhang, Y. Chang, Q. Zheng, R. Zhang, C. Hu, W. Jia, Altered intestinal microbiota associated with colorectal cancer, Front. Med. 13 (4) (2019) 461–470, https://doi.org/10.1007/s11684-019-0695-7.
[12] Q. Yao, M. Tang, L. Zeng, Z. Chu, H. Sheng, Y. Zhang, Y. Zhou, H. Zhang, H. Jiang, M. Ye, Potential of fecal microbiota for detection and postoperative surveillance of colorectal cancer, BMC Microbiol. 21 (1) (2021) 156, https://doi.org/10.1186/s12866-021-02182-6.
[13] F. Chen, X. Dai, C.-C. Zhou, K. Li, Y. Zhang, X.-Y. Lou, Y.-M. Zhu, Y.-L. Sun, B.-X. Peng, W. Cui, Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome-associated metabolites in the detection of colorectal cancer and adenoma, Gut 71 (7) (2022) 1315–1325, https://doi.org/10.1136/gutjnl-2020-323476.
[14] H. Shuwen, W. Yinhang, Z. Xingming, Z. Jing, L. Jinxin, W. Wei, D. Kefeng, Using whole-genome sequencing (WGS) to plot colorectal cancer-related gut microbiota in a population with varied geography, Gut Pathog. 14 (1) (2022) 50, https://doi.org/10.1186/s13099-022-00524-x.
[15] J. Yang, A. McDowell, E.K. Kim, H. Seo, W.H. Lee, C.-M. Moon, S.-M. Kym, D. H. Lee, Y.S. Park, Y.-K. Jee, Y.-K. Kim, Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis, Exp. Mol. Med. 51 (10) (2019), https://doi.org/10.1038/s12276-019-0313-4. Article 10.
[16] Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, L. Su, X. Li, X. Li, J. Li, L. Xiao, U. Huber-Schönauer, D. Niederseer, X. Xu, J. Y. Al-Aama, J. Wang, Gut microbiome development along the colorectal adenoma–carcinoma sequence, Nat. Commun. 6 (1) (2015), https://doi.org/10.1038/ncomms7528. Article 1.
[17] X. Fan, Y. Jin, G. Chen, X. Ma, L. Zhang, Gut microbiota dysbiosis drives the development of colorectal cancer, Digestion 102 (4) (2021) 508–515, https://doi.org/10.1159/000508328.
[18] A. Artemev, S. Naik, A. Pougno, P. Honnavar, N.M. Shanbhag, A. Artemev, S. Naik, A. Pougno, P. Honnavar, N.M. Shanbhag, The association of microbiome dysbiosis with colorectal cancer, Cureus 14 (2) (2022), https://doi.org/10.7759/cureus.22156.
[19] L.J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. Loncar Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, T. Klammsteiner, M. Kolev, L. Lahti, M.B. Lopes, V. Moreno, I. Naskinova, E. Org, I. Paciência, G. Papoutsoglou, J. Truu, Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment, Front. Microbiol. 12 (2021). https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511.
[20] M. Bose, H.M. Wood, C. Young, P. Van Nang, M. Van Doi, C. Vaccaro, T.A. Piñero, J. Arguero, L.C. Melendez, C.T. Valladares, P. Quirke, R.A. Seshadri, International C. R. C. Microbiome Network (AMS/CRUK), Analysis of an Indian colorectal cancer faecal microbiome collection demonstrates universal colorectal cancer-associated patterns, but closest correlation with other Indian cohorts, BMC Microbiol. 23 (1) (2023) 52, https://doi.org/10.1186/s12866-023-02805-0.
[21] J. Zhen, C. Liu, F. Liao, J. Zhang, H. Xie, C. Tan, W. Dong, The global research of microbiota in colorectal cancer screening: a bibliometric and visualization analysis, Front. Oncol. 13 (2023). https://www.frontiersin.org/articles/10.3389/fonc.2023.1169369.
[22] C. Yu, Z. Zhou, B. Liu, D. Yao, Y. Huang, P. Wang, Y. Li, Investigation of trends in gut microbiome associated with colorectal cancer using machine learning, Front. Oncol. 13 (2023). https://www.frontiersin.org/articles/10.3389/fonc.2023.1077922.
[23] F. Beghini, L.J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A.M. Thomas, M. Valles-Colomer, G. Weingart, Y. Zhang, M. Zolfo, C. Huttenhower, E.A. Franzosa, N. Segata, Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, Elife 10 (2021) e65088, https://doi.org/10.7554/eLife.65088.
[24] G. Ditzler, R. Polikar, G. Rosen, Multi-layer and recursive neural networks for metagenomic classification, IEEE Trans. NanoBioscience 14 (6) (2015) 608–616, https://doi.org/10.1109/TNB.2015.2461219. IEEE Transactions on NanoBioscience.
[25] B. Bakir-Gungor, M. Temiz, A. Jabeer, D. Wu, M. Yousef, microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach, Front. Microbiol. 14 (2023) 1264941, https://doi.org/10.3389/fmicb.2023.1264941.
[26] M. Yousef, J. Allmer, Y. İnal, B.B. Gungor, G-S-M: a comprehensive framework for integrative feature selection in omics data analysis and beyond. https://doi.org/10.1101/2024.03.30.585514, 2024.
[27] M. Yousef, L. Abdallah, J. Allmer, maTE: discovering expressed interactions between microRNAs and their targets, Bioinformatics 35 (20) (2019) 4020–4028, https://doi.org/10.1093/bioinformatics/btz204.
[28] M. Yousef, G. Goy, B. Bakir-Gungor, miRModuleNet: detecting miRNA-mRNA regulatory modules, Front. Genet. 13 (2022) 767455, https://doi.org/10.3389/fgene.2022.767455.
[29] E. Qumsiyeh, L. Showe, M. Yousef, GediNET for discovering gene associations across diseases using knowledge based machine learning approach, Sci. Rep. 12 (1) (2022) 19955, https://doi.org/10.1038/s41598-022-24421-0.
[30] M. Yousef, G. Goy, R. Mitra, C.M. Eischen, A. Jabeer, B. Bakir-Gungor, miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking, PeerJ 9 (2021) e11458, https://doi.org/10.7717/peerj.11458.
[31] M. Unlu Yazici, J.S. Marron, B. Bakir-Gungor, F. Zou, M. Yousef, Invention of 3Mint for feature grouping and scoring in multi-omics, Front. Genet. 14 (2023) 1093326, https://doi.org/10.3389/fgene.2023.1093326.
[32] N.S. Ersoz, B. Bakir-Gungor, M. Yousef, GeNetOntology: identifying affected gene ontology groups via grouping, scoring and modelling from gene expression data utilizing biological knowledge based machine learning, Front. Genet. 14 (2023) 1139082.
[33] M. Yousef, D. Voskergian, TextNetTopics: text classification based word grouping as topics and topics' scoring, Front. Genet. 13 (2022) 893378, https://doi.org/10.3389/fgene.2022.893378.
[34] D. Voskergian, B. Bakir-Gungor, M. Yousef, TextNetTopics Pro, a topic model-based text classification for short text by integration of semantic and document-topic distribution information, Front. Genet. 14 (2023) 1243874, https://doi.org/10.3389/fgene.2023.1243874.
[35] E. Qumsiyeh, Z. Salah, M. Yousef, miRGediNET: a comprehensive examination of common genes in miRNA-Target interactions and disease associations: insights from a grouping-scoring-modeling approach, Heliyon 9 (12) (2023) e22666, https://doi.org/10.1016/j.heliyon.2023.e22666.
[36] A. Jabeer, M. Temiz, B. Bakir-Gungor, M. Yousef, miRdisNET: discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning, Front. Genet. 13 (2023) 1076554, https://doi.org/10.3389/fgene.2022.1076554.
[37] Ü.G. Söylemez, M. Yousef, B. Bakir-Gungor, AMP-GSM: prediction of antimicrobial peptides via a grouping–scoring–modeling approach, Appl. Sci. 13 (8) (2023) 5106, https://doi.org/10.3390/app13085106.
[38] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, M. Yousef, Review of feature selection approaches based on grouping of features, PeerJ 11 (2023) e15666, https://doi.org/10.7717/peerj.15666.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: machine learning in Python, Mach. Learn. Phython 12 (2011) 2825–2830.
[40] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) 785–794, https://doi.org/10.1145/2939672.2939785.
[41] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, Knime - the Konstanz information miner: version 2.0 and beyond, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 26–31, https://doi.org/10.1145/1656274.1656280.
[42] N. LaPierre, C.J.-T. Ju, G. Zhou, W. Wang, MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction, Methods 166 (2019) 74–82, https://doi.org/10.1016/j.ymeth.2019.03.003.
[43] N. Trivieri, R. Pracella, M.G. Cariglia, C. Panebianco, P. Parrella, A. Visioli, F. Giani, A.A. Soriano, C. Barile, G. Canistro, T.P. Latiano, L. Dimitri, F. Bazzocchi, D. Cassano, A.L. Vescovi, V. Pazienza, E. Binda, BRAFV600E mutation impinges on gut microbial markers defining novel biomarkers for serrated colorectal cancer effective therapies, J. Exp. Clin. Cancer Res. 39 (1) (2020) 285, https://doi.org/10.1186/s13046-020-01801-w.
[44] M.A. Osman, H. Neoh, N.-S. Ab Mutalib, S.-F. Chin, L. Mazlan, R.A. Raja Ali, A. D. Zakaria, C.S. Ngiu, M.Y. Ang, R. Jamal, Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer, Sci. Rep. 11 (1) (2021), https://doi.org/10.1038/s41598-021-82465-0. Article 1.
[45] L. Zhao, X. Zhang, Y. Zhou, K. Fu, H.C.-H. Lau, T.W.-Y. Chun, A.H.-K. Cheung, O. O. Coker, H. Wei, W.K.-K. Wu, S.H. Wong, J.J.-Y. Sung, K.F. To, J. Yu, Parvimonas

micra promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients, Oncogene 41 (36) (2022), https://doi.org/10.1038/s41388-022-02395-7. Article 36.
[46] C.-W. Png, Y.-K. Chua, J.-H. Law, Y. Zhang, K.-K. Tan, Alterations in co-abundant bacteriome in colorectal cancer and its persistence after surgery: a pilot study, Sci. Rep. 12 (1) (2022), https://doi.org/10.1038/s41598-022-14203-z. Article 1.
[47] C.C. Wong, J. Yu, Gut microbiota in colorectal cancer development and therapy, Nat. Rev. Clin. Oncol. (2023) 1–24, https://doi.org/10.1038/s41571-023-00766-x.
[48] K.B. Laupland, F. Edwards, L. Furuya-Kanamori, D.L. Paterson, P.N.A. Harris, Bloodstream infection and colorectal cancer risk in Queensland Australia, 2000-2019, Am. J. Med. (2023), https://doi.org/10.1016/j.amjmed.2023.05.003.
[49] Y. Shimomura, L. Zha, S. Komukai, N. Narii, T. Sobue, T. Kitamura, S. Shiba, S. Mizutani, T. Yamada, N. Sawada, S. Yachida, Mediation effect of intestinal microbiota on the relationship between fiber intake and colorectal cancer, Int. J. Cancer 152 (9) (2023) 1752–1762, https://doi.org/10.1002/ijc.34398.
[50] B.J. Parker, P.A. Wearsch, A.C.M. Veloo, A. Rodriguez-Palacios, The genus Alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health, Front. Immunol. 11 (2020), https://doi.org/10.3389/fimmu.2020.00906.
[51] J.-E. Lee, J. Lee, J.H. Kim, N. Cho, S.H. Lee, S.B. Park, B. Koh, D. Kang, S. Kim, H. M. Yoo, Characterization of the anti-cancer activity of the probiotic bacterium Lactobacillus fermentum using 2D vs. 3D culture in colorectal cancer cells, Biomolecules 9 (10) (2019), https://doi.org/10.3390/biom9100557. Article 10.
[52] W. Gou, C. Ling, Y. He, Z. Jiang, Y. Fu, F. Xu, Z. Miao, T. Sun, J. Lin, H. Zhu, H. Zhou, Y. Chen, J.-S. Zheng, Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes, Diabetes Care 44 (2) (2020) 358–366, https://doi.org/10.2337/dc20-1536.
[53] E. Odin, A. Sondén, G. Carlsson, B. Gustavsson, Y. Wettergren, Folate pathway genes linked to mitochondrial biogenesis and respiration are associated with outcome of patients with stage III colorectal cancer, Tumor Biol. 41 (6) (2019) 1010428319846231, https://doi.org/10.1177/1010428319846231.
[54] V. Lacombe, G. Lenaers, G. Urbanski, Diagnostic and therapeutic perspectives associated to cobalamin-dependent metabolism and transcobalamins' synthesis in solid cancers, Nutrients 14 (10) (2022), https://doi.org/10.3390/nu14102058. Article 10.
[55] M. Wyatt, K.L. Greathouse, Targeting dietary and microbial tryptophan-indole metabolism as therapeutic approaches to colon cancer, Nutrients 13 (4) (2021), https://doi.org/10.3390/nu13041189. Article 4.
[56] J.-W. Huh, M.J. Kim, J. Kim, H.G. Lee, S.-B. Ryoo, J.-L. Ku, S.-Y. Jeong, K.J. Park, D. Kim, J.F. Kim, J.W. Park, Enterotypical Prevotella and three novel bacterial biomarkers in preoperative stool predict the clinical outcome of colorectal cancer, Microbiome 10 (1) (2022) 203, https://doi.org/10.1186/s40168-022-01388-8.
[57] E. Russo, L.D. Gloria, G. Nannini, G. Meoni, E. Niccolai, M.N. Ringressi, S. Baldi, R. Fani, L. Tenori, A. Taddei, M. Ramazzotti, A. Amedei, From adenoma to CRC stages: the oral-gut microbiome axis as a source of potential microbial and metabolic biomarkers of malignancy, Neoplasia 40 (2023) 100901, https://doi.org/10.1016/j.neo.2023.100901.
[58] F. Bellerba, D. Serrano, H. Johansson, C. Pozzi, N. Segata, A. NabiNejad, E. Piperni, P. Gnagnarella, D. Macis, V. Aristarco, C.A. Accornero, P. Manghi, A. Guerrieri-Gonzaga, R. Biffi, L. Bottiglieri, C. Trovato, M.G. Zampino, F. Corso, R. Bellocco, S. Gandini, Colorectal cancer, Vitamin D and microbiota: a double-blind Phase II randomized trial (ColoViD) in colorectal cancer patients, Neoplasia 34 (2022) 100842, https://doi.org/10.1016/j.neo.2022.100842.
[59] P.J. Lee, S.J. Woo, H.M. Yoo, N. Cho, H.P. Kim, Differential mechanism of ATP production occurs in response to succinylacetone in colon cancer cells, Molecules 24 (19) (2019), https://doi.org/10.3390/molecules24193575. Article 19.
[60] M. Yousef, A. Kumar, B. Bakir-Gungor, Application of biological domain knowledge based feature selection on gene expression data, Entropy 23 (1) (2021), https://doi.org/10.3390/e23010002.
[61] M. Yousef, F. Ozdemir, A. Jaaber, J. Allmer, B. Bakir-Gungor, *PriPath: identifying dysregulated Pathways from differential gene Expression via grouping, Scoring and Modeling with an embedded machine learning approach* [preprint], Review (2022), https://doi.org/10.21203/rs.3.rs-1449467/v1.
[62] M. Yousef, E. Ülgen, O. Uğur Sezerman, CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis, PeerJ Computer Science 7 (2021) e336, https://doi.org/10.7717/peerj-cs.336.