Identifying Taxonomic Biomarkers of Colorectal Cancer in Human Intestinal Microbiota Using Multiple Feature Selection Methods

Amhar Jabeer
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
amhar.jabeer@agu.edu.tr

Ayşegül KOÇAK
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
aysegul.kocak@agu.edu.tr

Hüseyin AKKAŞ
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
huseyin.akkas@agu.edu.tr

Ferhan Yenisert
Department of Bioengineering
Abdullah Gul University
Kayseri, Turkey
ferhan.yenisert@agu.edu.tr

Özkan Ufuk NALBANTOĞLU
Department of Computer Engineering
Erciyes University
Kayseri, Turkey
nalbantoglu@erciyes.edu.tr

Malik YOUSEF
Department of Information System
Zefat Academic College
Zefat, Israel
malik.yousef@gmail.com

Burcu BAKIR GÜNGÖR
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
burcu.gungor@agu.edu.tr

2022 Innovations in Intelligent Systems and Applications Conference (ASYU) | 978-1-6654-8894-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASYU56188.2022.9925551

***Abstract*— A variety of bacterial species called gut microbiota work together to maintain a steady intestinal environment. The gastrointestinal tract contains tremendous amount of different species including archaea, bacteria, fungi, and viruses. While these organisms are crucial immune system stabilizers, the dysbiosis of the intestinal flora has been related to gastrointestinal disorders including Colorectal cancer (CRC), intestinal cancer, irritable bowel syndrome and inflammatory bowel disease. In the last decade, next-generation sequencing (NGS) methods have accelerated the identification of human gut flora. CRC is a deathly condition that has been on the rise in the last century, affecting half a million people each year. Since early CRC diagnosis is critical for an effective treatment, there is an immediate requirement for a classification system that can expedite CRC diagnosis. In this study, via analyzing the available metagenomics data on CRC, we aim to facilitate the CRC diagnosis via finding biomarkers linked with CRC, and via building a classification model. We have obtained the metagenomic sequencing data of the healthy individuals and CRC patients from a metagenome-wide association analysis and we have classified this data according to the disease stages. Conditional Mutual Information Maximization (CMIM), Fast Correlation Based Filter (FCBF), Extreme Gradient Boosting (XGBoost), min redundancy max relevance (mRMR), Information Gain (IG) and Select K Best (SKB) feature selection algorithms were utilized to cope with the complexity of the features. We observed that the SKB, IG, and XGBoost techniques made significant contributions to decrease the microbiota in use for CRC diagnosis, thereby reducing cost and time. We realized that our Random Forest classifier outperformed Adaboost, Support Vector Machine, Decision Tree, Logitboost and stacking ensemble classifiers in terms of CRC classification performance. Our results reiterated some known and some potential microbiome associated mechanisms in CRC, which could aid the design of new diagnostics based on the microbiome.**

Keywords— Feature selection, Metagenomics, Human gut microbiome, Classification, Biomarker discovery.

I. INTRODUCTION

The vast number of bacteria that live in our digestive tract are a part of the intricate microbial ecosystem, that is known as the human gut microbiota. Latest studies showed that commensal bacteria are critical for human physiology and disease [1]. Disruption of the ongoing connection between intestinal epithelial cells and the intestinal microbiota, often known as dysbiosis, has been linked to several diseases [1]. The rapidly expanding research on disease-associated microbiota indicates that microbiota variety loss is a typical characteristic of dysbioses [3]. One of the diseases in which dysbiosis greatly affects the pathogenesis is colorectal cancer (CRC), which seriously threatens human health and is the third most common malignancy [2]. In this regard, for CRC patients, it is crucial to comprehend the constituents of the human intestinal microbiome, which is made up of the collective genomes of the microbes living there and their possible roles [3].

CRC, which causes for almost 500,000 deaths yearly and which is the leading cause of death globally, presents a significant healthcare problem [4]. Changes in the dysbiotic

978-1-6654-8894-5/22/$31.00 ©2022 IEEE

gut's bacterial abundances have been demonstrated to be crucial in modulating cell proliferation, DNA damage, and immunological responses in CRC patients. In this respect, the human intestine microbiome, which is the cumulative DNA of the bacterial community found in human digestive tract, has been proposed as a potential diagnostic tool for CRC. *Enterotoxigenic Bacteroides fragilis, Streptococcus bovis, Fusobacterium nucleatum, Peptostreptococcus anaerobius, Escherichia coli and Enterococcus faecalis* have been identified as CRC potential pathogens in animal studies and in observational studies [5]. Despite the fact that some studies have shown how crucial the gut microbiota is for the pathophysiology of CRC, this field is quite young. Although many classification standards and methodologies have been employed in earlier investigations, the major strains implicated in the development and progression of CRC remain unknown [2]. Recently, the creation of billions of reads in a single run is now possible thanks to the rapid progress of next-generation sequencing (NGS) technology. Along this line, the gastrointestinal microbiota was identified using metagenomic NGS techniques, which allow for the examination of a sample's whole genetic content and display biological and behavioral characteristics of the microbial communities. Hence, the metagenomic examination of the gastrointestinal system sheds light on the impact of the human gut microbiota on human physiology and diseases [6]. The microbial abundance profiles are frequently employed in disease prediction since the microbiome makeup of cases and controls differs from that of controls. Feature selection techniques may help to clarify disease mechanisms in the metagenome-based disease prediction challenge. Therefore, researches in this area are important. In order to reduce the diversity of species, that is, to select informative attributes, mRMR [7], Lasso [8], Elastic Net and Iterative sure select algorithms [9] were previously utilized. Bakir-Gungor et al. [10] obtained high performance metrics in the models that they developed by applying CMIM, FCBF, mRMR, SKB feature selection in the metagenomic dataset associated with type 2 diabetes. Similarly, Bakir-Gungor et al. [11] analyzed the metagenomic dataset associated with Inflammatory Bowel Disease using six different feature selection methods and identified 14 taxonomic biomarkers. They reported that the model they created with these features achieved an Area Under the Curve (AUC) score of 0.93. The developed method has been tested on an independent data set, and it has been reported that successful performance metrics have been obtained. Although various feature selection strategies perform well in various contexts, they are just recently starting to garner attention in this field. A recently published review paper [1] noted that some of the different feature selection methods have achieved good results in human microbiome researches. There is no agreement on the best feature selection techniques to use for metagenomic studies, as pointed out in [1]. It is possible to identify subgroups of bacteria that are extremely distinctive by using supervised learning on data from the human gut microbiome. As a result, disease diagnosis can be performed using disease prediction algorithms which can categorize unlabeled samples. Lately, Marcos-Zambrano et al. [2] [3] analyzed 89 research papers and summarized the popular applications of machine learning (ML) in microbiota research. They noted that Random Forest (RF), Logistic Regression (LR), k-NN (k nearest neighbor) and Support Vector Machines (SVM) are the most often utilized supervised learning algorithms for microbiome investigation. They concluded that when choosing the ML algorithm, more than one factor should be considered like the number of features, number of observations, etc. They advised using and comparing multiple approaches, then picking the one with the best performance value. Using a metagenomic data set connected to CRC, in this work we suggest to build a strong ML model that could improve the diagnostic precision of CRC. To cope with the large complexity of the number of features, we want to use sturdy feature selection approaches. By utilizing advanced feature selection techniques including CMIM, FCBF, mRMR, select K Best (SKB), Extreme Gradient Boosting (XGBoost), and Information Gain (IG) we primarily seek to i) identify candidate biomarkers of CRC, and ii) detect which subgroup of the microbiota is more informative for this disease. We plan to calculate a significance value for a species and discover distinctive taxonomic types for CRC. The originality of this research effort can be streamlined as follows: i) to improve CRC diagnosis via developing a model for classification, and ii) to identify CRC biomarkers. The following sections are structured as follows. We provide the dataset we utilized in this study, and outline our methodology in the section titled "Materials and Methods". We describe our findings in the "Results" section after applying feature selection techniques and classification methods to CRC-associated metagenomic data. As potential biomarkers of CRC, the species we identified through our analysis are evaluated in the "Discussions" section, where they are compared with the gold standard features which previously identified as having a link to CRC. In the "Conclusion" section, the manuscript is concluded with future prospects.

II. MATERIALS AND METHODS

We used 108 human samples' unprocessed microbiome DNA sequencing data within this study. Referring to the associated metadata, the raw sequencing data of the samples were categorized into disease states after being retrieved from Zeller's [15] repository and from the European Nucleotide Archive with the accession codes ERP005534 and PRJEB6070. According to the Human Microbiome Project Consortium's Standard Operating Procedures [16], quality filtering techniques were applied to the raw sequences.

After following the above-mentioned preprocessing steps, metagenome samples were classified according to the taxa that represented the microbial species that they belonged to. During this procedure, MetaPhlAn2 taxonomic classification tool was used to infer the relative abundance composition of each taxon within a sample. The features that will be used in the ML algorithms were produced by these species and their relative abundances. The data includes 108 samples and 1,455 microbial species, as shown in Fig. 1. Forty eight of the samples are CRC patients, and 60 of them are healthy individuals. Fig. 1 displays sample lines from the CRC metagenomics dataset after the raw data has been preprocessed. For each sample, this dataset displays the relative abundance values for each taxonomic species. The characteristics relate with different species of bacteria,

978-1-6654-8894-5/22/$31.00 ©2022 IEEE

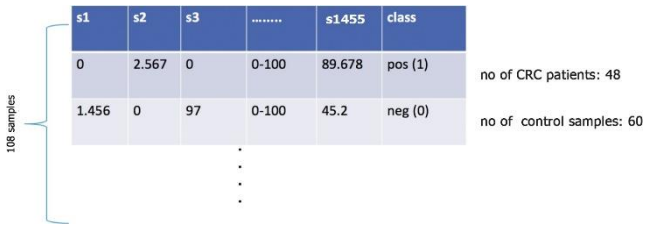

Figure 1. Colorectal cancer-associated metagenomics dataset representation.

viruses, and archea. Healthy (shown with 0) and CRC patient (shown with 1) are the two class labels of the samples.
The following steps summarize the methodology used in this study: (i) feature selection to discover the most effective CRC-related biomarkers; (ii) creating a model to categorize CRC patients and control samples, then evaluating the model using a range of evaluation measures. Fig. 2 presents the workflow of our methodology.

A. Feature Selection

The effectiveness and efficiency of feature selection as a data preparation technique, particularly for high-dimensional data, has been demonstrated for numerous data mining and machine learning challenges [17]. Regarding their ability to cope with problems like redundancy and correlation, nonlinearity in data, having a significantly higher number of features than samples, noise in the input features and the target class, various feature selection techniques were examined in studies [18] and [19]. Some recent feature selection approaches incorporate the biological information into the machine learning process *[12] [13]*. The data that we used has high dimension (1,455 microbial species), which could affect the effectiveness of the classification algorithm. As a result, a feature selection procedure is required to minimize the model's dimensionality and to make it easier to classify and interpret. Min Redundancy Max Relevance (mRMR) *[14]*, Lasso *[15]*, Elastic Net *[16]*, and iterative sure select method *[17]* have been used in literature to choose the most important features or to decrease the number of species.

In this study, in order to identify nominee taxonomic biological markers by boosting the accuracy of classification, and by lowering the number of attributes, we use Peng's mRMR [7], Fleuret's CMIM [18], FCBF [19], XGBoost, IG and SKB [20] feature selection algorithms. The goal of the mRMR [9] algorithm is to detect the features with the maximum correlation to a class for prediction (max relevance), and with the least link between them (min redundancy). This approach begins with an empty set, weighs the features using mutual information, and then combines sequential search with forward selection to find the ideal subset of traits. It is a polyvariate feature selection technique used to identify class relevance as well as the dependency between each feature pair. CMIM [25] calculates mutual information and conditional entropy with the class, and the features are evaluated for their importance. It chooses the feature if it contains extra information. FCBF [26] sorts the features according to their shared information with the class. Subsequently, the features that have mutual information below a preset threshold are eliminated. It employs the concept of "predominant correlation". Following a classifier-

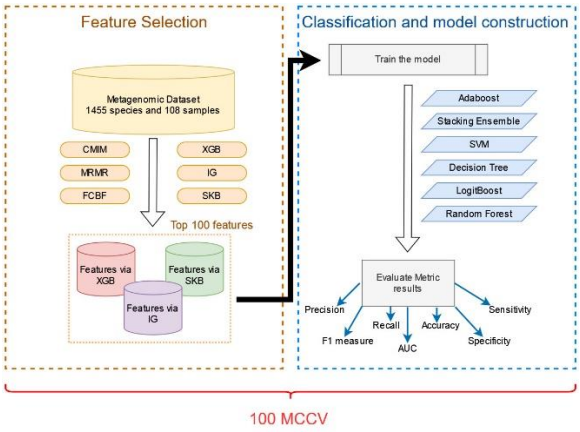

Figure 2. Workflow of methodology. (i) The most important species of CRC-associated metegenome are identified using feature selection techniques (highlighted in brown); (ii) The chosen features are utilized to construct models, which are subsequently applied for classification (shown in blue) using 100 fold MCCV (Monte Carlo Cross Validation).

independent approach, features are picked that have a high correlation to the target variable but with low connection with others. To put it another way, FCBF seeks to reduce redundancy among chosen features. FCBF is open to interpretation and a sturdy approach that produces typically satisfactory outcomes. In addition to enhancing classification performance, the adoption of filter-based feature choices for big data analysis in the biomedical domain can hasten the discovery of biomarkers and produce fascinating biological interpretations. In select K best, the features with the k highest scores are the features that are selected after implementing a method to score the features against the class label. In XGBoost feature selection, the more a feature is employed in decision trees to make significant key decisions, the more important it becomes. In this study, Python 3 skfeature and sklearn modules are used to implement the above-mentioned feature selection techniques.

B. Classification Model Construction

In our preliminary analysis, we used RF, Decision Tree, AdaBoost, LogitBoost, ensemble of LogitBoost with kNN, and SVM with kNN (k nearest neighbor) to examine the effects of various classification algorithms. We proceeded with RF in our subsequent studies since the tree model is simple for interpretation and it can be quickly converted into a rule set. Additionally, one of the most often used algorithms in the study of the gut microbiota is RF [29]. We employed 100-fold MCCV (Monte Carlo cross-validation), which entails selecting part of the data at random (without exchanging them) to build the training set and after that allocating the rest of the data to the test set [30]. Hence, different training and test partitions are created at random over many iterations of this process. We decided to use 90% for training and 10% for testing. Our technique is implemented using the Konstanz Information Miner (KNIME) platform. In KNIME, we utilized the RF predictor node from the H20 package [21].

C. Model Performance Evaluation

The accuracy, AUC measurements, and F1 Scores were used to evaluate the prediction performances of the

978-1-6654-8894-5/22/$31.00 ©2022 IEEE

constructed models. Accuracy is a reliable and common performance metric for balanced datasets. Alternative metrics, including the F1 score and AUC, were utilized to assess the performance of the proposed models because the dataset used in this work had an unbalanced distribution of classes. Precision is a condition that denotes performance in an anticipated condition. However, recall displays the proportion of real examples that the classifier successfully identified. The F1 metric denotes the harmonic mean of recall and precision. When there is an unequal class distribution and when someone wants to strike a compromise between precision and recall, the F1 score is a viable option. AUC measure approximates the probability that a randomly chosen positive example will get a better score than a randomly chosen negative example. We report all the performance measures using the average of the 100-fold MCCV results.

III. RESULTS

A. Feature selection and classification

The ongoing relation between the mucosal immune system and the gut microbiota is a distinctive feature of CRC pathogenesis. Therefore, the microbiome contains species that can serve as a biomarker. A CRC-related metagenomics dataset containing relative abundance values for 1455 distinct species and 108 samples is analyzed to identify those organisms with biomarker capabilities. By using above-mentioned feature selection procedures, this study aims to discard useless and superfluous features. Features are initially examined using each feature selection method. LogitBoost, Decision Tree, SVM, AdaBoost, Random Forest, and ensemble classifiers such as Logitboost and kNN, SVM and kNN are utilized to assess the effects of various classification algorithms. The number of trees in Random Forest, c and gamma parameters for SVM, and n estimators for Adaboost and Logitboost are all tuned. The performance of several classifiers are evaluated using various metrics, utilizing the top 100 features selected using the above-mentioned 6 feature selection methods (as presented in Fig. 3). As seen in this figure, XGBoost, IG, and SKB methods improved the sensitivity, recall, F1 score, AUC, and accuracy values for multiple classifiers that are used to analyze the metagenomic data related to CRC. Using the same data, it can be seen that mRmR, FCBF, and CMIM feature selection algorithms all had low accuracy and high recall along with indicators of unsatisfactory fitting across tested models. Three chosen feature selection algorithms (IG, SKB, and XGBoost) were utilized to determine the scaled relevance values for each feature across various classifiers in order to choose the most pertinent and beneficial features. The scaled important value threshold of 0.5 is used to eliminate the top 100 features with the lowest rankings. As shown in Fig. 4, this technique selected 16, 19, and 28 features using the SKB, XGBoost, and IG algorithms, respectively. All of the three feature selection methods identified the same 8 features. Using the 8 features that were chosen, the RF classifier performed better than the individual feature selection techniques. The findings of this research revealed that the designed RF model with only 8 features generated satisfactory accuracy values that could be used for diagnostic purposes (as demonstrated in Fig. 3). As shown in Fig. 3, the final RF model had a 0.83 accuracy, 0.86 F1 score and 0.81

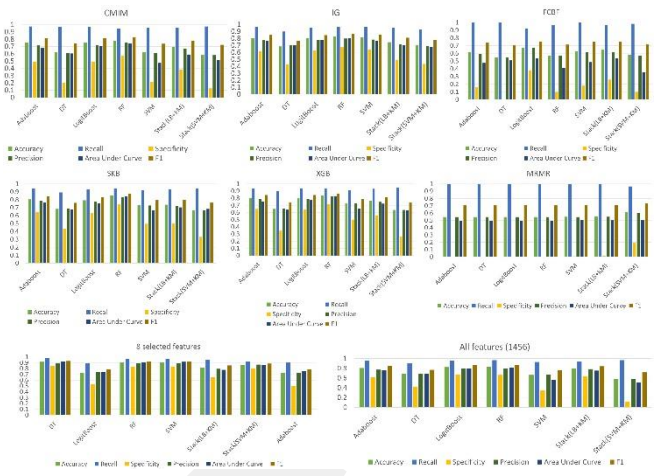

Figure 3. Performance assessment of various classifiers on CRC metagenomics dataset using 100-fold MCCV, and using above-mentioned feature selection techniques, 8 selected features, and all features.

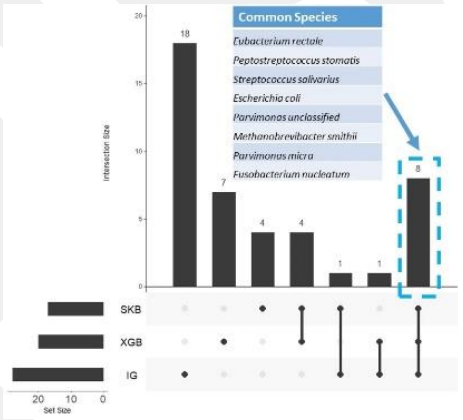

Figure 4. The number of species that were chosen using various feature selection algorithms, and the details of commonly selected species.

AUC when all 1455 features were included (feature selection not applied). The RF model reached 0.91 accuracy, 0.92 F1-score and 0.92 AUC values by utilizing only the 100 features that are often recognized in six promising feature selection approaches. As shown in Fig. 3, the model with the chosen 100 features performed 18% better in terms of specificity metrics, 11% better in terms of AUC value, 8% better in terms of accuracy, 9% better in terms of precision metrics, and 6% better in terms of F1-score metrics when crosschecked with the performance of the RF model including all features.

The feature selection techniques XGBoost, SKB, and IG were applied to eliminate gratuitous and inessential features (species) from CRC-associated metagenomics dataset which included the sequence reads from 1455 taxa for 108 samples. 42 of 1455 features are selected by at least one of the three different feature selection algorithms, 8 features are selected by all of the tested feature selection algorithms. As a result, we found *Eubacterium rectale, Peptostreptococcus stomatis, Streptococcus salivarius, Methanobrevibacter smithii, Escherichia coli, Parvimonas, Parvimonas micra,* and *Fusobacterium nucleatum* as main biomarkers of CRC.

978-1-6654-8894-5/22/$31.00 ©2022 IEEE

IV. DISCUSSION

The second-highest cancer-related worldwide death rates is associated with colorectal cancer (CRC). It is the third most prevalent malignancy [33]. The major causes of CRC are genetic factors as well as environmental factors, including lifestyle and food that disrupt the balance of gut bacteria [22] [23]. As recent research articles report, the gut microbiota's imbalance is related to many disorders such as inflammatory bowel disease and CRC [24]. By controlling metabolites and interacting with the host intestinal epithelial cells, the gut microbiota plays a significant role in CRC [37]. Researchers attempt to develop new ways for CRC diagnosis using gut microbiota as a biomarker. In our study, we identified *Eubacterium rectale, Peptostreptococcus stomatis, Streptococcus salivarius, Methanobrevibacter smithii, Escherichia coli, Parvimonas micra, and Fusobacterium nucleatum* as main CRC biomarkers. *Eubacterium rectal*e, one of the gut bacteria, has a role in reducing intestinal inflammation by generating butyrate, which upregulates the secretion of IL-10, an anti-inflammatory cytokine [25] [26]. *Peptostreptococcus stomatis and Parvimonas micra* are associated with the abundance of enterotoxigenic *Bacteroides fragilis (ETBF)*, an operator bacteria in CRC development. *P. stomatis, P. micra, and ETBF* were relevant to malignant laterally spreading tumors (LST) by increasing the production of IL-6, an inflammatory cytokine [27] [28]. Increases in IL-6 trigger intestinal inflammation and it is a critical molecule in the microenvironment of colorectal tumors [28]. The combination of *P. micra- P. stomatis–ETBF* is reported to generate high precision in identifying adenoma recurrence after LST resection [28]. The same study noted that *P. stomatis–P. micra–ETBF* combination demonstrated stronger diagnostic capability than the single biomarker *P. stomatis* [28]. The overrepresentation of one of the most well-known gut bacteria, *Escherichia coli*, triggers CRC by producing colibactin. The *Escherichia* family of intestinal pathogens produces genotoxins such as Cytolethal distending toxin, a carcinogen that causes double-strand DNA breaks via its deoxyribonuclease activity [29]. Known as one of the carcinogenic taxa, *Fusobacterium*, were discovered ranging from the early stages of carcinogenesis to the late stages [30] [31]. *Fusobacterium nucleatum* stimulates enhancement of β-catenin signaling, generates Fap2 (the autotransporter protein) that hinders CRC progression by suppressing immune cell activation and induces the development of CRC cell proliferation through the FadA adhesion virulence factor by inducing oncogenic gene expression [30]. *M. smithii* is a member of the methanogenic archaea, which is one of the three main groups of organisms in the human gut (sulfate-reducing bacteria, acetogenic bacteria, and methanogenic archaea). *M. smithii* can consume hydrogen. Since *M. smithii* utilizes H2 at a lower threshold than acetogens do, it is efficient in removing H2 from the gut environment [46]. Suppressing pathogen-activated inflammatory pathways, *S. salivarius* was discovered to be capable of affecting immune responses and had regulatory effects on the NF-B pathway in human intestinal epithelial cells [32].

V. CONCLUSION

Around 200 typical bacterial species and another 1,000 uncommon ones make up the multicellular organ known as the human gut microbiota. The human immune system, which is crucial for regulating a number of host activities, can be influenced by gut microbiota [49]. As a result, metagenomic research of the human gut microbiome provides new insights on a variety of diseases, including CRC. It has been repeatedly indicated that people with CRC exhibit an imbalanced intestinal microbiota concentration and a disease-linked limiting of diversification, and that their complex etiology combines immunological instability, genetics, and ecological factors [50]. Since the human microflora regulates the microbiome, metagenomic analysis of the intestinal flora reveals significant morphological and physiological indicators, indicating disease states [51]. It is necessary to develop a classification system that can speed up CRC diagnosis since accurate diagnosis is essential for advancing effective treatment in CRC. In order to increase diagnostic precision and pinpoint probable CRC pathobionts, this work applies a number of supervised machine learning techniques to metagenomics data linked with CRC. Overall, this research supports the application of finely tuned machine learning and feature selection techniques on metagenomics data associated to disease. We believe that our research will clarify the roles of the gut microbiome in maintaining a healthy intestinal flora and hasten the discovery of CRC related promising targets for developing diagnostic and therapeutic techniques.

VI. REFERENCES

[1] L. J. Marcos-Zambrano *et al.*, "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment," *Front. Microbiol.*, vol. 12, 2021, Accessed: Jun. 16, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2021.634511

[2] "Gut Microbiota Dysbiosis Drives the Development of Colorectal Cancer - FullText - Digestion 2021, Vol. 102, No. 4 - Karger Publishers." https://www.karger.com/Article/FullText/508328 (accessed Jun. 16, 2022).

[3] J. Halfvarson *et al.*, "Dynamics of the human gut microbiome in inflammatory bowel disease," *Nat. Microbiol.*, vol. 2, p. 17004, Feb. 2017, doi: 10.1038/nmicrobiol.2017.4.

[4] K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2016," *CA. Cancer J. Clin.*, vol. 66, no. 4, pp. 271–289, Jul. 2016, doi: 10.3322/caac.21349.

[5] Y. Cheng, Z. Ling, and L. Li, "The Intestinal Microbiota and Colorectal Cancer," *Front. Immunol.*, vol. 11, 2020, Accessed: Jun. 16, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2020.615056

[6] "The gut microbiota in IBD | Nature Reviews Gastroenterology & Hepatology." https://www.nature.com/articles/nrgastro.2012.152 (accessed Jun. 16, 2022).

[7] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," p. 40.

[8] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," 1996, doi: 10.1111/J.2517-6161.1996.TB02080.X.

[9] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nat. Commun.*, vol. 8, no. 1, p. 1784, Dec. 2017, doi: 10.1038/s41467-017-01973-8.

[10] B. Bakir-Gungor, O. Bulut, A. Jabeer, O. U. Nalbantoglu, and M. Yousef, "Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods," *Front. Microbiol.*, vol. 12, 2021, Accessed: Jun. 16, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2021.628426

978-1-6654-8894-5/22/$31.00 ©2022 IEEE

[11] B. Bakir-Gungor, H. Hacılar, A. Jabeer, O. U. Nalbantoglu, O. Aran, and M. Yousef, "Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods," *PeerJ*, vol. 10, p. e13205, 2022, doi: 10.7717/peerj.13205.

[12] M. Yousef, A. Kumar, and B. Bakir-Gungor, "Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data," *Entropy*, vol. 23, no. 1, p. 2, Dec. 2020, doi: 10.3390/e23010002.

[13] M. Yousef, A. Sayıcı, and B. Bakir-Gungor, "Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis," in *Database and Expert Systems Applications - DEXA 2021 Workshops*, Cham, 2021, pp. 205–214. doi: 10.1007/978-3-030-87101-7_20.

[14] "Brown et al. - Conditional Likelihood Maximisation A Unifying Fr.pdf." Accessed: Jun. 13, 2022. [Online]. Available: https://www.jmlr.org/papers/volume13/brown12a/brown12a.pdf

[15] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[16] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[17] "Duvallet et al. - 2017 - Meta-analysis of gut microbiome studies identifies.pdf." Accessed: Jun. 13, 2022. [Online]. Available: https://www.nature.com/articles/s41467-017-01973-8.pdf

[18] F. Fleuret and E. Ch, "Fast Binary Feature Selection with Conditional Mutual Information," p. 25.

[19] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," in *2008 23rd International Symposium on Computer and Information Sciences*, Oct. 2008, pp. 1–4. doi: 10.1109/ISCIS.2008.4717949.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn. PYTHON*, p. 6.

[21] "KNIME - the Konstanz information miner." https://dl.acm.org/doi/epdf/10.1145/1656274.1656280 (accessed Jun. 15, 2022).

[22] I. Kahouli, C. Tomaro-Duchesneau, and S. Prakash, "Probiotics in colorectal cancer (CRC) with emphasis on mechanisms of action and current perspectives," *J. Med. Microbiol.*, vol. 62, no. Pt 8, pp. 1107–1123, Aug. 2013, doi: 10.1099/jmm.0.048975-0.

[23] S. Il, P. Je, and A. J, "[Genetic and environmental factors in colorectal cancer. Mutations in the familial adenomatous polyposis gene]," *Tidsskr. Den Nor. Laegeforening Tidsskr. Prakt. Med. Ny Raekke*, vol. 117, no. 14, May 1997, Accessed: Jun. 17, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9235686/

[24] J. Wang *et al.*, "Novel Regulatory Roles of Wnt1 in Infection-Associated Colorectal Cancer," *Neoplasia N. Y. N*, vol. 20, no. 5, pp. 499–509, Apr. 2018, doi: 10.1016/j.neo.2018.03.001.

[25] K. I, T.-D. C, and P. S, "Probiotics in colorectal cancer (CRC) with emphasis on mechanisms of action and current perspectives," *J. Med. Microbiol.*, vol. 62, no. Pt 8, Aug. 2013, doi: 10.1099/jmm.0.048975-0.

[26] A. Rivière, M. Selak, D. Lantin, F. Leroy, and L. De Vuyst, "Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut," *Front. Microbiol.*, vol. 7, 2016, Accessed: Jun. 17, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2016.00979

[27] Y. Bao *et al.*, "Long noncoding RNA BFAL1 mediates enterotoxigenic Bacteroides fragilis-related carcinogenesis in colorectal cancer via the RHEB/mTOR pathway," *Cell Death Dis.*, vol. 10, no. 9, p. 675, Sep. 2019, doi: 10.1038/s41419-019-1925-2.

[28] S. Zamani, R. Taslimi, A. Sarabi, S. Jasemi, L. A. Sechi, and M. M. Feizabadi, "Enterotoxigenic Bacteroides fragilis: A Possible Etiological Candidate for Bacterially-Induced Colorectal Precancerous and Cancerous Lesions," *Front. Cell. Infect. Microbiol.*, vol. 9, p. 449, Jan. 2020, doi: 10.3389/fcimb.2019.00449.

[29] G. Cuevas-Ramos, C. R. Petit, I. Marcq, M. Boury, E. Oswald, and J.-P. Nougayrède, "Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 25, pp. 11537–11542, Jun. 2010, doi: 10.1073/pnas.1001261107.

[30] M. Castellarin *et al.*, "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Genome Res.*, vol. 22, no. 2, pp. 299–306, Feb. 2012, doi: 10.1101/gr.126516.111.

[31] L. C *et al.*, "Dysbiosis of fungal microbiota in the intestinal mucosa of patients with colorectal adenomas," *Sci. Rep.*, vol. 5, Jan. 2015, doi: 10.1038/srep07980.

[32] I. Sliepen, J. Van Damme, M. Van Essche, G. Loozen, M. Quirynen, and W. Teughels, "Microbial interactions influence inflammatory host cell responses," *J. Dent. Res.*, vol. 88, no. 11, pp. 1026–1030, Nov. 2009, doi: 10.1177/0022034509347296.

978-1-6654-8894-5/22/$31.00 ©2022 IEEE