Machine learning: a powerful tool for identifying key microbial agents associated with specific cancer types

Jia Feng[1,*], Kailan Yang[1,*], Xuexue Liu[1], Min Song[1], Ping Zhan[2], Mi Zhang[1], Jinsong Chen[1] and Jinbo Liu[1]

[1] Department of Laboratory Medicine, The Affiliated Hospital of Southwest Medical University, Sichuan Province Engineering Technology Research Center of Molecular Diagnosis of Clinical Diseases, Molecular Diagnosis of Clinical Diseases Key Laboratory of Luzhou, Sichuan, China
[2] Department of Obstetrics, The Affiliated Hospital of Southwest Medical University, Luzhou, Sichuan, China
* These authors contributed equally to this work.

ABSTRACT

Machine learning (ML) includes a broad class of computer programs that improve with experience and shows unique strengths in performing tasks such as clustering, classification and regression. Over the past decade, microbial communities have been implicated in influencing the onset, progression, metastasis, and therapeutic response of multiple cancers. Host-microbe interaction may be a physiological pathway contributing to cancer development. With the accumulation of a large number of high-throughput data, ML has been successfully applied to the study of human cancer microbiomics in an attempt to reveal the complex mechanism behind cancer. In this review, we begin with a brief overview of the data sources included in cancer microbiomics studies. Then, the characteristics of the ML algorithm are briefly introduced. Secondly, the application progress of ML in cancer microbiomics is also reviewed. Finally, we highlight the challenges and future prospects facing ML in cancer microbiomics. On this basis, we conclude that the development of cancer microbiomics can not be achieved without ML, and that ML can be used to develop tumor-targeting microbial therapies, ultimately contributing to personalized and precision medicine.

Subjects Microbiology, Oncology, Data Mining and Machine Learning
Keywords Machine learning, Cancer microbiomics, Host-microbe interaction, High-throughput data

Submitted 8 February 2023
Accepted 26 September 2023
Published 23 October 2023

Corresponding author
Jinbo Liu, liujb7203@swmu.edu.cn

Academic editor
Mohd Adnan

Additional Information and Declarations can be found on page 18

DOI **10.7717/peerj.16304**

INTRODUCTION

With the popularization of the concept of "Holobiont" in life and medical science, individual phenotypes are seen as the result of complex interactions resulting from the joint expression of host and related microbial genomes (*Simon et al., 2019*). In short, almost all functions of macroscopic organisms, including their development, growth, and health, are influenced by the complex microbial communities in which they reside (*Foster et al., 2017*). It is well known that humans are a host of multiple microbial symbionts, and microbes have co-evolved with human bodies by inhabiting them thereby creating individual habitats that are complex and are specific according to the host physiology

(*Nallanchakravarthula, Amruta & Ramamurthy, 2021*). The efforts of human microbiome projects and advances in diverse omics technologies have revolutionized our understanding of the relationship between humans and microbial symbionts (*Parida & Sharma, 2020*). Microbiota and host maintain a dynamic equilibrium referred to as eubiosis that actively influences many physiological processes and plays an important role in keeping human health. However, imbalance in the number and types of microbial population leads to ecological imbalance. In this case, the dominant strain can induce chronic inflammation, toxin and carcinogenic metabolites through multiple mechanisms, thereby affecting the host microenvironment and homeostasis (*Vimal, Himal & Kannan, 2020*). Thus, dysbiosis may directly or indirectly contribute to carcinogenesis in human beings.

The human associated microbiome consists of members from different phylogenetic groups such as bacteria, viruses, fungi, protozoa, archaea and others dominated by bacteria that live symbiotically and nonsymbiotically (*Lloyd-Price, Abu-Ali & Huttenhower, 2016*). The commensal microbiota colonizes any surface exposed to the surrounding factors, including mucosa and skin (respiratory, gastrointestinal, and urogenital), with the gut being the most densely colonized and widely colonized organ (*Loganathan & Priya Doss, 2022*). More recent developments show the existence of potentially harboring low-biomass microbial populations in other body sites, initially considered 'sterile', such as breast, lung, prostate, bladder, liver, pancreas and blood circulatory systems (*Geller et al., 2017*; *Meng et al., 2018*; *Nejman et al., 2020*; *Parhi et al., 2020*; *Poore et al., 2020*; *Riquelme, Maitra & McAllister, 2018*). Cancer has long been believed to be a complex illness brought on by interactions between the host and its microenvironment (*Elinav et al., 2013*). Cancer and its associated host microbes are collectively referred to as a Cancer microbiomics (*Nallanchakravarthula, Amruta & Ramamurthy, 2021*). Numerous studies have demonstrated that each microbial niche may influence cancer promotion through community-level interactions mediated by altered microbiome dysbiosis (*Farrell et al., 2012*; *Sobhani et al., 2011*; *Xuan et al., 2014*), direct interaction of individual members (*Koshiol et al., 2016*; *Pereira-Marques et al., 2019*), or *via* secreted or modulated metabolites. Although the causal relationship between microbes and cancer is not absolute, specific microbiome signatures and diversity may be a favorable biomarker for diagnosis and prognosis in patients with cancer.

Increasing evidence suggested that the composition of microbiota changes during the carcinogenesis or development, which gives rise to the promising diagnostic value of microbiome based signatures (*Chen et al., 2022*; *Zhang et al., 2022*). Next generation sequencing (NGS) is emerging as a powerful microbiome investigation method, allowing characterization of microbial communities at unprecedented resolution, without prior culturing (*Escobar-Zepeda, Vera-Ponce de Leon & Sanchez-Flores, 2015*). Due to difficulties identifying biomarkers with standard statistical methods for disease diagnosis, the field has moved to applying predictive ML models for classification of patient phenotypes. ML has great ability to detect informative patterns in the data with limited prior knowledge of the underlying system, which has shown to be useful for identifying key molecular signatures, discovering potential patient stratifications, and particularly for generating models that

can accurately predict phenotypes (*Li et al., 2022a*). Meanwhile, potential biomarkers associated with human disease can be identified through interpretable models (*Carrieri et al., 2021*; *Gou et al., 2021*; *Wilmanski et al., 2019*). Thus, ML has been increasingly applied to cancer microbiomics data to classify samples and predict various outcomes (*Fukui et al., 2020*; *Topcuoglu et al., 2020*).

The difficulties in accurate and efficient data analysis have become the immediate challenge that must be tackled for further investigation of the cancer microbiomics. The recently proliferated field of ML sheds new light on such tasks. In the following sections, we will provide an overview of the data sources, methods, potential application, and challenges of ML in cancer microbiomics. This narrative review aims to provide a theoretical basis for promoting the development of cancer microbiomics. At the same time, microbiology researchers and clinicians can quickly understand the current status of microbiology research in cancer and the application of ML ideas.

To assess an unbiased comprehensive analysis of ML, microbiomics and malignancy correlation studies, we used PubMed and Web of Science to identify relevant literature published before June 2023 to identify and include studies in our analysis. The search terms used were "deep learning or predictive modeling or artificial intelligence or machine learning" AND "tumor or cancer" AND "microflora or microbiome or microbiology or microbial or germ or microorganism or microbe or metagenomics or metagenome". Only studies published in English or with official English translations were included in this study. The reference lists of eligible studies were manually screened to identify additional relevant literature. As shown in Fig. 1, after an initial screening of titles and abstracts, full-text articles were carefully verified. Included studies met the following criteria: (1) research articles related to microbiomics, (2) the disease studied included malignancy, (3) the study method used ML-related algorithms. Duplicates and studies with unclear ML algorithms were excluded. Twenty-three studies assessing the relevance of microbes, malignancies, and ML were identified by the inclusion and exclusion criteria described above.

NEXT GENERATION SEQUENCING TECHNOLOGY IN MICROBIOME

Currently, microbial community analysis by multiple NGS methods mainly includes amplicon sequencing, metagenomic NGS (mNGS), RNA sequencing and omics-based sequencing (See Fig. 2A). Taxa abundance is the most commonly used feature, as microbiome profiles are assumed to be different in healthy and disease states. Amplicon sequencing is the most common NGS method that could survey almost all bacteria and sample types by targeting the 16s/18s ribosomal RNA (rRNA) housekeeping marker gene to quantify the microbiome composition (*Liu et al., 2021*). The operational taxonomic units (OTUs) is often used for microbial community analysis, which group sequences into a consensus sequence (the OTU) at a defined sequence similarity threshold by some pipelines (*e.g.*, QIIME-uclust, MOTHUR and USEARCH-UPARSE). For the 16S rRNA gene, a threshold of 97% sequence identity is generally used to define OTU to the species level (OTU-97%) (*Edgar, 2018*), and be widely used in studies. As an alternative to OTUs,

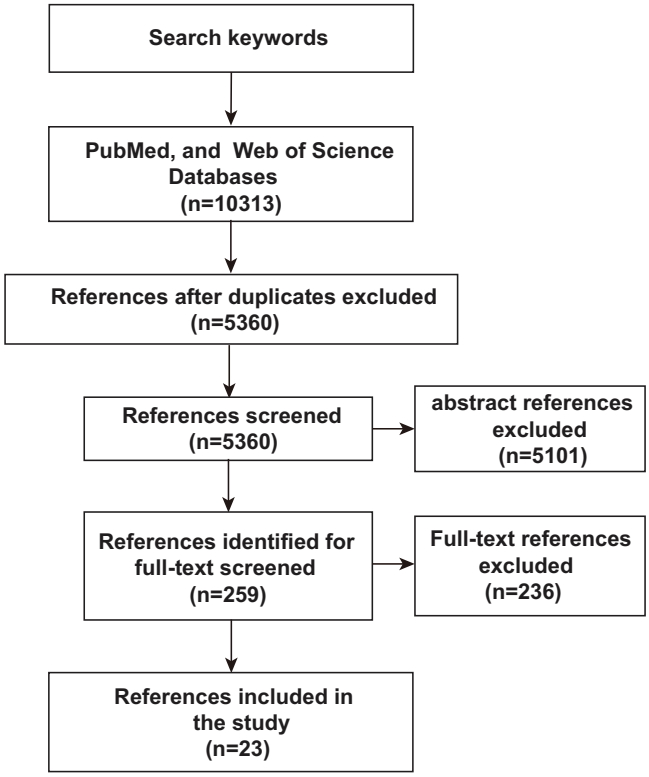

Figure 1 Schematic diagram of literature search. PubMed and Web of Science were searched for keywords, and then after an initial screening of titles and abstracts, the full text was finally scrutinized for inclusion of 23 relevant studies. Full-size DOI: 10.7717/peerj.16304/fig-1

amplicon sequencing variants (ASVs) attempt to remodel the exact biological sequences present in the sample using some pipelines (*e.g.*, Qiime2-Deblur, DADA2, and USEARCH-UNOISE3) (*Callahan, McMurdie & Holmes, 2017*). That is, sequences are grouped into the same OTU with a threshold of 100% sequence identity. Compared with OTUs, ASVs has the potential to improve both the sensitivity and specificity of 16S rRNA gene sequence inference. Finally, OTUs/ASVs table is obtained as a helpful and important starting point for most ML prediction approaches to better understand taxonomic variation within microbial communities and its prediction of host traits (*Loganathan & Priya Doss, 2022*; *Zhang, Chen & Wong, 2021*). In conclusion, Amplicon sequencing is technically mature enough to obtain sufficient information about microbial community composition in addition to being affordable for large-scale studies (*Laudadio et al., 2018*). Compared with amplicon sequencing, mNGS could simultaneously examine all genes in all organisms contained in the sample. it can result in species or strain level and has potential to perform metagenomic population functional analysis. However, it is costly and time-consuming. After mNGS, taxonomic table is obtained by comparing to the genome. Then, the taxonomic table can be transferred into input features for ML classifiers and perform prediction for diseases (*Liu et al., 2021*).

To capture the real-time functional activities of the microbiome, metatranscriptomics, metaproteomics sequencings, and metabolome detection assess the microbial RNA,

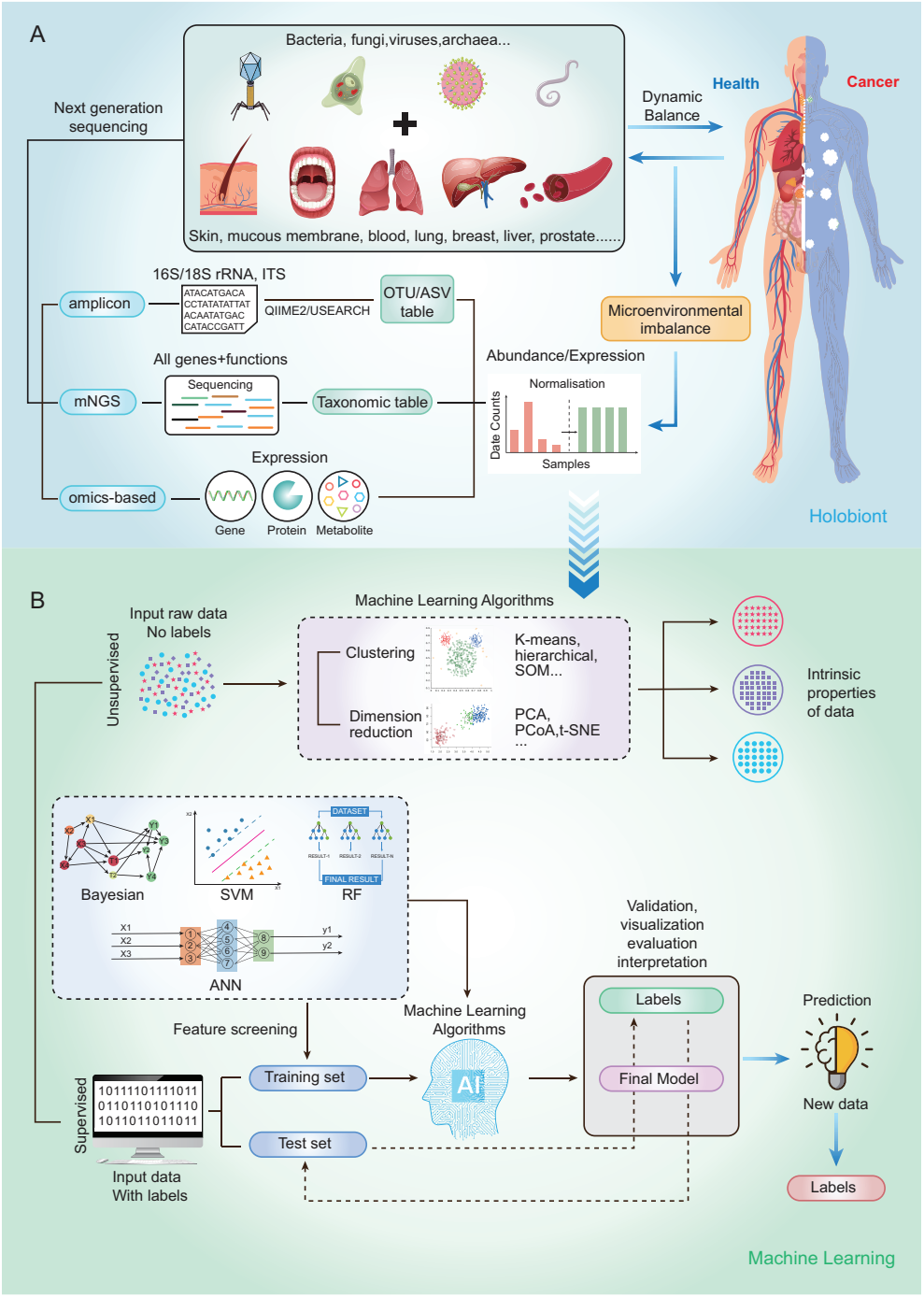

Figure 2 Schematic diagram of disease prediction modeling with microbiome data. (A) The health of the human body is inextricably linked to its microbiota. By subjecting the microbiota of a specific site to next-generation sequencing (amplicon, metagenome, omics-based), it may be possible to make predictions about diseases through the microbiota. (B) Machine learning is usually categorized into supervised (*e.g.*, RF, SVM, ANN) and unsupervised (*e.g.*, clustering and dimensionality reduction), with very different steps and outputs. The basic steps to build a disease prediction model are: 1. Sequencing data are reasonably divided into training and testing sets, and the data are preprocessed (*e.g.*, missing value interpolation, outlier elimination, normalization, or standardization); 2. Parameters are optimized, and appropriate ML algorithms are selected to build the model; and 3. The model needs to be validated, visualized, evaluated, and interpreted. Full-size DOI: 10.7717/peerj.16304/fig-2

protein, and metabolite respectively, to extract information on gene/protein/metabolite expression. Thus, gene/protein/metabolite expression is obtained as the input features for downstream analysis and classification by ML.

OVERVIEW OF MACHINE LEARNING

In the past 5–6 years, tremendous strides have been made towards using ML for predicting specific phenotypes, from the microbiome data available in public databases (*Seneviratne et al., 2020*). It is noteworthy that a large volume of high-throughput data has been generated in the microbiome, prompting the association between microbiome and cancer to become increasingly clear, and pushing researchers to use ML-based approaches to predict systematic trends in the cancer microbiomics.

By using a variety of statistical, probabilistic and optimization methods to learn from past experience, ML can build appropriate models from large, unstructured and complex datasets to solve medically relevant problems (*Choi et al., 2020*). Generally, ML algorithms can be categorized as supervised learning and unsupervised learning. Supervised learning is the most common approach to ML, models in which are optimized (*i.e.*, fitted) with a training set of manually labeled input and output data to predict the values of future inputs. Thus, supervised learning algorithms are mainly used to build ML models. For some modeling algorithms (*e.g.*, random forest (RF)), the importance of each variable (*e.g.*, species and gene) for the model prediction may be estimated for model interpretation or further analysis (*Salim et al., 2023*). Supervised learning methods are further categorized into regression and classification. Regression returns scores to reflect the possible outcome (a.k.a. class label), while classification directly returns possible outcome the sample can belong to (a.k.a. class label). Among the conventional ML methods, logistic regression (LR), RF (*Man et al., 2019*), and support vector machine (SVM) (*Lyashenko et al., 2020*) are the most frequently used supervised learning algorithms. Although ML algorithms such as RFs can handle a large number of features, their accuracy can still be limited in complex datasets (*Statnikov et al., 2013*). Deep learning (DL) belongs to a subclass of supervised learning, which has attracted increasing attention in microbiome research in recent years due to its unique advantages in processing large amounts of high and complex data. For microbiome data, the input features are relative abundances instead of images, which can be used to build DL models that classify the outcomes into presence or absence of disease state (*Xu et al., 2023*). The most commonly used form of DL is artificial neural networks (ANN) (*Guo et al., 2020*).

On the other hand, the unsupervised learning method is essentially used for processing data with no predefined labels. Because no labeled input was provided, the algorithm may explore the intrinsic properties of the data that are not apparent to human. Unsupervised learning optimizes a model by learning the entire dataset and representing it in a more compact way, or by clustering samples together based on the similarity of their features (*i.e.*, clustering and dimension reduction) (*Raza & Singh, 2021*). Dimension reduction methods include t-distributed stochastic neighbor embedding (t-SNE), principal components analysis (PCA) and principal coordinate analysis (PCoA), which have been widely used for omics data visualization and feature extraction by extracting a set of

principal variables from high-dimensional feature space (*Goodswen et al., 2021*). The clustering algorithms, including self-organizing map (SOM), hierarchical clustering and k-means clustering, are frequently implemented to partition or stratify a set of objects into multiple clusters based on similarities or differences.

The ML methods widely used in microbiome research are well-described by Namkung as published in 2020 (*Namkung, 2020*). Here, we focus on the implementation of supervised learning methods in human cancer microbiomics studies. So, we briefly sort out the basic steps of building disease prediction models based on microbiome data by analyzing relevant research results in recent years (Fig. 2B). Firstly, the sequencing data were reasonably divided into training and test sets, and the data were preprocessed (such as missing value interpolation, outlier elimination, normalization or standardization). Then the parameters are optimized and the appropriate ML algorithm is selected to build the model. Finally, the model needs to be validated, visualized, evaluated, and interpreted.

APPLICATION OF ML FOR CANCER MICROBIOMICS

The immense potential of the human microbiome grasped huge attention from the research community of cancer. There has been an explosion in the study of the microbiome in cancer research. In view of the correlation between non-invasive samples and a variety of cancers, as well as the existing microbial multi-omics big data, the current research direction of ML in cancer microbiomics is mainly focused on the development of non-invasive tools for cancer diagnosis, prediction and monitoring. We summarize the main literature on the application of microbial data to build cancer diagnostic models, and the specific research methods and strategies are shown in Table 1. As can be seen from Table 1, a total of 23 studies used ML algorithm to predict cancer phenotypes, among which 12 studies applied RF algorithm, accounting for 52.2%, suggesting that RF algorithm is one of the most widely used ML algorithms in cancer microbiomics.

ML algorithms in pan-cancer

Based on the extensive association between microbiome and cancer, and to characterize the cancer-associated microbiome, *Poore et al. (2020)* re-examined microbial reads from 18,116 samples and 33 cancer types from The Cancer Genome Atlas (TCGA) compendium of whole genome sequencing (WGS; n = 4,831) and whole transcriptome sequencing (RNA-Seq; n = 13,285) studies. They built the stochastic gradient boosting ML model that was able to distinguish well between and within cancer types and stages. They then performed deep metagenomic sequencing of plasma samples from cancer individuals with prostate, lung, and skin cancer (n = 100) and non-cancer, HIV-, healthy controls (n = 69). The ML model also could distinguish well healthy from cancer and cancer from cancer in the cell-free microbial profiles. This finding has already shown promise for a universal strategy for cancer diagnosis based on the microbiome.

Coincidentally, *Xu et al. (2023)* applied a large amount of microbial data from 21 cancer types, including 11,819 tissue samples (metagenomic data and 16S rRNA sequencing data) and 1,845 blood samples (metagenomic data), to construct DeepMicroCancer. The DeepMicroCancer is a set of tissue/blood microbiome based RF and tissue-blood

Table 1 The representative applications of machine learning in human cancer microbiomics in recent years.

ML algorithm	Sample type	Data type	Input features	Sample size	External verification	Groups	AUC	Disease	Ref.
GBM	Tissue and blood	Whole genome and transcriptome	Taxonomic table	18,116	169	CA *vs.* BD *vs.* HC	0.891	Pan-cancer	*Poore et al. (2020)*
RF	Tissue and blood	Whole genome and transcriptome	Taxonomic table	13,664	/	CA *vs.* HC	0.90	Pan-cancer	*Xu et al. (2023)*
NB	Fecal	16S rRNA	OTUs	90	/	CRC *vs.* CRA *vs.* HC	0.969	CRC	*Zackular et al. (2014)*
LR	Fecal	Metagenome	Taxonomic table	156	335	CRC *vs.* CRA	0.87	CRC	*Zeller et al. (2014)*
RF	Fecal	16S rRNA	OTUs	404	/	CRC *vs.* CRA *vs.* HC	0.853	CRC	*Baxter et al. (2016)*
SVM	Fecal	Metagenome	Taxonomic table	2,424	903	CRC *vs.* HC	0.809	CRC	*Pasolli et al. (2016)*
NB	Fecal	16S rRNA	OTUs	141	141	CRC *vs.* CRA *vs.* HC	0.994	CRC	*Ai et al. (2017)*
RF	Oral and fecal	16S rRNA	OTUs	234	/	CRC *vs.* CRA *vs.* HC	0.94	CRC	*Flemer et al. (2018)*
RF	Fecal	Metagenome	Taxonomic table	156	218	CRC *vs.* HC	0.91	CRC	*Koohi-Moghadam et al. (2019)*
RF	Fecal	Metagenome	Taxonomic table	60	231	CRC *vs.* HC	1	CRC	*Gupta et al. (2019)*
LR	Fecal	Metagenome	Taxonomic table	768	203	CRC *vs.* BD	0.92	CRC	*Wirbel et al. (2019)*
RF	Fecal	Metagenome	Taxonomic table	141	/	CRC *vs.* HC	0.76	CRC	*Kishk et al. (2018)*
LogitBoost	Fecal	Metagenome	Taxonomic table	696	/	CRC *vs.* BD	0.968	CRC	*Bang et al. (2019)*
RF	Fecal	16S rRNA	OTUs	242	/	CRC *vs.* HC	0.999	CRC	*Qu et al. (2019)*
RF	Fecal	16S rRNA	OTUs	46	/	CRC *vs.* HC	0.85	CRC	*Trivieri et al. (2020)*
LR	Fecal	Metagenome	Metabolites	192	156	CRC *vs.* CRA *vs.* HC	0.98	CRC	*Chen et al. (2022)*
GBM	Fecal	16S rRNA	OTUs	615	/	CRC *vs.* IBD *vs.* HC	0.80	CRC	*Seo et al. (2022)*
RF	Lower respiratory tract	Metagenome	Taxonomic table	150	85	CA *vs.* BD *vs.* HC	0.882	Lung cancer	*Jin et al. (2019)*
SVM	Gut	16S rRNA	OTUs	107	74	CA *vs.* HC	0.976	Lung cancer	*Zheng et al. (2020)*
RF	Lower respiratory tract	Metagenome	Taxonomic table	60	/	CA *vs.* BD	0.959	Lung cancer	*Chen et al. (2023)*
LR	Serum	Metagenome	Taxonomic table	350	/	CA *vs.* HC	0.99	Brain tumor	*Yang et al. (2020)*

Table 1 (continued)

ML algorithm	Sample type	Data type	Input features	Sample size	External verification	Groups	AUC	Disease	Ref.
RF	Peritoneum	16S rRNA	OTUs	30	/	CA *vs.* BD	0.94	Ovarian cancer	*Miao et al. (2020)*
RF	Gastric juice	16S rRNA	OTUs	139	/	CA *vs.* HC	1	Gastric cancer	*Wei et al. (2023)*

Note:
CA, cancer; BD, benign diseases; HC, health controls; AUC, area under the curve; operational taxonomic units; CRC, colorectal cancer; CAR, colorectal adenoma; IBD, inflammatory bowel disease; ML, machine learning; RF, random forest; NB, Naïve Bayesian; LR, logistic regression; SVM, support vector machine; GBM, gradient boosting machine; Ref., reference.

microbiome based transfer learning models that can be used to diagnose a wide range of cancer types. Thereinto, the tissue RF model demonstrated superior predictive performance, with area under the curve (AUC) >0.9 for the prediction of all other cancers except for lymphoid neoplasm diffuse large B-cell lymphoma (only seven samples). The top contributing bacteria *Ralstonia, Tetrasphaera, Nitrospira*, and *Luteimonas* of this RF model were generally distributed in natural environments like soil and water habitats or plant surfaces, while some of which such as *Ralstonia* were proved to cause infection when transferred to hospital settings (*Ryan & Pembroke, 2018*). In a study by *Higuchi et al. (2021)*, it was also shown that some bacteria such as *Ralstonia* constitute a mesothelioma-specific microbiota that promotes the process of cancer progression. Although the blood RF model has decent performance, which is unsurprisingly not comparable with that of the tissue RF model. The top contributor in the blood RF model is *Herbaspirillum*, which has been reported to be isolated from multiple cancer patients (*Chemaly et al., 2015*; *Gungor et al., 2020*; *Suwantarat et al., 2015*). In addition, the tissue–blood transfer learning (TL) model has a precision of 0.5 or higher for most cancers, and the multiclassification averaged AUC reached 0.89, which performed better than the RF model in predicting cancer types in small samples (*Xu et al., 2023*). These findings suggest a new class of microbial-based cancer diagnostics, providing a unique opportunity to develop cancer diagnostics.

ML algorithms in colorectal cancer (CRC)

As routine screening programs for CRC, non-invasive methods such as fecal occult blood testing (FOBT) and carcinoembryonic antigen (CEA) testing, and invasive methods such as high-quality colonoscopy can reduce the risk of death from CRC to some extent (*Medical Advisory Secretariat, 2009*; *Forones & Tanaka, 1999*; *Waldmann et al., 2021*). However, the large-scale use of these methods is limited by the low accuracy of non-invasive methods and the damage caused by invasive methods. The researchers therefore focused on the gut microbiome, a novel non-invasive method of CRC detection, in an attempt to find new potential biomarkers. There is no doubt that CRC is one of the most widely applied cancer types for ML algorithms. This has been summarized in detail in a review published by *Zhang, Chen & Wong (2021)*, and we have found some new investigations regarding CRC diagnosis based on that review as well (see Table 1). It is

worth mentioning that almost all CRC related literature in Table 1 used fecal microbiota to establish diagnostic models. With the exception of *Kishk et al. (2018)* almost all ML models have AUC values of 0.8 and above. This not only reflects the advantages of non-invasiveness and accessibility of stool types, but also reflects the huge representation of CRC characteristics of stool samples. In the following, we will supplement some recent new research in detail.

Due to the abundance of microbial species and quantity, it is particularly important to select suitable dimensionality reduction methods for feature selection prior to modeling. Based on two public data sets, *Qu et al. (2019)* used single method (correlation-based feature selection and maximum relevance-maximum distance) and multiple dimension-reduction methods (single method combination) to select features. Finally, multiple ML algorithm are utilized for modeling and diagnostic performance evaluation. The results show that the model with RF combined with multiple dimension reduction method has the best performance (AUC = 0.999). In additional, Koohi-Moghadam introduced a novel concept—MetaMarker (https://bitbucket.org/mkoohim/metamarker under GPLv3 license) (*Koohi-Moghadam et al., 2019*). Unlike the traditional OTU tables, it is effective in identifying unknown bacterial sequences that may be important for disease. Based on the whole-metagenome sequencing of fecal samples, they compared the performance evaluation of RF modeling in multiple national datasets with metaMarker or linear discriminant analysis of effect sizes (LEfSe) markers (*Segata et al., 2011*). The results showed that MetaMarker has better performance than LEFSe in distinguishing CRC samples from healthy individuals in a multiracial population. Upon pooling biomarkers from MetaMarker and LEfSe, the ML model improved classification performance beyond that of either model and the AUC could reach 0.91.

The colorectal adenoma (CRA) and inflammatory bowel disease (IBD) are the two high-risk diseases for CRC at early screening. CRA is a precancerous lesion, and early diagnosis can effectively prevent the development of CRC (CRA to CRC) (*Bray et al., 2017*). Furthermore, the increase of some specific harmful bacteria, such as enterotoxigenic *Bacteroides fragilis* and *Escherichia coli*, is associated with the chronic tissue inflammation, release of pro-inflammatory mediators and oncogenic mediators, which may increase the chance of CRC in patients with IBD (inflammation-anisoplasia-cancer) (*Quaglio et al., 2022*). *Zackular et al. (2014)* first used the Kruskal-Wallis test to identify OTUs with significant differences, and then performed linear discriminant analysis (LDA) to determine the effect sizes of these specific attributes. The largest differences between adenoma and healthy groups included age, race and five OTUs (including *Clostridiales*, *Clostridium*, *Lachnospiraceae* and *Bacteroides*). The authors used the above screened feature variables to build a Naïve Bayesian (NB) model. It was found that modeling with the NB algorithm could increase the probability of adenoma detection by more than 50 times, with an AUC as high as 0.969. The model showed that when at least one of these five OTUs was not detected, it was a signal for the presence of adenoma. The differential microorganisms identified by their studies are also supported by previous evidence (*Castellarin et al., 2012*; *Kostic et al., 2013*; *Rubinstein et al., 2013*). For example, *Bacteroides fragilis*, a pathogenic variant that is a common commensal, has been shown to directly

influence the development of CRC in a mouse genetic model by producing a metalloproteinase toxin (*Sears et al., 2008*). Similarly, *Zeller et al. (2014)* developed a CRC diagnostic model using metagenomic data from feces of CRC patients *vs*. CRA patients (*Zeller et al., 2014*). First, the authors assessed the predictive ability of microorganisms that differed significantly between the two groups and found that individual species already had some predictive ability (AUC up to 0.75). Next, the authors selected the top 22 species to build the last absolute shrinkage and selection operator (LASSO)-LR model, whose AUC could be raised to 0.84. Among those significantly enriched in CRC were *Fusobacterium*, *Porphyromonas asaccharolytica* and *Peptostreptococcus stomatis*. Although it is not clear which microorganisms are directly associated with CRC, recent evidence has shown that Fusobacterium species are prevalent CRC-associated (*Castellarin et al., 2012*) and tumor-accelerating (*Kostic et al., 2012*) microorganisms. Meanwhile, when the fecal metagenome was combined with FOBT, the accuracy of the model could be improved (AUC of 0.87). The patients included in the control group of this study were patients with small colorectal adenomas (diameter <10 mm), and thus the model developed in this study is a good predictor of microadenomas that are not easily detected by colonoscopy. In other words, for smaller colorectal adenomas, microbiomics can be used as a complementary test to colonoscopy to reduce the probability of their being missed. In addition, *Seo et al. (2022)*. developed a gradient boosting machine (GBM) model for CRC and IBD differential diagnosis. Model evaluation showed that the IBD risk model (AUC of 0.84, specificity of 74%, and sensitivity of 91%) performed slightly better than the CRC risk model (AUC of 0.80, specificity of 84%, and sensitivity of 71%) for disease classification. In the IBD risk model, the most important genera were *Dorea*, *Blautia*, *Enhydrobacter*. Meanwhile, in the CRC risk model, the most important genera were *Parvimonas*, *Peptostreptococcus*, *Psychrilyobacter*.

Some oral-associated microorganisms have been reported to be present in the fecal and colonic mucosal microbiota associated with CRC (*Baxter et al., 2016*; *Flemer et al., 2017*; *Zeller et al., 2014*), which may also be candidates for CRC biomarkers. *Flemer et al. (2018)* analyzed microbiota in oral, colonic mucosa, and feces from CRC patients (n = 99), colorectal polyp patients (n = 32), and healthy controls (n = 103). The authors first filtered out features that appeared in less than 5% of the individuals. Then in each iteration of the 10-fold cross-validation, LASSO algorithm was used to select features for 90% of the data set. Finally, The highest accuracy, with an AUC of 0.94, specificity of 95%, and sensitivity of 76%, was achieved when combining fecal and oral microbiota data to create a diagnostic model for RF. This study shows that simultaneous analysis of oral and fecal microbiomes can enhance the diagnostic efficacy of CRC.

CRC carrying the *BRAF*^{V600E} mutation has been reported to have a low response to conventional therapy and a poor prognosis (*De Sousa et al., 2013*; *Ursem, Atreya & Van Loon, 2018*). *Trivieri et al. (2020)* found that unique microbiota signatures can distinguish cases of *BRAF* mutations in CRC. They first studied fecal microbiota characteristics in CRC-carrying mice, a cohort of human CRC subjects, and a tumor-free control group by performing bacterial 16S rRNA sequencing. All of these data confirmed that a distinctive microbiota's fingerprint could be distinguished between serrated *BRAF*^{V600E} and *BRAF* wt

CRC's patients, with the former strongly resembling healthy subjects. They then detected 10 candidate species that were differentially expressed in the two CRC groups with good predictive potential to distinguish patients with *BRAF*^{V600E} from those with *BRAF* wt. Finally, the authors constructed an RF model based on these bacterial markers, which performed well in discriminating serrated CRC driven by *BRAF* mutation from *BRAF* wild-type CRC cases (AUROC = 0.85, 95% confidence interval [0.69–1.01]), with *Prevotella enoeca* and *Ruthenibacterium lactatiformans* contributing the most.
In conclusion, this RF model can differentiate the *BRAF* status of CRC patients, and thus microbiomics may be a potential non-invasive test for it.

However, focusing on one disease may not detect biomarkers specific to that disease because the host's gut microbial community may be susceptible to a variety of diseases simultaneously (*Kinross, Darzi & Nicholson, 2011*). Therefore, using the gut microbiota to explore multiple diseases simultaneously may provide a more comprehensive diagnostic value for clinical purposes. *Bang et al. (2019)* developed diagnostic models for CRC and other diseases (multiple sclerosis, idiopathic arthritis, myalgic myelitis, acquired immunodeficiency syndrome and stroke) from five classification levels (phylum, class, order, family and genus), four classifiers (LogitBoost, SVM, K nearest neighbor and logistic model tree) and two feature selection methods (forward selection and backward elimination). The LogitBoost model performed best when built using genus level and backward elimination, where the diagnostic accuracy for CRC was 96.84%. However, the subset of features selected in this study may not contain all microorganisms associated with the six diseases, which could introduce errors into the study results. The characteristics selected for the study can distinguish these six diseases simultaneously, and PSBM3 is one of the important biomarkers. PSBM3 belongs to a family of bacteria known as *Erysipelotrichaceae* associated with the immune system, and its abundance is positively correlated with tumor necrosis factor alpha levels (*Dinh et al., 2015*; *Kaakoush, 2015*; *Palm et al., 2014*). Not coincidentally, *Wirbel et al. (2019)* developed a cross-regional CRC diagnostic model by analyzing eight geographically and technologically distinct CRC fecal metagenomic studies. At first, the authors identified 94 differentially abundant microorganisms from the 849 species. Then, among these markers, the analysis was focused on the most important core group consisting of 29 species such as *Fusobacterium*, *Porphyromonas*, *Parvimonas*, *Peptostreptococcus*, *Gemella*, *Prevotella*, and *Solobacterium*. Most of these species are from the *orphyromonas* and *Dialister* genera and the *Clostridiales* order, which is strongly enriched in the CRC group and usually undetectable in controls (including type two diabetes, Parkinson's disease, and inflammatory bowel disease). Next, the authors applied differential microbial species to the LASSO-LR classifier for CRC diagnostic modeling, which can reach an AUC of 0.92. The results show that this polymicrobial CRC classifier can overcome technical and geographic research differences and has the potential to expand its applicability through validation in other regions.

Metabolic changes can be more directly observed in the state of caner cells than genomic and proteomic changes, and therefore can be used as a promising cancer marker (*Patel & Ahmed, 2015*). *Chen et al. (2022)* developed a panel of gut microbiome-associated serum metabolites (GMSM) that can accurately predict CRC and colorectal adenoma from

normal healthy population (N) through a combined analysis of serum metabolomic and faeces metagenomic. In this study, they first performed a PCA of differential metabolites (CRC *vs*. CRA, CRA *vs*. N, CRC *vs*. N) in the discovery dataset and found that the pattern was similar in both CRA and CRC patients, while the healthy population could be clearly distinguished from both populations. Subsequently, through an integrated microbiome–metabolome associated analysis, they found that these correlated species-metabolite pairs included bacterial species that were reported to be associated with CRC initiation and progression (such as CRC-promoting *P. micra*, *F. nucleatum*, *Odoribacter splanchnicus* and *Alistipes finegoldii*) and probiotics (such as *Parabacteroides distasonis* and *B. longum*). Based on these differential GMSM, 32 metabolite characteristics were initially screened after 200 LASSO experiments. Of these, eight metabolites were reliably identified in both non-targeted and targeted metabolomics assays, and both assays could accurately distinguish between normal and abnormal colorectal patients in the discovery cohort, with an AUC of 0.95 (95% CI [0.85–1.00]). PCA analysis based on the abundance of eight GMSM was still able to clearly separate normal individuals from abnormal colorectal patients. Then, a GMSM panel-based logistic regression model to predict CRC and CRA was built and yielded an AUC of 0.98 (95% CI [0.94–1.00]) in the modelling cohort. In the validation cohort, the GMSM model was significantly better than the clinical marker CEA (AUC = 0.92 *vs*. 0.72, sensitivity = 83.5% *vs*. 35.8%, specificity = 84.9% *vs*. 86.4%), and in CRA (AUC = 0.84 sensitivity = 63.2%) and early CRC (AUC = 0.93 sensitive = 88.2%) also performed well. In addition, the GMSM model was superior to the FOBT/FIT test (65.2% sensitivity) in detecting CRC. This study confirms the potential application of GMSM in CRC and CRA.

ML algorithms in lung cancer (LC)

Although the lung was previously thought to be sterile, the presence of a diverse microbiome in the lower respiratory tract has been confirmed (*Wang et al., 2021*). Based on metagenomic data, *Jin et al. (2019)* and *Chen et al. (2023)* analyzed lower respiratory tract microbiome profiles of LC patients, non-malignant lung disease patients, or healthy individuals. They all found that microbiota abundance was significantly lower in the lower respiratory tract of LC patients compared to controls. Furthermore, *Jin et al. (2019)* developed a RF model based on age, years of smoking, and 11 bacterial species (including *Streptococcus sp. I-P16*, *Prevotella melaninogenica*, *Acidovorax sp. KKS102*, *Corynebacterium urealyticum*, *Streptococcus sanguinis*, *Pseudomonas aeruginosa*, *Streptococcus pseudopneumoniae*, *H. influenzae*, *Campylobacter concisus*, *Bacteroides salanitronis*, and *B. japonicum*) with an AUC of 0. 882. Particularly, *Bradyrhizobium japonicum* was only present in patients with LC, yet *Acidovorax* was only present in patients with lung disease. *Bradyrhizobium japonicum* has been associated with inflammatory bowel disease in previous studies (*Bhatt et al., 2013*), further suggesting the relevance of inflammation to cancer. Then, *Chen et al. (2023)* based on five key genera (*Prevotella*, *Klebsiella*, *Pedobacter*, *Mycobacterium*, and *Xanthomonas*) and one tumor marker (neuron-specific enolase) as variables, the best diagnostic model for LC was constructed using the RF classifier (AUC = 0.959). These model-related genera are

speculated to be related to LC. For example, *Klebsiella* is usually associated with lung infections (*Paudel et al., 2020*). *Mycobacterium* is a well-defined pathogen that causes tuberculosis, and a history of tuberculosis has been clinically shown to be strongly associated with an increased risk of LC (*Hadifar et al., 2021*). *Prevotella* was significantly more abundant in LC patients than in patients with benign lung disease, which is consistent with previous studies (*Tsay et al., 2018*). *Xanthomonas* has also been reported to have antitumor effects (*Ma et al., 2020*). These key microorganisms are of great value in the study of cancer mechanisms and have the potential to become new targets for LC therapy.

In addition, the gut microbiota can also be used to develop a diagnostic model for lung cancer. Based on 16S rRNA sequencing data from early-stage LC patients and healthy individuals, *Zheng et al. (2020)* screened 13 OTUs for SVM modeling using minimum-redundancy maximum-relevancy (mRMR), which had a high diagnostic accuracy for LC (AUC = 0.976) (*Alshamlan, Badr & Alohali, 2015*; *Zheng et al., 2020*). Among them, *Roseburia spp*. has been shown to influence immune maintenance and anti-inflammatory effects through multiple metabolic pathways (*Louis & Flint, 2009*), *Ruminococcus bromii*-related microorganisms promote intestinal health by degrading starch and short-chain fatty acids (*Abell et al., 2008*), *Streptococcus infantis* modulates lung intrinsic immunity by detoxifying polycycles (*Hosgood et al., 2014*), and *Veillonella* is associated with Th17-mediated immunity in the lung (*Mur et al., 2018*). However, this study is only a preliminary study and the detailed link between the enteropulmonary axis and lung immunity or cancer development remains to be further elucidated.

ML algorithms in other cancers

In previous studies, brain cancer was a disease that could not be linked to the microbiome because microorganisms usually cannot cross the blood-brain barrier, which was present throughout various regions of the brain. However, some nanoscale extracellular vesicles released by microorganisms can cross the blood-brain barrier into the brain (*Cattaneo et al., 2017*), and thus extracellular vesicle microbiome data in the blood may be a powerful biomarker for assessing brain disease. Yang collected serum specimens from 152 brain cancer patients and 198 healthy controls, and performed genomic analysis of microbial extracellular vesicle components (*Yang et al., 2020*). This study used the relative abundance of OTUs at the genus level as a model variable. The highest diagnostic performance was achieved when GBM was combined with Logistic regression to build a model with an AUC of 0.99. It was found that *Dialister* and *E. rectale* were significantly decreased in the blood of brain cancer patients, while *Lachnospiraceae NK4A136* was significantly increased. This is consistent with the findings of several previous studies, which suggested that the abundance of *Dialister*, *E. rectale*, and *Lachnospiraceae NK4A136* in the intestine was associated with several neurological disorders (*Cattaneo et al., 2017*; *Jiang et al., 2015*; *Strati et al., 2017*; *Vogt et al., 2017*). For example, *E. rectale* abundance is lower in amyloid-positive (or negative) patients with mild cognitive impairment (*Cattaneo et al., 2017*), yet *Dialister* abundance is lower in patients with Alzheimer's disease, autism, and depression, compared to healthy controls (*Jiang et al., 2015*; *Strati et al., 2017*; *Vogt et al., 2017*). This study is the first to analyze extracellular vesicles in blood as a biomarker

for brain cancer diagnosis, and its findings are relevant as a new and accurate method for brain cancer diagnosis.

Due to the absence of effective screening methods for early detection of ovarian cancer at this stage and the fact that early ovarian cancer symptoms are usually reported as non-specific (stomach upset, bloating and constipation), this results in more than 60% of patients being diagnosed at an advanced stage (*Gaona-Luviano, Medina-Gaona & Magaña-Pérez, 2020*). Miao subjected peritoneal fluid from patients with ovarian cancer and benign ovarian masses to next-generation sequencing assays, and the obtained OTUs were evaluated by a series of feature selection methods, including Lasso coefficients, RF variable significance, distance correlation, t-test, and Mann-Whitney test (*Miao et al., 2020*). when using the top 18 OTUs in combination with age, BMI and serum tumor markers (cancer antigen 125 and human epididymis protein 4) to build the RF diagnostic model can greatly improve the accuracy of ovarian cancer diagnosis (AUC of 0.94). Through the above ML algorithm analysis, three different microbial characteristics were identified and may be related to the diagnosis and pathogenesis of ovarian cancer. They are anti-inflammatory properties (*Akkermansia* and *muciniphila*), estrogenic response (*Rikenellaceae*) and vascular permeability (*Alphaproteobacteria*). Reaserch has comfirmed that *Akkermansia* has a clear anti-inflammatory effect and may increase immunotherapy effectiveness in cancer patients (*van Passel et al., 2011*). Rikenellaceae is considered to be fecal microbiota related to ESR1 function, mainly involved in breast cancer prediction, prognosis and metastasis (*Carausu et al., 2019*; *Javurek et al., 2016*; *Li et al., 2022b*).

It has been reported that the development of gastric cancer is associated with microorganisms, and in addition to *H. pylori*, which is currently known to be a high risk factor for gastric cancer, other microorganisms may also serve as potential biomarkers for gastric cancer (*Eun et al., 2014*; *Lertpiriyapong et al., 2014*; *Sha et al., 2020*). *Wei et al. (2023)* collected gastric juice from healthy individuals (n = 61) and gastric cancer patients (early stage n = 48, late stage n = 30) and sequenced the 16S rRNA V1–V4 region. They adopted six ML algorithms, SVM, RF, LR, Catboost, neural network, and gradient boosting tree, to construct a risk prediction model for gastric cancer, among which the RF prediction model had the highest accuracy of 82.73%. In this model, the bacteria with the highest predictive value were *Streptococcus*, *Lactobacillus* and *Ochrobactrum*. Among them, *Streptococcus* is the bacteria with the highest proportion in early and advanced gastric cancer, which is consistent with the results of *Zhou et al. (2022)*. Study has confirmed that *Streptococcus* can promote the development of gastric cancer in a number of ways, including increasing N-nitroso compounds that cause DNA damage, and regulating the expression of key molecules important in cancer development (*San-Millán & Brooks, 2017*). In contrast, *Lactobacillus* induces anti-cancer effects by enhancing cancer cell apoptosis and protecting against oxidative stress (*Badgeley et al., 2021*).

THE CHALLENGES AND FUTURE OUTLOOK OF ML IN CANCER MICROBIOMICS APPLICATIONS

Although ML has made great achievements in microbiological research, its application in cancer microbiomics still faces many difficulties. Only through in-depth research and

analysis of these issues can we facilitate the application of ML in cancer microbiomics and accelerate the pace of its clinical translation. These difficulties are mainly manifested in the following aspects. The first aspect is the amount of data. The accuracy of a ML model is highly dependent on the amount of data it is trained on. Due to some objective factors, the sample size included in modeling is very small, such as expensive genomic sequencing data.. Therefore, in microbiome research, the current situation of reducing sequencing costs and increasing the amount of sequencing data is still worthy of strong appeal. The second aspect is the quantity and quality of input features. In a typical microbiome study, there are often thousands of features from the otus, and the number of features greatly exceeds the number of included samples, making it difficult to accurately include key features, leading to overfitting of the model and reducing the accuracy of the model classification. Although a large number of efficient feature selection methods are used to select the most relevant features as input to the model, it is difficult to overcome this dilemma. However, overfitting can often be prevented by fitting multiple models and using multiple validation sets or cross-validation to compare their predictive accuracy on the test data. Alternatively, overfitting can be reduced by training the model on a larger number of data points. The third challenge is the generalization ability of the ML model, that is, the model developed should apply not only to the training data set, but also to other similar data sets. At present, relatively few studies on the application of ML in cancer microbiomics involve validation of training models with external data and multi-center datasets. Therefore, model validation in multiple external data sets or multi-center data sets will be an effective means to avoid the generation of overfitting models and evaluate the generalization ability of models efficiently. Finally, due to the limitation of detection technology, the microbial information in a large number of solid tumors has not been deeply mined, so high-precision detection technology is still expected.

In addition to the challenges mentioned above, there is a greater requirement to focus on analyzing the quality of the data and the processing of research results with high AUC. In a study by *Poore et al. (2020)*, the authors demonstrated the ability to differentiate between multiple cancer types using only nucleic acids from blood and tissue microbes by using ML algorithms, with most of their models exceeding 95% accuracy. Due to the large size of the study, its methodology was fairly rigorous from the description of the article, while obtaining very satisfactory results, thus making it susceptible to imitation by a large number of scholars. However, the study did make some errors in the analysis of the data. First, there were significant errors in the normalization of the raw data read counts. *Poore et al. (2020)* used normalization rather than raw data to construct their machine learning classifier to eliminate batch effects. However, during the standardization process, many cancer types were labeled with incorrect values, and these incorrect manual labels were used to create highly accurate classifiers. The Voom-SNM normalization approach used by *Poore et al. (2020)* would inadvertently append *a priori* information about tumor type to the normalized data. In order to address such false-positive events, it is necessary to ensure the credibility of the results by choosing the appropriate standardized approach for different data types, as well as appropriate pre-experiments to verify their feasibility. Secondly, *Poore et al. (2020)* incorrectly categorized human reads as bacteria. These human

readings matched to bacteria were unrelated to the actual presence of bacteria in the tumor samples, leading to millions of false-positive findings for bacterial readings. To eliminate this interference, reads not mapped to the human genome can be compared to the complete CHM13 human genome using a sequence comparison tool such as Bowtie2 to exclude additional reads with human matches. Beyond that, a great deal of the data and findings in this article need to be confirmed one by one. Similar major data analysis errors need to be reduced in future studies of the microbiome and ML to avoid invalid findings.

In cancer and microbiota-related studies, more attention should be paid to regions other than the gut microbiota, such as the oral cavity, skin, breast, abdominal cavity, urine, and reproductive tract, because studies of these specific microbiota are still scarce. Also, there is a need to think about how to use ML tools to process microbiome data and how to describe the relationship between training data, test data, overfitting and generalization. In the future, researchers should actively embrace the artificial intelligence revolution and become ML practitioners, leveraging research findings in computer science to achieve the use of microbiota analysis as a powerful tool for disease prevention, prediction, diagnosis and treatment.

CONCLUSION

Despite the many challenges, ML remains an important tool in the study of cancer and microbiota, and it represents a critical step in the search for sensitive, specific and non-invasive methods of cancer diagnosis. In short, researchers have used machine learning to excavate a large number of microorganisms with potential links to the development of a variety of tumors, which will promote the development of new tumor markers.

ABBREVIATIONS

ML	Machine learning
NGS	Next generation sequencing
mNGS	Metagenomic NGS
OTU	Operational taxonomic unit
RF	Random forest
LR	Logistic regression
SVM	Support vector machine
DL	Deep learning
ANN	Artificial neural networks
SOM	Self-organizing map
PCA	Principal components analysis
PCoA	Principal coordinate analysis
t-SNE	T-distributed stochastic neighbor embedding
TCGA	The Cancer Genome Atlas
WGS	Whole genome sequencing
TL	Transfer learning
CRC	Colorectal cancer

FOBT	Fecal occult blood testing
CEA	Carcinoembryonic antigen
LEfSe	Linear discriminant analysis of effect sizes
CRA	Colorectal adenoma
IBD	Inflammatory bowel disease
LDA	Linear discriminant analysis
NB	Naïve Bayesian
LASSO	Last absolute shrinkage and selection operator
GBM	Gradient boosting machine; FDR, false discovery rate
GMSM	Gut microbiome-associated serum metabolites
N	Normal healthy population
LC	Lung cancer
mRMR	Minimum-redundancy maximum-relevancy
rRNA	Ribosomal RNA
ROC	Receiver operating characteristic
AUC	Area under the curve

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Sichuan Science and Technology Program, China (2021YFS0332, 2022NSFSC1426, 2022YFS0312, 2022ZHYZ0012) and the Scientific Research Program of Southwest Medical University (2022QN071). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Sichuan Science and Technology Program, China: 2021YFS0332, 2022NSFSC1426, 2022YFS0312, 2022ZHYZ0012.
Scientific Research Program of Southwest Medical University: 2022QN071.

Competing Interests

The authors declare no conflict of interest.

Author Contributions

- Jia Feng conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Kailan Yang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Xuexue Liu conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Min Song performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Ping Zhan performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Mi Zhang conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jinsong Chen conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jinbo Liu conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review and hence did not utilize raw data.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.16304#supplemental-information.

REFERENCES

Abell GC, Cooke CM, Bennett CN, Conlon MA, McOrist AL. 2008. Phylotypes related to Ruminococcus bromii are abundant in the large bowel of humans and increase in response to a diet high in resistant starch. *FEMS Microbiology Ecology* **66(3)**:505–515 DOI 10.1111/j.1574-6941.2008.00527.x.

Ai L, Tian H, Chen Z, Chen H, Xu J, Fang JY. 2017. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget* **8**:9546–9556 DOI 10.18632/oncotarget.14488.

Alshamlan H, Badr G, Alohali Y. 2015. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *BioMed Research International* **2015(9)**:604910 DOI 10.1155/2015/604910.

Badgeley A, Anwar H, Modi K, Murphy P, Lakshmikuttyamma A. 2021. Effect of probiotics and gut microbiota on anti-cancer drugs: mechanistic perspectives. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1875(1)**:188494 DOI 10.1016/j.bbcan.2020.188494.

Bang S, Yoo D, Kim SJ, Jhang S, Cho S, Kim H. 2019. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Scientific Reports* **9**:10189 DOI 10.1038/s41598-019-46249-x.

Baxter NT, Ruffin MT IV, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine* **8**:37 DOI 10.1186/s13073-016-0290-3.

Bhatt AS, Freeman SS, Herrera AF, Pedamallu CS, Gevers D, Duke F, Jung J, Michaud M, Walker BJ, Young S, Earl AM, Kostic AD, Ojesina AI, Hasserjian R, Ballen KK, Chen YB, Hobbs G, Antin JH, Soiffer RJ, Baden LR, Garrett WS, Hornick JL, Marty FM, Meyerson M.

2013. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. *New England Journal of Medicine* **369(6)**:517–528 DOI 10.1056/NEJMoa1211115.

Bray C, Bell LN, Liang H, Collins D, Yale SH. 2017. Colorectal cancer screening. *WMJ* **116**:27–33.

Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11(12)**:2639–2643 DOI 10.1038/ismej.2017.119.

Carausu M, Bidard FC, Callens C, Melaabi S, Jeannot E, Pierga JY, Cabel L. 2019. ESR1 mutations: a new biomarker in breast cancer. *Expert Review of Molecular Diagnostics* **19(7)**:599–611 DOI 10.1080/14737159.2019.1631799.

Carrieri AP, Haiminen N, Maudsley-Barton S, Gardiner LJ, Murphy B, Mayes AE, Paterson S, Grimshaw S, Winn M, Shand C, Hadjidoukas P, Rowe WPM, Hawkins S, MacGuire-Flanagan A, Tazzioli J, Kenny JG, Parida L, Hoptroff M, Pyzer-Knapp EO. 2021. Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific Reports* **11**:4565 DOI 10.1038/s41598-021-83922-6.

Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA. 2012. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Research* **22(2)**:299–306 DOI 10.1101/gr.126516.111.

Cattaneo A, Cattane N, Galluzzi S, Provasi S, Lopizzo N, Festari C, Ferrari C, Guerra UP, Paghera B, Muscio C, Bianchetti A, Volta GD, Turla M, Cotelli MS, Gennuso M, Prelle A, Zanetti O, Lussignoli G, Mirabile D, Bellandi D, Gentile S, Belotti G, Villani D, Harach T, Bolmont T, Padovani A, Boccardi M, Frisoni GB. 2017. Association of brain amyloidosis with pro-inflammatory gut bacterial taxa and peripheral inflammation markers in cognitively impaired elderly. *Neurobiology of Aging* **49**:60–68 DOI 10.1016/j.neurobiolaging.2016.08.019.

Chemaly RF, Dantes R, Shah DP, Shah PK, Pascoe N, Ariza-Heredia E, Perego C, Nguyen DB, Nguyen K, Modarai F, Moulton-Meissner H, Noble-Wang J, Tarrand JJ, LiPuma JJ, Guh AY, MacCannell T, Raad I, Mulanovich V. 2015. Cluster and sporadic cases of herbaspirillum species infections in patients with cancer. *Clinical Infectious Diseases* **60(1)**:48–54 DOI 10.1093/cid/ciu712.

Chen F, Dai X, Zhou CC, Li KX, Zhang YJ, Lou XY, Zhu YM, Sun YL, Peng BX, Cui W. 2022. 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)**:1315–1325 DOI 10.1136/gutjnl-2020-323476.

Chen Q, Hou K, Tang M, Ying S, Zhao X, Li G, Pan J, He X, Xia H, Li Y, Lou Z, Zhang L. 2023. Screening of potential microbial markers for lung cancer using metagenomic sequencing. *Cancer Medicine* **12(6)**:7127–7139 DOI 10.1002/cam4.5513.

Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. 2020. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology* **9**:14 DOI 10.1167/tvst.9.2.14.

De Sousa EMF, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, Rodermond H, van der Heijden M, van Noesel CJ, Tuynman JB, Dekker E, Markowetz F, Medema JP, Vermeulen L. 2013. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine* **19(5)**:614–618 DOI 10.1038/nm.3174.

Dinh DM, Volpe GE, Duffalo C, Bhalchandra S, Tai AK, Kane AV, Wanke CA, Ward HD. 2015. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *Journal of Infectious Diseases* **211(1)**:19–27 DOI 10.1093/infdis/jiu409.

Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34(14)**:2371–2375 DOI 10.1093/bioinformatics/bty113.

Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. 2013. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nature Reviews Cancer* **13(11)**:759–771 DOI 10.1038/nrc3611.

Escobar-Zepeda A, Vera-Ponce de Leon A, Sanchez-Flores A. 2015. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics* **6(104)**:348 DOI 10.3389/fgene.2015.00348.

Eun CS, Kim BK, Han DS, Kim SY, Kim KM, Choi BY, Song KS, Kim YS, Kim JF. 2014. Differences in gastric mucosal microbiota profiling in patients with chronic gastritis, intestinal metaplasia, and gastric cancer using pyrosequencing methods. *Helicobacter* **19(6)**:407–416 DOI 10.1111/hel.12145.

Farrell JJ, Zhang L, Zhou H, Chia D, Elashoff D, Akin D, Paster BJ, Joshipura K, Wong DT. 2012. Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut* **61(4)**:582–588 DOI 10.1136/gutjnl-2011-300784.

Flemer B, Lynch DB, Brown JM, Jeffery IB, Ryan FJ, Claesson MJ, O'Riordain M, Shanahan F, O'Toole PW. 2017. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66(4)**:633–643 DOI 10.1136/gutjnl-2015-309595.

Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW. 2018. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67(8)**:1454–1463 DOI 10.1136/gutjnl-2017-314814.

Forones NM, Tanaka M. 1999. CEA and CA 19-9 as prognostic indexes in colorectal cancer. *Hepatogastroenterology* **46**:905–908.

Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. 2017. The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548(7665)**:43–51 DOI 10.1038/nature23292.

Fukui H, Nishida A, Matsuda S, Kira F, Watanabe S, Kuriyama M, Kawakami K, Aikawa Y, Oda N, Arai K, Matsunaga A, Nonaka M, Nakai K, Shinmura W, Matsumoto M, Morishita S, Takeda AK, Miwa H. 2020. Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. *Journal of Clinical Medicine* **9(8)**:2403 DOI 10.3390/jcm9082403.

Gaona-Luviano P, Medina-Gaona LA, Magaña-Pérez K. 2020. Epidemiology of ovarian cancer. *Chinese Clinical Oncology* **9(4)**:47 DOI 10.21037/cco-20-34.

Geller LT, Barzily-Rokni M, Danino T, Jonas OH, Shental N, Nejman D, Gavert N, Zwang Y, Cooper ZA, Shee K, Thaiss CA, Reuben A, Livny J, Avraham R, Frederick DT, Ligorio M, Chatman K, Johnston SE, Mosher CM, Brandis A, Fuks G, Gurbatri C, Gopalakrishnan V, Kim M, Hurd MW, Katz M, Fleming J, Maitra A, Smith DA, Skalak M, Bu J, Michaud M, Trauger SA, Barshack I, Golan T, Sandbank J, Flaherty KT, Mandinova A, Garrett WS, Thayer SP, Ferrone CR, Huttenhower C, Bhatia SN, Gevers D, Wargo JA, Golub TR, Straussman R. 2017. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357(6356)**:1156–1160 DOI 10.1126/science.aah5043.

Goodswen SJ, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. 2021. Machine learning and applications in microbiology. *FEMS Microbiology Reviews* **45(5)**:232 DOI 10.1093/femsre/fuab015.

Gou W, Ling CW, He Y, Jiang Z, Fu Y, Xu F, Miao Z, Sun TY, Lin JS, Zhu HL, Zhou H, Chen YM, Zheng JS. 2021. Interpretable machine learning framework reveals robust gut

microbiome features associated with type 2 diabetes. *Diabetes Care* **44(2)**:358–366 DOI 10.2337/dc20-1536.

Gungor AA, Demirdag TB, Dinc B, Azak E, Yazal Erdem A, Kurtipek B, Ozkaya Parlakay A, Sari N. 2020. A case of infective endocarditis due to Herbaspirillum huttiense in a pediatric oncology patient. *The Journal of Infection in Developing Countries* **14(11)**:1349–1351 DOI 10.3855/jidc.13001.

Guo S, Zhang H, Chu Y, Jiang Q, Ma Y. 2020. A neural network-based framework to understand the type 2 diabetes (T2D)-related alteration of the human gut microbiome. *Cold Spring Harbor Laboratory* **1(2)**:e20 DOI 10.1002/imt2.20.

Gupta A, Dhakan DB, Maji A, Saxena R, PV PK, Mahajan S, Pulikkan J, Kurian J, Gomez AM, Scaria J, Amato KR, Sharma AK, Sharma VK. 2019. Association of flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *mSystems* **4(6)**:e00438-19 DOI 10.1128/mSystems.00438-19.

Hadifar S, Mostafaei S, Behrouzi A, Fateh A, Riahi P, Siadat SD, Vaziri F. 2021. Strain-specific behavior of Mycobacterium tuberculosis in A549 lung cancer cell line. *BMC Bioinformatics* **22**:154 DOI 10.1186/s12859-021-04100-z.

Higuchi R, Goto T, Hirotsu Y, Otake S, Oyama T, Amemiya K, Mochizuki H, Omata M. 2021. Streptococcus australis and Ralstonia pickettii as major microbiota in mesotheliomas. *Journal of Personalized Medicine* **11(4)**:297 DOI 10.3390/jpm11040297.

Hosgood HD III, Sapkota AR, Rothman N, Rohan T, Hu W, Xu J, Vermeulen R, He X, White JR, Wu G, Wei F, Mongodin EF, Lan Q. 2014. The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. *Environmental and Molecular Mutagenesis* **55(8)**:643–651 DOI 10.1002/em.21878.

Javurek AB, Spollen WG, Ali AMM, Johnson SA, Lubahn DB, Bivens NJ, Bromert KH, Ellersieck MR, Givan SA, Rosenfeld CS. 2016. Discovery of a novel seminal fluid microbiome and influence of estrogen receptor alpha genetic status. *Scientific Reports* **6**:23027 DOI 10.1038/srep23027.

Jiang H, Ling Z, Zhang Y, Mao H, Ma Z, Yin Y, Wang W, Tang W, Tan Z, Shi J, Li L, Ruan B. 2015. Altered fecal microbiota composition in patients with major depressive disorder. *Brain, Behavior, and Immunity* **48(2102–2112)**:186–194 DOI 10.1016/j.bbi.2015.03.016.

Jin J, Gan Y, Liu H, Wang Z, Yuan J, Deng T, Zhou Y, Zhu Y, Zhu H, Yang S, Shen W, Xie D, Wu H, Liu D, Li W. 2019. Diminishing microbiome richness and distinction in the lower respiratory tract of lung cancer patients: a multiple comparative study design with independent validation. *Lung Cancer* **136(4)**:129–135 DOI 10.1016/j.lungcan.2019.08.022.

Kaakoush NO. 2015. Insights into the role of Erysipelotrichaceae in the human host. *Frontiers in Cellular and Infection Microbiology* **5**:84 DOI 10.3389/fcimb.2015.00084.

Kinross JM, Darzi AW, Nicholson JK. 2011. Gut microbiome-host interactions in health and disease. *Genome Medicine* **3(3)**:14 DOI 10.1186/gm228.

Kishk A, Elzizy A, Galal D, Razek EA, Fawzy E, Ahmed G, Gawish M, Hamad S, El-Hadidi M. 2018. A hybrid machine learning approach for the phenotypic classification of metagenomic colon cancer reads based on kmer frequency and biomarker profiling. In: *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*. Piscataway: IEEE, 11–21.

Koohi-Moghadam M, Borad MJ, Tran NL, Swanson KR, Boardman LA, Sun H, Wang J. 2019. MetaMarker: a pipeline for de novo discovery of novel metagenomic biomarkers. *Bioinformatics* **35(19)**:3812–3814 DOI 10.1093/bioinformatics/btz123.

Koshiol J, Wozniak A, Cook P, Adaniel C, Acevedo J, Azócar L, Hsing AW, Roa JC, Pasetti MF, Miquel JF, Levine MM, Ferreccio C, The Gallbladder Cancer Chile Working Group. 2016.

Salmonella enterica serovar Typhi and gallbladder cancer: a case-control study and meta-analysis. *Cancer Medicine* **5(11)**:3235–3310 DOI 10.1002/cam4.915.

Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL, El-Omar EM, Brenner D, Fuchs CS, Meyerson M, Garrett WS. 2013. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe* **14(2)**:207–215 DOI 10.1016/j.chom.2013.07.007.

Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M. 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Research* **22(2)**:292–298 DOI 10.1101/gr.126573.111.

Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. 2018. Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *OMICS: a Journal of Integrative Biology* **22(4)**:248–254 DOI 10.1089/omi.2018.0013.

Lertpiriyapong K, Whary MT, Muthupalani S, Lofgren JL, Gamazon ER, Feng Y, Ge Z, Wang TC, Fox JG. 2014. Gastric colonisation with a restricted commensal microbiota replicates the promotion of neoplastic lesions by diverse intestinal microbiota in the Helicobacter pylori INS-GAS mouse model of gastric carcinogenesis. *Gut* **63(1)**:54–63 DOI 10.1136/gutjnl-2013-305178.

Li P, Luo H, Ji B, Nielsen J. 2022a. Machine learning for data integration in human gut microbiome. *Microbial Cell Factories* **21**:241 DOI 10.1186/s12934-022-01973-4.

Li Z, Wu Y, Yates ME, Tasdemir N, Bahreini A, Chen J, Levine KM, Priedigkeit NM, Nasrazadani A, Ali S, Buluwela L, Arnesen S, Gertz J, Richer JK, Troness B, El-Ashry D, Zhang Q, Gerratana L, Zhang Y, Cristofanilli M, Montanez MA, Sundd P, Wallace CT, Watkins SC, Fumagalli C, Guerini-Rocco E, Zhu L, Tseng GC, Wagle N, Carroll JS, Jank P, Denkert C, Karsten MM, Blohmer JU, Park BH, Lucas PC, Atkinson JM, Lee AV, Oesterreich S. 2022b. Hotspot ESR1 mutations are multimodal and contextual modulators of breast cancer metastasis. *Cancer Research* **82(7)**:1321–1339 DOI 10.1158/0008-5472.CAN-21-2576.

Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell* **12(5)**:315–330 DOI 10.1007/s13238-020-00724-8.

Lloyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. *Genome Medicine* **8**:51 DOI 10.1186/s13073-016-0307-y.

Loganathan T, Priya Doss CG. 2022. The influence of machine learning technologies in gut microbiome research and cancer studies—a review. *Life Sciences* **311(10)**:121118 DOI 10.1016/j.lfs.2022.121118.

Louis P, Flint HJ. 2009. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiology Letters* **294(1)**:1–8 DOI 10.1111/j.1574-6968.2009.01514.x.

Lyashenko C, Herrman E, Irwin J, James A, Strauss S, Warner J, Khor B, Snow M, Ortiz S, Waid E, Nasry B, Chai J, Choong C, Palmer E, Kutsch K, Forsyth A, Choi D, Maier T, Machida CA. 2020. Adjunctive dental therapies in caries-active children: shifting the cariogenic salivary microbiome from dysbiosis towards non-cariogenic health. *Human Microbiome Journal* **18**:100077 DOI 10.1016/j.humic.2020.100077.

Ma J, Gnanasekar A, Lee A, Li WT, Haas M, Wang-Rodriguez J, Chang EY, Rajasekaran M, Ongkeko WM. 2020. Influence of intratumor microbiome on clinical outcome and immune processes in prostate cancer. *Cancers* **12(9)**:2524 DOI 10.3390/cancers12092524.

Man WH, van Houten MA, Mérelle ME, Vlieger AM, Chu M, Jansen NJG, Sanders EAM, Bogaert D. 2019. Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. *The Lancet Respiratory Medicine* **7(5)**:417–426 DOI 10.1016/S2213-2600(18)30449-1.

Medical Advisory Secretariat. 2009. Fecal occult blood test for colorectal cancer screening: an evidence-based analysis. *Ontario Health Technology Assessment Series* **9(10)**:1–40.

Meng S, Chen B, Yang J, Wang J, Zhu D, Meng Q, Zhang L. 2018. Study of microbiomes in aseptically collected samples of human breast tissue using needle biopsy and the potential role of in situ tissue microbiomes for promoting malignancy. *Frontiers in Oncology* **8**:318 DOI 10.3389/fonc.2018.00318.

Miao R, Badger TC, Groesch K, Diaz-Sylvester PL, Wilson T, Ghareeb A, Martin JA, Cregger M, Welge M, Bushell C, Auvil L, Zhu R, Brard L, Braundmeier-Fleming A. 2020. Assessment of peritoneal microbial features and tumor marker levels as potential diagnostic tools for ovarian cancer. *PLOS ONE* **15(1)**:e0227707 DOI 10.1371/journal.pone.0227707.

Mur LA, Huws SA, Cameron SJ, Lewis PD, Lewis KE. 2018. Lung cancer: a new frontier for microbiome research and clinical translation. *Ecancermedicalscience* **12**:866 DOI 10.3332/ecancer.2018.866.

Nallanchakravarthula S, Amruta N, Ramamurthy C. 2021. Cancer microbiome; opportunities and challenges. *Endocrine, Metabolic & Immune Disorders-Drug Targets* **21(2)**:215–229 DOI 10.2174/1871530320999200818134942.

Namkung J. 2020. Machine learning methods for microbiome studies. *Journal of Microbiology* **58(3)**:206–216 DOI 10.1007/s12275-020-0066-8.

Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, Rotter-Maskowitz A, Weiser R, Mallel G, Gigi E, Meltser A, Douglas GM, Kamer I, Gopalakrishnan V, Dadosh T, Levin-Zaidman S, Avnet S, Atlan T, Cooper ZA, Arora R, Cogdill AP, Khan MAW, Ologun G, Bussi Y, Weinberger A, Lotan-Pompan M, Golani O, Perry G, Rokah M, Bahar-Shany K, Rozeman EA, Blank CU, Ronai A, Shaoul R, Amit A, Dorfman T, Kremer R, Cohen ZR, Harnof S, Siegal T, Yehuda-Shnaidman E, Gal-Yam EN, Shapira H, Baldini N, Langille MGI, Ben-Nun A, Kaufman B, Nissan A, Golan T, Dadiani M, Levanon K, Bar J, Yust-Katz S, Barshack I, Peeper DS, Raz DJ, Segal E, Wargo JA, Sandbank J, Shental N, Straussman R. 2020. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368(6494)**:973–980 DOI 10.1126/science.aay9189.

Palm NW, de Zoete MR, Cullen TW, Barry NA, Stefanowski J, Hao L, Degnan PH, Hu J, Peter I, Zhang W, Ruggiero E, Cho JH, Goodman AL, Flavell RA. 2014. Immunoglobulin A coating identifies colitogenic bacteria in inflammatory bowel disease. *Cell* **158(5)**:1000–1010 DOI 10.1016/j.cell.2014.08.006.

Parhi L, Alon-Maimon T, Sol A, Nejman D, Shhadeh A, Fainsod-Levi T, Yajuk O, Isaacson B, Abed J, Maalouf N, Nissan A, Sandbank J, Yehuda-Shnaidman E, Ponath F, Vogel J, Mandelboim O, Granot Z, Straussman R, Bachrach G. 2020. Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nature Communications* **11(1)**:3259 DOI 10.1038/s41467-020-16967-2.

Parida S, Sharma D. 2020. Microbial alterations and risk factors of breast cancer: connections and mechanistic insights. *Cells* **9(5)**:1091 DOI 10.3390/cells9051091.

Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLOS Computational Biology* **12**: e1004977 DOI 10.1371/journal.pcbi.1004977.

Patel S, Ahmed S. 2015. Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. *Journal of Pharmaceutical and Biomedical Analysis* **107**:63–74 DOI 10.1016/j.jpba.2014.12.020.

Paudel KR, Dharwal V, Patel VK, Galvao I, Wadhwa R, Malyla V, Shen SS, Budden KF, Hansbro NG, Vaughan A, Yang IA, Kohonen-Corish MRJ, Bebawy M, Dua K, Hansbro PM. 2020. Role of lung microbiome in innate immune response associated with chronic lung diseases. *Frontiers in Medicine* **7**:554 DOI 10.3389/fmed.2020.00554.

Pereira-Marques J, Ferreira RM, Pinto-Ribeiro I, Figueiredo C. 2019. Helicobacter pylori infection, the gastric microbiome and gastric cancer. *Advances in Experimental Medicine and Biology* **1149**:195–210 DOI 10.1007/978-3-030-21916-1.

Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, McKay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579(7800)**:567–574 DOI 10.1038/s41586-020-2095-1.

Qu K, Gao F, Guo F, Zou Q. 2019. Taxonomy dimension reduction for colorectal cancer prediction. *Computational Biology and Chemistry* **83(5)**:107160 DOI 10.1016/j.compbiolchem.2019.107160.

Quaglio AEV, Grillo TG, De Oliveira ECS, Di Stasi LC, Sassaki LY. 2022. Gut microbiota, inflammatory bowel disease and colorectal cancer. *World Journal of Gastroenterology* **28(30)**:4053–4060 DOI 10.3748/wjg.v28.i30.4053.

Raza K, Singh NK. 2021. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging Formerly Current Medical Imaging Reviews* **17(9)**:1059–1077 DOI 10.2174/1573405617666210127154257.

Riquelme E, Maitra A, McAllister F. 2018. Immunotherapy for pancreatic cancer: more than just a gut feeling. *Cancer Discovery* **8(4)**:386–388 DOI 10.1158/2159-8290.CD-18-0123.

Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. 2013. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host & Microbe* **14(2)**:195–206 DOI 10.1016/j.chom.2013.07.012.

Ryan MP, Pembroke JT. 2018. Brevundimonas spp: emerging global opportunistic pathogens. *Virulence* **9(1)**:480–493 DOI 10.1080/21505594.2017.1419116.

Salim F, Mizutani S, Zolfo M, Yamada T. 2023. Recent advances of machine learning applications in human gut microbiota study: from observational analysis toward causal inference and clinical intervention. *Current Opinion in Biotechnology* **79**:102884 DOI 10.1016/j.copbio.2022.102884.

San-Millán I, Brooks GA. 2017. Reexamining cancer metabolism: lactate production for carcinogenesis could be the purpose and explanation of the Warburg Effect. *Carcinogenesis* **38(6 Pt 1)**:119–133 DOI 10.1093/carcin/bgw127.

Sears CL, Islam S, Saha A, Arjumand M, Alam NH, Faruque AS, Salam MA, Shin J, Hecht D, Weintraub A, Sack RB, Qadri F. 2008. Association of enterotoxigenic bacteroides fragilis infection with inflammatory diarrhea. *Clinical Infectious Diseases* **47**:797–803 DOI 10.1086/591130.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. *Genome Biology* **12(6)**:R60 DOI 10.1186/gb-2011-12-6-r60.

Seneviratne CJ, Balan P, Suriyanarayanan T, Lakshmanan M, Lee DY, Rho M, Jakubovics N, Brandt B, Crielaard W, Zaura E. 2020. Oral microbiome-systemic link studies: perspectives on current limitations and future artificial intelligence-based approaches. *Critical Reviews in Microbiology* **46(3)**:288–299 DOI 10.1080/1040841X.2020.1766414.

Seo H, Kwon CO, Park JH, Kang CS, Shin TS, Yang EY, Jung JW, Moon BS, Kim YK. 2022. Dietary efficacy evaluation by applying a prediction model using clinical fecal microbiome data of colorectal disease to a controlled animal model from an obesity perspective. *Microorganisms* **10(9)**:1833 DOI 10.3390/microorganisms10091833.

Sha S, Ni L, Stefil M, Dixon M, Mouraviev V. 2020. The human gastrointestinal microbiota and prostate cancer development and treatment. *Investigative and Clinical Urology* **61(Suppl 1)**:S43–S50 DOI 10.4111/icu.2020.61.S1.S43.

Simon JC, Marchesi JR, Mougel C, Selosse MA. 2019. Host-microbiota interactions: from holobiont theory to analysis. *Microbiome* **7(1)**:5 DOI 10.1186/s40168-019-0619-4.

Sobhani I, Tap J, Roudot-Thoraval F, Roperch JP, Letulle S, Langella P, Corthier G, Tran Van Nhieu J, Furet JP. 2011. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLOS ONE* **6(1)**:e16393 DOI 10.1371/journal.pone.0016393.

Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV. 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1**:11 DOI 10.1186/2049-2618-1-11.

Strati F, Cavalieri D, Albanese D, De Felice C, Donati C, Hayek J, Jousson O, Leoncini S, Renzi D, Calabrò A, De Filippo C. 2017. New evidences on the altered gut microbiota in autism spectrum disorders. *Microbiome* **5**:24 DOI 10.1186/s40168-017-0242-1.

Suwantarat N, Adams LL, Romagnoli M, Carroll KC. 2015. Fatal case of Herbaspirillum seropedicae bacteremia secondary to pneumonia in an end-stage renal disease patient with multiple myeloma. *Diagnostic Microbiology and Infectious Disease* **82(4)**:331–333 DOI 10.1016/j.diagmicrobio.2015.04.011.

Topcuoglu BD, Lesniak NA, Ruffin MT IV, Wiens J, Schloss PD. 2020. A framework for effective application of machine learning to microbiome-based classification problems. *mBio* **11(3)**:e20 DOI 10.1128/mBio.00434-20.

Trivieri N, Pracella R, Cariglia MG, Panebianco C, Parrella P, Visioli A, Giani F, Soriano AA, Barile C, Canistro G, Latiano TP, Dimitri L, Bazzocchi F, Cassano D, Vescovi AL, Pazienza V, Binda E. 2020. BRAF(V600E) mutation impinges on gut microbial markers defining novel biomarkers for serrated colorectal cancer effective therapies. *Journal of Experimental & Clinical Cancer Research* **39**:285 DOI 10.1186/s13046-020-01801-w.

Tsay JJ, Wu BG, Badri MH, Clemente JC, Shen N, Meyn P, Li Y, Yie TA, Lhakhang T, Olsen E, Murthy V, Michaud G, Sulaiman I, Tsirigos A, Heguy A, Pass H, Weiden MD, Rom WN, Sterman DH, Bonneau R, Blaser MJ, Segal LN. 2018. Airway microbiota is associated with upregulation of the PI3K pathway in lung cancer. *American Journal of Respiratory and Critical Care Medicine* **198(9)**:1188–1198 DOI 10.1164/rccm.201710-2118OC.

Ursem C, Atreya CE, Van Loon K. 2018. Emerging treatment options for BRAF-mutant colorectal cancer. *Gastrointestinal Cancer: Targets and Therapy* **8**:13–23 DOI 10.2147/gictt.S125940.

van Passel MW, Kant R, Zoetendal EG, Plugge CM, Derrien M, Malfatti SA, Chain PS, Woyke T, Palva A, de Vos WM, Smidt H. 2011. The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLOS ONE* **6(3)**:e16876 DOI 10.1371/journal.pone.0016876.

Vimal J, Himal I, Kannan S. 2020. Role of microbial dysbiosis in carcinogenesis & cancer therapies. *Indian Journal of Medical Research* **152(6)**:553–561 DOI 10.4103/ijmr.IJMR_1026_18.

Vogt NM, Kerby RL, Dill-McFarland KA, Harding SJ, Merluzzi AP, Johnson SC, Carlsson CM, Asthana S, Zetterberg H, Blennow K, Bendlin BB, Rey FE. 2017. Gut microbiome alterations in Alzheimer's disease. *Scientific Reports* **7**:13537 DOI 10.1038/s41598-017-13601-y.

Waldmann E, Kammerlander AA, Gessl I, Penz D, Majcher B, Hinterberger A, Bretthauer M, Trauner MH, Ferlitsch M. 2021. Association of adenoma detection rate and adenoma characteristics with colorectal cancer mortality after screening colonoscopy. *Clinical Gastroenterology and Hepatology* **19(9)**:1890–1898 DOI 10.1016/j.cgh.2021.04.023.

Wang D, Cheng J, Zhang J, Zhou F, He X, Shi Y, Tao Y. 2021. The role of respiratory microbiota in lung cancer. *International Journal of Biological Sciences* **17(13)**:3646–3658 DOI 10.7150/ijbs.51376.

Wei Q, Zhang Q, Wu Y, Han S, Yin L, Zhang J, Gao Y, Shen H, Zhuang J, Chu J, Liu J, Wei Y. 2023. Analysis of bacterial diversity and community structure in gastric juice of patients with advanced gastric cancer. *Discover Oncology* **14**:7 DOI 10.1007/s12672-023-00612-7.

Wilmanski T, Rappaport N, Earls JC, Magis AT, Manor O, Lovejoy J, Omenn GS, Hood L, Gibbons SM, Price ND. 2019. Blood metabolome predicts gut microbiome alpha-diversity in humans. *Nature Biotechnology* **37(10)**:1217–1228 DOI 10.1038/s41587-019-0233-9.

Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine* **25(4)**:679–689 DOI 10.1038/s41591-019-0406-6.

Xu W, Wang T, Wang N, Zhang H, Zha Y, Ji L, Chu Y, Ning K. 2023. Artificial intelligence-enabled microbiome-based diagnosis models for a broad spectrum of cancer types. *Briefings in Bioinformatics* **24(3)**:57 DOI 10.1093/bib/bbad178.

Xuan C, Shamonki JM, Chung A, Dinome ML, Chung M, Sieling PA, Lee DJ. 2014. Microbial dysbiosis is associated with human breast cancer. *PLOS ONE* **9(1)**:e83744 DOI 10.1371/journal.pone.0083744.

Yang J, Moon HE, Park HW, McDowell A, Shin TS, Jee YK, Kym S, Paek SH, Kim YK. 2020. Brain tumor diagnostic model and dietary effect based on extracellular vesicle microbiome data in serum. *Experimental and Molecular Medicine* **52(9)**:1602–1613 DOI 10.1038/s12276-020-00501-x.

Zackular JP, Rogers MAM, Ruffin MT IV, Schloss PD. 2014. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research* **7(11)**:1112–1121 DOI 10.1158/1940-6207.Capr-14-0129.

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10(11)**:766 DOI 10.15252/msb.20145645.

Zhang W, Chen X, Wong KC. 2021. Noninvasive early diagnosis of intestinal diseases based on artificial intelligence in genomics and microbiome. *Journal of Gastroenterology and Hepatology* **36(4)**:823–831 DOI 10.1111/jgh.15500.

Zhang J, He Y, Xia L, Yi J, Wang Z, Zhao Y, Song X, Li J, Liu H, Liang X, Nie S, Liu L. 2022. Expansion of colorectal cancer biomarkers based on gut bacteria and viruses. *Cancers* **14(19)**:4662 DOI 10.3390/cancers14194662.

Zheng Y, Fang Z, Xue Y, Zhang J, Zhu J, Gao R, Yao S, Ye Y, Wang S, Lin C, Chen S, Huang H, Hu L, Jiang GN, Qin H, Zhang P, Chen J, Ji H. 2020. Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* **11(4)**:1030–1042 DOI 10.1080/19490976.2020.1737487.

Zhou CB, Pan SY, Jin P, Deng JW, Xue JH, Ma XY, Xie YH, Cao H, Liu Q, Xie WF, Zou XP, Sheng JQ, Wang BM, Wang H, Ren JL, Liu SD, Sun YW, Meng XJ, Zhao G, Chen JX, Cui Y, Wang PQ, Guo HM, Yang L, Chen X, Ding J, Yang XN, Wang XK, Qian AH, Hou LD, Wang Z, Chen YX, Fang JY. 2022. Fecal signatures of Streptococcus anginosus and Streptococcus constellatus for noninvasive screening and early warning of gastric cancer. *Gastroenterology* **162(7)**:1933–1947.e1918 DOI 10.1053/j.gastro.2022.02.015.