

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Machine Learning-based approaches for microbiome characterization in cancer

Marco André Carvalho Teixeira

Mestrado em Bioengenharia

Supervisor: Helder P. Oliveira, PhD

Co-supervisor: Tânia Pereira, PhD

Co-supervisor: Céu Figueiredo, PhD

July 3, 2023

Machine Learning-based approaches for microbiome characterization in cancer

Marco André Carvalho Teixeira

Mestrado em Bioengenharia

July 3, 2023

Resumo

O microbioma humano – a comunidade de microrganismos no nosso corpo – reflete o estado de saúde do hospedeiro. Alterações do microbioma intestinal e oral foram associadas a várias doenças, entre as quais o cancro. Recentemente foram identificados biomarcadores baseados na microbiota intestinal, oral e tumoral para o diagnóstico e caracterização de cancro. Uma vez que estes biomarcadores são altamente complexos, métodos de *Machine Learning* (ML) facilitam a sua descoberta. O material genético de microrganismos nos tecidos tumorais e noutros locais pode entrar no sangue, formando DNA microbiano circulante (mbDNA). Isto motivou o desenvolvimento de métodos de ML para diagnóstico e caracterização de cancro com base em mbDNA. Tais métodos podem constituir uma forma de rastreio não invasivo, permitindo um diagnóstico precoce e aumentando a probabilidade de sobrevivência dos pacientes. No entanto, a performance dos modelos atuais não é suficiente para uso clínico. Para além disto, o conhecimento sobre mbDNA é escasso, sobretudo sobre a sua origem e associação com o microbioma tumoral.

Esta dissertação explora métodos de ML para a identificação do tipo de cancro, com base em mbDNA e no microbioma tumoral, assim como a associação entre ambos. Foi efetuada uma revisão de literatura sobre ML para caracterização de cancro, tendo-se concluído que os métodos baseados em *Random Forest* levam frequentemente a melhores previsões que outros modelos; no entanto, a elevada variância e a validação inadequada dos modelos, bem como uma remoção insuficiente de variações técnicas levantam dúvidas sobre a generalização da maioria dos resultados. A performance de três modelos *um contra o resto* para a identificação do tipo de cancro foi comparada, tendo os melhores resultados sido obtidos com *Gradient Boosted Trees*. Um classificador multiclasse demonstrou uma performance semelhante a múltiplos classificadores *um contra o resto*, sendo capaz de previsões não ambíguas e exigindo um menor custo computacional durante o treino. Recorrendo a modelos treinados com ambos os tipos de dados, o potencial do mbDNA para a identificação do tipo de cancro mostrou-se ligeiramente inferior ao do microbioma tumoral.

Numa investigação da contribuição do microbioma do tumor para o mbDNA, os perfis de microbioma tumoral e derivados do sangue revelaram-se semelhantes, sugerindo uma elevada mobilidade dos microrganismos entre o tumor e o sangue. Poucos géneros possuíam abundâncias diferenciais entre os dois locais, e a maioria destes estavam sub-representados no mbDNA. Foram desenvolvidos cinco modelos de ML para estimar as abundâncias de microrganismos no tumor, com base no mbDNA (Regressão Linear e quatro Redes Neurais), de acordo com diferentes pressupostos sobre a origem do mbDNA. A maioria resultou em previsões melhores que aleatórias, com a Regressão Linear a atingir os melhores resultados, seguida de um Regressor Semi-paramétrico. Estes resultados sugerem que a maioria dos géneros atravessam a interface entre o sangue e o tumor e que as contribuições do microbioma tumoral e de outros locais do corpo para o mbDNA não podem ser separadas. Este trabalho foi o primeiro a usar ML para investigar a origem do mbDNA e a sua associação com o microbioma tumoral, abrindo caminho para outros métodos. Uma análise focada num único tipo de cancro e a medição de outras variáveis, como o microbioma intestinal, poderá ajudar a esclarecer estas questões.

Abstract

The human microbiome – the community of microorganisms living in the human body – is associated with the health of its host. Changes in the gut and oral microbiomes have been linked to many diseases, including cancer. Recent advances have uncovered microbial signatures for diagnosis and characterization of several types of cancer, based on the gut, oral, and tumor microbiomes. Machine Learning (ML) methods are useful when searching for such signatures, as these are highly complex. The genetic material of microorganisms in the gut, tumor, and other reservoirs, may leak into the bloodstream, forming microbial circulating DNA (mbDNA). This has motivated the development of ML-based methods for cancer diagnosis and cancer type identification using mbDNA, which have shown promising performance. Such methods could be a non-invasive screening approach for cancer, enabling early diagnosis and improving the chance of survival. However, the performance of current ML models for cancer diagnosis using mbDNA is not sufficient for clinical usage. Furthermore, little is known about mbDNA, its origins, and association with the tumor microbiome.

This work focused on ML methods for cancer type identification using blood-derived and tumor microbiome data, as well as the association between the tumor microbiome and mbDNA. Firstly, a review of ML approaches for cancer characterization was conducted; methods based on Random Forest were found to often outperform other models; however, high model variance, the inadequate removal of technical variations and insufficient model validation raise doubts on the generalization of most findings. The performance of three one-versus-rest state-of-the-art ML models for cancer type prediction from mbDNA was compared, with Gradient Boosted Trees achieving the best results. Based on this model, a multiclass classifier was developed; this model had a comparable performance to multiple one-versus-rest classifiers, while being non-ambiguous and less costly to train. Through models trained with both types of data, the potential of mbDNA for cancer type identification proved to be slightly inferior to that of the tumor microbiome.

The contribution of the tumor microbiome to mbDNA was also investigated. The blood-derived and tumor microbiome profiles of the same patients were found to be similar, suggesting a high mobility of microorganisms between the tumor tissues and the bloodstream. Few genera had differential abundances between the two sites, and most of the ones that did were underrepresented in the blood-derived profiles. Five ML regressors – Linear Regression and four models based on Artificial Neural Networks – estimating the microbial abundances in the tumor from mbDNA were evaluated; all were developed under different assumptions regarding the origins of mbDNA. Most achieved better than random predictions, with Linear Regression providing the best estimates, followed by a Semiparametric Regressor. These results suggest that most genera cross the blood-tumor interface and that the contributions of the tumor and gut (or other reservoirs) microbiomes to mbDNA cannot be isolated. This work was the first to employ ML to investigate the origins of mbDNA and its association with the tumor microbiome and paves the way for other analyses. Further work on specific types of cancer and the assessment of other variables, such as the composition of the gut microbiome, is expected to help answer these questions.

Agradecimentos

Ao Francisco Silva, por toda a ajuda ao longo da dissertação, pelas sugestões geniais, mas sobretudo por me acalmar e relembrar do valor deste trabalho várias vezes ao longo dos últimos meses.

Aos meus orientadores, Hélder Oliveira e Tânia Pereira, por terem confiado em mim, por me terem dado liberdade para explorar as minhas próprias ideias e por todas as contribuições para este trabalho. Embora ainda tenha muito para aprender, ensinaram-me como fazer investigação, e não poderia desejar melhores professores.

À Céu Figueiredo e ao Rui Ferreira, pela inesgotável paciência para explicar-me o que fazem as bactérias no nosso corpo, e como as podemos estudar – este trabalho teria sido impossível sem a vossa contribuição. Espero que esta dissertação ajude a vossa investigação no futuro.

A todas os restantes membros do VCMi por me terem recebido de braços abertos.

A todos os amigos com quem desabafei sobre as frustrações deste trabalho, sobretudo ao Rodrigo, ao Gonçalo, à Mafalda, à Raquel, ao Zé Pedro e ao Pedro por, para além de ouvirem, tornarem estes dias mais divertidos.

Aos meus pais e avós, não só por me apoiarem durante este último ano, mas sobretudo por me darem as oportunidades que nunca tiveram. Espero ter mostrado, e continuar a mostrar, que o vosso esforço valeu a pena.

Marco

Funding

This work was supported by Fundação para a Ciência e Tecnologia by national funds through Project PTDC/BTM-TEC/0367/2021.

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	2
1.3	Objectives and main contributions	3
1.4	Structure of the dissertation	4
2	Literature review	5
2.1	Obtaining microbiome data	7
2.2	Feature transformations for microbiome data	7
2.3	Feature selection methods	9
2.4	Machine Learning models for microbiome data	13
2.4.1	Logistic Regression	17
2.4.2	Support Vector Machines	18
2.4.3	Tree-based classifiers	20
2.4.4	Artificial Neural Networks	24
2.5	Current limitations	30
2.6	Summary	33
3	The human microbiome as a biomarker for cancer diagnosis	35
3.1	Dataset	36
3.2	A multiclass classifier for cancer type identification using blood-derived microbiome profiles	38
3.2.1	Materials and methods	38
3.2.2	Metrics to evaluate classifier performance	40
3.2.3	Results and discussion	41
3.3	Tumor and blood microbial signatures for cancer type identification	44
3.3.1	Materials and methods	44
3.3.2	Results and discussion	44
3.4	Summary	52
4	Uncovering the association between the tumor and blood microbiomes	53
4.1	Dataset	54
4.2	Pre-processing microbiome abundance data	54
4.2.1	Removing batch effects	55
4.2.2	Normalizing microbiome profiles	57
4.3	Assessing the similarity between blood-derived and tumor microbiome profiles	58
4.3.1	Materials and methods	58
4.3.2	Results and discussion	58

4.4	Predicting the tumor microbiome from blood-derived profiles	62
4.4.1	Machine Learning regressors	62
4.4.2	Hyperparameter search and global parameters	70
4.4.3	Evaluation metrics for regression	70
4.4.4	Results and discussion	72
4.5	Summary	78
5	Conclusions and future work	79
A	Publications found in the literature survey	83
	References	97

List of Figures

2.1	Diagram of the number of articles found and included in the literature review. . .	6
2.2	Pipeline for ML-based cancer identification from microbial abundance profiles. .	8
2.3	Overview of feature selection with LEfSe for cancer diagnosis from genera abundances.	12
2.4	Margins and decision boundary for an SVM classifier.	19
2.5	Example of a Decision Tree classifier for cancer diagnosis using two input features.	21
2.6	Diagram of an MLP for cancer diagnosis from microbial abundance profiles. . . .	26
2.7	Diagram of the DeepEn-Phy model, based on MLPs.	27
2.8	Diagram of an Autoencoder for microbial abundance profiles.	29
2.9	Overview of needed improvements in microbiome-based identification of cancer.	31
3.1	Boxplots of test classification metrics per class and training/test split of the dataset for multiclass classifiers.	43
3.2	ROC curves per target class for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles.	47
3.3	Precision-Recall curves per target class for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles.	48
3.4	Test confusion matrix for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles.	50
4.1	UMAP and t-SNE 2D projections of absolute abundance profiles before batch effect removal, colored according to sequencing center.	56
4.2	UMAP and t-SNE 2D projections of absolute abundance profiles after batch effect removal, colored according to sequencing center and sample type.	57
4.3	Proposed models for the mobility of microorganisms between the bloodstream and tumor tissues.	61
4.4	Histograms of the Pearson's r of the abundance of each genus between matched blood-derived and tumor microbiome profiles.	62
4.5	Volcano plot of blood-derived and tumor microbial abundances.	63
4.6	Architecture of the MLP regressor.	66
4.7	Architecture of the SPR.	67
4.8	Architecture of the PNN model.	68
4.9	Architecture of the TopoPNN model.	69
4.10	10 genera of highest and lowest bias parameters in the Linear Regression model trained with absolute microbial abundances and <i>likely contaminants removed</i> . . .	74

List of Tables

2.1	Methods for microbiome feature selection.	10
2.2	Examples of applications of ML methods for cancer characterization from microbiome data.	15
2.3	Advantages and disadvantages of ML methods for cancer characterization from microbiome data.	16
3.1	Number of matched blood and tumor biopsy sample pairs per cancer type in the dataset.	37
3.2	Classifier hyperparameters and search spaces.	39
3.3	Evaluation metrics for all evaluated models, in a multiclass and OvR setting. . . .	41
3.4	Evaluation metrics for GBT classifiers with and without oversampling, and using either blood-derived or tumor biopsy microbial abundance profiles.	45
3.5	Test AUPRC per target class for GBT classifiers with SMOTE data augmentation using either tumor biopsy microbiome profiles or blood-derived abundances. . . .	49
4.1	Number of genera in both tumor and blood microbiome profiles with each level of decontamination.	54
4.2	Average Pearson's r between matched and non-matched blood-derived and tumor microbiome profiles, grouped by cancer type, together with p -values from Mann-Whitney U tests.	59
4.3	Regressor hyperparameters and search spaces.	71
4.4	Performance metrics for ML regressors using relative abundance profiles.	72
4.5	Performance metrics for ML regressors using absolute abundance profiles, without decontamination.	75
4.6	Performance metrics for ML regressors using absolute abundance profiles, with <i>likely contaminants removed</i>	75
4.7	Performance metrics for ML regressors using absolute abundance profiles, with the <i>most stringent</i> decontamination.	76
4.8	Test performance metrics per cancer type for the Linear Regression model trained with absolute abundance profiles and <i>likely contaminants removed</i>	77
A.1	Publications found in the literature survey.	84

Abbreviations

AdaBoost	Adaptative Boosting
ANOVA	One-way analysis of variance
ASV	Amplicon sequence variant
AUC	Area under the Receiver Operating Characteristic curve
AUPRC	Area under the Precision-Recall curve
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CART	Classification and Regression Trees
CESC	Cervical Squamous Cell Carcinoma And Endocervical Adenocarcinoma
CNN	Convolutional Neural Network
COAD	Colon Adenocarcinoma
CV	Cross-validation
DLBC	Lymphoid Neoplasm Diffuse Large B-Cell Lymphoma
DNA	Deoxyribonucleic acid
EBM	Explainable Boosting Machines
ESCA	Esophageal Carcinoma
FN	False negatives
FOBT	Fecal occult blood tests
FP	False positives
FPR	False positive rate
GBM	Glioblastoma Multiforme
GBT	Gradient Boosted Trees
GRNN	General Regression Neural Networks
HNSC	Head And Neck Squamous Cell Carcinoma
IBD	Inflammatory bowel disease
IG	Information Gain
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LASSO	Least Absolute Shrinkage and Selection
LDA	Linear Discriminant Analysis
LEfSe	Linear Discriminant Analysis Effect Size
LGG	Brain Lower Grade Glioma

LIHC	Liver Hepatocellular Carcinoma
logCPM	Log-counts per million
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
mbDNA	Microbial circulating DNA
ML	Machine Learning
MLP	Multilayer Perceptron
mRMR	Minimum Redundancy – Maximum Relevance
MSE	Mean squared error
MSE _{Norm}	Normalized mean squared error
NN	Artificial Neural Network
OTU	Operational taxonomic unit
OV	Ovarian Serous Cystadenocarcinoma
OvR	One-versus-rest
PCA	Principal Component Analysis
PNN	Parallel Neural Network
PRAD	Prostate Adenocarcinoma
r^{Norm}	Normalized Pearson's r
RBF	Radial basis function
READ	Rectum Adenocarcinoma
ReLU	Rectified Linear Unit
RFE	Recursive Feature Elimination
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
SMOTE	Synthetic Minority Oversampling Technique
SPR	Semiparametric Regressor
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine with Recursive Feature Elimination
TCGA	The Cancer Genome Atlas
THCA	Thyroid Carcinoma
TMM	Trim Mean of M Values
TN	True negatives
TopoPNN	Topological Parallel Neural Network
TP	True positives
t-SNE	t-Distributed Stochastic Neighbor Embedding
UCEC	Uterine Corpus Endometrial Carcinoma
UMAP	Uniform Manifold Approximation and Projection
UVM	Uveal Melanoma
WGS	Whole genome sequencing
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Context

Besides our own cells, the human body contains microorganisms living inside it. The community of microorganisms, including bacteria, virus, and fungi, that inhabit the human body constitutes the human microbiome [1]. Recent advances in DNA and RNA sequencing technologies [2] and *omics* analysis [3, 4] have enabled a better understanding of the human microbiome and its association with host health. The composition of the microbiome is different between body sites and individuals [5], and is influenced by the host characteristics, diet, and intake of antibiotics, among other factors [6]. Diseases may also impact the composition and metabolic functions of the human microbiome: changes in the gut microbial content have been linked to obesity [7, 8], colorectal cancer [9], inflammatory bowel disease (IBD) [7], and anxiety and depression [10], while cardiovascular disease [11], diabetes [12], and pancreatic, colorectal, and oral cancer [13, 14, 15, 16] were all found to impact the oral microbiome. This correlation between the characteristics of the human microbiome and diseased states has led to several proposed methods for diagnosis based on microbial biomarkers [17, 18, 19, 20] and therapies based on microbiome modulation [21, 22].

Cancer is one of the diseases for which its links with the human microbiome have been more widely studied. Cancer is estimated to have been responsible for 10M deaths worldwide in 2020 alone, making it the second most common cause of death [23, 24]. Despite encompassing a heterogeneous set of diseases, guided by complex interactions between the genotype of the host and the environment, cancer and several of its characteristics have been found to be associated with the microbiome composition and metabolic function of several locations in the human body. As examples, changes in the gut microbiome were linked to colorectal [9, 25, 26], lung [27, 28], cervical [29], and breast cancer [30], as well as melanoma [31]. Furthermore, the microbiome of cancer patients was found to be predictive of their response to chemotherapy and cancer immune checkpoint therapy [32, 33, 34]. Most of the research on the association between cancer and the human microbiome has focused on colorectal cancer and the gut [35, 36, 37, 38]; however, recent publications have uncovered associations between the oral, tumor, and gut microbiomes and cancer presence, prognosis, and response to treatment in other cancer types [18, 39, 40, 41, 42, 43, 44].

Efforts to derive microbiome profiles from The Cancer Genome Atlas (TCGA) [45] database have also uncovered microbiome-based signatures statistically associated with multiple types of cancer [44, 46].

Tumor tissues also host a community of microorganisms within themselves, which was found to be associated with cancer type [47, 41, 44, 48, 49] and survival times [3, 39, 42]. The tumor microenvironment enables the existence of a microbiome as it is a nutritious, highly vascularized space; furthermore, cancer cells evade recognition by immune cells, thus forming a safe environment for microorganisms [50, 51]. Genetic material from microorganisms in the tumor tissue may also enter the bloodstream, thus forming part of the microbial circulating DNA (mbDNA) [52]. However, mbDNA may not be inside living, viable, microorganisms [53, 44], and part of it may also originate from other reservoirs, such as the gut and mouth [54, 55].

1.2 Motivation

The microbial signatures associated with cancer and cancer characteristics are often complex and involve a large number of species. Therefore, Machine Learning (ML) approaches are useful to uncover the links between cancer and the human microbiome [56]. Machine Learning models are algorithms that make predictions given a dataset of learnable examples and can identify complex patterns in high-dimensional data more easily than humans [57]. Many publications on the links between cancer and the human microbiome include ML-based analysis [58, 39, 59, 44, 36, 60].

As early detection of cancer improves the chances of survival and the quality of life of patients, methods for cancer screening are highly sought after [61]. Screenings are a reliable way to diagnose cancers and precursor lesions in an apparently healthy population using widely available tests [61]. These are usually specific to a cancer type and include immunochemical fecal occult blood tests (FOBT) for colorectal cancer [62] and mammography for breast cancer [63]. More recently, blood-based tests have been proposed as a non-invasive screening approach [64, 65]. All these blood tests seek biomarkers indicative of cancer, potentially allowing the screening of multiple cancer types using the same test [66]. Due to the association between the human microbiome and cancer, complex mbDNA biomarkers identified by ML models have been proposed as methods for cancer screening [67, 44, 68, 69]. Such blood-derived biomarkers could lead to non-invasive pan-cancer screening approaches [44].

Despite the promising results of mbDNA biomarkers found using ML models for cancer type identification and cancer diagnosis, these approaches require further improvement in performance before widespread clinical use. For instance, in [44], the ML models using mbDNA for cancer type prediction achieved an average Area Under the Precision-Recall Curve of only 77.3% across all classes. Furthermore, it is unknown if ML models using mbDNA are capable of better predictions than those based on the tumor or gut microbiomes, and vice-versa. This increase in performance may be achieved by developing ML methods better suited for microbiome data and taking the underlying biological and microbial dynamics into account. Likewise, all studies applying ML models for cancer type identification and diagnosis have so far used one model per cancer type;

therefore, it is also unknown if a single ML classifier capable of identifying multiple cancer types at once could achieve comparable performance to models trained for the diagnosis of a specific type of cancer. Such a model could enable a pan-cancer screening method based on mbDNA.

Besides the need for an improvement in the performance of ML models using mbDNA for cancer diagnosis, little is known about and how the blood microbiome in diseased states relates to the microbiome in other body sites, as well as the contribution of the microbiome in each of these sites to mbDNA: as previously mentioned, mbDNA may originate in reservoirs other than the tumor, most likely the gut or the mouth [54, 55]. There is also no clear understanding of how microorganisms enter the bloodstream in healthy individuals, let alone under diseased states [53]. This is also true for the tumor microbiome, despite hypotheses on how microorganisms first reach the tumor microenvironment [50, 51]. Therefore, it is unknown whether the tumor and blood-derived microbiome profiles are similar or not, and if the microbial signatures uncovered by ML models using either of these are identical. Understanding the origins and similarities between the tumor and blood-derived microbiome profiles is essential to validating mbDNA as a potential biomarker for cancer diagnosis and it would help assess if the mbDNA signatures associated with cancer are caused by biologically relevant phenomena or a consequence of technical artifacts.

1.3 Objectives and main contributions

This dissertation focuses on ML-based methods for cancer type identification using mbDNA and the tumor microbiome. It attempts to answer important questions surrounding the use of mbDNA for cancer type identification and diagnosis, paving the way for a non-invasive blood screening method for multiple cancer types. Firstly, it aims at assessing the potential of mbDNA for cancer type diagnosis, comparing it to the discriminatory power of the tumor microbiome. Besides this, this dissertation also seeks to develop an ML model for the identification of multiple cancer types at once. Another large goal of this dissertation revolves around better understanding the blood microbiome and the origins of mbDNA in cancer patients. It attempts to assess the similarities and differences between the tumor and blood-derived microbiome profiles of the same patient, and present novel insights into how microorganisms enter the bloodstream and the tumor tissues, all the while relying on ML models to do so.

With this in mind, the main contributions of this thesis are:

- A review of ML methods for cancer characterization based on the microbiome, highlighting the most widely used and promising ML approaches from sample collection to model validation, along with current limitations and future perspectives. This resulted in an article submitted to the journal *IEEE Reviews in Biomedical Engineering*;
- The development of a multiclass ML classifier for the identification of multiple types of cancer;
- A comparison between state-of-the-art ML models for cancer type identification with blood-derived microbiome data;

- The assessment of the potential of mbDNA biomarkers for cancer type identification and the comparison of ML using mbDNA against models using data from the tumor microbiome;
- An analysis of the similarity between blood-derived and tumor microbiome abundance profiles;
- The development of ML models to predict the tumor microbiome composition using blood-derived microbial abundance data;

1.4 Structure of the dissertation

The structure of this dissertation is as follows: in Chapter 2, state-of-the-art ML approaches for cancer identification and characterization through microbiome data are reviewed. Methods for data transformation and feature selection are also discussed. Then, in Chapter 3, the potential of mbDNA for cancer diagnosis is assessed and compared to that of the tumor microbiome. Furthermore, a multiclass ML classifier for multi-cancer identification is developed and evaluated. Chapter 4 focuses on the differences and similarities between the tumor microbiome and the blood-derived microbiome data, as well as the origins of mbDNA. This analysis included the development of ML regressors to estimate the tumor microbial abundances using the blood-derived profiles. Chapters 3 and 4 begin with a small introduction, which expands the concepts approached in Chapter 1 and highlights the goals of the experiments described in that Chapter. Then, the dataset used in those experiments is discussed, followed by a description of the methods and results of each experiment, along with a discussion about the results obtained. The Chapters end with a summary of their main conclusions. Finally, the main conclusions of this dissertation are summarized in Chapter 5, along with future perspectives in ML-based cancer characterization using microbiome data and how to better understand the origins of mbDNA.

Chapter 2

Literature review

As previously discussed, the changes in the human microbiome mediated by cancer are highly complex and discovering microbial biomarkers for cancer diagnosis or characterization may benefit from the capabilities of ML models to recognize patterns. This Chapter presents and discusses the most widely used and promising novel ML methods for cancer characterization using microbiome data. It is based on a review article entitled *A Review of Machine Learning Methods for Cancer Characterization from Microbiome Data*, written as part of this dissertation and submitted to *IEEE Reviews in Biomedical Engineering*. In Section 2.1, it is briefly described how to obtain data describing the human microbiome. Then, transformations for microbiome data to improve the performance of ML models are discussed in Section 2.2, followed by a review of dimensionality reduction approaches most suited for the ML-based identification of microbial biomarkers in Section 2.3. After this, in Section 2.4, ML models are discussed, and their strengths and weaknesses are enumerated. Finally, Section 2.5 addresses current limitations on the use of ML models for cancer characterization with microbiome data and how these can be overcome.

In this literature review, articles published in PubMed, Scopus, and IEEEExplore between January 1st 2015 and May 1st 2023 are included. These articles were searched with the query *(microbiota OR microbiome OR bacteria) AND machine learning AND (tumor OR neoplasia OR cancer)* and those with relevant abstracts and manuscripts were kept. Full-text articles and abstracts were considered relevant if they mentioned the application of an ML model for cancer diagnosis or characterization (including, for example, cancer stage, response to treatment, or survival time). Studies that did not use ML for cancer characterization but proposed novel methods for microbiome data transformation and feature selection were also included. Reviews and perspective articles were excluded. The number of publications resulting from this search are shown in Figure 2.1, while Appendix A holds all the full-text articles analyzed during this review, along with information on the ML models, learning tasks, feature selection methods, input features, and sample types used in each.

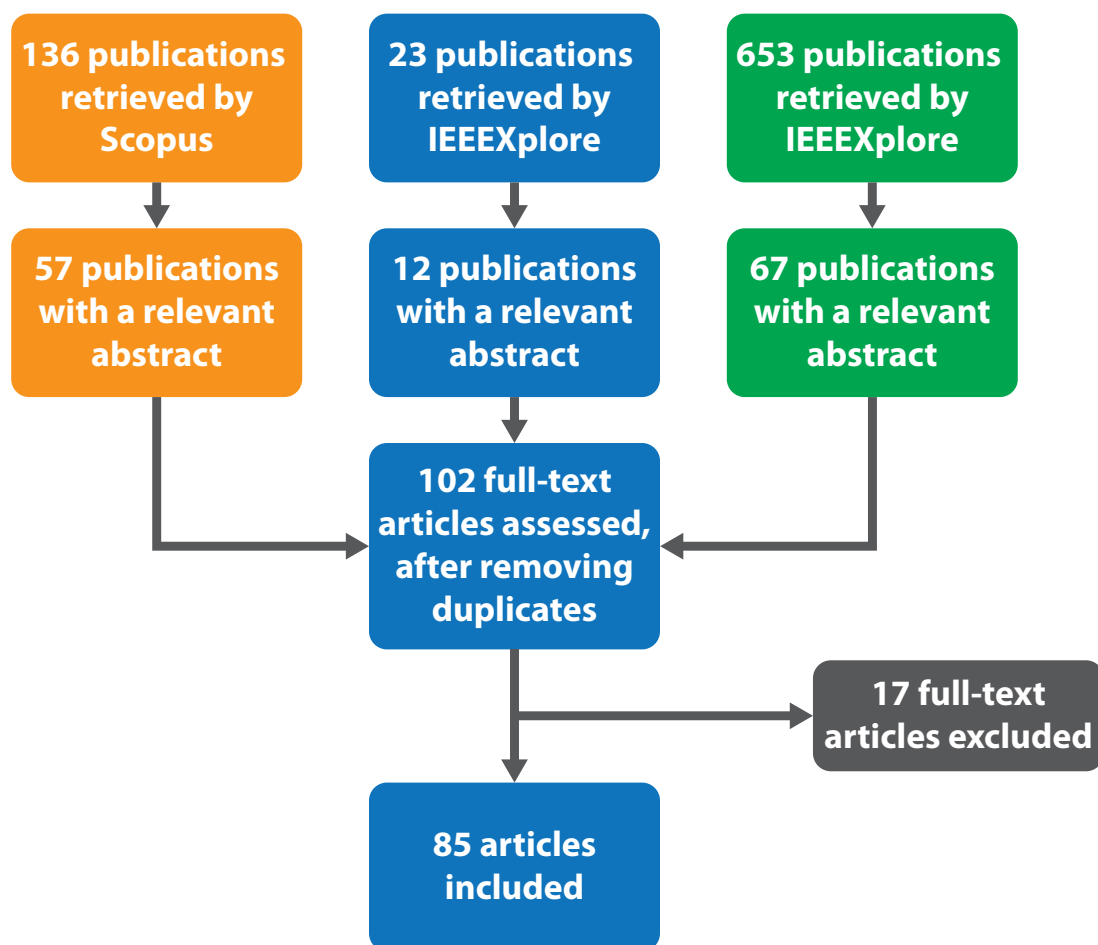


Figure 2.1: Diagram of the number of articles found and included in this literature review.

2.1 Obtaining microbiome data

The analysis of the human microbiome entails the isolation of DNA or RNA from a sample. These samples may be fecal [56], mucosal swabs [70, 49], biopsies [18, 44], blood [44, 68], urine [71], or bone marrow [68]. Care must be taken to avoid the contamination of samples with genetic material from microorganisms in the collection environment and during the isolation of genetic material, and cross-contamination during sample processing. *In silico* approaches for decontamination after sample processing and data collection also help mitigate the influence of contaminants [44].

The microbiome is most often characterized according to the abundance of each taxon in a sample. Sequenced fragments of genetic material, called reads, are assigned to their taxonomic units of origin by comparing their genetic sequence with those in genetic databases [72]. A taxonomic unit is a set of related organisms, whose degree of relatedness depends on the type of unit chosen. As an example, when characterizing the microbiome through species-level abundances, the data for one sample is a vector of non-negative integers; each element holds the number of reads assigned to a given species [73]. This data is often converted to relative abundances, expressing the proportion of organisms of a given taxonomic unit in a sample [74]. The most commonly used taxonomic levels when deriving abundance profiles are the genus level [44, 68, 69], amplicon sequence variants (ASVs), and operational taxonomic units (OTUs), the latter of which do not strictly correspond to traditional phylogenies [41, 75]. The ASV and OTU-level profiles are less computationally costly to obtain than traditional taxonomic classifications [76]; furthermore, current sequencing technologies make a species-level characterization of the microbiome difficult, leading to a preference for genus-level profiles [77]. These abundances are then further processed to remove the influence of technical variations and batch effects [44, 78, 79] and used as input features for an ML model for microbial signature detection. Figure 2.2 shows the pipeline for microbiome-based ML model development, from sample collection to model validation. Although less commonly used, the abundances of multiple taxonomic levels can also be combined as input features to an ML model [80, 81], or be incorporated into phylogenetic trees [82].

Besides abundance profiles, the microbiome can be characterized via other variables. For instance, transcriptomics and metabolomics may be used to derive functional profiles, and these functional profiles may also be obtained with sequencing data as well [83, 39, 27]. These functional profiles quantify the relative abundance of groups of organisms, genes, and metabolites of similar metabolic functions [77, 83]. Other types of microbiome data include single nucleotide polymorphisms in bacterial DNA [25] and gene expression [42].

2.2 Feature transformations for microbiome data

The characteristics of microbiome data depend on the biological variable measured; however, because most ML models using microbiome data in the literature use microbial abundances as input features, this type of data will be discussed in greater detail. The raw read count profiles follow a zero-inflated negative binomial distribution, with many taxa having no reads assigned to it [84, 79],

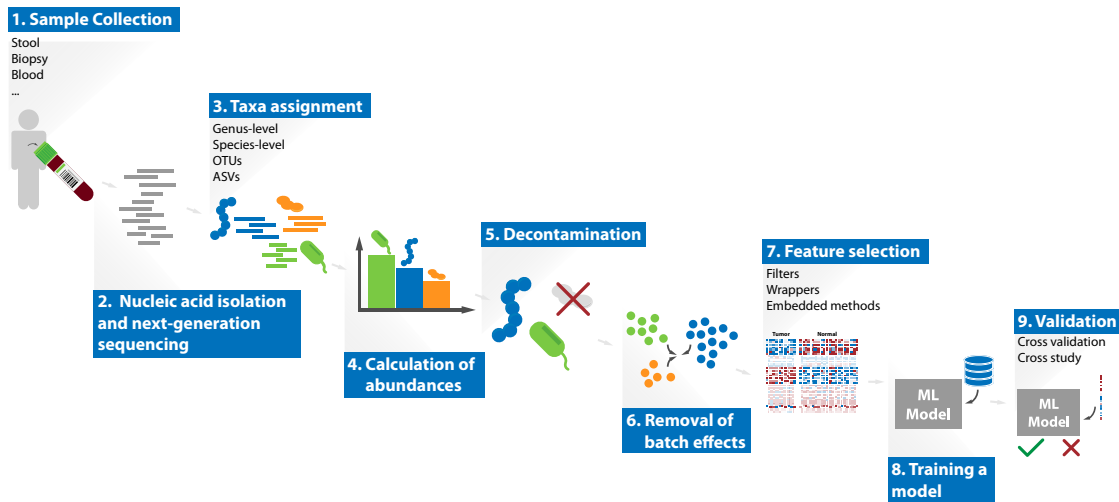


Figure 2.2: Pipeline for ML-based cancer identification from microbial abundance profiles. The nucleic acid in the collected samples is sequenced and the resulting reads are assigned taxonomic classifications. Likely contaminants and batch-effects are removed. Feature selection methods, discussed in Section 2.3, can be used to select the most relevant features, which are used to train an ML model. This model should be evaluated using an independent test set and cross-study validations.

which hinders the performance of ML and statistical models, leading to overdispersion [85]. However, these profiles are often transformed into relative abundances and processed to remove the variability linked to technical phenomena before being used to train ML models, as discussed in the previous Section. This transformation also usually includes a log transformation, approximating the abundance profiles to a Normal distribution and reducing their skewness [86, 74, 44]. Even so, the log-transformed relative abundance microbiome profiles still have an abundance of zero-valued entries and are heterogeneous [84]. Therefore, multiple feature transformation approaches to alleviate these issues have been proposed.

Data transformations of microbiome data decrease the influence of outliers when training ML models, improving their performance [87]. Most studies on the association between cancer and the microbiome which employ ML models transform the relative abundances of each taxon across all samples into zero mean and unit variance [88, 80, 44]; sometimes, although rarely, these values are discretized [83]. However, other alternatives have also been used, such as the cube-root normalization proposed in [89] and re-scaling taxa distributions to a unitary scale [81]. Mulenga et al. [90] also proposed applying two normalization methods to microbiome profiles and concatenating the resulting profiles with the original. This approach improved the classification performance of a state-of-the-art ML model when combined with data augmentation using a Variational Autoencoder [90].

There are not enough publications comparing different normalization approaches for microbiome data and their impact on ML models to confidently conclude which normalization method leads to a better performance. Despite the claims that some methods can better reflect the true

microbiome composition [84] or improve the accuracy of models [89], the most suitable normalization method is likely dependent on the characteristics of the original dataset and the type of data used: some transformations, such as those resulting in a homogeneous variance across features, are often better at decreasing the influence of outliers, while approaches that preserve the variance of taxa abundances, as the cube-root normalization, may be more effective in the absence of such outliers [90, 87].

2.3 Feature selection methods

Microbiome profiles contain information on a large number of taxa, leading to data with a number of features much larger than the number of available samples. This high dimensionality of the data negatively impacts the performance of ML algorithms, as it gives rise to many possible relationships between taxa abundances, most of them spurious [91]. This leads to models with high variance that overfit the training data [91]. Furthermore, high-dimensional data increases the computational cost of ML model training [92]. Therefore, methods to decrease the dimensionality of the data prior to training ML models are often used when studying the association between cancer and the microbiome.

When developing ML models to identify or characterize cancer based on the microbiome, researchers are mainly interested in finding microbial biomarkers, such as species whose increased abundance is associated with a particular type of cancer. Thus, it is important to have the ability to analyze the rules applied by ML models to derive predictions. Because of this, methods for dimensionality reduction that transform the original feature space, such as Principal Component Analysis (PCA) [98], are very rarely used. Rather, approaches that select a subset of the features in the dataset to be used as input for an ML model are preferred. Table 2.1 contains an overview of the methods discussed in this Section, along with their advantages and disadvantages. These feature selection methods can be divided into filters, wrappers, and embedded approaches [92], of which filters are most commonly used with microbiome data.

Filter feature selection approaches are independent of the ML model used, and rank features according to a criterion, selecting the highest-ranked before passing it to the model [93]. Most often, filters independently assess the predictive capability of each feature, disregarding the patterns formed by a combination of multiple features. Furthermore, because filters are independent of the downstream ML model, they do not take the learning strategies employed by these models into account. Thus, the features selected by filter methods may not be the optimal subset for the ML model used.

Some of the most used and simpler filter feature selectors use statistical tests to rank features according to their relevance in classification tasks. These tests identify features with different values across classes (e.g., cancer types) and include the Student's t test for independent samples [91] and its non-parametric version, the Mann-Whitney U test [99]. Both tests were used in [75] to identify epithelial ovarian cancer from peritoneal fluid microbiota. The Student's t -test and the Mann-Whitney U test are only capable of comparing the values of features between two classes;

Table 2.1: Methods for microbiome feature selection.

Type	Method	Advantages	Disadvantages	Examples
Filters	Statistical tests	Fast computation Selects relevant features	Selects redundant features Assumes feature independence Model agnostic	[93, 83]
	mRMR	Avoids redundancy Selects relevant features	Assumes feature independence Model agnostic	[94, 47, 95]
	Surrogate model	Selects relevant features	Features may not be the best for the predictor Limited choice of surrogate models	[95]
	LEfSe	Takes biological relevance into account	Selects redundant features Assumes feature independence Model agnostic Requires biological covariates	[69]
Wrappers	RFE	Takes estimator performance into account	High computational cost Limited choice of ML estimators Greedy search	[96]
Embedded	LASSO	Fast computation Model-specific	Limited choice of ML estimators	[97, 86, 75, 80]

in multiclass classification tasks, when multiple classes ought to be compared, one-way analysis of variance (ANOVA) tests, which uses the F-distribution to compare the mean value of a feature between all classes, or its non-parametric alternative the Kruskal-Wallis test [100], can be used [101]. For instance, in [83], ANOVA was used to select features for colorectal cancer prediction using the gut microbiome. These statistical tests are already frequently used to identify differentially abundant microorganisms between cohorts, regardless of any ML-based analysis [102, 103].

Statistical test-based feature selection has a tendency to result in highly correlated feature sets: if multiple correlated features have a large discriminative power, they will all be selected despite their redundancy [104]. One of the most used filters for feature selection in microbiome data, Minimum Redundancy - Maximum Relevance (mRMR) [105], addresses this issue [47, 95, 94, 28]. This algorithm was developed for microarray gene expression data and greedily searches for features that are both informative and uncorrelated with those previously selected [105]. In mRMR, the relevance of features is quantified with information gain (IG) or an F-test between a feature and the target classes, for discrete and continuous variables, respectively. The first selected feature is that with the highest relevance. For each subsequent iteration, the sum of Pearson's correlation coefficients or IG between a feature and all those previously selected is used to estimate the redundancy of the remaining features; that with the lowest redundancy is added to the set of selected features [105].

Another popular feature selection method for microbiome data in classification tasks is Linear Discriminant Analysis Effect Size (LEfSe) [106]. This method not only considers the discriminatory power of a feature, through a Kruskal-Wallis test [100], but it also assesses its biological relevance by testing if this discriminatory capability holds across cohorts of different characteristics [106]. To do this, LEfSe divides the dataset into sub-classes reflecting biological characteristics of samples (e.g. the sex of a patient, as done in [69], or other clinical covariates in [41]). Then, for each feature with significant differences between target classes, its values between samples of the same sub-class but different target classes are compared with a Wilcoxon signed-rank test [107], as seen in Figure 2.3. The feature is deemed biologically irrelevant and excluded from the dataset if any comparison results in a p value outside a predefined confidence interval, or if there is a difference in sign among the comparisons. Lastly, Linear Discriminant Analysis (LDA) is used to estimate the effect size of the resulting features and rank them [106]. The effect size is quantified by averaging the difference in class means against the difference in class means along the first linear discriminant axis [106]. Published applications of LEfSe for feature selection include the identification of colorectal polyps using the gut microbiome [70], oral squamous cell carcinoma with the oral microbiome [108], and breast cancer through mbDNA [67], as well as predicting the recurrence and metastasis of lung cancer based on the tumor microbiome and host gene expression [109].

Surrogate ML models can also be used as feature selection filters. The most commonly used surrogate models used for this end are those based on Decision Trees, such as Random Forests and Boosted Trees, which can inexpensively attribute an importance score (such as Gini importance) to the input features [110, 111, 97]. The best-ranked features are then selected as input for a more

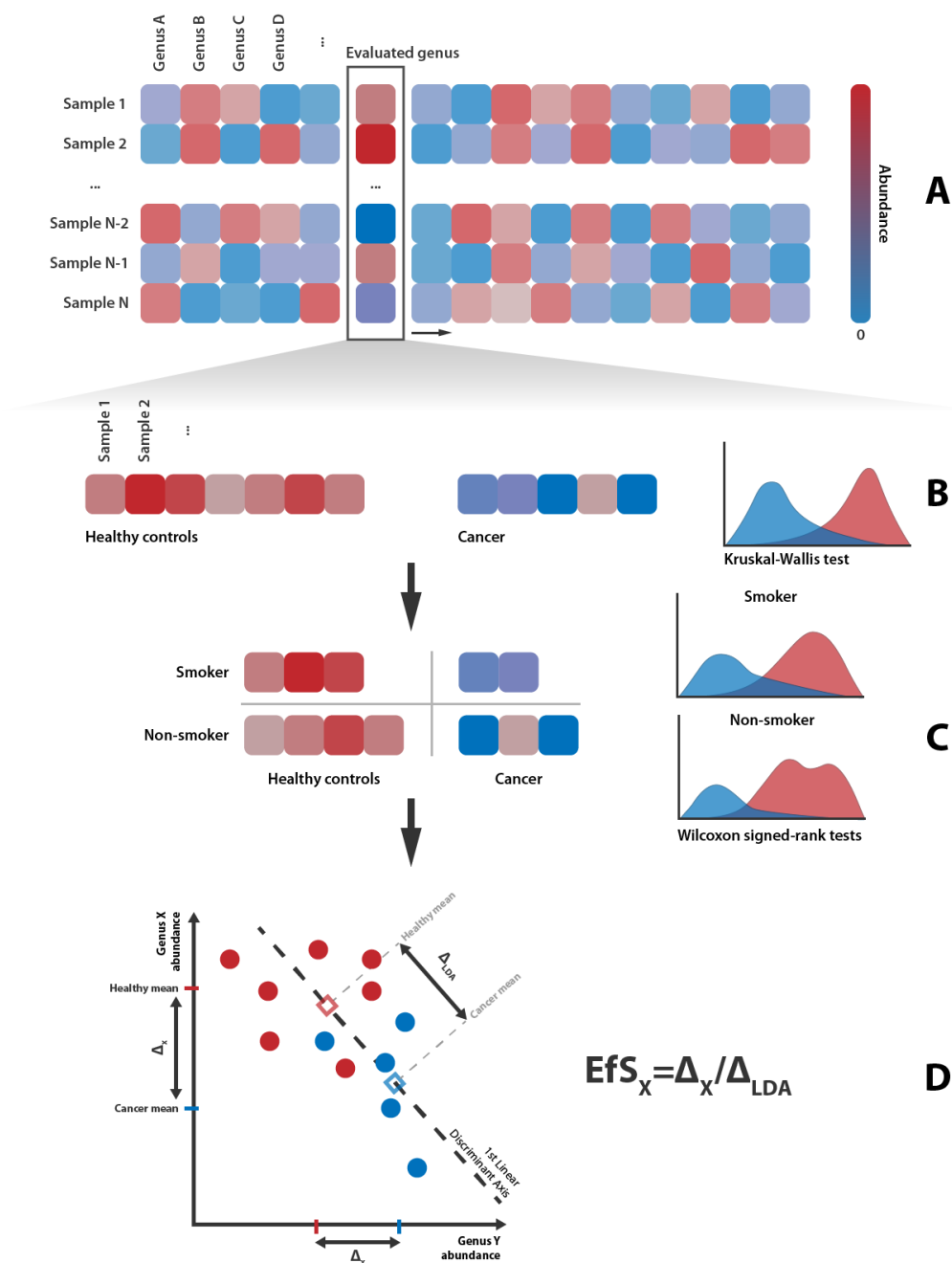


Figure 2.3: Overview of feature selection with LEfSe for cancer diagnosis from genera abundances. **A** – Features, corresponding to the abundances of a given genus across the samples in the training subset, are sequentially evaluated. **B** – A Kruskal-Wallis test is used to determine if the genus has differential abundances between the classes (in this case, healthy controls and cancer). Genera without significantly different distributions are discarded. **C** – The classes are further divided according to clinical covariates (e.g., smoking status). For each subgroup, the abundance distributions are again compared between classes with Wilcoxon signed-rank tests. If there is no difference in distributions or if the sign of this different is not the same for all comparisons, the genus is excluded. **D** – The effect size (EfS_X) of the feature is estimated according to the difference (or average difference, in multiclass settings) between the mean abundance of each class (Δ_X), divided by the difference in class means along the first linear discriminant axis (Δ_{LDA}).

powerful downstream model, which yields the final classification [95]. Logistic Regression is also often used as a surrogate model for feature selection, with the importance of each feature being given by their associated parameter in the model [86, 13, 75, 89, 112]. This will be discussed in greater detail in Section 2.4.1. Note that the features selected by these surrogate models may not be the most adequate for the final estimator, as these are subject to different constraints and use other optimization approaches.

Dimensionality reduction of metagenome data can also be performed with embedded or wrapper methods, which are part of or dependent on the predictor model, respectively [113]. Wrappers in particular are used less frequently than filters and embedded methods as they entail a greater computational cost. The most used embedded method is L1-regularized Logistic Regression, also known as Least Absolute Shrinkage and Selection (LASSO) regression [114, 29, 40, 75], which forces some of the model parameters to take zero values [115]. Besides being used as surrogate models for filter feature selection, LASSO regression can also be considered an embedded feature selection method if it is used as the final estimator. This approach was used for colorectal cancer identification coupled with a Multilayer Perceptron [89] and Generalized Neural Networks [86]. LASSO was also included in the ensemble used for feature selection in [75] for ovarian cancer identification from taxonomical profiles and serum tumor marker levels. Recursive Feature Elimination (RFE), in which the less important features for a ML model are iteratively removed, is a popular feature selection wrapper, being mostly used with tree-based estimators [116, 25, 117, 103, 60, 30, 48]. The combination between RFE and Support Vector Machines is also common and known as SVM-RFE [118, 119, 96, 120, 28].

It is unclear what feature selection approach is better for microbiome data and most ML models. One study compared the efficacy of statistical test-based methods, mRMR, a Fast Correlation Based Filter [121], Conditional Mutual Information, and prior selection with Gradient Boosted Trees for dimensionality reduction in colorectal cancer prediction from microbial abundance profiles [95]. Methods based on statistical tests and selection by Gradient Boost Trees achieved the best results when coupled with a Random Forest classifier [95]. However, this may not hold true for other datasets and ML models. It is not even clear if feature selection is at all necessary, as many authors have reported good model performance without employing any feature selection mechanism [122, 88, 18, 123, 124]. Ideally, feature selection approaches should be selected with cross-validation; alternatively, ML models capable of performing feature selection, such as deep learning approaches, may be used.

2.4 Machine Learning models for microbiome data

Learning tasks can be divided into classification or regression, when the target variable is categorical or numerical, respectively [115]. Most ML models for cancer characterization are developed to perform classification tasks (for instance, cancer type identification or classifying patients as responsive or non-responsive to treatments) [95, 125, 30, 126]. In these tasks, ML models take a feature vector $\mathbf{x}_n \in \mathbb{R}^M, n \in \mathbb{N} : n \leq N$ as input, where M is the number of features and N the

number of samples in the dataset, and predict a target value $\hat{y}_n \in \mathcal{C}$, where $\mathcal{C} = \{c_1, \dots, c_K\}$ is the set of K possible classes. This is done by applying a decision function ϕ such that $\hat{y}_n = \phi(\mathbf{x}_n)$. As an example, when using microbial abundance profiles as input features, each element of the n -th feature vector, $x_n^{(m)}$, gives the abundance of the taxa m in sample n . The ML predictions $\hat{y}_n$ should be as close as possible to the ground truth classes of the samples, denoted by $y_n \in \mathcal{C}$, and this is often achieved by training models to minimize a loss function that quantifies the difference between the predicted and ground-truth variables [127]. Most ML classifiers can estimate conditional probabilities for each class, leading to probabilistic predictions $\hat{\mathbf{p}}_n = [p_n^{(1)}, \dots, p_n^{(K)}]$, in which $p_n^{(k)}$ gives the posterior probability for class c_k , i.e., $Pr(y_n = c_k | \mathbf{x}_n)$. The final predicted class is taken as that with the maximum posterior [127]. When there are only two possible target classes, such as in colorectal cancer identification, where a sample is either from a colorectal cancer patient or a healthy control, the learning task is a binary classification task; when there are multiple target classes, as in cancer type identification, the learning task is considered multiclass classification [127].

Regression tasks are much rarer in cancer characterization using the microbiome. These have so far been restricted to survival time prediction and estimating the number of tumors formed in murine models [39, 26]. Therefore, this section will focus mainly on ML models for classification, highlighting those most popular and most promising novel approaches. Table 2.2 contains an overview of the ML models discussed in this Section and their applications, while Table 2.3 summarizes their strengths and weaknesses.

Table 2.2: Examples of applications of ML methods for cancer characterization from microbiome data.

Learning Task	SVM	Decision Trees	Random Forest	Boosted Trees	Logistic Regression	NNs
Cancer type identification	[47]	-	[47]	[44]	-	-
Identification of colorectal cancer	[122, 25, 58, 123, 94, 81]	[122, 128, 123]	[122, 114, 25, 58, 128, 37, 123, 81]	[122, 25, 123]	[122, 114, 123]	[86, 86, 25, 90, 89, 81]
Identification of other cancer types	[31, 129, 28]	[31]	[108, 116, 88, 29, 41, 75, 31, 130, 30, 49]	[67, 131, 69]	[67, 13, 71, 75, 31, 69, 28]	[31]
Multi-disease identification (including cancer)	[118, 97, 132, 133]	-	[125, 134, 97, 132, 133]	[59]	[118, 97]	[80, 59, 132, 133]
Prediction of treatment outcomes and survival time	[40, 126]	[135, 40]	[39, 40, 136, 126]	-	[40, 112]	[40, 136]
Identification of cancer precursors and complications	[70]	[70]	[70, 43, 27, 68]	-	[70]	[70]

Table 2.3: Advantages and disadvantages of ML methods for cancer characterization from microbiome data.

Type	Method	Advantages	Disadvantages
SVM	SVM	Can be adapted for wrapper feature selection Good generalization performance Good performance with few samples Widely available implementations	Difficult interpretation of rules High sensitivity to hyperparameter tuning Not adequate for multiclass classification Computationally expensive
Tree-based	Decision Tree	Can be adapted for wrapper feature selection Easy to interpret Widely available implementations	High variance
	Random Forest	Can be adapted for wrapper feature selection Improved interpretability over SVMs and NNs Low variance Widely used Usually outperforms other methods Widely available implementations	Reduced interpretability over Decision Trees and Logistic Regression
	Boosted Trees	Can be adapted for wrapper feature selection Improved interpretability over SVMs and NNs Low variance	Computationally expensive Reduced interpretability over Random Forests and Logistic Regression
	Logistic Regression	Can be adapted for wrapper and embedded feature selection (LASSO) Easy to interpret Widely available implementations	Linear decision boundary
Artificial Neural Networks		High complexity Large variety of architectures Prior feature selection is not necessary Can generate new data	Very difficult to interpret Computationally expensive to train Few works using them so far Require large numbers of samples
		Promising results for improved performance over other methods Aid data integration	

2.4.1 Logistic Regression

Logistic regression is a linear model for classification that, due to its simplicity, has been widely used in studies investigating the tumor-specific microbiome, mainly for feature selection and benchmarking [86, 80]. This model assumes that the microbiome data originated from class-conditional Gaussian distributions of equal covariance matrices [115]. In binary classification, the estimated posterior probability of a sample $\mathbf{x}_n$ belonging to the positive class ($\hat{p}_n$) is given by Equation 2.1 [137, 115]. The model parameters $\beta \in \mathbb{R}^M$ and β_0 are learned using iterative processes, usually gradient descent with binary cross entropy as a loss function [115].

$$\hat{p}_n = \frac{1}{1 + e^{\mathbf{x}_n^T \beta + \beta_0}} \quad (2.1)$$

Despite having linear decision boundaries, Logistic Regression classifiers can, in some cases, achieve similar results to non-linear counterparts in microbiome cancer characterization: in [123], L2 regularized Logistic Regression achieved a comparable Area Under the Receiver Operating Characteristic Curve (AUC) to Random Forest and Boosted Trees for colorectal cancer identification. However, this is seldom observed [122, 82, 31]. For example, Logistic Regression failed to improve colorectal polyp identification over a Multilayer Perceptron and Naïve Bayes classifier [70], and colorectal cancer identification over a Multimodal Neural Network [80]. It is likely because of this decreased performance that Logistic Regression has mainly been used for feature selection and benchmarking.

L1 and L2 regularization are approaches to decrease model variance by constraining its parameters [138]. L2 regularization penalizes the squared value of the model parameters (β and β_0), decreasing their absolute values [139]. L1 regularization, or LASSO, constrains the model parameters so that the sum of their absolute values is less than a number and has been briefly discussed in Section 2.3 [138]. LASSO regression forces some of the parameters to take null values, effectively disregarding the corresponding features during inference [91]. Combining L1 and L2 regularization results in a model known as Elastic Net [140]. Therefore, LASSO and Elastic Net are embedded forms of feature selection and have been used to select microbiome features for a more powerful downstream classifier, as previously mentioned. As the decision boundary is linear, feature contribution for classification can more intuitively be quantified by the corresponding model parameter. The most relevant features are chosen accordingly [86, 75].

The same properties that make Logistic Regression adequate for feature selection also explain its improved interpretability over methods such as Random Forests and Artificial Neural Networks: the contribution of each feature can be quantified by the parameters of the model [123]. This allows researchers to more easily identify potential microbial biomarkers associated with specific types of cancer [123]. This strategy was used to identify seven genera potentially associated with invasive cervical cancer [29] and to access feature contributions when identifying colorectal cancer using taxonomical profiles [123].

2.4.2 Support Vector Machines

Support Vector Machines (SVMs) [141] are models that are commonly used with microarray data [142]. Therefore, it is not surprising that these are also often employed to identify microbiome-related biomarkers for cancer. In their simplest form, SVMs are binary classification models that define the decision boundary as a hyperplane. When the data is linearly separable, there are infinite hyperplanes that perfectly separate the samples according to their class [115]; SVMs select the hyperplane with the largest margin to the training data [143]. In the context of SVMs, the target variables y_n can only be 1 or -1. The decision function ϕ is given by Equation 2.2, where $\beta \in \mathbb{R}^M$ and β_0 are the model parameters. Because not all problems are linearly separable, the SVM allows for some misclassifications of training samples. Each sample beyond the margin is assigned a slack variable $\xi_n \in \mathbb{R}$, $\xi_n \geq 0$ that quantifies its distance to the margin, and the model penalizes the sum of all slack variables according to a hyperparameter C , such that increasing C leads to tighter margins [144, 145, 91]. Figure 2.4 provides further clarification on the meaning of these slack variables. Training the SVM classifier thus entails solving the optimization task in Equation 2.3. In SVMs, only some of the training samples define the decision boundary, and these are called the support vectors [127].

$$\phi(\mathbf{x}_n) = \text{sign}[\mathbf{x}_n^T \beta + \beta_0] \quad (2.2)$$

$$\begin{aligned} \min_{\beta, \beta_0} & \frac{1}{2} \|\beta\|^2 + C \sum_{n=1}^N \xi_n \\ \text{subject to} & \xi_n \geq 0 \wedge y_n(\phi(\mathbf{x}_n)) \geq 1 - \xi_n \\ & , \forall n \in \mathbb{N} : n \leq N \end{aligned} \quad (2.3)$$

The SVM described leads to linear decision boundaries in the input space. Such linear SVMs were found to achieve comparable performance to non-linear models in colorectal cancer identification from microbial abundance profiles [123]. However, SVMs can be adapted to non-linear decision boundaries by extending the original feature space with feature transformations. This results in non-linear decision boundaries in the original feature space and may improve the separability of the data [91]. Feature extension can be done without increasing the computational load with the kernel trick: for a feature transformation $h(\mathbf{x}_n)$, its kernel is a function $\kappa : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$ such that $\kappa(\mathbf{x}_n, \mathbf{x}_l) = \langle h(\mathbf{x}_n), h(\mathbf{x}_l) \rangle$, where $\langle \cdot, \cdot \rangle$ represents the inner product [146]. If the kernel for a transformation is known, there is no need to compute the mappings of the original samples to the new feature space, decreasing computational cost [146]. The most used kernel in cancer-related microbiome analysis is the radial basis function (RBF) [31, 123]. However, when kernels are used, the prediction for a sample is based on its kernel distance to the support vectors [147]. Therefore, the decisions of kernel SVMs are more difficult to interpret than with other ML models such as Random Forests or Logistic Regression [147].

SVMs have been used to predict colorectal patient survival time [58, 126], cancer prognosis, and drug responses from microbiome and gene expression data [96], and to identify colorectal

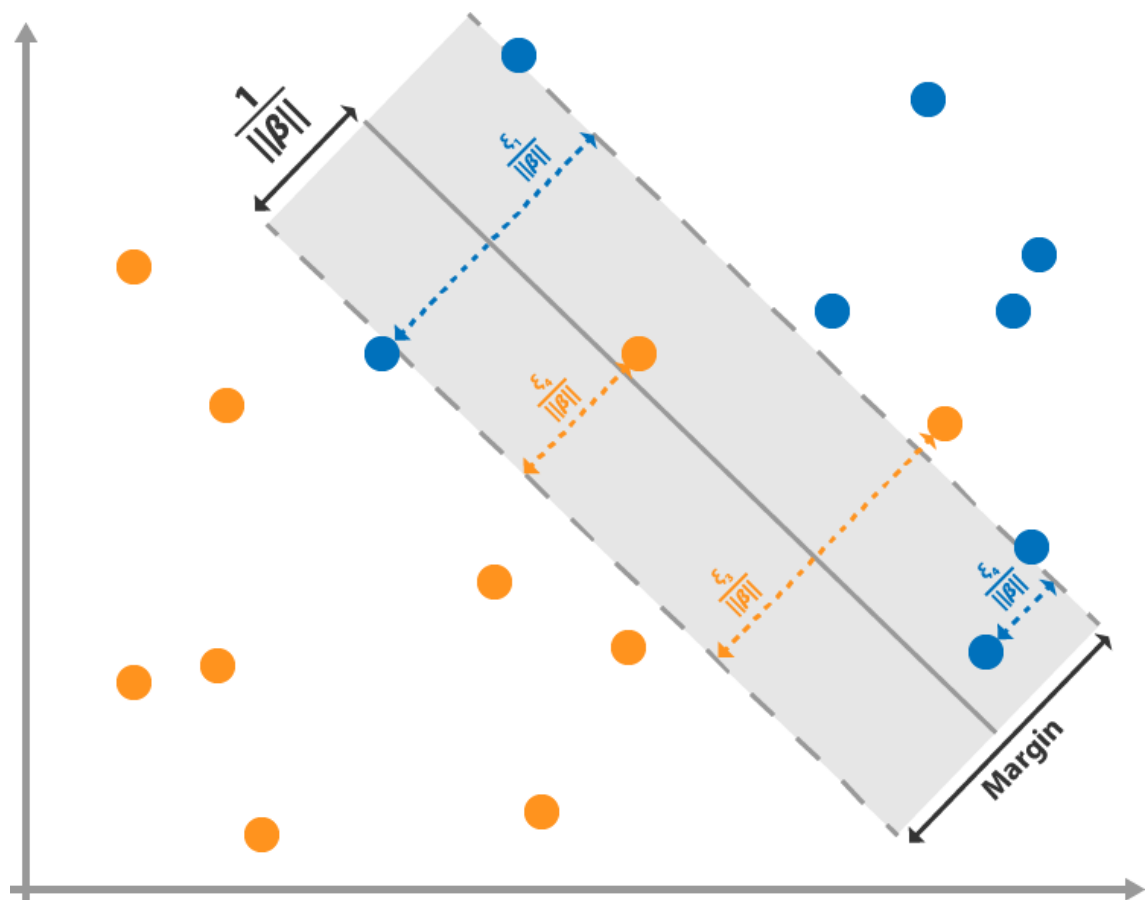


Figure 2.4: Margins and decision boundary (solid line) for an SVM classifier. Colored dashed lines show the values of the slack variables ξ_n , which are proportional to the distance between a sample and the margin for its class.

cancer patients using taxonomic profiles [123, 94]. Furthermore, SVMs are frequently used as benchmarks when evaluating other methods [122, 80, 31]. The popularity of SVMs for cancer-related host-trait prediction from the microbiome can be attributed to their widespread availability in ML libraries and the good performance demonstrated by these models in multiple works. As an example, in [97], SVMs outperformed Logistic Regression when predicting various diseases using taxonomical profiles. In a few cases, SVMs achieved better or comparable results to Random Forest classifiers [97, 81]. However, in most of the experimental settings, SVMs failed to surpass the performance of Random Forests, namely in survival time prediction [126], cancer types identification [47], and the identification of colorectal polyps [70].

Because SVMs are inherently binary classifiers, extending them to multiclass classification comes with multiple problems. Usually, a classifier is trained for each class, in a one-versus-rest approach, with the final prediction taken as the one yielding the highest value for the decision function. Because each classifier is trained with different targets, there is no assurance that these predictions are scaled equally [127]. Furthermore, this approach is computationally expensive, as it requires training a potentially large number of classifiers. These issues make SVMs rarely used for microbiome-based multicancer prediction. In such an application, an SVM achieved a much worse performance than a Random Forest, a k-Nearest Neighbors, and a Decision Tree classifier [47], all models that can be directly used in multiclass settings.

Despite all the shortcomings, SVMs may prove useful when only a small training set is available. These models are less prone to overfit the training data and can show good generalization performance when the number of features far exceeds that of training samples [148]. This has motivated its usage in oral carcinoma identification from taxonomical and functional profiles in a study comprising only 38 samples [129]. However, it is unclear if other models could have achieved comparable performance, as no other approaches were tested. Furthermore, this good generalization performance is dependent on an adequate choice of the regularization parameter C , which requires costly cross-validation [115].

2.4.3 Tree-based classifiers

Models based on Decision Trees – Random Forests in particular – are the most frequently used ML models when studying the links between cancer and the microbiome. Decision Trees are sequential models that apply successive rules to yield a final prediction. Multiple identical implementations of Decision Trees exist, with the most popular being Classification and Regression Trees (CART), ID3 [149], C4.5 [150], and C5.0. In this Section, Decision Trees will be described in the context of multiclass classification with continuous data, the most common setting in microbiome-based cancer characterization. However, these can also be used with categorical features and, under some implementations, in regression tasks. As seen in Figure 2.5, at each level of the tree, starting from the root, the value of a feature is compared to a threshold learned from the training data. This threshold and the feature used in each comparison are selected by exhaustively searching for those leading to the greatest purity in the child nodes. Purity, expressing the homogeneity in target classes of the samples in a leaf, is commonly quantified using the Gini index or entropy [151].

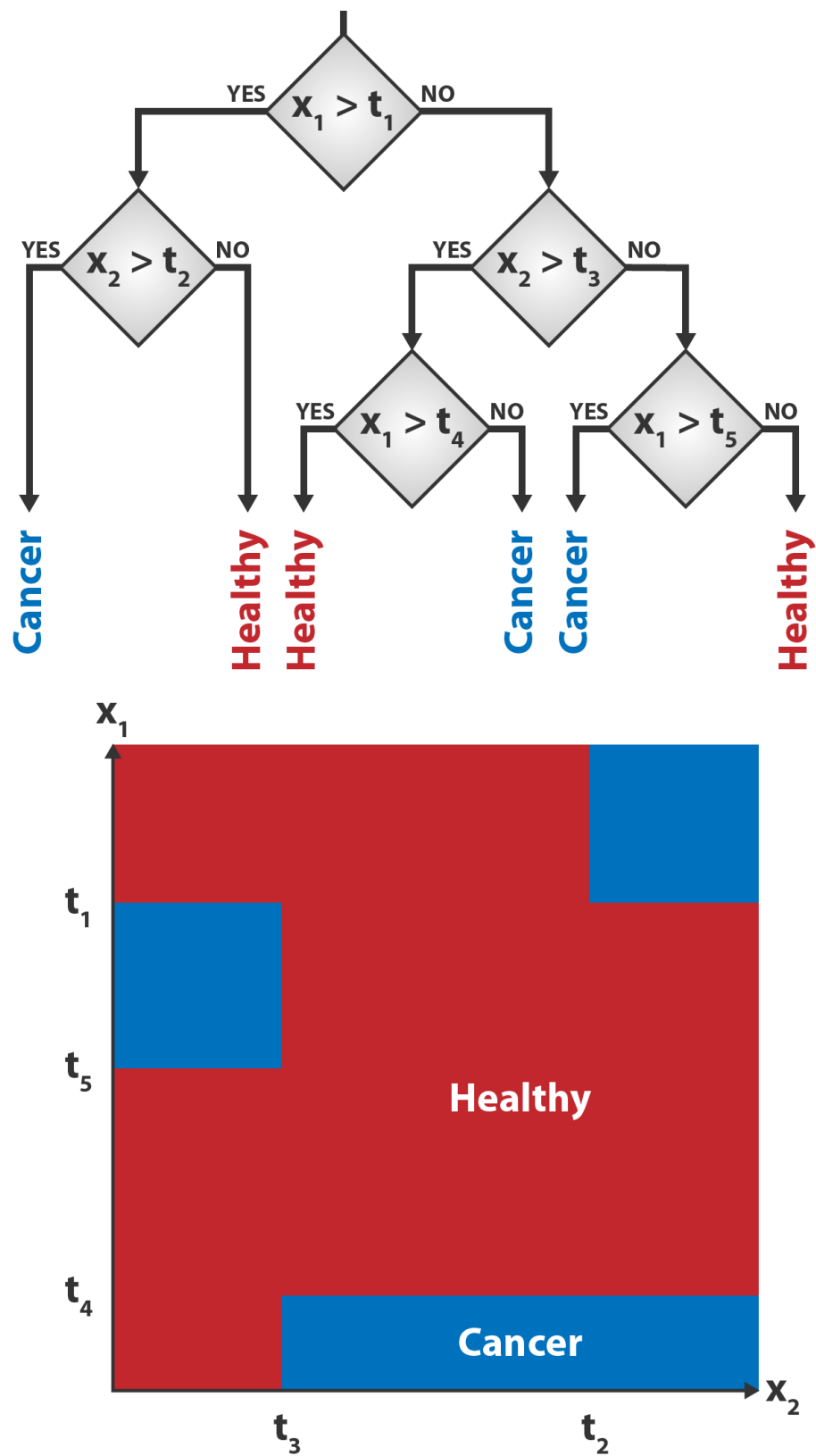


Figure 2.5: Example of a Decision Tree classifier for cancer diagnosis using two input features (x_1 and x_2). A sequence of comparisons between the values of the input features and learned thresholds t_n results in the division of the feature space into cuboids. Each of these regions is associated with a target class (healthy or cancer), according to the most prevalent class of training samples inside that region.

The decision path followed by the model is defined by the result of each comparison and the final prediction is given by the class associated with the leaf reached [152]. Because every node (comparison) imposes a constant decision boundary, the feature space is partitioned into cuboids. The model predicts the class of a sample falling into a given partition of the feature space according to the most frequent class in that partition [127].

Although it is possible to split the nodes until all leaves have samples of only one class, this will likely lead to overfitting [152]. Decision Trees are prone to high variability, as small changes in the training data may result in different splitting thresholds: changes in the first nodes propagate into the subsequent levels, greatly impacting the overall classification rules [91]. To combat this, the growth can be stopped when no split increases the purity by more than a small amount [127]. Alternatively, pruning approaches reduce the number of nodes in the tree after its construction by collapsing internal nodes. Pruning methods try to balance the complexity of the Decision Tree with its performance on the training data [152].

2.4.3.1 Random Forest

Even with the aforementioned regularization approaches, Decision Trees are prone to high variance, which helps explaining why these models are rarely used in studies applying ML for cancer characterization from microbiome data. Another approach to reduce the variance shown by Decision Trees is bagging, in which multiple models are trained using samples and features randomly sampled with replacement from the training set [153, 91]. Bagging Decision Trees results in a Random Forest model, in which the different trees have slightly different structures, as they are trained on different data [152]. The final prediction is the average of the predictions of the individual trees [152]. In a work on multicancer identification from microbial abundances, a Random Forest classifier achieved improved generalization performance over a Decision Tree [47]. A similar result was found in [123] for colorectal cancer identification.

Unlike Decision Trees, Random Forest classifiers are highly popular when studying the association between cancer and the microbiome. This is supported by several published benchmarking experiments in which this model outperformed all other tested algorithms in tasks including identifying colorectal cancer [123], melanomas in mice [31], cancer subtypes [47], and other host traits [97]. Random Forests have been applied to predicting the survival time of colorectal cancer patients from gene expression and microbiome taxonomic profiles [126] and identifying several tumor types such as epithelial ovarian cancer [75], tonsillar squamous cell carcinoma [88], lung adenocarcinoma [41], colorectal cancer [37], oral squamous cell carcinoma [49], and in a multi-class classification setting [47].

2.4.3.2 Boosted Trees

Boosting is another approach to decrease model variance and operates on a similar idea to bagging; however, instead of training the standalone estimators independently, it builds them sequentially [127]. The complexity of individual estimators is low and, if using Decision Trees as the

base models, these are kept shallow [154]. In boosted models based on Decision Trees, each tree is trained with samples drawn from the dataset, weighted according to the performance of the previous classifier on that data point. This way, the next individual model is trained to improve the performance on data that the previous tree failed to predict [127]. The final prediction is taken as a weighted average of the prediction of all the trees, or as a majority voting [127]. Specific details on boosting vary according to implementation, of which the most popular include Adaptive Boosting (AdaBoost) [155], Extreme Gradient Boosting (XGBoost) [156], LightGBM [157], and CatBoost [158]. Some of these implementations use gradient boosting, which, when combined with Decision Trees as a base model, leads to so-called Gradient Boosted Trees (GBT) [156].

The improved generalization performance of boosting methods over Decision Trees was exemplified in [123]. However, it is unclear whether boosting can achieve better performance over Random Forests in cancer-related host-trait prediction from the microbiome. In this same publication, a boosting approach and a Random Forest demonstrated comparable performance [123], while in [31] a boosting model showed a decreased AUC but improved Precision and Recall over Random Forests. Boosting methods have also been proposed to predict tissue malignancy in breast cancer using bacterial taxonomic profiles from biopsies [18] and to identify cancer types from microbiome data [44].

2.4.3.3 Interpretability and Explainable Boosting Machines

Besides the excellent performance demonstrated by Random Forests and boosted Decision Trees in cancer-related microbiome host-trait prediction, these models are often used due to their improved interpretability over other *black box* algorithms such as non-linear SVMs and Artificial Neural Networks. For sufficiently shallow Decision Trees, decision rules can be easily analyzed and understood, making finding putative biomarkers easier. Decision tree flowcharts are already widely used in medicine for diagnosis, so they are familiar to doctors and health professionals [159]. As Random Forests and boosting combine the predictions of several trees, this direct interpretation of feature contribution is lost. However, it is possible to quantify the importance of each feature by analyzing the number of times a feature is used in a node or the average increase in purity achieved with it [111]. As an example, in [68], Random Forests were used to identify acute myeloid leukemia from taxonomical profiles, and the contribution of each taxon for classification was assessed using the average increase in purity. However, these scores do not fully reflect the complex decision rules of the classifier, which depend on the association of multiple features.

With boosting, the link between importance scores and decision rules is even more unclear than for Random Forests, as different trees have a different impact on the final prediction. Explainable Boosting Machines (EBM) [160] try to make this association clearer so that the decision of a model is better understood through feature importance scores. EBM is an implementation of Generalized Additive Models [161] that also models pairwise feature interactions, and in which the feature transformations are modeled using a boosting algorithm. When training EBMs, each tree can only use one feature. The set of M features is iterated multiple times, with each tree contributing by a small amount to the final classification. Because of this, a large number of trees

is needed in EBMs, making training times longer than for other boosting approaches [162, 163]. The importance of each feature can be easily assessed by analyzing the importance scores and using heatmaps for pairwise interactions [164, 163]. Because each tree can use only one feature and is a weak classifier, these scores provide better insight into the classification rules [160]. Explainable Boosting Machines have been used to identify colorectal cancer and adenoma from microbial gut functional profiles [83], allowing greater interpretability without sacrificing accuracy. More transparent methods such as EBM may help identify novel microbial biomarkers for cancer characterization.

2.4.4 Artificial Neural Networks

Artificial Neural Networks (NNs) are complex non-linear ML models that usually achieve better performance than other simpler estimators [165]. An NN is formed of multiple layers, each with stacked activation units. Most often, activation units take a vector as input and output a scalar, applying an activation function on the linear combination of inputs and a bias parameter, as per Equation 2.4, in which $\beta_{u,l} \in \mathbb{R}^M$ and $\beta_{0,u,l} \in \mathbb{R}$ are the learnable parameters for the u -th activation unit of layer l ; the function $\psi_l : \mathbb{R} \mapsto \mathbb{R}$ is the activation function of layer l , which is usually a Rectified Linear Unit (ReLU) or one of its variations [166]. After each layer, its outputs are concatenated and used as input for the following layer, until the last is reached. This last layer is known as the output layer, as it gives the final prediction of the model. Thus, NNs learn a predictor function that is a chain of functions modeled by each layer. These are estimated according to the parameters of each activation unit. The layers between the input and output are called hidden layers [167]. Architectures with more than one hidden layer are considered Deep Neural Networks [168]. In NNs, the model parameters are learned using stochastic gradient descent. This computation is often referred to as back-propagation, as the gradient of the error with respect to the parameters is computed using the chain rule from the output to the input layer [167].

$$\mathbf{x}'_{u,l} = \psi_l(\mathbf{x}^T \beta_{u,l} + \beta_{0,u,l}) \quad (2.4)$$

Deep NNs do not require extensive previous feature selection and transformation, unlike the other ML models discussed so far. These deep architectures are complex enough to do the necessary processing of input features by themselves [169]. The output vectors of hidden layers can be perceived as mappings of a sample to a transformed feature space, or latent space, in which class separability is enhanced, in the case of classification tasks [167]. Even so, multiple works on the tumor-specific microbiome have used Deep NNs trained with previously selected features [70, 31, 81]. Doing so may hinder performance, as these features may not be ideal given the model.

The decision rules of NNs are difficult to interpret because of their complexity and non-linearity. Because of this, NNs are seen as *black box* models which make biomarker identification more difficult than Random Forests or linear models. In cancer characterization from microbiome

data, works assessing feature importance usually resort to model-agnostic methods to better understand the decisions made by NNs: Mulenga et al. [90] used L2 regularization, while Arabameri et al. [86] employed Derivative-based Feature Selection. These methods do not provide a complete understanding of the model decisions, nor of the interdependencies between features, as provided by Decision Trees, for instance.

2.4.4.1 Multilayer Perceptrons

The Multilayer Perceptron (MLP) is an NN in which consecutive layers are fully connected, that is, all the outputs of a hidden layer contribute to the input of the next layer, as seen in Figure 2.6. One drawback of Multilayer Perceptrons is that these can easily overfit the training data if no regularization is adopted, such as dropout and batch normalization, due to their large number of parameters [170, 171]. Because of this, MLPs often show a decreased performance when compared against other models discussed in this Chapter. For instance, in [81], an MLP failed to improve the performance of an RBF-kernel SVM for colorectal cancer prediction using microbiome taxonomic profiles. However, the MLP used features pre-selected using Linear Regression, which may not be ideal for NN classification. Similar findings have also been reported for melanoma identification in mice while employing feature selection [31]. On the other hand, the MLP with prior feature selection used in [86] achieved higher accuracy but worse sensitivity when compared to a k -Nearest Neighbors classifier and an SVM. Therefore, it is unclear if MLPs can achieve better predictions than those provided by simpler models such as Random Forests and non-linear SVMs.

DeepEn-Phy [82] is a notable approach using MLPs. This model takes microbial abundances incorporated into a phylogenetic tree as input features and uses a set of cascaded MLPs trained end-to-end. In DeepEn-Phy, the phylogenetic trees are divided into levels, from leaves to roots; each MLP operates in one level, with a single level having multiple MLPs. The first MLPs use abundances in the leaf nodes and their outputs are fed into downstream MLPs, until reaching the root estimator. Figure 2.7 provides a diagram of this approach. In smoking status and Body Mass Index prediction from the gut microbiome, DeepEn-Phy outperformed Random Forest, Gradient Boosted Trees, and PopPhy-CNN, another deep learning model operating on phylogenetic trees [172].

2.4.4.2 Multimodal Artificial Neural Networks

Multimodal NNs combine MLPs to integrate features from different modalities. In a study, a multimodal NN was used to identify several diseases, including colorectal cancer, using both taxonomic and functional profiles from the gut microbiome [80]. MLPs were trained using features from each modality alone and embeddings were derived by concatenating the outputs from the last hidden layers of each modality. This multimodal NN achieved improved accuracy over MLPs that used just either taxonomic or functional profiles; it also outperformed Random Forest, an SVM, GBT, and LASSO regression for colorectal cancer prediction [80]. This shows the potential of combining multiple types of data in increasing the accuracy of tumor diagnosis and introduces

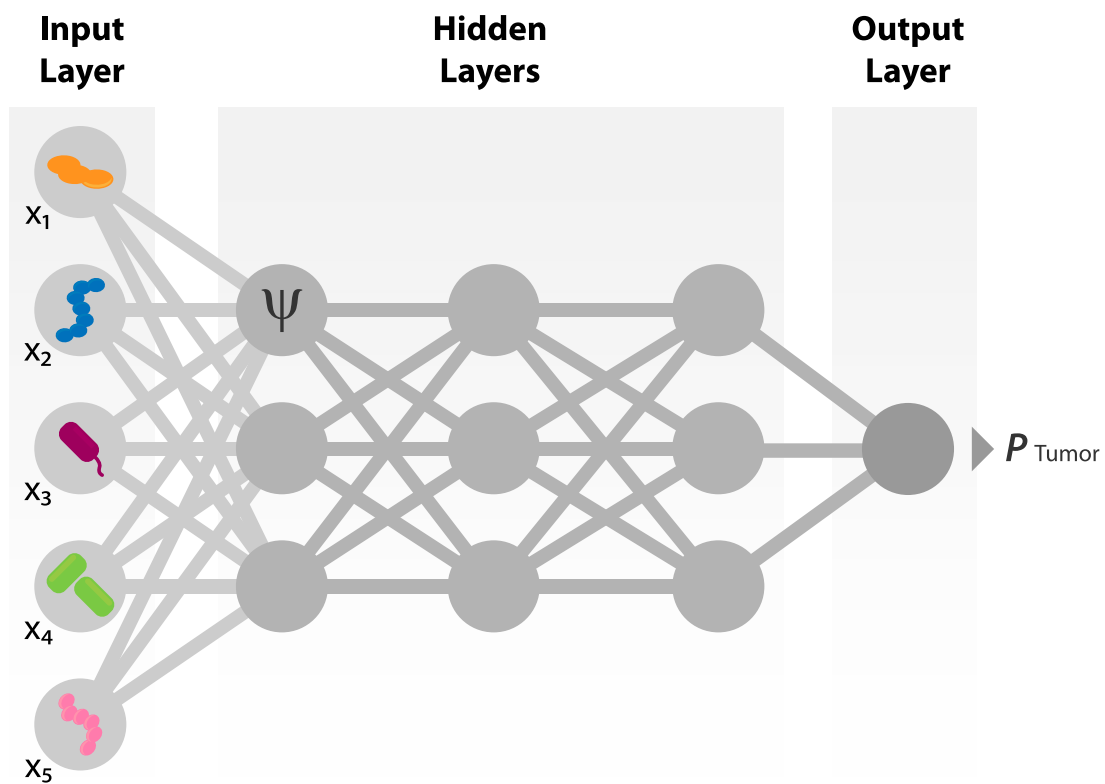


Figure 2.6: Diagram of an MLP for cancer diagnosis from microbial abundance profiles. The input layer contains the abundance of each taxon. These act as inputs ($\mathbf{x}$) for the activation units in the hidden layers, which apply a function ψ . The output layer returns the probability of the presence of cancer.

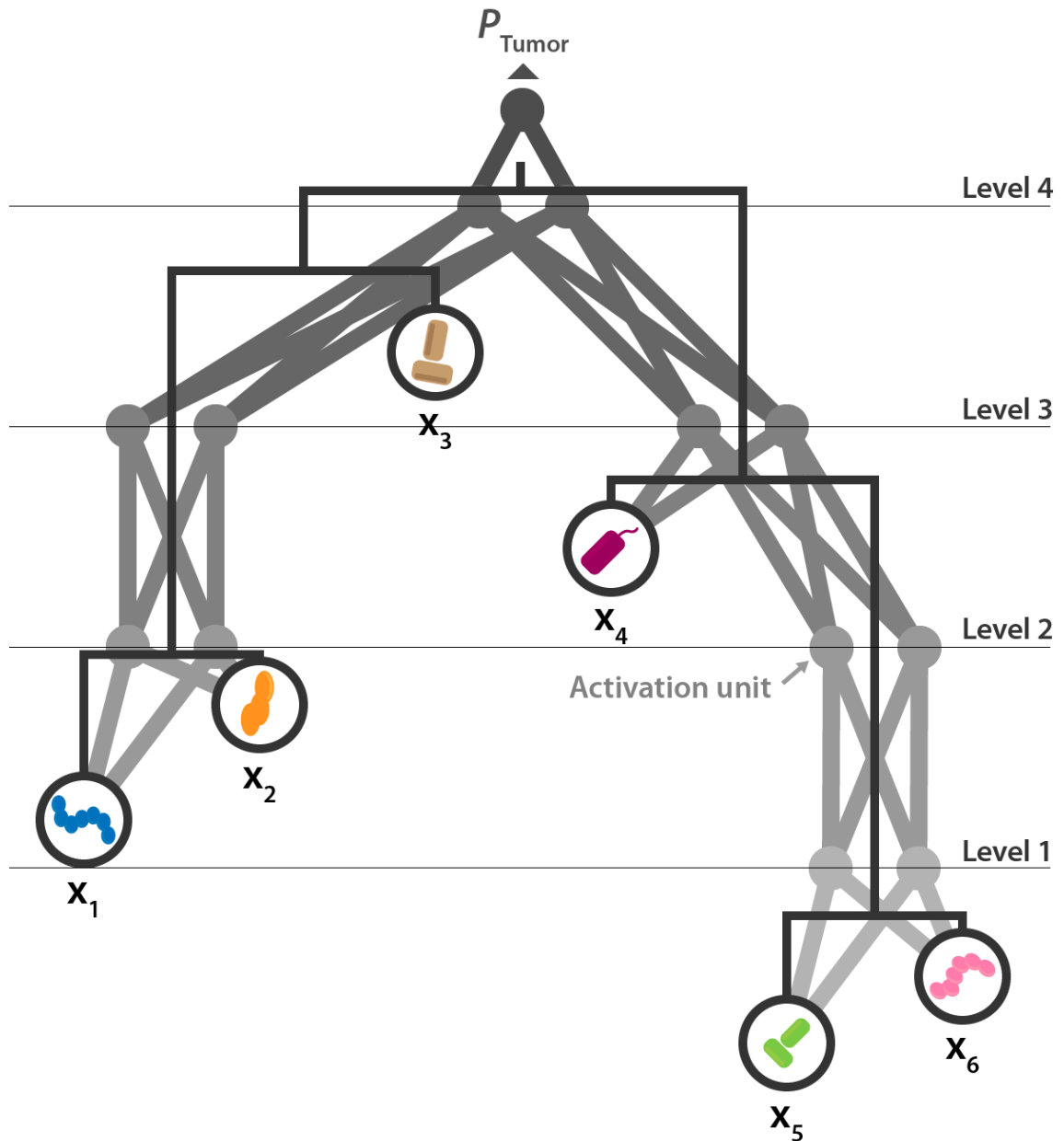


Figure 2.7: DeepEn-Phy diagram. Microbial abundances are incorporated into a phylogenetic tree. This tree is divided into levels (four, in this example) which are traversed from the leaves to the root. At each level, MLPs transform groups of abundances (x_n) and representations of the previous level that are connected in the phylogenetic tree. All MLPs are trained end-to-end. Figure adapted from [82].

multimodal NNs as a promising method to integrate various modalities of features that complementarily capture the complex mechanisms that underly cancer development.

2.4.4.3 Autoencoders

As previously mentioned, the outputs of hidden layers of NNs can be interpreted as mappings of the input vectors into a latent space. If a layer has less than M activation units, with M the number of input features, these latent representations can be used as compressed representations of samples [115]. Autoencoders are NNs that try to reconstruct the input data: not to achieve useful predictions, but to learn representations of the input vectors in a latent space [167], as seen in Figure 2.8.

Due to the high dimensionality of microbiome data, Autoencoders can be useful in avoiding the curse of dimensionality and improving the performance of ML models. However, despite having been proposed for dimensionality reduction of microbiome data [173], Autoencoders are rarely used in microbiome-based cancer characterization. In [174], Autoencoders coupled with an MLP failed to improve the predictive performance of MetAML [97], a tool which uses Random Forests, SVMs or Logistic Regression, when identifying colorectal cancer from the gut microbiome [174]. However, it is unclear if this is due to the classifier used or the dimensionality reduction strategy.

Autoencoders can also be adapted to perform data augmentation, generating synthetic samples. Because data augmentation techniques increase the number of training samples, they decrease the variance of ML models, assuming that the synthetic samples are derived from a distribution similar to that of the original data [175]. This is particularly important for models that are prone to overfit the training data, such as deep NN. While studying the cancer-related microbiome, this issue is made worse due to the scarcity of data and its high dimensionality. Therefore, data augmentation with Autoencoders can be a powerful aid in the development of more accurate ML models acting on microbiome data. In [90], data augmentation using Variational Autoencoders improved the predictions of an MLP classifier for colorectal cancer identification using gut microbial abundances [90], with an increase in the AUC of up to 30%.

2.4.4.4 Other Artificial Neural Networks

Despite being more rarely used than the NNs presented in the previous sections, other variations of NNs have been used for cancer characterization using microbiome data. General Regression Neural Networks (GRNN) [176] are one such alternative, and are variations of kernel regression networks [177]. In a GRNN, there is one activation unit per training sample; all activation units apply an RBF centered in each of the training samples during inference. The output of the model is a weighted sum of the activation units [176]. One advantage of GRNN over MLPs is the smaller number of hyperparameters: only the spread parameter needs to be specified, while in MLPs one must choose the number of layers and activation units, and other parameters for optimization [86]. However, the pattern layer grows with the training set, making learning and inference computationally expensive [178]. This model was used to detect colorectal cancer from taxonomic profiles

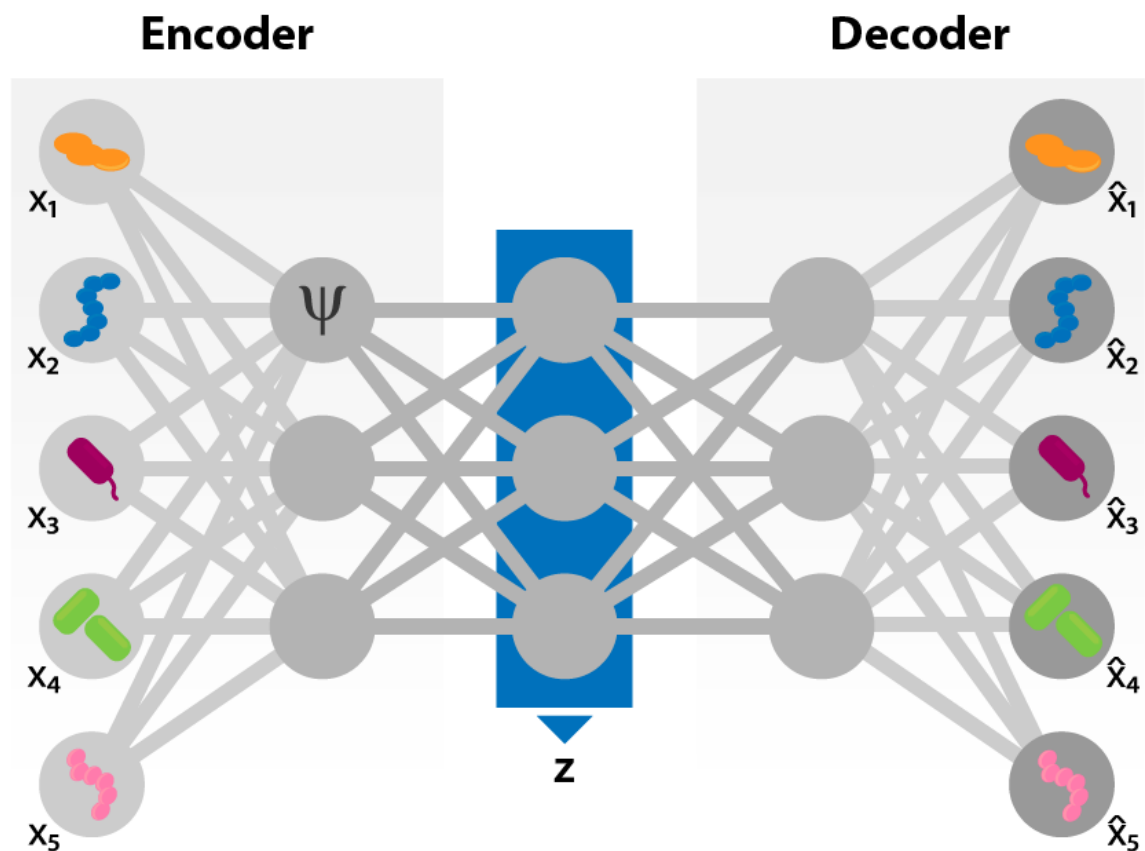


Figure 2.8: Diagram of an Autoencoder for microbial abundance profiles. The model predicts the abundance of each taxon ($\hat{x}$), which also act as inputs (x). Activation units apply a function ψ to these inputs and the output of a hidden layer provides a latent representation of the original profile (z).

of the gut microbiome, having outperformed an MLP and an RBF kernel SVM [86]. According to the authors, GRNNs can achieve better accuracy than MLPs, even with a small number of training samples.

Convolutional Neural Networks (CNNs) are NNs that use a convolution in any of their layers, instead of linearly combining all inputs weighted by the parameters of the activation units [167]. These types of NNs are most adequate for data with a known grid-like structure, being mostly used with images and time-series [167]. Microbiome data does not usually follow this topology, instead being tabular data with no inherent positional or temporal information. However, it can be transformed to one- or two-dimensional grid-like representations, enabling the applications of CNNs. One way to achieve this is by incorporating microbial abundances into phylogenetic trees and then converting these trees into matrix representations. This approach was proposed in [172], coupled with a CNN nicknamed PopPhy-CNN. PopPhy-CNN was used to predict smoking status, Body Mass Index, and identify type 2 diabetes, but it has not been applied to cancer-related problems yet [82, 172]. Doing so may enable the application of state-of-the-art CNN architectures, improving the predictive performance of ML models for cancer characterization from the microbiome. Another approach to convert microbial abundance profiles into matrices was proposed in [179]. This method, named Microbial Embedding, Grouping, and Mapping Algorithm (MEGMA), transforms the taxonomical profiles of each sample into color images, which can then be used as input for a CNN [179]. This approach has been used to identify colorectal cancer, among other diseases, from the gut microbiome, having outperformed Random Forests, SVMs, Logistic Regression, MLPs, and PopPhy-CNN [179].

Graph Neural Networks (GNNs), more specifically Convolutional GNNs, have also been applied to identify colorectal cancer from gut microbial abundances [132, 180]. These NNs operate on graphs, with convolutional GNNs aggregating the information in a node together with that of its neighboring nodes, similarly to convolution in CNNs [181]. In one study, a GNN achieved worse classification performance in colorectal cancer identification than a Random Forest model and an MLP [132]; however, it is not clear how the microbial abundance profiles were transformed into graph representations. In [180], the authors built a graph from microbial abundance profiles by considering each sample a node, and edges weighted by the similarity between samples. This representation was then used to train boosted GNNs in a semi-supervised learning setting [180]. These boosted GNNs outperformed an SVM, an MLP, a Random Forest, and GBT [180].

2.5 Current limitations

This Section contains a discussion on the current limitations of research on ML methods for microbiome-based cancer characterization, which are summarized in Figure 2.9, along with how to address them.

Many publications on ML for cancer characterization from microbiome data present contradictory results, with questionable generalization due to data leakage or confounding technical

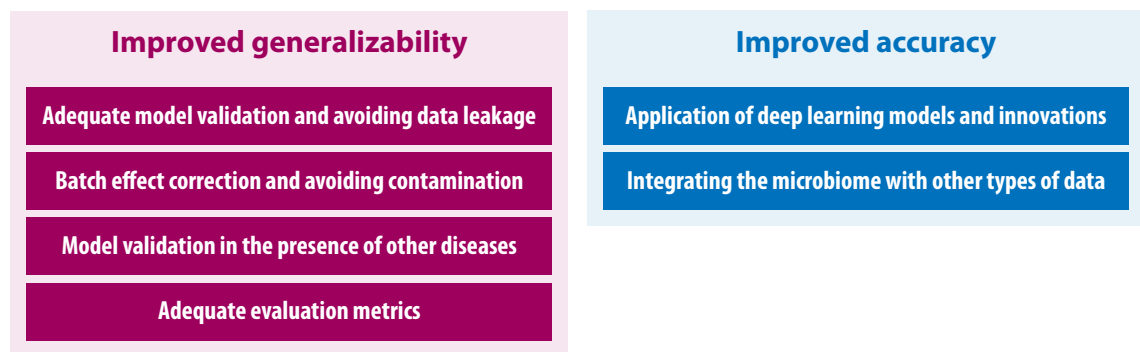


Figure 2.9: Overview of needed improvements in microbiome-based identification of cancer. Adequate model validation can improve the generalizability of the proposed biomarkers; data integration and improvements in deep learning architectures may improve their accuracy, enabling clinical usage.

variables. Every ML model should be trained with a training subset of the data, and its generalization performance should be evaluated on a hold-out test subset. If any inference about the structure of the data is made using the test set or the entire dataset, the generalization of the results may become compromised. Often, promising bacterial biomarkers are chosen with statistical tests on the whole dataset; these biomarkers are then used as input features to an ML model. This feature selection based on all samples may lead to an overfitting to the dataset. In other cases, the publications are not clear about the validation strategies employed and the possibility of data leakage, making it difficult to assess the generalization of their conclusions.

The lack of correction for batch effects or testing of models in different datasets to those used for training further raises doubts about how the findings of most works can be generalized. Not all differences in the data are linked to biologically relevant processes: rather, some may be due to technical variation or clinical covariates. As an example, the TCGA dataset is known to show strong batch effects linked to sequencing centers [182, 183, 44, 78]. When possible, the data should be normalized to remove the effects of technical variations. However, in some cases, the biological variable of interest is perfectly correlated with another covariate. Under these conditions, it is almost impossible to determine which changes are biologically relevant or not. Therefore, this ambiguity should be avoided, or, at least, control samples enabling the normalization of the data should be acquired. Biological characteristics of the sampled population may also induce variations in the data. These confounding variables, such as geographical location, age, and sex, greatly affect the composition of the microbiome, and the results obtained for a population with insufficient variability are unlikely to generalize to other settings [184]. To better assess the generalization of ML models, these should be tested under cross-study validation and researchers should use a dataset with samples from different contexts. Usually, this means including multiple datasets from different geographical locations [86, 83]. At each iteration, models are trained using some of these datasets and tested on the remaining ones, similarly to cross-validation. Many published works, however, lack this cross-study validation and no fair assessment of generalization is provided.

Overall, improper, non-generalizable assessments of model performance and the failure to correct batch effects have led to a lack of agreement in the results of several publications. As an example, in [94] and [37], the genus *Fusobacterium* was found to be more abundant in the gut microbiome of colorectal cancer patients; however, this genus was not identified as a possible biomarker in [35, 44] nor [86]. Zhou et al. [38] aggregated the proposed microbial biomarkers for colorectal cancer from various publications and found inconsistent results. Nevertheless, the overall findings show a difference between colorectal cancer and the healthy microbiome, although the microorganisms responsible may differ across studies. Besides the lack of agreement in microbial biomarkers, these shortcomings make it difficult to compare the impact of different ML algorithms and pre-processing approaches in model performance: an improvement may not be a result of a different methodology, but due to overfitting. Recently, some studies have focused on improving the generalization of ML models to detect colorectal cancer, with multiple datasets from different regions and a consistent processing pipeline [35, 44]. However, this should be extended to all works using ML for cancer characterization.

Finally, it is still unclear if bacterial biomarkers for cancers found through ML hold specificity in the presence of other diseases. Most of the published research on the tumor-specific microbiome compares the microbiome profiles of cancer patients to those of healthy individuals, excluding those with other diseases [94, 49]. Works developing models capable of identifying other diseases, such as type 2 diabetes or IBD, only include colorectal cancer [80, 97]. Excluding confounding diseases that may affect the microbiome is likely to make it simpler to find alterations linked to cancer. However, these will not be adequate for widespread clinical usage in case they lose their predictive power in the presence of comorbidities. For instance, type 2 diabetes increases the risk of colorectal cancer [185, 186], so employing biomarkers that cannot distinguish between the two diseases may yield incorrect results for a large fraction of the human population. Several publications have also found that type 2 diabetes is linked to disruptions in the gut microbiome [187, 188]. Therefore, it is unlikely that the patterns differentiating the tumor and healthy microbiomes will be similar to those found between cancer and type 2 diabetes. Further research into the differences between the human microbiota in the presence of tumors and other confounding diseases is required to better understand the limitations of using microbiome-derived biomarkers for cancer diagnosis and analysis.

The predictive capability of ML models applied to the tumor-specific microbiome is still unclear. For colorectal cancer identification from gut taxonomic profiles, classifiers can currently achieve a promising performance: the GRNN in [86] had an AUC of up to 85% in an independent test set, while the SVM in [94] achieved 81% accuracy while using microbial abundance features. These metrics are indicative of good performance, but not robust enough for widespread use yet. Care should also be taken regarding the evaluation metrics used, to provide representative estimates of model performance. Analyzing the performance of models is difficult due to the small datasets used for development [44, 69], especially for cancer types other than colorectal, which in some cases have as few as 23 patients [18]. Furthermore, models are often evaluated using a one-versus-rest approach, which exacerbates class imbalance. Evaluation metrics such as accuracy are

not adequate for highly imbalanced datasets. Even the AUC may hide the low precision or recall of a classifier, with a heavily skewed class distribution and a small number of positive instances [189]. For instance, despite having an average AUC of over 82%, in many cases reaching 100%, the ML classifiers to distinguish between cancer subtypes using the tumor microbiome in [44] had an Area Under the Precision-Recall Curve (AUPRC) of as little as 36%. To better understand the potential of ML approaches in identifying tumors based on the microbiome, these should be evaluated using metrics adequate for imbalanced datasets such as AUPRC and the F-score [189].

Improving the predictive power of ML models applied to cancer microbiome analysis will likely require the implementation of state-of-the-art methods such as NNs, increased regularization, and the development of larger bespoke datasets. Despite the increasing popularity of NNs for this purpose, many advances in deep ML architectures are still to be applied. For instance, Autoencoders or attention mechanisms [190] are seldom used, despite having achieved improved performance over other methods in biological tasks [191, 192, 193]. These approaches are expected to better model the complex relations between cancer and the human microbiome. However, when using these methods, special care should be taken to avoid overfitting. As such, the current preference for simpler ML models may be explained by a lack of data. Therefore, there is a need for larger cancer-related microbiome datasets. Ideally, these should be created with the goal of analyzing the microbiome, avoiding contamination, and keeping technical variations at a minimum. Increasing the number of available samples with which to train ML and deep learning models should lead to a better performance. Furthermore, new ways to constrain the latent representations of the input vectors are needed to avoid overfitting.

Besides enabling the usage of deep learning models and decreasing model variance, new cancer-related microbiome datasets could include additional variables other than taxonomic abundance profiles. Other modalities, such as functional profiles, were shown to improve the performance of ML models when coupled with microbial abundance data, as previously discussed. These other modalities are not limited to variables related to the microbiome; rather, information regarding the host can also be used. For instance, [125] combined microbiome abundance data and information on the host DNA content found in fecal samples to identify colorectal cancer. However, few approaches integrate multiple types of data – most likely due to the small number of available datasets with data other than abundance profiles. Doing so could not only lead to an improvement in the performance of estimators, but also to a better understanding of the human microbiome, how it is influenced by cancer, and its interactions with the host.

2.6 Summary

The human microbiome holds great potential for cancer diagnosis, prognosis, and therapies, offering a promising avenue for research. However, the biomarkers derived from this complex system present a challenge in their analysis. To address this, ML methods have emerged as crucial tools and have been widely used in recent breakthroughs to unravel cancer-microbiome relationships. Most approaches have focused on gut-derived taxonomic profiles, mainly for colorectal cancer

prediction. Despite promising, models developed for other cancer types do not yet achieve sufficient performance for widespread usage. Due to their increased interpretability, tree-based models are the most used; however, recent works suggest that deep learning architectures such as Autoencoders and multimodal NNs can lead to a better performance. As many published works do not ensure the generalization of their results, avoiding data leakage and confounding technical variables, it is difficult to make comparisons between methods and draw meaningful conclusions. Therefore, future research should be careful to implement adequate generalization assessments.

There is a need for customized microbiome datasets derived from cancer patients, ensuring meticulous processes to prevent contamination and account for batch effects. Additionally, further research is warranted to integrate microbial taxonomy profiles with other data types, including gene expression and host characteristics. This approach will be critical for a better understanding of the complex interplay between the microbiome and the human host in the context of cancer. It is also necessary to evaluate the precision of ML models for cancer identification in the presence of other diseases, to increase their generalization.

Chapter 3

The human microbiome as a biomarker for cancer diagnosis

The use of mbDNA for cancer diagnosis has been proposed by numerous authors [194, 195, 59, 44, 69]. In 2020, Poore et al. [44] determined the metagenomic composition of TCGA [45] samples, including blood samples. This allowed for a multi-cancer investigation of the links between the human microbiome and its cancer-related diseased states on a larger scale when compared to the small datasets previously available. In this same work, a GBT model was trained to distinguish between tumor types using mbDNA. The authors argue that these models successfully discriminated between most cancer types [44], a claim mainly supported by the high AUC achieved by the models; however, this metric may not be adequate to evaluate the classification performance in a dataset with an imbalanced number of samples per class, such as the one used. The reported values for the Area Under the Precision-Recall Curve, which takes class support into account, are much lower [196]. This suggests that the performance of ML models may be, as of yet, insufficient for the clinical usage of microbiome-based biomarkers in blood for cancer type prediction.

This Chapter aims to answer some of the questions raised by the latest publications on ML for cancer type prediction based on the human microbiome. The experimental setup used in [44] was followed broadly, taking advantage of the publicly available TCGA-derived microbiome profiles. The Chapter begins with a description of the dataset used in these experiments (Section 3.1). After this, the classification performance of three ML models for cancer type classification using blood microbiome profiles is compared in Section 3.2. Despite there being published comparisons between classifiers for single-type cancer diagnosis from microbiome profiles [122, 25, 123], this comparison is seldom made for multi-cancer classification and the blood microbiome. Then, in the same Section, a multiclass learning framework is compared against a one-versus-rest setting, assessing if the latter can be replaced by the less computationally demanding multiclass approach. Lastly, in Section 3.3, to investigate if the mbDNA is indicative of cancer type with comparable predictive power to the tumor microbiome, the best classifier according to the previous experiments is trained with the microbiome profiles of either blood or tumor biopsy samples from the same patients.

3.1 Dataset

The Kraken-derived microbiome profiles from [44] were downloaded and used in the experiments described in this chapter. These profiles correspond to the relative abundance of microorganisms (virus and bacteria) in TCGA [45] samples, on the genus level. In [44], the authors proposed and implemented four *in silico* decontamination approaches. The profiles with the *most likely contaminants removed* decontamination level were used as input features in the experiments outlined here. This level of decontamination was chosen as it provides a middle-ground between no decontamination, which may lead to models relying on non-biologically relevant and non-generalizable variations in the microbiome, and the most stringent decontamination level, which could remove differentially abundant genera across tumor types. The microbial abundance profiles derived in [44] contain information about genera artificially added as contaminants. These spiked genera were removed from the dataset. This resulted in microbiome profiles with information about 1553 genera. The associated metadata, which aggregates metadata from the original TCGA [45] samples, was also downloaded from [44].

The microbiome data used in this work included the pre-processing done in [44]. Despite not having been implemented as part of this dissertation, this pre-processing pipeline is relevant to the methodology and affects the conclusions derived in this work. Therefore, an overview of the pre-processing steps employed to derive the microbiome profiles used in this Chapter is provided here. The raw read counts obtained with Kraken [197] were normalized using Trim Mean of M Values (TMM) [198] and transformed with *Voom* [74] to resemble microarray data. After this transformation, the genus abundances are expressed in log-counts per million (logCPM) and are thus relative abundances. Finally, variations in the data linked to technical covariates, namely sequencing center, were estimated and removed using Supervised Normalization [199].

To fairly compare the capability to predict the tumor type from microbiome profiles derived from tumor biopsies against that of profiles obtained using blood samples, *primary tumor* and *blood derived normal* samples were grouped into pairs from the same patient. To obtain these pairs of matched biopsy and blood samples, the profiles were grouped according to their TCGA case identifier. However, some cases have more than one blood or biopsy sample. For those cases, samples were first grouped on sequencing center; if there still were more than one blood and biopsy samples in a group, these were further associated according to sequencing strategy (either Whole Genome Sequencing (WGS) or RNA-Seq). Further ambiguous associations were resolved by taking a blood and tumor sample at random; blood samples with no matching biopsy samples, and vice versa, were discarded. This resulted in a total of 1648 tumor and 1648 blood-derived microbiome profiles, from 1515 patients. The number of pairs of samples per disease type is shown in Table 3.1.

Table 3.1: Number of matched blood and tumor biopsy sample pairs per cancer type in the dataset.

Disease Type	Number of pairs of matched samples
Bladder Urothelial Carcinoma	115
Brain Lower Grade Glioma	91
Breast Invasive Carcinoma	107
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	64
Colon Adenocarcinoma	103
Esophageal Carcinoma	47
Glioblastoma Multiforme	54
Head and Neck Squamous Cell Carcinoma	140
Kidney Chromophobe	9
Kidney Renal Clear Cell Carcinoma	5
Kidney Renal Papillary Cell Carcinoma	34
Liver Hepatocellular Carcinoma	32
Lung Adenocarcinoma	94
Lung Squamous Cell Carcinoma	18
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	7
Ovarian Serous Cystadenocarcinoma	55
Prostate Adenocarcinoma	121
Rectum Adenocarcinoma	55
Sarcoma	36
Skin Cutaneous Melanoma	18
Stomach Adenocarcinoma	104
Thyroid Carcinoma	134
Uterine Corpus Endometrial Carcinoma	156
Uveal Melanoma	50

3.2 A multiclass classifier for cancer type identification using blood-derived microbiome profiles

Most published works on the use of ML methods for microbiome-based cancer diagnosis focus on classifiers for a single cancer type [58, 30, 129]. Even when experimenting with multiple cancer types, models are often developed and tested independently for each type of tumor, in a one-versus-rest (OvR) approach [44, 132]. In this setting, K models for binary classification are trained, with the k -th classifier predicting the presence or absence of class c_k . These models are also evaluated independently. A classifier for multi-cancer diagnosis, predicting the presence of multiple cancer types at the same time, on the other hand, should be developed and evaluated under a multiclass framework, predicting the posterior of all classes simultaneously, given a sample $\mathbf{x}_n$. This results in a prediction vector $\hat{\mathbf{p}}_n = [\hat{p}_n^{(1)}, \dots, \hat{p}_n^{(K)}]$, with $\hat{p}_n^{(k)}$ being the estimated $Pr(c_k|\mathbf{x}_n)$, which sums up to one. The predicted class is the one with the greatest posterior probability.

A multiclass classifier for cancer identification is preferable to multiple binary models for at least three reasons. Firstly, in an OvR framework, a sample can be assigned multiple or no cancer types, thus making the predictions potentially ambiguous. This does not occur with a multiclass model. Secondly, OvR exacerbates class imbalances for individual classifiers, inducing bias towards the majority class if care is not taken. Finally, assuming a comparable complexity between ML models, the OvR approach is computationally more expensive and requires longer training times, as it entails training K classifiers. It is likely that some of the mechanisms influencing the composition of the microbiome are similar for more than one cancer type; thus, a single multiclass model may be sufficient to capture microbial signatures for multiple types of cancer.

In this section, three state-of-the-art ML classifiers for cancer type prediction from mbDNA are compared. This allowed for the selection of the best performing classifier in this task, which was used in all subsequent experiments. Then, multiclass models were compared against K individual binary classifiers for cancer type prediction using blood-derived microbiome profiles, to assess if these achieve comparable performance. If OvR classifiers fail to improve the classification quality over multiclass models for multi-cancer identification, the latter would be preferred due to the reduced training duration, reduced computational cost, and non-ambiguity.

3.2.1 Materials and methods

In this Section, three ML classifiers were used: Logistic Regression, Random Forest, and GBT. These ML models usually lead to the best results in cancer identification from the microbiome and have been described in Section 2.4 of this thesis. The Logistic Regression and Random Forest implementations in the *scikit-learn* version 1.1.1 [200] package for Python were used, while the GBT classifier employed was based on the XGBoost [156] version 1.7.4 Python implementation.

All models were evaluated using 5-fold cross-validation (CV): the dataset was randomly partitioned into five groups, preserving the global proportion of tumor types inside each group. Samples from the same patient were included in the same group to avoid data leakage. Iteratively, one

Table 3.2: Classifier hyperparameters and search spaces.

Model	Hyperparameter	Search space
Logistic Regression	L1 regularization parameter	$\{0.01, 0.001\}$
Random Forest	Maximum tree depth	$\{6, 8, 10, 15, 20\}$
GBT	Minimum decrease in loss to partition a node	$\{0.0, 0.1\}$
	L1 regularization parameter	$\{0, 0.1, 1\}$

group was used as the test subset $\mathcal{X}_{Test}$, while all others constituted the training subset $\mathcal{X}_{Train}$. The CV partition was repeated five times.

The tunable hyperparameters of each model are shown in Table 3.2. These were selected using grid search by maximizing the F1 score (see Section 3.2.2) in a validation subset created from the training subset following 5-fold CV. The parameters of the Logistic Regression classifier were updated using Stochastic Gradient Descent, implemented through the *SGDClassifier* class in the *scikit-learn* [200] package. The learning rate option was set to *optimal*, which uses an adaptative learning rate following the heuristic proposed in [201], and L1 regularization was employed. A validation subset was constructed with 10% of the samples in the training subset. Parameter updating was interrupted after five rounds without improvement of the criterion in this validation subset. The number of Decision Trees in the Random Forest classifier was set to 50, with nodes split according to the Gini criterion. The Decision Trees in the GBT classifier were grown to a depth of up to five levels. To decrease the variance of the model, each tree was trained with a random subset of 90% of the training data and 60% of the features. The Cross Entropy, given by Equation 3.1 was used as the loss function. The number of trees grown was set through early stopping: 10% of the training subset was used as a second validation subset and training was interrupted after five rounds of no decrease in the Cross Entropy in this second validation subset.

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K \left(y_n^{(k)} \log \hat{y}_n^{(k)} + (1 - y_n^{(k)}) \log (1 - \hat{y}_n^{(k)}) \right) \quad (3.1)$$

Before passing the data to the ML models, the mean and standard deviation of the feature distributions were computed from the training subset and used to transform features to zero mean and unit variance. Furthermore, the contribution of each sample of class c_k to the loss function of the classifiers was set to the inverse of the number of samples of class c_k in the training subset. This decreases the bias for underrepresented classes.

To compare the performance of ML models in predicting tumor types from microbiome profiles in a multiclass and one-versus-rest (OvR) setting, the classifiers were developed and evaluated under these two frameworks. The classifiers were trained to predict tumor type using the microbiome profiles derived from blood samples. In the multiclass setting, one model simultaneously predicted posterior probabilities for all tumor types given the genus-level abundances in a sample n , resulting in a prediction $\hat{p}_n = [Pr(c_1|\mathbf{x}_n), Pr(c_2|\mathbf{x}_n), \dots, Pr(c_K|\mathbf{x}_n)]$. In the OvR setting, K models were trained, each estimating the posterior for a tumor type c_k . The learning task for each

individual model is a binary classification task: to train model k , the target classes were transformed according to $y_n^{(k)} = I(y_n = c_k)$, where $I(\mathcal{A})$ is a step function returning 1 if the condition $\mathcal{A}$ is true, and 0 otherwise. Each model then returns an estimate of the posterior $\hat{p}_n^{(k)} = Pr(c_k | \mathbf{x}_n)$.

3.2.2 Metrics to evaluate classifier performance

The models developed in this Chapter were evaluated according to five metrics. Two of these metrics were Precision and Recall. Given a binary prediction $\hat{y}_n^{(k)} \in \{0, 1\}$ for sample n having tumor type c_k and the corresponding ground-truth target $y_n^{(k)} \in \{0, 1\}$, n is said to be: a true positive if $\hat{y}_n^{(k)} = 1 \wedge y_n^{(k)} = 1$; a true negative if $\hat{y}_n^{(k)} = 0 \wedge y_n^{(k)} = 0$; a false positive if $\hat{y}_n^{(k)} = 1 \wedge y_n^{(k)} = 0$; a false negative if $\hat{y}_n^{(k)} = 0 \wedge y_n^{(k)} = 1$. Let TP , TN , FP , FN be the number of true positives, true negatives, false positives, and false negatives, respectively. Precision for class k is given by Equation 3.2, while Recall is defined according to Equation 3.3 [202]. Another metric assessed was the F1 score, corresponding to the harmonic mean between Precision and Recall. The F1 score for class k is thus given by Equation 3.4 [203].

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.4)$$

The classification performance was also evaluated in terms of the Area Under the Precision-Recall Curve (AUPRC). The AUPRC for a class c_k is computed from the posterior probability estimate for that class, $Pr(c_k | \mathbf{x}_n)$. This probabilistic estimate could be converted to a binary prediction by thresholding at 0.5; however, choosing a different threshold would lead to different values of Precision and Recall. The AUPRC is threshold-independent and calculated as follows: firstly, iterate over thresholds $\tau \in [0, 1]$ and compute the Precision and Recall by converting posterior estimates to binary predictions at τ , i.e., $\hat{y}_n^{(k)} = 1$ if $\hat{p}_n^{(k)} \geq \tau$ and $\hat{y}_n^{(k)} = 0$ otherwise. Then, plot pairs of Precision and Recall values at the same threshold, with Recall on the x axis and Precision on the y axis. Finally, calculate the area under the curve [204]. The last evaluation metric used was the AUC, which is also threshold-independent. It is calculated in a similar fashion to AUPRC; however, the Recall at threshold τ is plotted in the y axis, with the x axis holding the corresponding False Positive Rate (FPR), as given by Equation 3.5 [205]. A perfect classification yields an AUC and AUPRC of 1, while a totally incorrect set of predictions leads to an AUC and AUPRC of 0. The AUC is commonly used to evaluate the performance of classifiers; nevertheless, the AUPRC and the F1 score were also analyzed as these provide a better representation of classification performance in highly imbalanced datasets than other metrics, including AUC [189].

$$FPR = \frac{FP}{FP + TN} \quad (3.5)$$

Table 3.3: Evaluation metrics for all evaluated models, in a multiclass and OvR setting. The average (μ) and standard deviation (σ) across the training/test splits of the dataset are shown. These are averaged across target classes and shown in percentages.

Strategy	Model	Subset		AUPRC	AUC	F1	Precision	Recall
Multiclass	Logistic Regression	Training	μ	68.9	95.6	65.4	64.0	74.1
			σ	8.6	1.6	7.6	8.8	7.3
		Test	μ	22.0	77.7	22.7	24.9	25.8
			σ	11.0	9.5	10.5	12.1	16.6
	Random Forest	Training	μ	87.1	97.9	77.6	77.2	84.1
			σ	2.6	0.5	1.7	1.6	1.3
		Test	μ	50.2	90.2	42.9	44.1	49.7
			σ	15.5	5.4	4.4	4.9	4.6
	Gradient Boosted Trees	Training	μ	100.0	100.0	100.0	100.0	100.0
			σ	0.0	0.0	0.0	0.0	0.0
		Test	μ	72.4	96.1	62.6	65.1	63.1
			σ	13.1	2.8	15.5	19.0	17.4
OvR	Logistic Regression	Training	μ	58.3	86.6	27.4	17.9	98.5
			σ	5.0	4.6	9.1	8.0	1.9
		Test	μ	48.1	77.5	18.2	11.0	82.0
			σ	7.3	6.8	5.8	4.5	16.3
	Random Forest	Training	μ	99.9	100.0	99.0	98.5	99.6
			σ	0.1	0.0	0.8	1.0	0.8
		Test	μ	67.0	93.8	33.4	61.8	25.3
			σ	15.7	5.3	11.1	19.1	9.6
	Gradient Boosted Trees	Training	μ	100.0	100.0	100.0	99.9	100.0
			σ	0.0	0.0	0.1	0.2	0.0
		Test	μ	73.6	95.2	60.9	73.6	54.5
			σ	13.5	5.1	15.9	18.6	17.1

When comparing models and training strategies, the statistical significance of differences in performance metrics was assessed using a Wilcoxon signed-rank test over the 25 training and test splits of the dataset.

3.2.3 Results and discussion

The training and test classification metrics are shown in Table 3.3. These results are expressed as an average across all training and test splits of the dataset, with the corresponding standard deviation. To aid interpretation, classification metrics were averaged across all cancer types.

The results in Table 3.3 show that classifiers with a high test AUC, such as GBT and Random Forest, also demonstrate low Precision and Recall. For instance, despite having a test AUC of

96.1% in the multiclass setting, suggesting a very high classification performance, the GBT classifier has a corresponding Recall of only 63.1%. This means that, on average, among 10 samples of patients with a given cancer type, the model fails to identify its presence in roughly four of them. This number contradicts the high-quality prediction suggested by the AUC values. Therefore, the F1 score, which combines the values for Precision and Recall, offers a more intuitive estimate of the performance of the classifiers.

This dissonance between the AUC and F1, Precision, and Recall, is likely a consequence of the imbalanced number of samples per class in the training set. On one hand, when converting multiclass probabilistic estimates to binary predictions, Precision and Recall account for the estimates for all classes, i.e., all $\hat{p}_n^{(k)}$, $\forall k \in \mathbb{N} : k \leq N$, by selecting the class with the highest posterior probability. The AUC, however, individually evaluates thresholds for the different classes; therefore, the optimal threshold for label assignment can differ between classes, making it easier to compensate estimates biased towards overrepresented classes. If the probability estimates were converted into binary predictions individually for each class, using the optimal threshold for each, this could lead to multiple classes assigned to the same sample, or samples being assigned no class, similarly to what occurs with OvR classifiers. This does not represent a real use-case of the classifiers evaluated here, as these should output only one predicted class per sample. On the other hand, the AUC is less sensitive to FP than the F1 score when there are few instances of the positive class, as is the case in this dataset. Under this imbalance, even if there are many FP relative to the number of TP, the FPR, used in the AUC, will be low, as the number of TN is likely much greater than the number of FP. This leads to high values for AUC. In this same situation, the value for Precision will be low, resulting in a low F1 score [206]. The AUPRC is, similarly to AUC, independent regarding the selected threshold for converting probabilistic into binary predictions, as previously mentioned. However, because the AUPRC and the F1 score are both based on Precision and Recall, the values for AUPRC are much closer to those of the F1 score than the values for AUC. In summary, this comparison between performance metrics demonstrates that the F1 score and the AUPRC offer quantifications of model performance which are intuitively more representative of the quality of the classification results. Therefore, most of the analyses performed in this work will focus on these two metrics.

The results presented in Table 3.3 show that the GBT model outperformed the Random Forest and Logistic Regression classifiers in both the multiclass and OvR settings, according to all the evaluation metrics assessed (all $p < 10^{-5}$). The Random Forest classifier in turn outperformed Logistic Regression (all $p < 10^{-5}$). This difference in the predictive performance of classifiers can be seen in Figure 3.1, which contains the box plots for the test evaluation metrics obtained for different classes and training/test splits of the dataset, in the multiclass setting. It is not surprising that the Logistic Regression model achieved the worst classification performance. This was the simplest model evaluated and, contrary to Random Forest and GBT, is a linear model. This comparably reduced complexity of Logistic Regression also explains its poor fitting capacity: it achieved a training AUPRC of 68.9%, much less than the 87.1% for the Random Forest classifier and 100.0% for Gradient Boosted Trees.

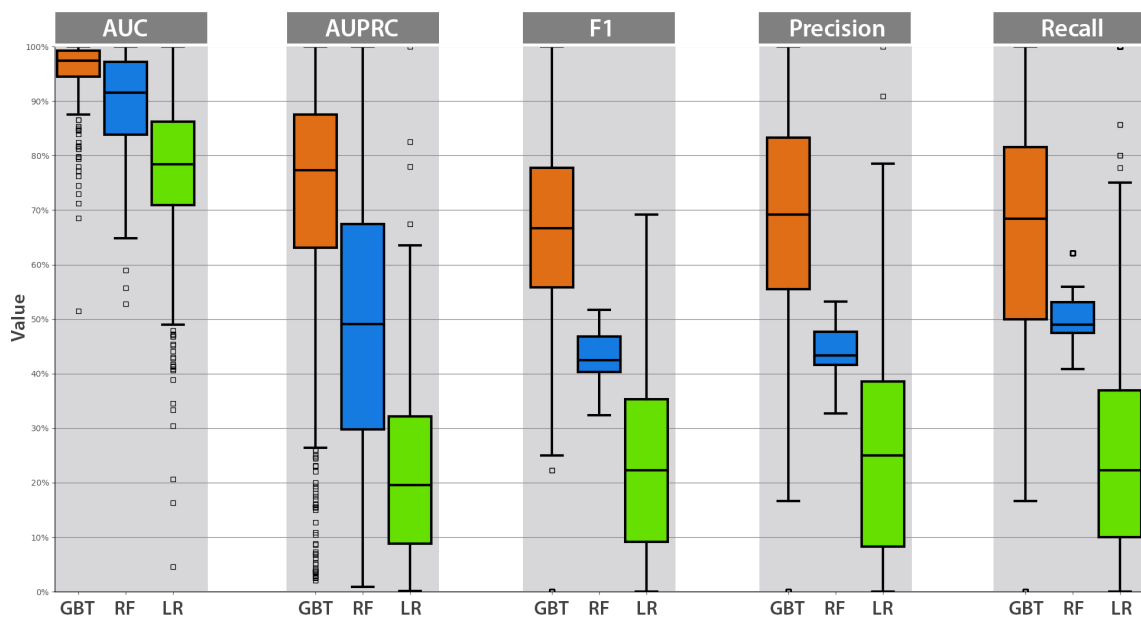


Figure 3.1: Boxplots of test classification metrics per class and training/test split of the dataset for multiclass classifiers. GBT – Gradient Boosted Trees. RF – Random Forest. LR – Logistic Regression.

The results obtained with the multiclass GBT classifier are close to those resulting from similar models in an OvR setting. The AUPRC of individual OvR models is greater than the one obtained with a multiclass classifier ($p = 0.004$); however, this gap in AUPRC is small, of 1.2% on average. Furthermore, there are no significant differences between the F1 score in both settings ($p = 0.43$). This similarity in classification performance between a multiclass and OvR classifiers is not seen for the Logistic Regression and Random Forest classifiers, which both exhibit a greater AUPRC in an OvR framework. However, the F1 score of these models under multiclass classification is greater than in OvR (all $p < 10^{-5}$). Note that Precision and Recall are calculated at a 50% probability threshold, which may not be ideal under the class imbalance seen in this dataset.

The results observed with the best-performing multiclass classifier (GBT) show that a multiclass classifier is capable of multi-cancer identification from blood-derived microbial abundance profiles with similar performance to individual models for single-cancer type prediction. Training this model comes at a lower computational cost than doing so for multiple individual classifiers and provides unambiguous predictions of cancer type. Following the results outlined in this section, all experiments involving cancer type classification from microbiome profiles described from now on were performed with a GBT classifier, which was best performing model, in a multiclass setting.

3.3 Tumor and blood microbial signatures for cancer type identification

There is mounting evidence that the tumor microbiome is distinct from that of healthy tissues [122, 58, 131, 48] and that blood-derived microbiome data contains changes associated with specific types of cancer [67, 59, 27, 44, 68, 69]. Such changes could allow for a non-invasive screening test, suitable for multiple types of cancer. Signatures in mbDNA were shown to discriminate between cancer types in patients for which circulating tumor DNA assays failed, suggesting their suitability for cancer type diagnosis [44]. However, it is necessary to compare the predictive power of mbDNA signatures to those found in the tumor microbiome, to better position the potential of liquid biopsy screening tests based on microbial biomarkers for cancer diagnosis. Although ML classifiers for cancer type prediction from both blood and tumor-derived microbial abundance profiles were developed and tested in [44], these were trained with an inconsistent number of samples, from slightly different cohorts. In the following experiments, microbiome data from blood and tumor biopsies from the same patients were used to train ML classifiers to predict cancer types. This allowed for a fair comparison of mbDNA against the tumor microbiome for cancer type diagnosis.

3.3.1 Materials and methods

To assess if the tumor and blood-derived microbiome profiles differ according to and are indicative of cancer type, multiclass GBT classifiers were trained to predict the tumor type of a pair of matched samples, using the microbiome profiles derived from either the tumor biopsy or the blood samples. Model training and evaluation followed the methodology outlined in Sections 3.2.1 and 3.2.2.

The training was repeated with and without oversampling with Synthetic Minority Oversampling Technique (SMOTE) [207]. This oversampling technique decreases model variance and class imbalance by creating synthetic samples. When oversampling a class, a sample of that class is chosen at random. Then, one of its k -nearest neighbors is randomly selected. A synthetic data point $\mathbf{x}_{New}$ is generated according to Equation 3.6, in which $\mathbf{x}_1$ and $\mathbf{x}_2$ are the selected samples, and λ is a random number in $[0, 1]$ [207]. The SMOTE implementation available in the Python package *imbalanced-learn* version 0.9.1 [208] was used. The Euclidean distance was used to find the k -nearest neighbors of a sample, with $k = 2$. The number of generated data points was such that all classes had the same number of samples in the training subset after oversampling.

$$\mathbf{x}_{New} = \mathbf{x}_1 + \lambda (\mathbf{x}_2 - \mathbf{x}_1) \quad (3.6)$$

3.3.2 Results and discussion

Oversampling the microbiome profiles with SMOTE improved the classification performance of Gradient Boosted Trees to predict cancer types from both tumor and blood-derived microbiome

Table 3.4: Evaluation metrics for GBT classifiers with and without oversampling, and using either blood-derived or tumor biopsy microbial abundance profiles. The average (μ) and standard deviation (σ) across the training/test splits of the dataset are shown. These are also averaged across target classes and shown in percentages.

Oversampling	Sample Type	Subset		AUPRC	AUC	F1	Precision	Recall
None	Blood	Training	μ	100.0	100.0	100.0	100.0	100.0
			σ	0.0	0.0	0.0	0.0	0.0
		Test	μ	72.4	96.1	62.6	65.1	63.1
			σ	13.1	2.8	15.5	19.0	17.4
	Tumor Biopsy	Training	μ	100.0	100.0	100.0	100.0	100.0
			σ	0.0	0.0	0.0	0.0	0.0
Test		μ	75.4	96.7	65.9	68.0	66.8	
		σ	12.7	2.9	14.5	17.8	17.0	
SMOTE	Blood	Training	μ	100.0	100.0	100.0	100.0	100.0
			σ	0.0	0.0	0.0	0.0	0.0
		Test	μ	73.1	96.4	64.6	67.1	65.3
			σ	13.3	2.7	15.0	18.3	17.6
	Tumor Biopsy	Training	μ	100.0	100.0	100.0	100.0	100.0
			σ	0.0	0.0	0.0	0.0	0.0
		Test	μ	76.6	96.9	67.3	69.0	68.4
			σ	12.0	2.4	14.6	17.9	16.9

profiles, according to all metrics (all $p < 10^{-5}$). These metrics are shown in Table 3.4, averaged across all tumor types. As the classifiers show high variance, given the large gap between the performance in the training and test subsets, data augmentation most likely improved the classification by avoiding an overfit to the training subset and thus reducing this variance [167].

The test AUPRC for the best performing model (GBT with SMOTE) is close to that reported in [44] (81.6%, versus the 76.6% obtained here). In [44], GBTs were also used, although in an OvR framework. Nevertheless, as the comparison described in this thesis only included matched blood and biopsy samples from the same patient, it used fewer tumor biopsy samples for training when compared to [44]. The performance of classifiers using blood-derived microbiome profiles is closer, as both sets of models were trained with a similar number of samples: [44] reported an average AUPRC of 77.3%, excluding tumor types with fewer than 20 samples in the dataset; the classifiers evaluated in this section achieved a corresponding AUPRC of 75.7% when applying SMOTE. Therefore, the classifiers developed in this work achieve comparable performance to state-of-the-art methods and can be used to derive meaningful conclusions on latter comparisons.

As previously stated, despite the high AUC of the best performing classifier, its AUPRC suggests that the classification performance of current models for cancer diagnosis from the microbiome is not sufficient for clinical usage. However, it is clear that all the evaluated models are capable of better than random predictions, as seen in the Receiver Operating Characteristic and Precision-Recall curves for classifiers using the blood-derived microbiome profiles with SMOTE in Figures 3.2 and 3.3: AUC values are greater than 50% and the AUPRC is generally greater than the ratio of samples of each cancer type. Thus, both the tumor and the blood-derived microbiome profiles exhibit changes reflective of the cancer type of a patient, as extensively reported [47, 88, 18, 41, 44, 69].

By associating blood and tumor biopsy samples from the same patient, it was possible to compare the potential of blood-derived and tumor microbiome profiles for multi-cancer diagnosis. Although the classifiers trained with blood-derived microbiome data are capable of better than random prediction of tumor types, their performance lags behind that of models using the tumor microbiome. When SMOTE was used, models using tumor profiles achieved a greater AUPRC than classifiers using blood data ($p < 10^{-5}$). The same was observed for AUC ($p = 0.001$), the F1 score ($p = 0.00009$), Precision ($p = 0.006$), and Recall ($p = 0.003$). Thus, the blood-derived metagenome is less indicative of cancer type than the tumor metagenome. This impacts the potential of mbDNA for cancer diagnosis, as it imposes a tradeoff between non-invasiveness and accuracy.

Tumor types are strongly correlated with their location in the body, and the human microbial content is different between organs [5]. Thus, the biopsy-derived microbiome profiles may preserve differences related to the underlying sampling location. This is less likely to happen for blood-derived profiles, as the sampling location is identical across all types of tumors, which might explain their inferior performance in predicting cancer types. Another factor likely impairing the discriminatory power of the blood metagenome is related to the abundance of microbial DNA. Blood samples contain less biomass and are thus more susceptible to contamination and technical

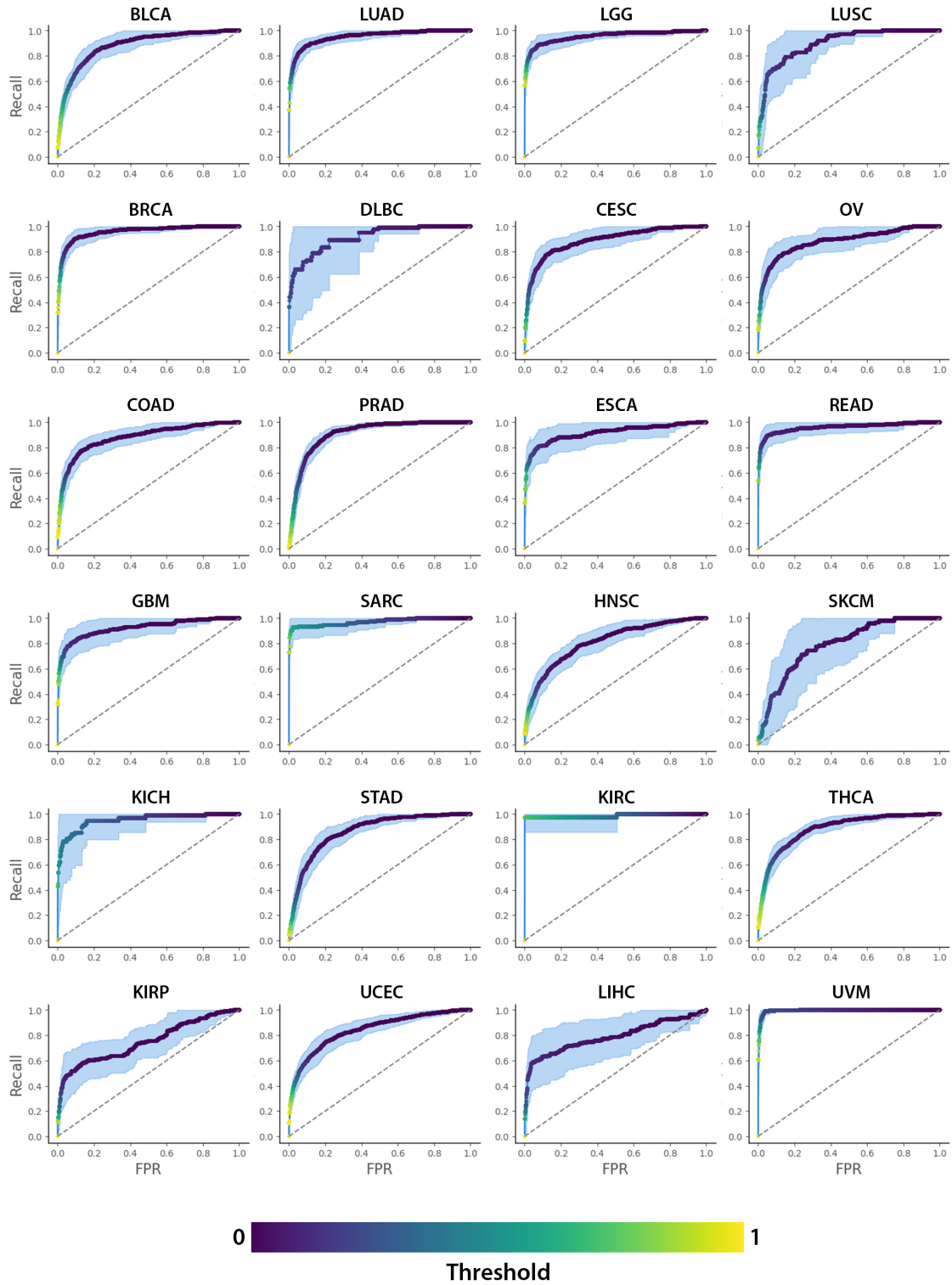


Figure 3.2: ROC curves per target class for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles. Lines show the average values over the 25 training/test dataset splits. The light blue area represents the standard deviation around the average curve. The dashed line represents the ROC for a random classifier.

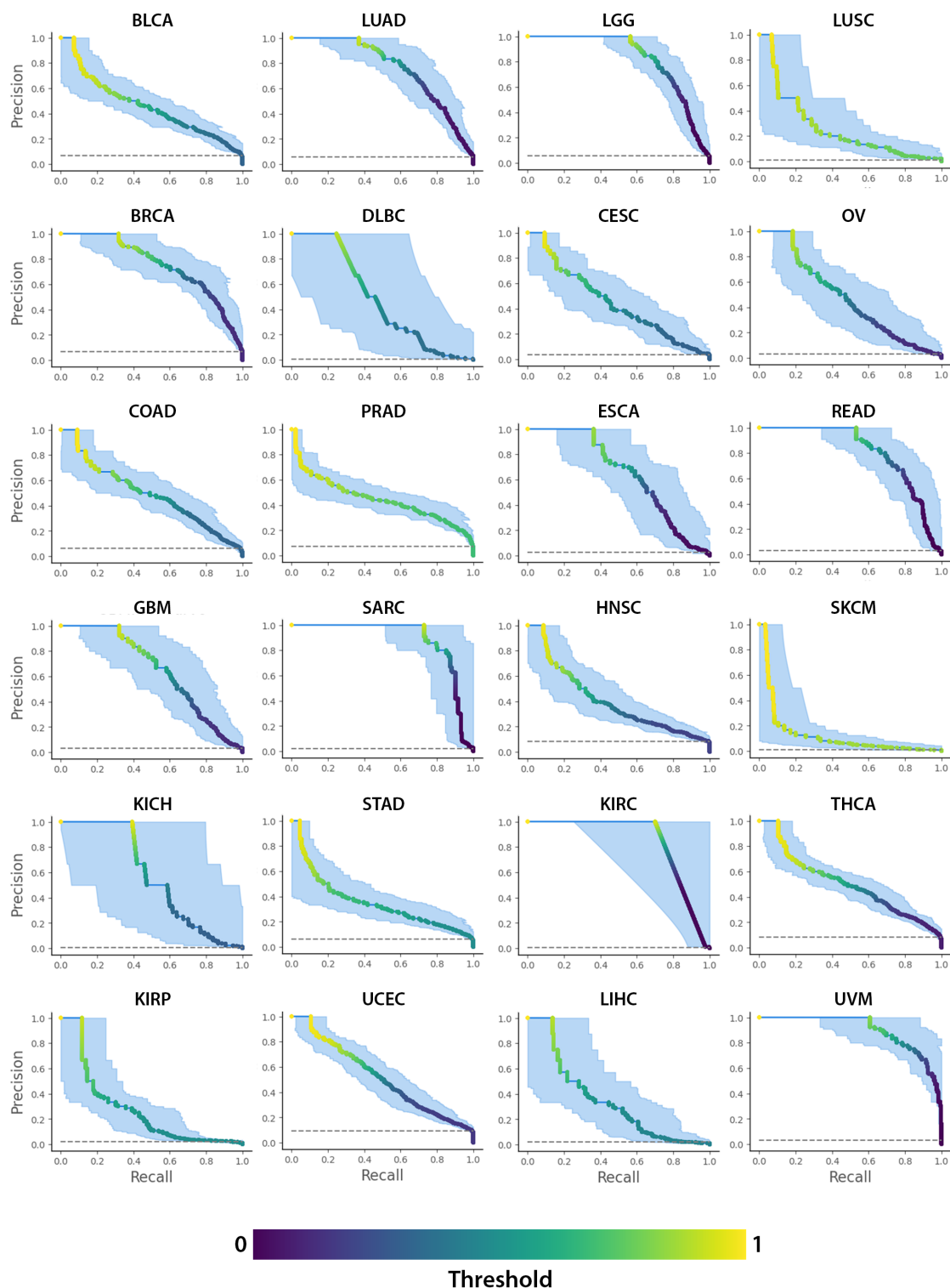


Figure 3.3: Precision-Recall curves per target class for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles. Lines show the average values over the 25 training/test dataset splits. The light blue area represents the standard deviation around the average curve. The dashed line represents the Precision-Recall curve for a random classifier.

Table 3.5: Average (μ) and standard deviation (σ) of test AUPRC per target class for GBT classifiers with SMOTE data augmentation using either tumor biopsy microbiome profiles or blood-derived abundances. Values shown in percentages and cancer types are sorted in descending order of average AUPRC with tumor microbiome profiles.

Cancer Type	Abbreviation	Tumor Biopsy		Blood	
		μ	σ	μ	σ
Uveal Melanoma	UVM	97.4	3.0	98.3	3.0
Prostate Adenocarcinoma	PRAD	90.9	4.4	85.2	7.0
Sarcoma	SARC	90.5	10.3	81.6	16.4
Rectum Adenocarcinoma	READ	89.6	7.1	86.6	10.6
Colon Adenocarcinoma	COAD	87.0	8.3	86.8	7.2
Kidney Renal Clear Cell Carcinoma	KIRC	86.0	22.9	73.4	32.8
Breast Invasive Carcinoma	BRCA	85.8	6.8	89.6	5.5
Brain Lower Grade Glioma	LGG	84.8	7.4	86.4	6.4
Lung Adenocarcinoma	LUAD	84.4	6.6	82.5	7.3
Kidney Renal Papillary Cell Carcinoma	KIRP	81.7	10.6	61.5	17.6
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	CESC	81.1	12.1	73.7	13.1
Head and Neck Squamous Cell Carcinoma	HNSC	81.1	9.9	76.5	7.5
Glioblastoma Multiforme	GBM	78.3	10.5	77.2	11.2
Kidney Chromophobe	KICH	77.3	26.7	67.2	32.1
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	76.2	24.1	72.9	26.6
Uterine Corpus Endometrial Carcinoma	UCEC	75.6	7.5	77.9	6.5
Esophageal Carcinoma	ESCA	75.6	14.5	75.6	13.0
Ovarian Serous Cystadenocarcinoma	OV	73.4	12.4	74.3	12.8
Thyroid Carcinoma	THCA	73.3	8.6	69.7	9.0
Liver Hepatocellular Carcinoma	LIHC	71.1	13.8	65.6	11.4
Stomach Adenocarcinoma	STAD	70.0	8.5	70.6	8.0
Bladder Urothelial Carcinoma	BLCA	65.7	9.9	61.4	8.5
Lung Squamous Cell Carcinoma	LUSC	38.8	20.9	26.2	19.6
Skin Cutaneous Melanoma	SKCM	23.1	20.3	33.9	26.8

variations than tumor tissue [53]. Although *in silico* decontamination was applied to the data used in this experiment, this is likely not sufficient to remove all contaminants and technical artifacts from the blood-derived microbial abundance profiles. Thus, the signal-to-noise ratio regarding the association between tumor type and the microbiome should be greater in primary tumor tissue than in blood. Lastly, some, if not most, of the mbDNA in diseased patients may have originated in the gut, resulting from a permeable intestinal barrier [209]. This phenomenon, known as “leaky gut”, is associated with diseased states, including colorectal cancer [210], and likely contributes to the noise in blood-derived profiles. Even the genera that is found in both tumor and blood-derived microbiome profiles does not necessarily translocate from the tumor tissue to blood. Therefore, the blood-derived microbial signatures of cancer types may not be identical to the ones based on the tumor microbiome.

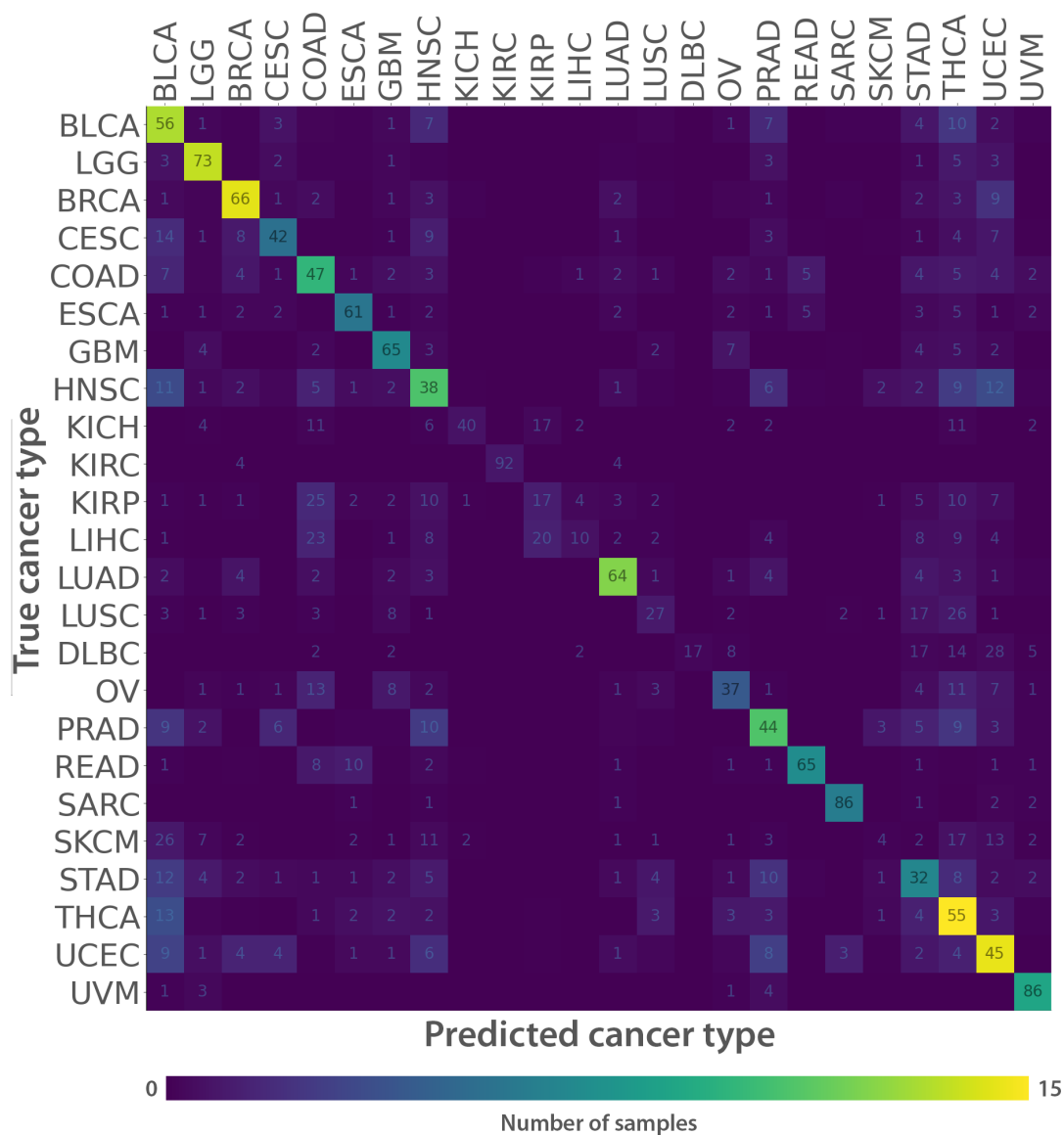


Figure 3.4: Test confusion matrix for the GBT classifier with SMOTE data augmentation based on blood-derived microbiome profiles. Cells are colored according to their absolute number of samples. Numbers represent the respective relative values, in percentage, with respect to the total number of samples in a row (number of samples per true class). Zeros are omitted. These results are an average over the 25 training/test splits of the dataset.

The classification performance of the evaluated models is heterogeneous across target classes. Table 3.5 contains the test AUPRC per cancer type of GBT trained with blood-derived and tumor microbiome profiles after SMOTE augmentation, while Figure 3.4 holds the test confusion matrix for the same model using mbDNA. The classifiers discriminate better uveal melanoma patients from those with other cancer types (average AUPRC of 97.4% using tumor profiles). Colon and rectum adenocarcinomas are also among the classes with better classification performance (average AUPRC of 87.0% and 89.6% with tumor profiles, respectively). Colorectal cancer is known to impact the composition of the gut microbiome [83, 94, 211] and a differential composition between the tumor microbiomes of colorectal adenomas and carcinomas has also been reported [58]. These results corroborate such altered microbiome profiles in colorectal cancer patients. Skin cutaneous melanoma and lung squamous cell carcinoma (average AUPRC of 23.1% and 38.8% with tumor profiles, respectively) have a below-average classification performance, which is likely a consequence of the small number of positive instances in the dataset (18 for both cancer types). Generally, classifiers trained with tumor microbiome data achieve a slightly improved or identical performance when compared to models using blood-derived profiles. Tumor types with a larger gap in AUPRC are those with a small number of instances in the dataset (see also Table 3.1), such as lung squamous cell carcinomas, skin cutaneous melanomas, kidney chromophobes, and kidney renal papillary cell carcinomas.

Overall, the results presented in this section demonstrate that both blood-derived and tumor microbiome abundance profiles have changes that discriminate between most cancer types. Therefore, microbial biomarkers derived from liquid biopsies could provide a non-invasive method for cancer type diagnosis. However, current ML approaches cannot capture microbial signatures of sufficient discriminatory power for clinical usage. Furthermore, the performance of blood-derived microbial biomarkers lags behind that of tumor microbial profiles.

A limitation of this study stems from the dataset used: TCGA blood samples were obtained through whole blood or buffy-coat extraction [212, 44], whereas liquid biopsy usually entails the analysis of plasma alone [213]. The microbial abundance profiles in the dataset may be partially based on mbDNA in white blood cells and other blood components, and it is unclear if the microbiome profiles acquired from plasma alone would follow the distributions seen here. Therefore, the microbial signatures found by blood-based models in this experiment may not be present in plasma-derived data, leading to an overestimation of the potential of microbial biomarkers obtained through liquid biopsy for cancer type identification. Furthermore, it has been suggested that many of the genera found in the TCGA-derived microbiome profiles used in this work are a consequence of the erroneous taxonomical classification of genetic material [196]. If this is true, the signatures uncovered by the classifiers may not be based on real data, making the perceived association between cancer types and the tumor and blood microbiome a result of spurious changes related to technical phenomena.

3.4 Summary

In this Chapter, ML models for cancer identification from blood-derived microbial abundance data were compared. The GBT models achieved improved classification performance when compared to Random Forest and Logistic Regression. Based on these findings, a ML classifier for multi-cancer identification from blood-derived microbial abundance data was developed and evaluated. This model, based on GBT, achieved comparable performance to multiple binary classifiers identifying a single cancer type each. This multiclass model allows for shorter training times and a reduced computational cost when compared to single-cancer type classifiers. Furthermore, it provides non-ambiguous predictions of disease types.

When training ML classifiers for cancer type prediction using matched blood and tumor biopsy samples from the same patients, those employing abundance profiles in tumor tissue achieved a slightly better classification performance. The results show that the tumor and blood-derived microbial profiles exhibit differences associated with cancer type, which can be captured by ML models; however, state-of-the-art ML models using tumor or blood-derived abundance profiles do not yet provide sufficient classification performance and generalization for widespread clinical usage.

Despite potentially providing a non-invasive means of diagnosis, it is unclear why blood-derived microbiome profiles are associated with a decreased diagnostic power when compared to tumor profiles, as little is known about the origins of the microbial genetic content found in blood. Low amounts of genetic material in blood samples and a compromised intestinal barrier may contribute to changes in blood microbiome profiles not directly linked to cancer, thus making them less indicative of cancer type. It is also possible that the ML models using tumor profiles rely on differences between the normal microbiome in different organs to identify cancer types, irrespectively of the presence or absence of tumor. Motivated by these questions, the similarities between blood-derived and tumor microbiome abundance profiles will be investigated in the next Chapter.

Chapter 4

Uncovering the association between the tumor and blood microbiomes

Up until recently, human blood was thought to be sterile. However, recent studies have proposed the existence of a healthy blood microbiome [214, 215], although this has been highly contested. In a large study, most healthy individuals were found to harbor no microbial species in their blood. Furthermore, when such microbiome is present, no common structure was found among the communities of microorganisms [55]. Being low-biomass, blood samples are prone to contamination, leading to an overestimation of their microbial contents; also, it is unclear if the bacteria detected in blood are viable or not [53]. However, under diseased states, microorganisms may enter the bloodstream. These microorganisms in blood were found to be indicative of the mortality of cardiovascular diseases [54] and cancer [44], as previously discussed.

Most microorganisms found in blood probably originate from other body sites, namely the skin, mouth, and gut [54, 55]. Some of these organisms may enter the tumor tissues, contributing to the tumor microbiome [216]. Other means for bacterial colonization of tumor tissues have also been proposed, including the migration from adjacent healthy tissue and mucosal sites [50, 216]. However, little is known about the dynamics of the blood and tumor microbiomes, and how these two relate to one another. It is possible that the tumor and blood microbiomes are highly similar, and that it is this similarity that makes the blood microbiome indicative of cancer type, as seen in the previous Chapter. It is also possible that the blood microbial content is dictated by permeable mucosal barriers, with different cancer types modulating this permeability and the composition of the gut and oral microbiomes in different ways, thus leading to an association between the blood microbiome and cancer types. In this latter scenario, the blood and tumor microbiomes are not necessarily identical.

Understanding the similarities and differences between the tumor and blood microbiomes is key to validating mbDNA as a biomarker for cancer diagnosis and characterization. It is important to discover the origins of the microorganisms found in blood and assess if the association between the blood microbiome and cancer types is caused by biologically relevant phenomena or technical artifacts. However, no analysis of the similarities between the blood and tumor microbial profiles

of the same patients has been published. In this Chapter, in Section 4.3, blood-derived and tumor microbial abundance profiles from the same patient are compared, allowing the quantification of their similarity and the identification of over and underrepresented genera in blood. Furthermore, in Section 4.4, ML regressors trained to predict the tumor microbiome profiles from matched blood-derived microbial abundances are analyzed, exploring different hypothesis regarding the origins and mobility of microorganisms across the blood-tumor interface.

4.1 Dataset

In the experiments described in this Chapter, the blood-derived and tumor microbial abundance profiles obtained with Kraken in [44] were used. Three levels of *in silico* decontamination were included: no decontamination, *likely contaminants removed*, and *most stringent decontamination*. After removing spiked genera, raw, non-decontaminated microbiome profiles contained information about 1993 genera. Profiles with *likely contaminants removed* had 1553 genera. Finally, the *most stringent decontamination* approach led to profiles with information for 173 genera. Pairs of blood-derived and tumor profiles from the same patient were grouped as described in Section 3.1. The TCGA metadata associated with the microbiome profiles was also downloaded from [44].

Table 4.1: Number of genera (corresponding to features) in both tumor and blood microbiome profiles with each level of decontamination.

Level of decontamination	Number of genera
No decontamination	1993
<i>Likely contaminants removed</i>	1553
<i>Most stringent decontamination</i>	173

Two alternative pre-processing pipelines were used. One yielded relative abundances per genus, and was implemented in [44] as described in Section 3.1. These relative abundances are expressed in logCPM and transformed to resemble microarray data [74, 44]. This transformation may aid the analysis of differential abundances by approximating the microbiome profiles to a Normal distribution [217]; however, representing microbial content in relative abundance makes analyzing the differences between the tumor and blood microbiomes challenging, as a change in the absolute abundance of one genus leads to a change in relative abundance for all other genera. Therefore, a pre-processing pipeline transforming the Kraken-derived raw read counts in absolute abundances was also implemented and is described in the next Section.

4.2 Pre-processing microbiome abundance data

This section describes the pre-processing pipeline applied to raw read counts to derive absolute abundance profiles.

4.2.1 Removing batch effects

Not all differential abundances in microbiome profiles are causally related to biologically relevant variables, such as tumor or sample type. Rather, some variability is due to sampling bias, biological covariates, and technical variation – the latter is also known as batch effects. The TCGA dataset is known to show strong batch effects associated with sequencing centers [183, 78]. Therefore, to achieve a greater generalization of the models and conclusions presented in this work, batch effects were removed using *ComBat-seq* [79] on the *sva* package [218] version 3.48.0 for the R programming language [219]. *ComBat-seq* uses a negative binomial regression model to remove batch effects and is thus suitable for metagenomic data [79]. Batches were set according to the sequencing center of each sample. Four covariates were defined, maintaining their associated variability in the adjusted profiles: disease type, sample type (blood or tumor biopsy), gender, and ethnicity.

To assess if the influence of batch effects was removed from the dataset through the pre-processing steps described in the previous section, the microbiome profiles were projected into two dimensions and visually analyzed. Before dimensionality reduction, the genera distributions were transformed to zero mean and unitary standard deviation. Two dimensionality reduction methods were used: Uniform Manifold Approximation and Projection (UMAP) [220] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [221]. For t-SNE, the implementation in the *scikit-learn* version 1.1.1 [200] package for Python was used; the perplexity parameter was set to 100, and an early exaggeration of 12 was used. The minimum point distance in UMAP was set to 0.1 and the number of neighbors to 15. Points, corresponding to samples, were colored according to their sequencing center, experimental strategy (WGS or RNA-seq), sequencing platform, tumor type, and sample type (blood or tumor biopsy). These plots for profiles before and after batch effect removal are shown in Figures 4.1 and 4.2, respectively.

The two-dimensional representations of the data before removing batch effects (see Figure 4.1) have four clusters, roughly corresponding to different sequencing centers. One is almost exclusively made up of samples from the Washington University School of Medicine; another is linked to the Broad Institute of MIT and Harvard; one other cluster is associated with samples from the Baylor College of Medicine. The remaining cluster has samples from MD Anderson Cancer Center, Harvard Medical School, and the University of North Carolina. None of the other evaluated covariates showed association with the clusters. After pre-processing the data, the scatter plots show two clusters (see Figure 4.2), with samples from the Washington University School of Medicine, the Broad Institute, and the Baylor College of Medicine now forming one community. No association between these two clusters and any covariates available in the metadata was found.

These results indicate that most of the variance in the data linked to technical variations, namely sequencing centers, was removed. The pre-processed data has similar distributions conditioned on sequencing center, something that did not occur before the removal of batch effects. Thus, the batch effect removal strategy was successful, ensuring the generalizability of the conclusions drawn in this work.

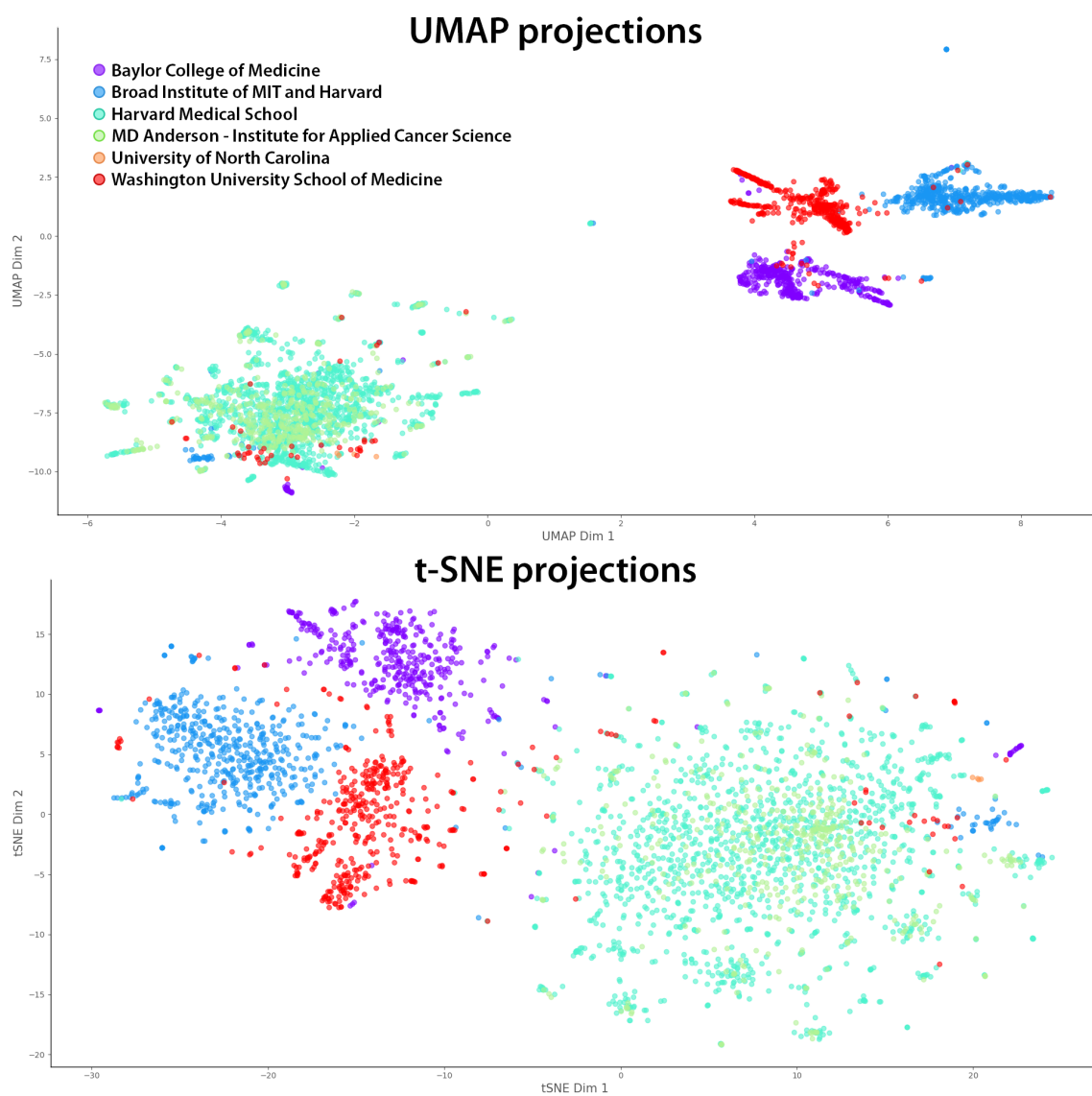


Figure 4.1: UMAP (top) and t-SNE (bottom) 2D projections of absolute abundance profiles before batch effect removal. Samples are colored according to their sequencing center.

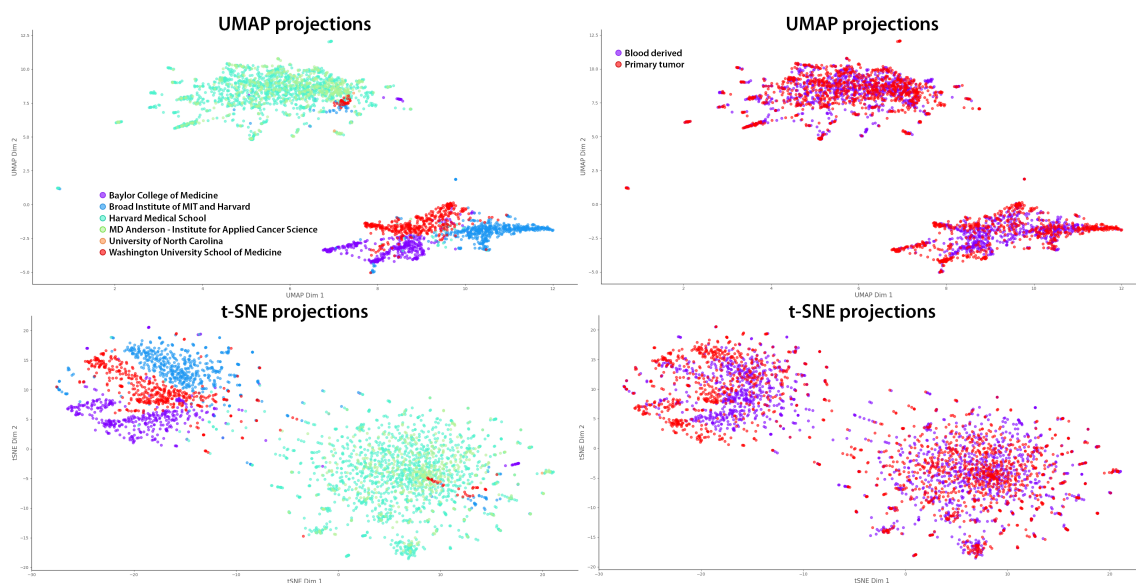


Figure 4.2: UMAP (top) and t-SNE (bottom) 2D projections of absolute abundance profiles after batch effect removal. Samples are colored according to their sequencing center (left) and sample type (right).

4.2.2 Normalizing microbiome profiles

The microbiome profiles should be adjusted according to library size (i.e., the total number of sequenced reads): as an example, take two samples, A and B, where A was sequenced with half the depth¹ of B. Even if the underlying microbiome profiles are equal, a genus in A will have half the absolute abundance when compared to that same genus B. Converting the profiles to relative abundances solves this issue; however, not only does it make the association between the tumor and blood microbiome more difficult to understand, as discussed previously, but it also decreases the magnitude of differentially abundant genera, making them more difficult to find [198]. To avoid this, the microbiome profiles after batch effects removal were normalized using TMM [198], with the implementation available in the Python package *conorm* version 1.2.0 [223]. Trim Mean of M Values discards genera with abnormally high or low abundances before estimating the total number of reads in each sample, thus preserving the signal of differentially abundant genera [198]. The default parameters of 0.3 and 0.05 quantile cutoff for log-fold changes and log-magnitude changes, respectively, were used.

After TMM normalization, the three decontamination levels (no decontamination, *likely contaminants removed*, and *most stringent decontamination*) were implemented by keeping only the genera present in the corresponding decontaminated profiles processed in [44].

¹The sequencing depth refers to the average number of times a nucleotide in the genetic content of a sample was sequenced. This affects the total number of produced reads and, consequently, the raw read counts per genus [222].

4.3 Assessing the similarity between blood-derived and tumor microbiome profiles

In this Section, the similarities and differences between the tumor and blood-derived microbiome profiles from the same patient are discussed. Genera with differential abundance between both types of samples are also identified.

4.3.1 Materials and methods

To evaluate the similarities between the blood-derived microbiome profiles and those from tumor tissues, the correlation between pairs of matched profiles was compared against the correlation between blood-derived and tumor profiles from different patients. The pairwise correlation between all tumor and blood-derived profiles was evaluated according to Pearson's r and these pairs were divided into two groups: matched pairs from the same patient, as outlined in Section 3.1, and non-matched pairs. The locations of the distributions of Pearson's r values for each group were compared using one-sided Mann-Whitney U tests [99]. The hypothesis that the correlation between blood-derived and tumor microbial abundance profiles from the same patient is greater than that of pairs of profiles from different patients was accepted at a 95% confidence threshold. This comparison was repeated for every cancer type, including only patients with that disease type, with relative and absolute abundances. The profiles with the *likely contaminants removed* level of decontamination were used.

The similarity between the distributions of each genus in blood-derived and tumor microbial abundance profiles was also quantified using Pearson's r . This comparison was done both for relative and absolute abundances, with the *likely contaminants removed* level of decontamination. Furthermore, to assess if any genus was differentially abundant in tumor and blood-derived profiles, the distributions of each genus in both samples were compared using Wilcoxon signed-rank tests [107]. This comparison was used to create a volcano plot, expressing the fold-change in abundance and p -value obtained for each genus. Differential abundance of a genus was considered statistically significant if the p -value of the Wilcoxon signed-rank test was less than 0.05. The relative abundance profiles derived in [44] are more adequate to study differential abundance than absolute abundance profiles, as the first were approximated to a Normal distribution in a way that preserves the mean-variance relationship, thus making statistical tests more robust [74, 217]. Therefore, this last comparison was done with the relative abundance profiles and *likely contaminants removed*.

4.3.2 Results and discussion

Globally, blood-derived and tumor microbiome profiles from the same patient had a Pearson's r greater than samples from different patients, both for relative and absolute abundances (both $p < 10^{-5}$). With relative abundances, the median value for pairs of samples from the same patient was 94.9%, compared to 92.9% for non-matched sample pairs. For absolute abundances, the

Table 4.2: Average Pearson's r between matched and non-matched blood-derived and tumor microbiome profiles, grouped by cancer type, together with p -values from Mann-Whitney U tests. Values for Pearson's r are shown in percentages.

Class	Relative abundances			Absolute abundances		
	Matched pairs	Non-matched pairs	p -value	Matched pairs	Non-matched pairs	p -value
Uveal Melanoma	97.0	96.6	$< 10^{-5}$	96.1	95.4	$< 10^{-5}$
Prostate Adenocarcinoma	95.9	94.8	$< 10^{-5}$	96.9	96.1	$< 10^{-5}$
Sarcoma	94.4	92.9	$< 10^{-5}$	96.5	95.7	$< 10^{-5}$
Rectum Adenocarcinoma	94.0	91.2	$< 10^{-5}$	93.7	89.0	$< 10^{-5}$
Colon Adenocarcinoma	93.1	89.7	$< 10^{-5}$	92.3	86.6	$< 10^{-5}$
Kidney Renal Clear Cell Carcinoma	94.3	93.2	0.0355	95.8	95.3	0.0145
Breast Invasive Carcinoma	94.5	91.3	$< 10^{-5}$	96.4	93.3	$< 10^{-5}$
Brain Lower Grade Glioma	95.4	93.6	$< 10^{-5}$	96.2	90.4	$< 10^{-5}$
Lung Adenocarcinoma	94.9	91.0	$< 10^{-5}$	95.7	90.3	$< 10^{-5}$
Kidney Renal Papillary Cell Carcinoma	93.7	92.7	0.0002	96.2	95.4	0.0002
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	95.6	93.3	$< 10^{-5}$	95.5	92.0	$< 10^{-5}$
Head and Neck Squamous Cell Carcinoma	92.7	91.2	$< 10^{-5}$	93.9	89.6	$< 10^{-5}$
Glioblastoma Multiforme	94.3	90.7	$< 10^{-5}$	96.5	94.3	$< 10^{-5}$
Kidney Chromophobe	94.7	94.1	0.0010	96.9	96.6	0.0032
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	94.4	93.1	0.0001	96.3	95.7	$< 10^{-5}$
Uterine Corpus Endometrial Carcinoma	95.0	93.4	$< 10^{-5}$	96.4	93.7	$< 10^{-5}$
Esophageal Carcinoma	96.6	93.6	$< 10^{-5}$	95.1	86.8	$< 10^{-5}$
Ovarian Serous Cystadenocarcinoma	94.3	90.0	$< 10^{-5}$	96.5	92.7	$< 10^{-5}$
Thyroid Carcinoma	96.3	94.5	$< 10^{-5}$	96.6	94.0	$< 10^{-5}$
Liver Hepatocellular Carcinoma	93.7	93.0	0.0001	96.1	95.5	0.0002
Stomach Adenocarcinoma	95.0	93.6	$< 10^{-5}$	96.4	93.2	$< 10^{-5}$
Bladder Urothelial Carcinoma	96.3	95.1	$< 10^{-5}$	96.4	95.0	$< 10^{-5}$
Lung Squamous Cell Carcinoma	94.4	92.8	$< 10^{-5}$	96.8	95.7	$< 10^{-5}$
Skin Cutaneous Melanoma	96.3	95.0	$< 10^{-5}$	96.6	95.2	$< 10^{-5}$

median Pearson's r of matched pairs was 96.2%, while pairs of samples from different patients had a median value of 91.0%. When this analysis was performed independently for each cancer type, the similarity between samples from the same patient was always significantly greater than the similarity between samples of different patients. Table 4.2 holds the median Pearson's r for matched and non-matched sample pairs, along with the Mann-Whitney U p -values.

These results show the tumor and blood-derived microbiome profiles contain similarities specific to an individual. This suggests that some of the microbial genetic material found in the blood and in tumor tissues has a common origin, giving rise to two possible simplified models for its origin, shown in Figure 4.3. In one possible scenario (A in Figure 4.3), the microorganisms present in tumor tissues translocate into the bloodstream, leading to a similar microbiome profile in both sites. As the microbiome composition in the tumor is indicative of the type of cancer, blood-derived data preserves some information about cancer type. The microorganisms in the tumor may have originated from normal adjacent tissue or neighboring mucosal organs [216]. As previously mentioned, the blood-derived data used in this work encompasses genetic material in blood compartments other than plasma. Thus, remains of microorganisms in leukocytes are also quantified when building the abundance profiles. It is therefore possible that the translocation of microorganisms from tumor tissues to blood is mediated by white blood cells. Although some authors have suggested that leukocytes play a major role in the spread of bacteria through the blood and across organs [214, 224], little is known about this.

Alternatively, bacteria may enter the tumor tissues through the bloodstream thanks to the increased vascularization in the tumor site, originating from another organ, most likely the gut [50] (B in Figure 4.3). In this latter case, cancer type-dependent changes in the gut microbiome and the permeability of the intestinal barrier might explain the information on cancer type present in both blood-derived and tumor microbiome profiles. These two scenarios, beyond being highly simplified hypotheses of the underlying biological phenomena, may happen in conjunction.

The analysis of the similarity between genera abundances in matched tumor and blood-derived profiles revealed that although profile pairs from the same patient are highly similar, not all organisms are systematically present in similar amounts in both profiles, as seen in Figure 4.4. When using relative abundances, the genera most similar between tumor and blood-derived profiles was *Lawsonia*, with a Pearson's r of 88.3%. Other highly similar genera are *Chlamydia* ($r = 87.6\%$), *Wolbachia* ($r = 87.5\%$), *Circovirus* ($r = 86.3\%$), and *Ichnovirus* ($r = 85.1\%$). On the other hand, *Peptostreptococcus* ($r = -5.2\%$), *Dorea* ($r = 3.3\%$), *Alphapapillomavirus* ($r = -2.8\%$), *Solobacterium* ($r = -2.6\%$), and *Fusobacterium* ($r = -2.2\%$) had the lowest similarity between matched profiles. For absolute abundance profiles, *Microvirus* ($r = 95.4\%$), *Lawsonia* ($r = 94.1\%$), *Wolbachia* ($r = 93.8\%$), *Plesiomonas* ($r = 93.7\%$), and *Brackiella* ($r = 93.6\%$) were the most similar genera, whereas *Pseudoflavonifractor* ($r = -3.0\%$), *Reinekea* ($r = -1.5\%$), *Subdoligranulum* ($r = -1.5\%$), *Methanomethylophilus* ($r = -1.5\%$), and *Methanofollis* ($r = -1.2\%$) had the lowest Pearson's r values. The distributions of Pearson's r values between relative and absolute abundance profiles are dissimilar: while relative profiles have comparably few genera with Pearson's r values close to zero and greater than 60%, the distribution of similarity values with absolute

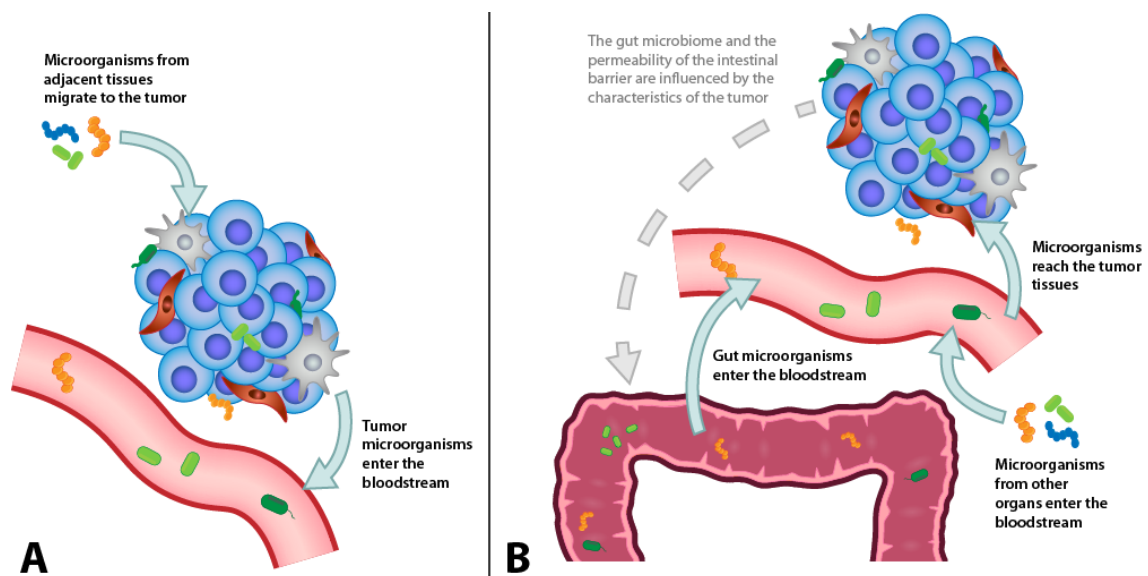


Figure 4.3: Proposed models for the mobility of microorganisms between the bloodstream and tumor tissues. A – Microorganisms reach the tumor tissues via adjacent normal and mucosal tissue, forming the tumor microbiome. Some of these organisms enter the bloodstream. B – Microorganisms in the gut and other organs enter the bloodstream. The gut microbiome composition and the permeability of the intestinal barrier is influenced by the characteristics of the tumor. Some of these microorganisms then migrate to the tumor tissue, forming the tumor microbiome.

profiles is roughly uniform, as seen in Figure 4.4. Despite this difference, these results show that not all genera are found in similar abundances in blood and in tumor tissues of the same patient. This suggests that there is differential permeability between the tumor and blood for different organisms.

The detection of systematic changes in relative abundance between the tumor and blood-derived profiles showed that although most genera have differential abundances with statistical significance, the magnitude of the difference is negligible: only 26 taxa have an absolute 2-fold change between its abundance in the blood and tumor greater than 0.5. Of these, most are under-represented in blood-derived profiles, as seen in Figure 4.5. All the genera of lowest similarity between tumor and blood-derived profiles from the same patient are also systematically found in lower abundance in blood. On the contrary, *Methylibium* is present in greater relative abundance in the blood than in tumor tissues. These results once again hint at a differential permeability of the blood-tumor interface between genera, with some, such as *Peptostreptococcus*, *Dorea*, *Alpha-papillomavirus*, *Solobacterium*, and *Fusobacterium* detected in lower amounts in the blood.

In summary, these results indicate that the blood-derived and the tumor microbiome profiles are highly similar, and contain signatures related to the host. This in part supports the usage of mbDNA to characterize the tumor microbiome. However, it is difficult to derive meaningful conclusions on the mobility of specific genera between the tumor and the bloodstream, as there are few genera systematically under or overrepresented in blood-derived profiles. Furthermore, few genera have similar abundances in both profiles across all patients. Hoping to better understand

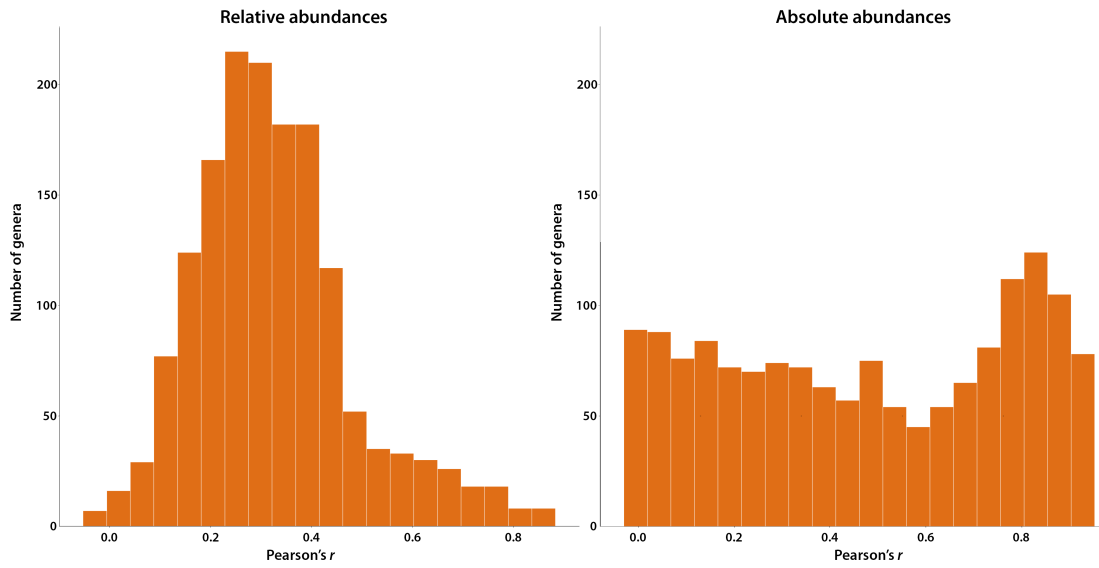


Figure 4.4: Histograms of the Pearson's r of the abundance of each genus between matched blood-derived and tumor microbiome profiles.

the dynamics of microbial translocation between the blood and tumor tissues, ML models were trained to estimate the tumor microbiome composition using blood-derived profiles from the same patient. These experiments are described in the next Section.

4.4 Predicting the tumor microbiome from blood-derived profiles

To better understand how the tumor and associated blood-derived microbiome profiles relate, and gain insights into how bacteria move between the bloodstream and tumor tissues, ML regressors to estimate the tumor microbial abundances from the blood-derived microbial profiles were developed. These models were designed following different assumptions about how microorganisms translocate between the tumor-blood interface and the origins of the organisms found in the blood. The models were trained and evaluated with both relative and absolute abundance profiles, with different levels of decontamination.

Regression tasks are rare when working with cancer-related microbiome: previous usages of ML regressors in this field are limited to survival time estimation [39] and predicting the number of tumors formed in murine models [26]. No ML-based analysis of the differences between the tumor and blood-derived microbiomes has been published, and little is known about how the microbiomes in these two sites are associated, as previously mentioned.

4.4.1 Machine Learning regressors

The ML regressors used in these experiments were designed according to slightly different assumptions on the dynamics of microbial translocation between the bloodstream and tumor tissues. As little is known about this subject, these assumptions are doomed to be oversimplifications of

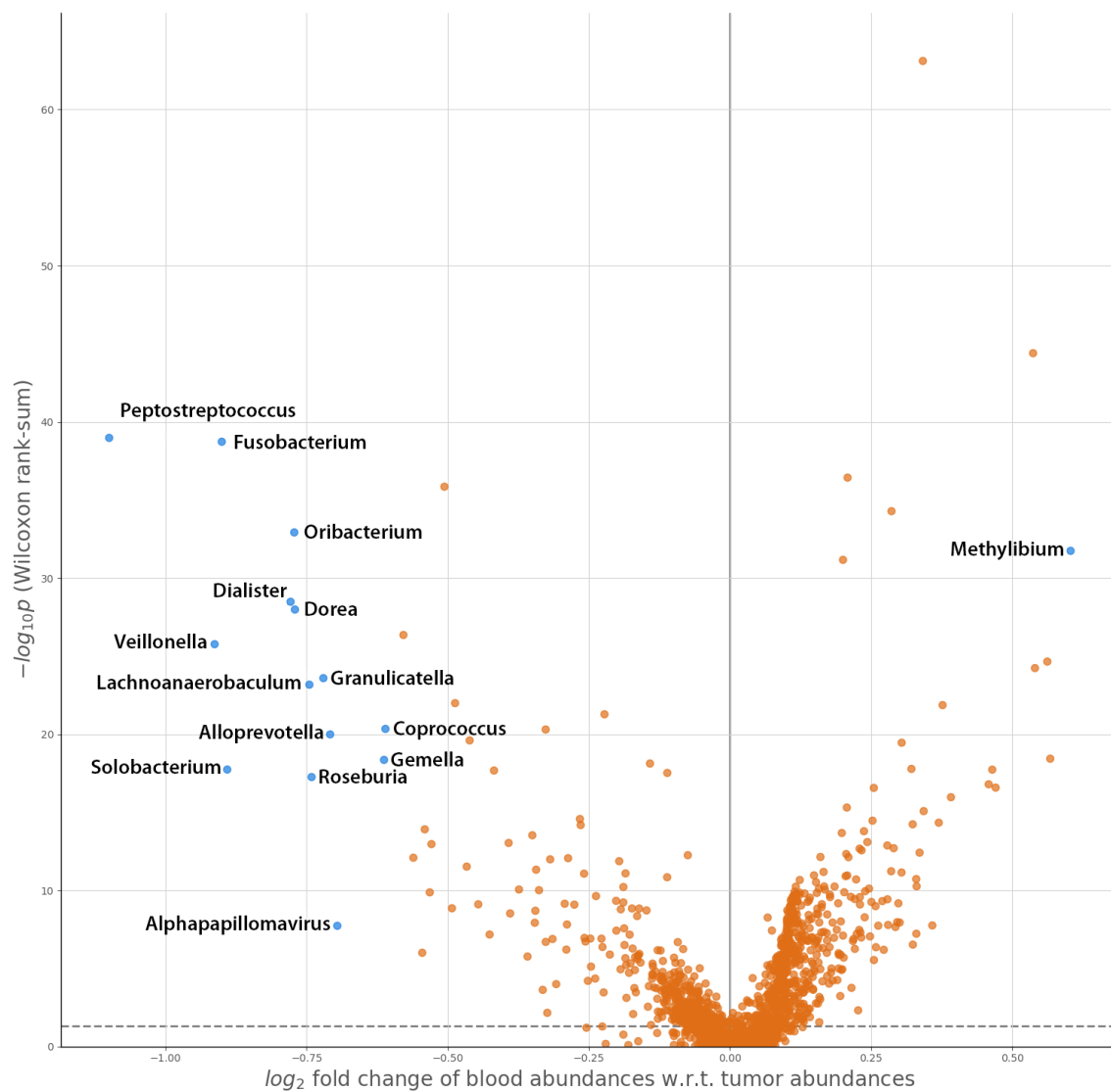


Figure 4.5: Volcano plot of blood-derived and tumor microbial abundances. Each marker represent a genus. The x axis holds the ratio between the blood and tumor abundances of a genus, in 2-fold changes. Genera with a ratio greater than 1.5 and less than -1.5 are highlighted in blue and labelled. The y axis gives the p -value of a Wilcoxon rank-sum test between the distributions of tumor and blood-derived abundances of a genus. The dashed line marks the 95% confidence threshold.

the real underlying biological processes; however, these models may improve our understanding of these dynamics, paving the way for more complex analysis. These motivations are presented here, along with a description of the five models tested: Linear Regression, an MLP, a Semiparametric Regressor, a Parallel NN, and a Topological Parallel NN.

An ML model for regression takes as input a feature vector $\mathbf{x} \in \mathbb{R}^{M_{In}}$, and predicts a vector $\hat{\mathbf{y}} \in \mathbb{R}^{M_{Out}}$ by applying a function ϕ , such that $\hat{\mathbf{y}} = \phi(\mathbf{x})$. The decision function is often learned by minimizing a loss function $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$, where $\mathbf{y} \in \mathbb{R}^{M_{Out}}$ is the ground-truth target vector. In the experiments carried out in this work, the ML regressors were used to estimate the tumor microbiome profiles using the blood-derived profile from the same patient. As both types of profiles have the same number of genera, $M_{In} = M_{Out}$, so both will be referred to as M from now on.

Training and test subsets ($\mathcal{X}_{Train}$ and $\mathcal{X}_{Test}$, respectively) were iteratively constructed using 5-fold CV, preserving the global proportion of tumor types inside each group. To avoid data leakage, samples from the same patient were included in the same subset. The CV partition was repeated five times, leading to 25 training/test splits of the dataset. When using absolute abundance profiles, both the input and target vectors were log-transformed according to Equation 4.1 to reduce skewness. For all types of profiles, the amplitudes of the target feature distributions were learned from the training subset and used to normalize the target values into the $[0, 1]$ range. This target transformation was reverted after inference to compute the evaluation metrics, whereas the loss function was calculated with the transformed values.

$$\mathbf{x}' = \log(\mathbf{x} + 1) \quad (4.1)$$

4.4.1.1 Linear Regression

Linear Regressors assume that the expected values of the target vectors are given by a linear combination of the input feature values [91]. The estimated abundance for genus $m \leq M$ in the tumor profile $n \leq N$, $\hat{y}_n^{(m)}$ is thus given by the sum of the blood abundances $x_n^{(k)}$, $\forall k \in \mathbb{N} : k \leq M$, weighted by linear coefficients $\{\beta_1^{(m)}, \dots, \beta_M^{(m)}\}$, and a bias term $\beta_0^{(m)}$, as per Equation 4.2 [127]. The parameters $\beta^{(m)} = [\beta_0^{(m)}, \beta_1^{(m)}, \dots, \beta_M^{(m)}]$ are set by minimizing the mean squared error (MSE, see Equation 4.3) between the predictions and the ground-truth targets [115]. To predict the abundances for all genera, M Linear Regressors were trained, each predicting the abundance of one genus.

$$\hat{y}_n^{(m)} = \beta_0^{(m)} + \sum_{k=1}^M x_n^{(k)} \beta_k^{(m)} \quad (4.2)$$

$$\mathcal{L}_{MSE} = \frac{1}{N M} \sum_{n=1}^N \sum_{m=1}^M \left(y_n^{(m)} - \hat{y}_n^{(m)} \right)^2 \quad (4.3)$$

Let X be the input matrix, formed by stacking the input abundance profiles in the training dataset with a one for multiplication with β_0 , as in Equation 4.4, and $\mathbf{y}^{(m)}$ the vector of target

values for genus m , as in Equation 4.5. The least squares estimate of the parameters of model m ($\beta^{(m)}$) is given by Equation 4.6 [115].

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_1^{(M)} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_N^{(1)} & \dots & x_N^{(M)} \end{bmatrix} \quad (4.4)$$

$$\mathbf{y}^{(m)} = [y_1^{(m)}, \dots, y_N^{(m)}]^T \quad (4.5)$$

$$\beta^{(m)} = (X^T X)^{-1} X^T \mathbf{y}^{(m)} \quad (4.6)$$

Linear Regression can be thought of as modelling the abundance of each genus in the tumor as proportional to the abundance of every genus in blood, so that an increase in the number of organisms of one genus in the blood-derived profiles always leads to an increase or decrease in the tumor abundance of another genus, depending on the sign of the corresponding model parameter. Biologically, this expresses a complex interaction between organisms, which influences the number of microorganisms of each genus that travel between the bloodstream and the tumor tissues. Thus, it can be perceived as the phenomena A and B in Figure 4.3 occurring simultaneously. Despite this complexity, by analyzing the parameters of Linear Regressors, it is possible to assess the contribution of each genus in blood to the estimated abundances in the tumor. This allows a better understanding of the dynamics dictating the translocation of organisms between the blood and the tumor tissues.

The Linear Regression implementation in *scikit-learn* version 1.1.1 [200] package for Python was used. To decrease the variance of the model, L2 regularization was employed. The mean and standard deviation of the feature distributions were learned from the training subset and used to transform features to zero mean and unit variance prior to model training and inference.

4.4.1.2 Multilayer Perceptron

Multilayer Perceptrons for classification were previously discussed in Section 2.4. Adapting MLPs to regression tasks only requires changing the loss function and, in some cases, the activations in the output layer to linear functions. The architecture of the MLP used in these experiments is shown in Figure 4.6. Prior to each hidden layer, the input values were scaled to zero mean and unit variance by Batch Normalization [225]. Residual connections were added to prevent vanishing gradients [226]. Furthermore, due to the high similarity between the blood and tumor microbiome profiles of the same patients, a residual connection between the input and output layers was added, forcing the hidden layers to estimate the difference between the input and target vectors [226]. Because target abundances were normalized to the $[0, 1]$ range, a sigmoid function was used as the activation function in the output layer. Mean squared error was used as the loss function.

The prediction function applied by MLPs can be very complex; thus, these models can be thought of as modelling the genera abundances in the tumor as being non-linearly dependent on the abundances of all genera in the blood. Once again, it could model both A and B in Figure 4.3,

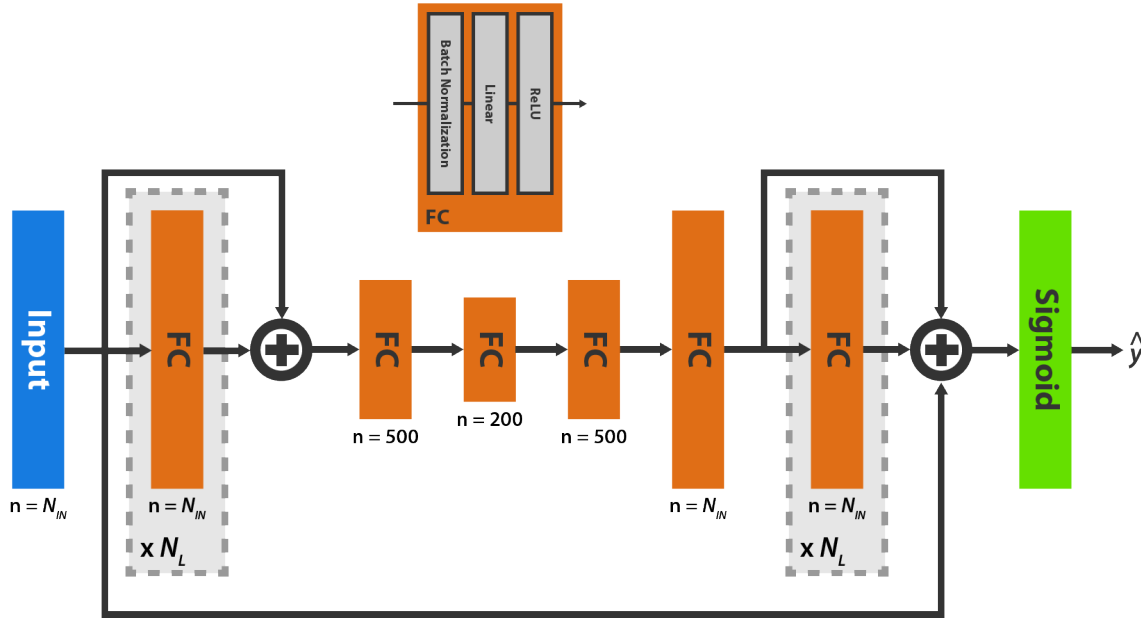


Figure 4.6: Architecture of the MLP regressor. n gives the number of activation units in each layer, with N_{In} being the number of input features. The number of layers between residual connections (N_L) was set as a tunable hyperparameter. FC – fully connected layer, constituted by batch normalization, a linear layer and ReLU as the activation function.

as well as their conjunction. This model is less constrained than any other regressor tested here and allows for a non-parametric estimation of the tumor microbiome profiles.

4.4.1.3 Semiparametric Regressor

The Semiparametric Regressor (SPR) models the microbial abundance of a genus in tumor tissues according to two components, as seen in Equation 4.7 and the model architecture in Figure 4.7. Firstly, there is a linear contribution of the abundance of the same genus in the input blood-derived profile ($x_n^{(m)}$), given by $\beta_L^{(m)} x_n^{(m)} + \beta_0^{(m)}$, where $\beta_L^{(m)}$ and $\beta_0^{(m)}$ are learnable parameters. This linear component captures the expected proportion of organisms translocating from the bloodstream to the tumor tissues, or vice versa, across all patients. The second component is an arbitrarily complex function $\psi^{(m)} : \mathbb{R}^M \rightarrow \mathbb{R}$ acting on the blood-derived abundances of all genera. The functions $\psi^{(m)}$ are modelled using an MLP and its contribution to the final estimate is given by the learnable parameter $\beta_{NL}^{(m)}$. Thus, besides the linear contribution of the same genus in the blood, a genus abundance in the tumor is said to depend on unique characteristics of a patient captured by the blood microbiome. These characteristics alter the permeability of the blood-tumor interface and the tumor microbial community in a unique way. This was inspired by model A in Figure 4.3. By analyzing the values of $\beta_0^{(m)}$, $\beta_L^{(m)}$, and $\beta_{NL}^{(m)}$, it should be possible to detect genera that are systematically over or underrepresented in blood-derived profiles, and genera for which this differential abundance across all patients does not happen.

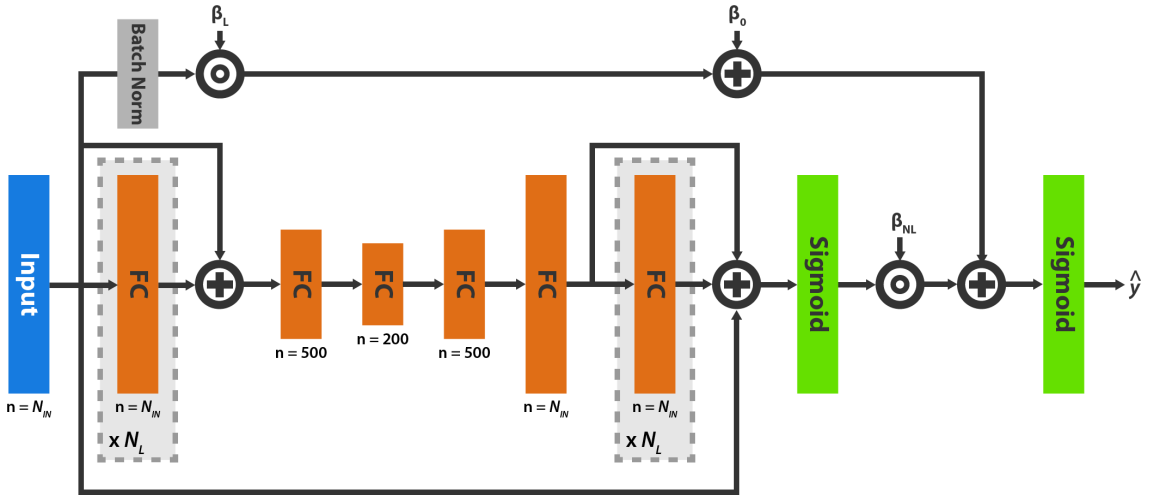


Figure 4.7: Architecture of the SPR. $\beta_L, \beta_0, \beta_{NL} \in \mathbb{R}^{N_{in}}$ are learnable parameters. The symbol $\circ$ represents element-wise multiplication.

$$\hat{y}_n^{(m)} = \beta_L^{(m)} x_n^{(m)} + \beta_0^{(m)} + \beta_{NL}^{(m)} \psi^{(m)}(\mathbf{x}_n) \quad (4.7)$$

The MLP used to model all the $\psi^{(m)}$ functions has a similar architecture to the MLP regressor presented in 4.4.1.2 and MSE was used as the loss function. The parameters $\beta_0^{(m)}$, $\beta_L^{(m)}$, and $\beta_{NL}^{(m)}$ were initialized with values sampled from a uniform distribution, with support given by the method described in [227].

4.4.1.4 Parallel Neural Network

The results shown in Section 4.3 indicate that not all genera have similar abundance between tumor and blood-derived profiles. This may be due to organisms from other tissues and organs, most likely the gut, entering the bloodstream, as previously shown in Figure 4.3B. Under this assumption, two components can be thought of as contributing to the abundances in blood-derived profiles: one is dependent on the organisms found in tumor tissues, while another depends on the genera originating in other organs. The model proposed here, nicknamed Parallel Neural Network (PNN) due to its parallel architecture, estimates the tumor microbiome profiles by attempting to separate these two components from the blood-derived profiles. In this PNN, it is assumed that there is small overlap between the genera originating in the tumor and those from other organs, and the model is incentivized to have little overlap between the features sent to each branch of the network. Thus, the microbial abundances in the tumor tissues are modeled according to Equation 4.8, where $\circ$ is an element-wise multiplication. The vectors $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^M$ are binary masks, with the m -th element equal to one if the genus m contributes to the corresponding component, and zero otherwise. These masks are learned during model training. The functions ψ_1 and ψ_2 are modeled using MLPs, as seen in Figure 4.8, allowing them to capture different tumor microbial communities between patients.

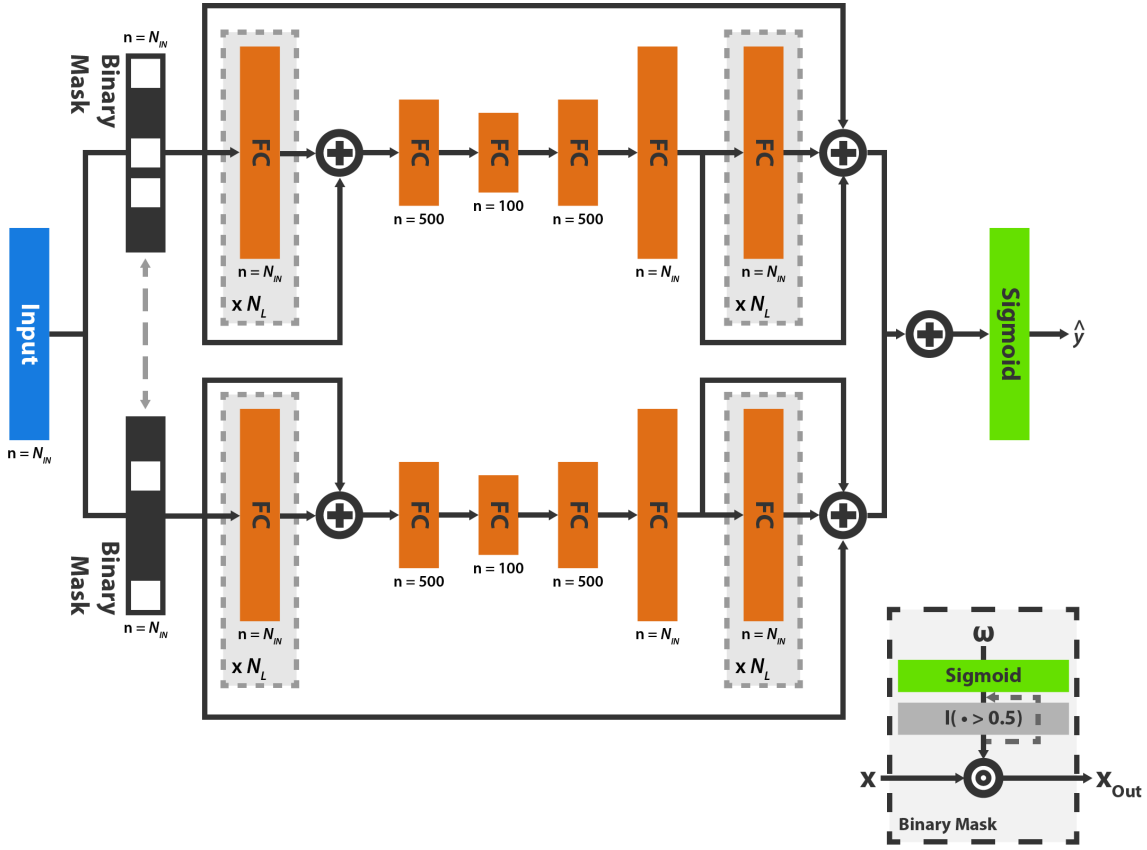


Figure 4.8: Architecture of the PNN model. The binary masks are learned by transforming the learnable parameters $\omega \in \mathbb{R}^{N_{in}}$ with a sigmoid function and thresholding the results at 0.5. $I(\bullet > 0.5)$ represents this thresholding operation, which is ignored during back-propagation. The binary masks are incentivized to be different by penalizing their cosine similarity.

$$\hat{\mathbf{y}}_n = \psi_1(\mathbf{m}_1 \circ \mathbf{x}_n) + \psi_2(\mathbf{m}_2 \circ \mathbf{x}_n) \quad (4.8)$$

The parameters of the masking layers ($\omega_k^{(m)}$, $\forall k \in \{1, 2\}$, $\forall m \in \mathbb{N} : m \leq M$) are transformed with a sigmoid function (σ) and then binarized using a step function, as outlined in Equation 4.9. However, this imposes a difficulty during backpropagation, as a step function is non-differentiable. Therefore, the step function is treated as a straight-through estimator [228]: during backpropagation, the gradient with respect to the input is set as equal to the gradient with respect to the output. These parameters are initialized by sampling from a uniform distribution, with support $[-10^{-3}, 10^{-3}]$. To force the two binary masks to be dissimilar, the loss function used to train the model was a sum between MSE and the cosine similarity between $\mathbf{m}_1$ and $\mathbf{m}_2$, as per Equation 4.10. The γ parameter, controlling the penalty assigned to the similarity between masks, was set to 0.05.

$$\mathbf{m}_k^{(m)} = \begin{cases} 0, & \sigma(\omega_k^{(m)}) < 0.5 \\ 1, & \sigma(\omega_k^{(m)}) \geq 0.5 \end{cases} \quad (4.9)$$

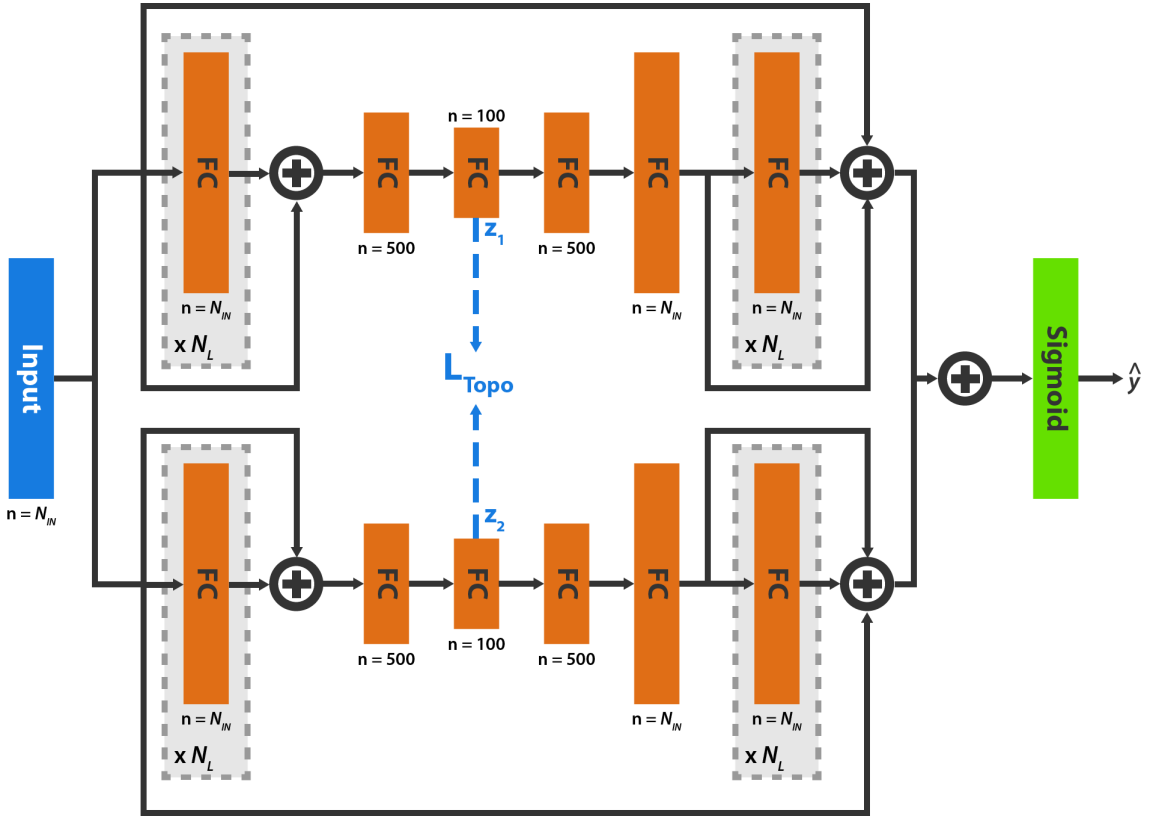


Figure 4.9: Architecture of the TopoPNN model. The latent representations of training samples, z_1 and z_2 , are forced to be topologically different by penalizing similar distances between topologically important points across both latent spaces.

$$\mathcal{L}_{PNN} = \gamma \frac{\mathbf{m}_1 \cdot \mathbf{m}_2}{\|\mathbf{m}_1\| \|\mathbf{m}_2\|} + \mathcal{L}_{MSE} \quad (4.10)$$

4.4.1.5 Topological Parallel Neural Network

The PNN described previously allows only a certain degree in overlap between the genera translocating between the bloodstream and tumor tissues, and those entering the bloodstream via other organs; however, this may not be the case: organisms from the same genus may originate in both tumor tissues and the gut and other reservoirs, and the contribution of each site to the blood abundances is governed by different mechanisms. Under these assumptions, the model for tumor microbiome profiles in Equation 4.8 can be modified into Equation 4.11. Like in the PNN, ψ_1 and ψ_2 are estimated using MLPs. To ensure that these two functions were different, a topological loss component was added to the MSE when training the model. Thus, the loss used to train this model, nicknamed Topological Parallel NN (TopoPNN), is given by Equation 4.12 and constitutes a novel approach to force a model to learn disentangled representations of the input data. Here, $\mathbf{z}_{1,n}$ and $\mathbf{z}_{2,n}$ are the latent representations of the input microbiome profile n in the MLP branches modelling ψ_1 and ψ_2 , respectively, as shown in Figure 4.9; $\mathcal{L}_{Topo}$ is the topological loss function proposed in [229].

This topological loss was originally proposed to train autoencoders; it penalizes discrepancies in the distance matrices of the most topologically important points between the input and latent space representations. Thus, the model is incentivized to place topologically relevant samples that are close in the input space closer together in the latent space as well [229]. By using the symmetrical of this topological loss in $\mathcal{L}_{TopoPNN}$, the TopoPNN is penalized if the distances between topological important points in both latent spaces are identical. Thus, the model is constrained to learn disentangled representations of the input space, ideally corresponding to different permeabilities of the blood-tumor interface and the intestinal barrier (and that of other organs). The most important training samples for the topology of the latent representations in each of the latent spaces are determined using Persistent Homology [230, 229]. Let D_1 and D_2 be the distance matrices between the latent representations of training samples in each latent space, π_1 and π_2 the sets of the most topologically important points in each latent space, and $D_k[\pi_j], k, j \in \{1, 2\}$ a vector with the elements in the distance matrix D_k between any two points in the set π_j . The topological loss $\mathcal{L}_{Topo}$ is given by Equation 4.13 [229].

$$\hat{\mathbf{y}}_n = \psi_1(\mathbf{x}_n) + \psi_2(\mathbf{x}_n) \quad (4.11)$$

$$\mathcal{L}_{TopoPNN} = -\tau \sum_{n=1}^N \mathcal{L}_{Topo}(\mathbf{z}_{1,n}, \mathbf{z}_{2,n}) + \sum_{m=1}^M \mathcal{L}_{MSE}(\hat{\mathbf{y}}^{(m)}, \mathbf{y}^{(m)}) \quad (4.12)$$

$$\mathcal{L}_{Topo} = \frac{1}{2} \left[\|D_1[\pi_1] - D_2[\pi_1]\|^2 + \|D_1[\pi_2] - D_2[\pi_2]\|^2 \right] \quad (4.13)$$

The parameter τ , determining the contribution of the topological loss to the total loss function, was set to 0.004. The topological loss was implemented with the *pytorch-topological* [231] Python package, version 0.1.7. Persistent Homology was calculated for the first dimension only.

4.4.2 Hyperparameter search and global parameters

Table 4.3 holds the hyperparameters of all models used in these experiments, along with their search space. These hyperparameters were selected through grid search, by minimizing the MSE in a validation subset of $\mathcal{X}_{Train}$. This validation subset was created through 5-fold CV when training Linear Regressors; for all other models, due to the long training times, it was built by sampling 10% of the training subset.

All NN models were trained up to 200 epochs, with training being interrupted after 10 epochs of no improvement in the loss for the validation subset. Dropout was used in all hidden layers and ReLU [166] was used as an activation function in all hidden layers.. These models were implemented using the *PyTorch* [232] version 1.12.1 Python package, and hyperparameter selection was performed with *Ray* [233] version 1.13.0.

4.4.3 Evaluation metrics for regression

The performance of the regressors trained to predict the microbiome profiles in the tumor tissues using the blood-derived profile from the same patients was evaluated according to two metrics:

Table 4.3: Regressor hyperparameters and search spaces.

Model	Hyperparameter	Search space
Linear Regression	L2 regularization parameter for relative abundance profiles	$\{2000, 3000, 4000\}$
	L2 regularization parameter for absolute abundance profiles	$\{1000, 5000, 10000\}$
MLP, SPR, PNN, and TopoPNN	Number of layers between residual connections (N_L in Figures 4.6, 4.7, 4.8, and 4.9)	$\{1, 2\}$
	Learning rate	$\{0.005, 0.001, 0.01\}$
	Dropout probability	$\{0.0, 0.3\}$

MSE and Pearson’s r . The statistical significance of differences in performance was assessed using a Wilcoxon signed-rank test over the 25 training and test splits of the dataset and a 95% confidence interval. The MSE for each genus, previously presented in Equation 4.3, was summed, yielding a global MSE. The Pearson’s r measures the linear correlation between two vectors, can take values between -1 and 1 , and is given by Equation 4.14. Here, $\hat{y}_n$ and $\bar{y}_n$ represent the average abundance in the n -th estimated and ground-truth tumor profile, respectively. Positive values of r indicate the two vectors increase and decrease together as we iterate through their elements; an r of zero dictates no association between the variables [234]. Unlike MSE, the Pearson’s r is offset and scale invariant, meaning that estimates that deviate from the ground truth microbiome profiles by a constant value and proportion of the estimated value are assigned an r of one. The Pearson’s r reported for a model is the average r across all samples in the subset.

$$r(\hat{\mathbf{y}}_n, \mathbf{y}_n) = \frac{\sum_{m=1}^M (\hat{y}_n^{(m)} - \bar{\hat{y}}_n) (y_n^{(m)} - \bar{y}_n)}{\sqrt{\sum_{m=1}^M (\hat{y}_n^{(m)} - \bar{\hat{y}}_n)^2} \sqrt{\sum_{m=1}^M (y_n^{(m)} - \bar{y}_n)^2}} \quad (4.14)$$

The ML regressors should provide abundance estimates that are more similar to the target tumor profiles than the input blood-derived profiles. Therefore, to aid the interpretation of the results obtained in this section, normalized variations of the previously mentioned performance metrics were also computed. The normalized version of the MSE is given by Equation 4.15, in which $MSE(\mathbf{x}_n, \mathbf{y}_n)$ denotes the MSE between the n -th input blood-derived microbiome profile and the corresponding target abundances. A redundant regressor yields a normalized MSE of 1, while a perfect estimate leads to a normalized MSE of 0. The normalized Pearson’s r is given by Equation 4.16; it leads to a score of 0 for regressors with equal performance to an identity function, and 1 for perfect estimates.

$$MSE_{Norm}(\hat{\mathbf{y}}_n, \mathbf{y}_n, \mathbf{x}_n) = \frac{MSE(\hat{\mathbf{y}}_n, \mathbf{y}_n)}{MSE(\mathbf{x}_n, \mathbf{y}_n)} \quad (4.15)$$

$$r_{Norm}(\hat{\mathbf{y}}_n, \mathbf{y}_n, \mathbf{x}_n) = \frac{r(\hat{\mathbf{y}}_n, \mathbf{y}_n) - 1}{1 - r(\mathbf{x}_n, \mathbf{y}_n)} + 1 \quad (4.16)$$

4.4.4 Results and discussion

As seen in Table 4.4, when relative abundance profiles were used, the Linear Regression model achieved the best test results of all the regressors evaluated ($p < 10^{-5}$ for all metrics and comparisons), with an MSE of 0.8879, or 61.8% when normalized against the input abundance profiles, and a Pearson’s r of 96.4% (37.8% normalized). Linear Regression can outperform complex deep NNs when the training dataset is small, as is the case in this work [91]. Furthermore as the loss functions used to train NNs are not always convex, it is possible that the parameters of the NNs evaluated here may lead to local optima of the loss functions and not global minima [91]. These computational shortcomings of NNs may explain their decreased performance when compared to Linear Regression. Nevertheless, the performance gap between Linear Regression and most NNs suggests that most assumptions about microbial mobility made during model development are wrong. The exception to this seems to be the SPR, which was the best performing NN evaluated, both in terms of MSE and Pearson’s r . The SPR has a linear component, which only accounts for the abundance of a target genus in the input blood-derived profiles. This may explain why this model is the only NN with a predictive performance close to Linear Regression.

Table 4.4: Performance metrics for ML regressors using relative abundance profiles. μ and σ are the average and standard deviation across training/test splits of the dataset.

Metric			Linear Regression	MLP	SPR	PNN	TopoPNN
MSE	Training	μ	0.6346	0.8511	0.8213	0.8812	0.8420
		σ	0.0053	0.0254	0.0536	0.0593	0.0425
	Test	μ	0.8879	0.9415	0.9162	1.0015	0.9440
		σ	0.0295	0.0298	0.0328	0.0986	0.0361
r^2	Training	μ	97.5	96.4	96.6	96.2	96.4
		σ	0.0	0.1	0.2	0.2	0.1
	Test	μ	96.4	96.2	96.3	96.0	96.2
		σ	0.1	0.1	0.1	0.3	0.1
MSE_{Norm}^2	Training	μ	44.1	59.2	57.1	61.3	58.6
		σ	0.3	1.7	3.6	4.7	3.1
	Test	μ	61.8	65.5	63.8	69.6	65.7
		σ	1.4	1.8	2.7	3.7	2.0
r_{Norm}^2	Training	μ	55.9	36.4	40.2	33.6	37.7
		σ	0.2	1.3	3.4	3.9	2.4
	Test	μ	37.8	33.5	35.5	29.6	33.7
		σ	0.9	1.4	2.4	4.1	1.6

The parameters of the proposed regressors provide insights into the interplay between the tumor and blood microbiomes. As an example, the Linear Regression coefficients directly quantify

²Values in percentages.

the contribution of one genus abundance to the estimated abundance of another. However, doing this analysis with models trained with relative abundance profiles is challenging, as a variation in the absolute abundance of a genus leads to a change in the relative abundances of all genera, as previously mentioned. Furthermore, for this same reason, ML regressors operating on relative abundances might yield worse predictions than models using absolute profiles. Thus, the five ML models presented in this Chapter were also trained and tested with absolute abundance profiles. This evaluation was repeated with three levels of decontamination to assess the impact of potential contaminants in the performance of models. The results are shown in Tables 4.5, 4.6, and 4.7 for profiles with no decontamination, with *likely contaminants removed*, and the *most stringent* decontamination, respectively. Similarly to what was seen with relative abundances, Linear Regression achieved the best predictive performance according to all metrics ($p < 10^{-5}$ for all metrics, comparisons, and decontamination levels), with SPR showing the best results among the NNs.

The normalized versions of the performance metrics showed little difference between decontamination levels; therefore, the analysis of the model parameters was done for models trained with absolute abundance profiles and *likely contaminants removed*. When predicting the abundance of a genus in the tumor tissues, the Linear Regression model generally attributes greater importance to the abundance of that same genus in the blood-derived profiles. For 465 genera (30% of the total number of genera in the profiles), the abundance of the same genus in blood was the feature with the highest coefficient. Furthermore, the parameters related to the blood abundance of the genus to predict, i.e., $\beta_m^{(m)}$, $\forall m \in \mathbb{N} : m \leq M$, are usually greater than those associated with other genera ($p < 10^{-5}$, Mann-Whitney U test). When ranking the model coefficients for each target genus in descending order, those relating to the blood abundances of the genus to predict have a median rank of eight out of 1553. This is further evidence of the importance given to the input abundance of the target genus. This high contribution of the blood abundance of the target genus to the estimated tumor abundance is expected, given the high similarity between matched blood and tumor microbiome profiles. Figure 4.10 shows the 10 highest and lowest values of the bias parameters of the Linear Regression model. Genera with a high bias, such as *Streptococcus*, *Mycobacterium*, and *Staphylococcus*, are likely underrepresented in the blood-derived profiles, relative to the tumor abundances.

On one hand, the tumor abundance of a microorganism estimated by Linear Regression usually depends mainly on the abundance of that same microorganism in the blood. This shows that most viruses and bacteria translocate between the tumor tissues and the bloodstream. On the other hand, Linear Regression was the best performing model among those evaluated. Setting computational constraints aside, this suggests that either the permeability of the blood-tumor interface changes from patient to patient, and can be estimated from mbDNA, or that part of the blood microbiome has its origin outside the tumor, leading to differences between the tumor and blood-derived microbiome profiles which vary between individuals. In any case, these results indicate a complex association between the tumor microbiome and mbDNA, in which a change in the blood abundance of one organism has a proportional effect on the tumor abundance of another. It is worth noting that the Linear Regression model shows high variance, with a larger gap between the train-

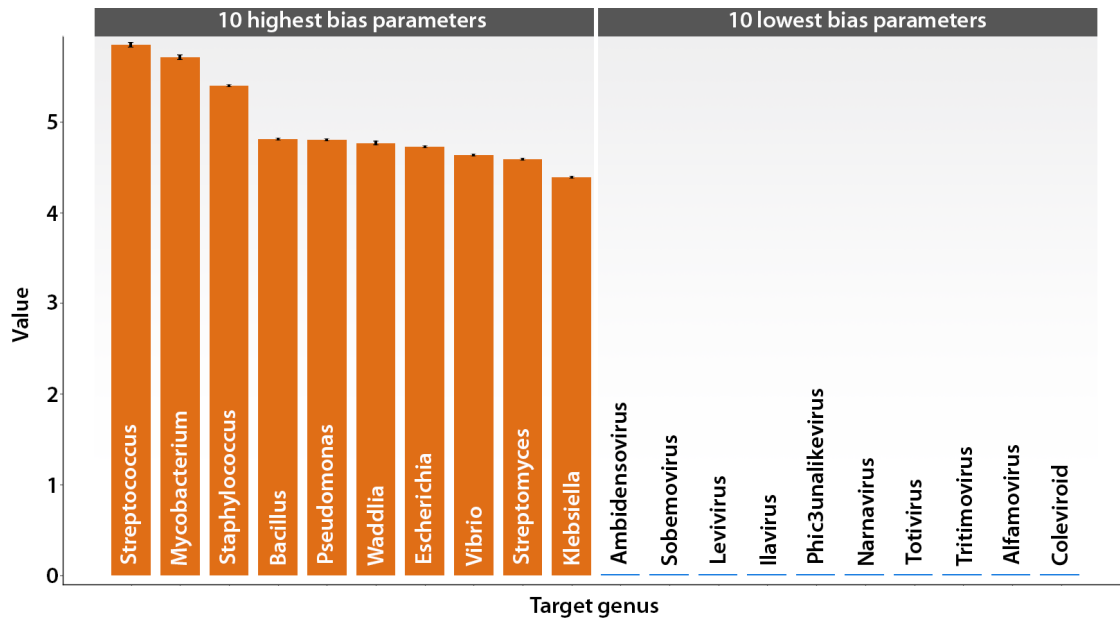


Figure 4.10: 10 genera of highest and lowest bias parameters in the Linear Regression model trained with absolute microbial abundances and *likely contaminants removed*.

ing and test metrics than all other models, for most levels of decontamination. Therefore, caution should be taken regarding the generalization of this model.

To assess the contribution of the linear and non-linear components when predicting a genus abundance in the tumor tissues, the values of the parameters $\beta_L^{(m)}$, $\beta_0^{(m)}$, and $\beta_{NL}^{(m)}$ of the SPR were analyzed. Only the parameters whose distribution across the 25 training/test splits of the dataset was not centered around zero according to a one-sample *t*-test with 95% confidence were included in this analysis. *Algiphilus*, *Natronomonas*, *Inovirus*, *Methyloceanibacter*, and *Anaerobaculum* were the five genera with the highest average $\beta_L^{(m)}$, while *Wolinella*, *Polerovirus*, *Sulfurihydrogenibium*, *Ferrovum*, and *Alloprevotella* had the smallest linear coefficients, on average. However, *Empedobacter*, *Elizabethkingia*, *Prevotella*, *Alphapartitivirus*, and *Acetobacter* had the highest ratio between the linear and non-linear parameters, i.e., $\beta_L^{(m)} / \beta_{NL}^{(m)}$, meaning that their predicted abundance in the tumor was almost independent of the abundances of other genera; this value was smallest for *Chilikevirus*, *Phietalikevirus*, *Aliagarivorans*, *Orthohepadnavirus*, and *Thermanaerothrix*. The highest bias parameters were assigned to *Haloterrigena*, *Ichnovirus*, *Muromegalovirus*, *Sanguibacteroides*, and *Exiguobacterium*, and the smallest to *Heliobacterium*, *Thermomicrobium*, *Caldivirga*, *Desulfurispirillum*, and *Bermanella*. These results contradict those obtained with Linear Regression, as none of these genera figures among those with highest and lowest bias in the Linear Regression model. Overall, there is a large variance of the values of the parameters across the training/test splits, even after filtering parameters with distributions symmetric around zero. For instance, *Algiphilus*, the genus with the highest linear coefficient on average, has values of β_L ranging from -0.50 to 1.56 across different dataset splits; the bias pa-

³Values in percentages.

Table 4.5: Performance metrics for ML regressors using absolute abundance profiles, without decontamination. μ and σ are the average and standard deviation across training/test splits of the dataset.

Metric			Linear Regression	MLP	SPR	PNN	TopoPNN
MSE	Training	μ	0.0517	0.0846	0.0812	0.0764	0.0898
		σ	0.0010	0.0091	0.0039	0.0149	0.0095
	Test	μ	0.0723	0.0957	0.0843	0.0874	0.1057
		σ	0.0048	0.0085	0.0057	0.0143	0.0132
r^3	Training	μ	97.6	96.3	96.6	96.6	96.0
		σ	0.0	0.2	0.1	0.4	0.3
	Test	μ	97.0	96.2	96.5	96.5	95.8
		σ	0.1	0.3	0.1	0.4	0.3
MSE_{Norm}^3	Training	μ	46.9	76.7	73.5	69.3	81.4
		σ	0.6	8.2	3.0	14.0	8.6
	Test	μ	65.7	87.0	76.6	79.1	95.9
		σ	3.6	7.2	3.4	9.0	9.1
r_{Norm}^3	Training	μ	46.7	18.5	24.2	24.4	11.9
		σ	0.3	5.0	1.5	8.3	6.0
	Test	μ	32.4	15.7	23.3	22.0	7.5
		σ	1.4	4.6	1.4	6.3	5.7

Table 4.6: Performance metrics for ML regressors using absolute abundance profiles, with *likely contaminants removed*. μ and σ are the average and standard deviation across training/test splits of the dataset.

Metric			Linear Regression	MLP	SPR	PNN	TopoPNN
MSE	Training	μ	0.0565	0.0891	0.0836	0.0763	0.0969
		σ	0.0011	0.0104	0.0031	0.0086	0.0184
	Test	μ	0.0757	0.0992	0.0860	0.0875	0.1084
		σ	0.0048	0.0106	0.0064	0.0082	0.0189
r^3	Training	μ	97.3	95.9	96.3	96.3	95.5
		σ	0.0	0.3	0.1	0.2	0.5
	Test	μ	96.6	95.8	96.2	96.2	95.4
		σ	0.1	0.4	0.2	0.2	0.7
MSE_{Norm}^3	Training	μ	48.8	77.0	72.3	65.9	84.1
		σ	0.5	9.3	2.8	6.8	18.2
	Test	μ	65.6	85.9	74.6	75.8	93.3
		σ	3.6	6.7	3.4	5.8	6.7
r_{Norm}^3	Training	μ	45.1	18.4	25.0	26.4	10.6
		σ	0.3	5.9	1.3	4.7	12.0
	Test	μ	32.1	16.5	24.2	23.8	8.9
		σ	1.4	5.1	1.2	3.7	8.8

Table 4.7: Performance metrics for ML regressors using absolute abundance profiles, with the *most stringent* decontamination. μ and σ are the average and standard deviation across training/test splits of the dataset.

Metric			Linear Regression	MLP	SPR	PNN	TopoPNN
MSE	Training	μ	0.0389	0.0484	0.0467	0.0589	0.0513
		σ	0.0009	0.0059	0.0024	0.0212	0.0069
	Test	μ	0.0433	0.0544	0.0480	0.0641	0.0574
		σ	0.0034	0.0066	0.0040	0.0222	0.0072
r^3	Training	μ	98.8	98.6	98.6	98.3	98.5
		σ	0.0	0.1	0.0	0.8	0.2
	Test	μ	98.8	98.5	98.6	98.3	98.4
		σ	0.1	0.1	0.1	0.8	0.2
MSE_{Norm}^3	Training	μ	62.7	78.0	75.1	94.9	82.6
		σ	1.3	9.8	3.2	33.8	10.5
	Test	μ	70.0	87.7	77.4	103.2	92.8
		σ	4.6	6.1	3.8	36.3	11.6
r_{Norm}^3	Training	μ	34.8	19.6	22.3	4.1	12.9
		σ	0.7	6.3	1.9	45.3	9.3
	Test	μ	29.3	15.8	21.3	0.7	8.6
		σ	3.0	5.2	2.0	50.5	9.9

parameter for *Haloterrigena*, the genus for which this parameter is greater on average, took values between -0.51 and 1.53 . This variance casts doubts on the interpretation of the model parameters presented here.

The PNN and TopoPNN models showed worse regression performance when compared to all other models. Furthermore, the PNN binary masks show a high variability between training/test splits of the dataset. Therefore, it is unclear if the contributions of the tumor and gut (or other reservoirs) microbiomes to the blood-derived abundances can be isolated, and if the assumption that the blood-derived abundances can be separated into these two components is true in the first place. There likely is interplay between the tumor, blood, and gut microbiomes, as microorganisms entering the bloodstream through the intestinal barrier or other organs may migrate to the tumor tissues [50], while organisms in the tumor may enter the bloodstream. Thus, deconvoluting the influence of each to the blood-derived microbial abundances may not be possible.

The models shown in this section were trained with samples from all cancer types. However, the mobility of microorganisms between the blood and the tumor tissues may not be the same for all locations in the body. The predictive performance of the models is different between cancer types, as seen in Table 4.8, which holds the performance metrics per cancer type for the Linear Regression model using absolute abundance profiles, with *likely contaminants removed*. Training models for specific cancer types may yield improved results and aid the interpretation of model parameters. Furthermore, developing models with a focus on a specific cancer type may allow more confident assumptions about how microorganisms travel between the bloodstream and tumor tissues, especially for cancers whose association with the blood microbiome is better known, such

Table 4.8: Test performance metrics per cancer type for the Linear Regression model trained with absolute abundance profiles and *likely contaminants removed*. μ and σ are the average and standard deviation across training/test splits of the dataset.

	MSE		r^4		MSE_{Norm}^4		r_{Norm}^4	
	μ	σ	μ	σ	μ	σ	μ	σ
Liver Hepatocellular Carcinoma	0.1007	0.0239	97.5	0.3	50.6	3.5	39.7	2.6
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0.0900	0.0137	97.5	0.3	50.8	6.0	32.6	5.5
Kidney Renal Papillary Cell Carcinoma	0.1021	0.0190	97.5	0.2	52.6	2.8	39.3	2.0
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0.0745	0.0310	96.4	1.0	55.6	17.2	37.4	6.3
Kidney Chromophobe	0.0614	0.0028	97.9	0.1	55.9	7.7	33.6	4.7
Uterine Corpus Endometrial Carcinoma	0.0609	0.0102	97.5	0.2	61.0	14.9	40.1	5.2
Thyroid Carcinoma	0.0494	0.0069	97.7	0.1	62.7	3.0	37.1	1.8
Lung Adenocarcinoma	0.0869	0.0192	96.2	0.4	63.0	10.4	30.7	6.0
Sarcoma	0.0652	0.0043	97.1	1.0	63.9	7.1	25.5	16.3
Prostate Adenocarcinoma	0.0310	0.0024	97.9	0.1	64.2	1.4	35.0	1.5
Breast Invasive Carcinoma	0.0731	0.0060	97.6	0.1	64.7	4.1	37.9	4.2
Skin Cutaneous Melanoma	0.0364	0.0108	97.7	0.3	64.9	10.5	37.3	5.7
Lung Squamous Cell Carcinoma	0.0668	0.0079	97.8	0.2	65.9	5.8	32.8	6.1
Rectum Adenocarcinoma	0.1048	0.0275	95.2	0.9	67.3	4.9	34.7	5.1
Glioblastoma Multiforme	0.0825	0.0178	97.6	0.2	68.0	7.3	37.0	5.4
Uveal Melanoma	0.0280	0.0045	97.4	0.2	69.6	5.5	34.5	3.1
Colon Adenocarcinoma	0.1400	0.0379	94.1	0.8	70.5	6.8	30.3	3.2
Bladder Urothelial Carcinoma	0.0377	0.0047	97.5	0.2	71.3	3.7	34.2	2.7
Head and Neck Squamous Cell Carcinoma	0.1216	0.0276	94.4	0.9	72.3	5.1	31.5	3.6
Stomach Adenocarcinoma	0.0581	0.0164	97.0	0.4	72.4	3.0	29.0	2.8
Brain Lower Grade Glioma	0.0649	0.0192	96.6	0.6	77.3	7.3	27.3	6.0
Ovarian Serous Cystadenocarcinoma	0.1288	0.0504	96.0	1.0	78.0	16.2	5.9	24.8
Esophageal Carcinoma	0.1116	0.0948	92.9	3.0	80.1	17.7	16.7	9.6
Kidney Renal Clear Cell Carcinoma	0.1166	0.0122	96.2	0.3	107.3	11.2	9.6	7.2

as colorectal cancer.

As previously mentioned, some authors have suggested that the microbiome profiles used in this work contain incorrect counts for many genera, due to read misclassification during metagenomic analysis [196]. If this is true, and some of the genera in the abundance profiles are not actually present in the tissues, it could hinder the predictive performance of the ML regressors. More importantly, it would invalidate any conclusion derived from these models, as these would not be supported by “true” data. However, the dataset used in this work contains the largest collection of microbiome profiles in cancer tissues and corresponding blood samples, and few alternatives exist. This raises the need for more high-quality cancer-related microbiome datasets to better understand

⁴Values in percentages.

the factors influencing the blood and tumor microbiomes in cancer patients.

4.5 Summary

In summary, the comparison between the blood and tumor microbiomes revealed that there is a high similarity between the blood-derived and tumor microbial abundance profiles of the same patient, which supports the usage of mbDNA to characterize the tumor microbiome. However, this similarity is not seen for all genera. Some genera (albeit few) have differential abundances in blood and tumor tissues; between these, most are underrepresented in the blood-derived profiles.

To better understand the biological dynamics between the tumor and blood microbiomes, five ML regressors were trained to predict the microbial abundances in the tumor tissues using blood-derived microbiome profiles. These models were developed considering different assumptions regarding the origins of the blood microbiome and the permeability of the tumor-blood interface. Most of the developed models could approximate the input abundances to the true microbial abundances in the tumor tissues, showing that the biological processes determining the spread of microorganisms throughout the bloodstream and the tumor tissues can be modelled with reasonable correctness. Linear Regression achieved the best results, followed by a semiparametric model. Overall, models have high variance, raising doubts on their generalization. An analysis of the model parameters showed that the tumor abundance of most genera is linearly dependent on the abundance of that genus in blood. The small number of training samples may explain why the Linear Regression model achieved improved performance over NNs; however, the biological assumptions used to develop the NNs, especially the PNN and TopoPNN models, are also likely to be incorrect.

Chapter 5

Conclusions and future work

Recent advances on the characterization of the human microbiome have uncovered associations between the microorganisms inhabiting the human body and health. There is mounting evidence that the changes in the gut microbiome are associated with colorectal cancer and its progression, and this knowledge is now being extended to other cancer types. This cancer-mediated modulation of the microbiome does not happen only in the gut, but in the tumor tissues as well, with the tumor microbiome exhibiting a distinct composition between cancer types. The genetic material of bacteria and other microorganisms in the tumor tissues, gut, mouth, and other organs may enter the bloodstream, leading to mbDNA. Thus, changes in the microbiome of either of these reservoirs could be reflected in the microbial genetic material circulating in the blood. This has motivated the usage of microbial biomarkers for cancer diagnosis, cancer type identification, survival time prediction, and overall characterization of cancer. As these microbiome-based biomarkers are highly complex, their discovery can greatly benefit from the application of ML models, capable of identifying intricate patterns.

In this work, a GBT classifier for cancer type identification using mbDNA and the tumor microbiome outperformed Random Forest and Logistic Regression. Furthermore, mbDNA led to slightly worse predictions than the tumor microbiome, thus having a weaker association with cancer type than the latter. To the best of my knowledge, it was the first time such a comparison between the potential of mbDNA and tumor microbiome for cancer type characterization was performed. However, the performance of GBT classifiers, despite being comparable to other state-of-the-art methods, is insufficient for clinical usage, a fact that is often masked by class imbalance and inadequate evaluation metrics. This corroborates the findings following a review on ML methods for cancer characterization from the microbiome.

An ML model that seeks multiple cancer types at the same time, in a multiclass setting, from mbDNA, was also developed for the first time. This model achieved comparable performance to multiple standalone classifiers trained to identify samples of only one type of cancer. Such a multiclass model can provide unambiguous predictions of cancer type, while being less computationally costly to train and store. This multiclass model thus aids the application of mbDNA biomarkers for cancer diagnosis.

This work also aimed at providing novel insights into the origin of mbDNA and the association between the tumor and blood-derived microbiome profiles. The comparison between both types of profiles revealed that these are similar, with the matched blood and tumor profiles of the same patients usually being more similar than those of different patients, suggesting that the blood-derived microbiome profiles capture unique characteristics of the tumor microbiome. However, not all genera reflect this high similarity between matched profiles; for some, their abundance in the bloodstream and in the tumor tissues varies greatly and shows little correlation. An analysis of the genera differentially abundant in the tumor and in the blood showed that only a few taxa have different abundances between these two locations. Most of these are underrepresented in the blood-derived profiles.

Almost all ML regressors trained to estimate the tumor microbiome profiles from mbDNA achieved better than random predictions. This corroborates the previous evidence that mbDNA contains information that partially reflects the tumor microbiome, and it further suggests that their association can be modelled. These models indicate a high mobility between the tumor-blood interface for most genera; however, they also suggest different permeabilities of this interface between patients, or a contribution of other reservoirs (the gut, mouth, or others) to the mbDNA. This follows the current most popular hypotheses regarding the origins of mbDNA. Nevertheless, these regressors provided conflicting assessments of the genera with highest mobility between the tumor tissues and blood and showed high variance. This means that, despite capturing associations between blood-derived and tumor microbiome profiles, these models have limited generalizability. Furthermore, the regressors failed at separating the contribution of the tumor microbiome to mbDNA from that of other reservoirs, assuming such multiple contributions exist.

The association between the host and its microbiome is highly complex, and microbiome data has characteristics that make the application of ML models challenging. For once, abundance taxonomical profiles have many zero values and small variations linked to technical phenomena. This leads to most common ML approaches overfitting the training data, something that was seen both in the literature review and the experiments undertaken in this thesis. Thus, the next steps in ML for cancer-related applications using microbiome data likely include the development and application of models which are more robust under zero-inflated data with high variability. Several models have already been proposed for this kind of data [235, 236, 237], some specifically designed for microbial count data [238]. Such models should find microbial signatures associated with cancer characteristics, while filtering out sporadic associations resulting from technical variations. In conjunction with this, Autoencoders and further work on feature selection and dimensionality reduction approaches for microbial count data are expected to improve the performance of ML models for cancer characterization from microbiome data by decreasing their variance. The development of larger cancer-related microbiome datasets with proper decontamination and correction for technical variations could also help decrease the variance of ML models, while also leading to a better generalization of the models developed and, consequently, biomarkers better suited for diagnosis.

The vast majority of past ML approaches, including those evaluated here, take on taxonomic

abundance profiles as input data. However, these may not be the most informative characteristics of the microbiome: for instance, gut microbiome functional profiles were shown to provide an improvement in performance over the taxonomic counterparts when distinguishing colorectal cancer from adenomas [83]. Data integration (for instance, using both taxonomic and functional features, as done in [80], or coupling the taxonomic profiles with other methods for cancer diagnosis such as fecal occult-blood testing [94]) may also improve the classification accuracy in cancer diagnosis, but few publications exploit this for tumors other than colorectal. Further work is expected to be undertaken to better understand how these different types of data and biological characteristics relate to one another, and if data integration can yield more powerful methods for cancer diagnosis and characterization.

Despite the complex dynamics of microbial mobility in the human body, and how little is known about the tumor and blood microbiomes, this dissertation successfully uncovered novel insights into how these two relate and the origins of mbDNA. The ML models used in this work proved essential to do so; however, there is still room for improvement in estimating the tumor microbiome profiles from blood-derived data. For once, it may be unreasonable to assume that the same prediction function results in good predictions for all cancer types, which usually correspond to different locations in the human body, with different microbiomes and potentially different microbial dynamics. Thus, future work should focus this analysis on and develop models for a specific type of cancer, which should lead to improved results. Furthermore, doing so for cancers in which there is more knowledge on the gut, tumor, and blood microbiomes, such as colorectal cancer, could allow the development of better ML models, guided by more accurate insights into the composition and origin of mbDNA. Including information on the gut microbiome is also crucial to better understand the origins of mbDNA, as only then it is possible to quantify the contribution of other reservoirs to the blood microbiome.

Appendix A

Publications found in the literature survey

Table A.1: Publications found in the literature survey.

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[58]	Random Forest and RBF-kernel SVM	Identifying colorectal cancer	Taxonomical Profiles and k-mer Frequencies	Biopsy	None
[82]	DeepEn-Phy (MLPs) and PopPhy-CNN* (Convolutional NN)	Predicting smoking status and Body Mass Index	Taxonomical Profiles as Phylogenetic Trees	Fecal	None
[239]	MLP	Identifying colorectal cancer, liver cirrhosis, obesity, HIV, type 2 diabetes, and IBD	Taxonomical Profiles	Fecal	None
[129]	SVM	Identifying oral carcinoma	Taxonomical and Functional Profiles	Oral swabs	None
[90]	MLP	Identifying colorectal cancer	Taxonomical Profiles	Fecal	VAE and Feature Extension
[122]	Logistic Regression, kernel SVM, Naïve Bayes, Decision Tree, Random Forest, Bagged Decision Trees, AdaBoost, Ensemble, Extra Tree, k-fold Random Forest, Gradient Boosted Trees	Identifying colorectal cancer	Taxonomical Profiles	Biopsy	None
[133]	Random Forest, SVM, MLP	Identifying colorectal cancer, liver cirrhosis, obesity, type 2 diabetes, and IBD	Taxonomical Profiles	Fecal	Autoencoder

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[95]	Random Forest	Identifying colorectal cancer	Taxonomical Profiles	Fecal	mRMR, Fleuret's CMIM, FCBF, XGBoost, Information Gain and other statistical tests
[89]	MLP	Identifying colorectal cancer	Taxonomical Profiles	Fecal	LASSO
[131]	Gradient Boosted Trees	Identifying lung adenocarcinoma and lung squamous cell carcinoma	Taxonomical Profiles	Biopsy	RFE
[240]	Random Forest	Identifying post-cholecystectomy patients	Taxonomical Profiles	Fecal	None
[108]	Random Forest	Identifying oral squamous cell carcinoma in patients with oral submucous fibrosis	Taxonomical Profiles and clinical covariates	Saliva	LEfSe
[124]	Random Forest	Identifying adenoma formers	Taxonomical Profiles	Fecal, oral, non-adenoma, and adnoma-adjacent mucosa	None
[25]	Random Forest, Gradient Boosted Trees, SVM, MLP, CatBoost	Distinguishing colorectal cancer from advanced adenoma	Viral abundances and bacterial SNPs	Fecal	RFE
[241]	MLP with input embedding	Identifying the association between bacteria and colon cancer	Taxa and disease types	None	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[118]	k-Nearest Neighbors, Logistic Model Trees, Additive Logistic Regression, and SVM	Identifying multiple sclerosis, juvenile idiopathic arthritis, myalgic encephalomyelitis/chronic fatigue syndrome, acquired immune deficiency syndrome, stroke, and colorectal cancer	Taxonomical Profiles	Fecal	RFE
[136]	Random Forest and MLP	Predicting the efficacy of immunotherapy in Non-Small Cell Lung Cancer patients	Taxonomical, functional, and protein profiles	Fecal	None
[132]	Random Forest, K-Nearest Neighbours, MLP, SVM, and Graph Convolutional NN	Identifying colorectal cancer, colorectal adenomas, Crohn's disease, ulcerative colitis, irritable bowel syndrome, obesity, cardiovascular disease, and post-acute COVID-19 syndrome	Taxonomical Profiles	Fecal	None
[125]	Random Forest	Identifying colorectal cancer and Crohn's disease, and predicting the response to therapy of Crohn's disease patients	Taxonomical and Functional Profiles and host DNA content	Fecal	None
[42]	Multivariate Cox Regression and CMS-ML	Predicting overall survival and progress-free survival rates	Taxonomical Profiles, Gene Expression, and Histological Images	Biopsy	Multivariate Cox Regression
[30]	Random Forest	Identifying breast cancer	Taxonomical Profiles	Fecal and breast tissue	RFE

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[211]	Random Forest and Bayesian Additive Regression Trees	Identifying colorectal cancer	Taxonomic Profiles	Fecal	Hierarchical feature engineering and AUC-RF
[242]	Binary/Ternary/Ratio, Linear and RBF-SVM*, Random Forest*, and Logistic Regression*	Identifying multiple diseases	Taxonomical and Functional Profiles, and Marker Genes	Variable	Statistical tests
[27]	Random Forest	Identifying cachexia in lung cancer patients	Taxonomical and Metabolic Profiles	Plasma	Statistical Tests
[26]	Random Forest Regressor	Predicting the number of tumors formed on a murine model after antibiotic treatment	Taxonomical Profiles	Fecal	None
[114]	Random Forest and LASSO	Identifying colorectal cancer and adenoma	Viral taxonomical profiles	Fecal	None
[134]	Random Forest	Identifying IBD, colorectal cancer, and ulcerative colitis	Taxonomical and Functional Profiles	Fecal	None and statistical tests
[40]	Logistic Regression, LASSO, ElasticNet, Decision Tree, Random Forest, MLP, SVM and an ensemble of all	Predicting the efficacy of immunotherapy in cancer patients	Taxonomical Profiles	Fecal	None
[43]	Random Forest	Predicting the occurrence of bacteremia in linfoma patients after chemotherapy	Taxonomical Profiles	Fecal	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[48]	Random Forest	Identifying oral squamous cell carcinoma	Taxonomical Profiles	Biopsy	RFE
[28]	SVM and Logistic Regression	Identifying lung cancer	Taxonomical Profiles	Fecal	Statistical Tests, mRMR, and RFE
[116]	Random Forest	Identifying oral cancer	Taxonomical and Functional Profiles	Saliva	RFE
[110]	Random Forest	Identifying hepatocellular carcinoma	Taxonomical Profiles	Liver and peritumor tissue	Random Forest
[13]	Logistic Regression	Identifying oral cancer	Taxonomical and Functional Profiles	Saliva	Variance Filter and Bootstrap Logistic Regression
[117]	Random Forest	Identifying cervical intraepithelial neoplasia	Taxonomical Profiles	Vaginal smear	RFE
[123]	Logistic Regression, L1, L2, and RBF-kernel SVM, Random Forest, Gradient Boosted Trees, Decision Tree	Identifying colorectal cancer	Taxonomical Profiles	Fecal	None
[126]	Random Forest, kernel SVM, and Naïve Bayes	Predicting the survival time of colorectal cancer patients	Taxonomical Profiles and Transcriptomics (differential gene expression)	Biopsy	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[81]	MLP, kernel SVM, Naïve Bayes, Random Forest, and k-Nearest Neighbors	Identifying colorectal cancer	Taxonomical Profiles and clinical covariates	Fecal	Logistic Regression assuming independent features
[75]	LASSO regression, Random Forest	Identifying ovarian cancer	Taxonomical Profiles and Serum Tumor Marker Levels	Peritoneal Fluid	Ensemble of Random Forest, LASSO regression, t-test, Distance Correlation, and Mann-Whitney U Test
[18]	Gradient Boosted Trees	Identifying breast cancer	Taxonomical Profiles	Biopsy	None
[59]	Gradient Boosting Machine (GBM), Generalized Linear Model, MLP and a GBM + MLP ensemble	Identifying chronic obstructive pulmonary disease, asthma, and lung cancer	Accumulated taxonomical profiles from extracellular vesicles	Serum	Statistical Tests
[68]	Random Forest	Identifying acute myeloid leukemia and distinguishing its precursors	Taxonomical Profiles	Blood and Bone Marrow	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[86]	General Regression Neural Network, SVM*, Random Forest*, k-Nearest Neighbors*, Naïve Bayes*, and MLP*	Identifying colorectal cancer	Taxonomical Profiles and Fecal Occult Blood Test results	Fecal	Derivative-based Feature Selection and LASSO Regression
[35]	Stacked Ensemble	Identifying colorectal cancer	Taxonomical Profiles, age, and body mass index	Fecal	None
[67]	Logistic Regression and Gradient Boosting Machine	Identifying breast cancer	Taxonomical Profiles	Serum	LEfSe, statistical tests, and RFE
[88]	Random Forest	Identifying tonsillar cell carcinoma	Taxonomical Profiles	Biopsy	None
[41]	Random Forest	Identifying lung adenocarcinomas and distinguishing between solid and subsolid nodules	Taxonomical Profiles	Biopsy	LEfSe
[29]	LASSO regression, Random Forest	Identifying invasive cervical cancer	Taxonomical Profiles	Fecal	Recursive Feature Elimination
[94]	Kernel SVM	Identifying colorectal cancer	Taxonomical Profiles	Fecal	mRMR

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[31]	MLP, RBF-kernel SVM, Random Forest, AdaBoost, CN2 Rule Inducer, k-Nearest Neighbors, Naïve Bayes, Decision Tree, Logistic Regression	Identifying melanoma in mice	Taxonomical Profiles	Fecal	PCA
[70]	Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, k-Nearest Neighbour, MLP, SVM	Identifying colorectal polyps	Taxonomical Profiles	Biopsy, fecal, rectal swabs	Kruskal-Wallis, MetaStats, LEfSe, Indicator
[84]	Bayesian Modeling	Normalizing taxonomical profiles	Taxonomical Profiles	Fecal	None
[38]	None	Identifying colorectal cancer	Taxonomical Profiles	Fecal	None
[47]	Random Forest, SVM, k-Nearest Neighbors, and Decision Tree	Identifying cancer subtypes	Taxonomical Profiles	Biopsy	mRMR and Recursive Feature Selection
[37]	Random Forest	Identifying colorectal cancer	Taxonomical Profiles	Fecal	None
[97]	Random Forest, SVM, LASSO regression, and ElasticNet	Identifying colorectal cancer, liver cirrosis, obesity, type 2 diabetes, and IBD	Taxonomical Profiles	Fecal and Swabs	Random Forest
[44]	Gradient Boosted Trees	Identifying cancer subtypes	Taxonomical Profiles	Biopsy and blood	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[80]	Multimodal Neural Network, Random Forest*, XGBoost*, LASSO regression*, Principal Component Regression*, and SVM*	Identifying IBD, type 2 diabetes, liver cirrhosis, and colorectal cancer	Taxonomical and Functional Profiles	Fecal	None
[135]	Decision Tree	Predicting cancer treatment outcomes	Taxonomical and Functional Profiles	Fecal	Statistical tests
[40]	Decision Tree, Logistic Regression, MLP, SVM	Predicting cancer treatment outcomes	Taxonomical Features	Fecal	None
[128]	Random Forest, Decision Trees, and Naïve Bayes	Identifying colorectal cancer	Taxonomical Profiles	Fecal	Correlation-based Feature Selection and Maximum Relevance Maximum Distance
[49]	Random Forest	Identifying oral squamous cell carcinoma	Taxonomical Profiles	Saliva, subgenival plaque, tumor surface, and mucosa swabs, and biopsy	None
[39]	Random Forest	Predicting the survival time of neuroblastoma patients	Taxonomical Profiles	Biopsy	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[83]	Explainable Boosting Machine	Identifying colorectal cancer	Taxonomical and Functional Profiles	Fecal	ANOVA
[243]	MLP	Identifying colorectal cancer, obesity, and underweight patients	Taxonomical Profiles	Fecal	None
[244]	Multi-task Neural Network, MLP*, SVM*, Random Forest*, Logistic Regression*, Naïve Bayes*	Predicting immunotherapy treatment outcomes in non-small cell lung cancer and renal cell carcinoma	Taxonomical Profiles	Fecal	None
[180]	Ensamble of Graph Neural Networks, SVM*, MLP*, Random Forest*, Deep Forest*, and Gradient Boosted Trees*	Identifying colorectal cancer and inflammatory bowel disease	Taxonomical Profiles	Fecal	None
[69]	Logistic Regression and Gradient Boosting Machine	Identifying brain cancer	Taxonomical Profiles	Serum and Biopsy	Akaike Information Criterion, LefSe, Gradient Boosting Machine
[36]	Random Forest	Identifying colorectal cancer	Taxonomical Profiles	Fecal	Abundance thresholds
[130]	Random Forest	Identifying breast cancer	Taxonomical Profiles	Fecal	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[71]	Logistic Regression	Identifying bladder cancer	Taxonomical Profiles	Urine	Abundance thresholds
[112]	Logistic Regression	Predicting the response to CAR-T cell therapy	Taxonomical Profiles, age, and gender	Fecal	Lasso
[120]	Logistic Regression and SVM	Identifying colorectal cancer	Taxonomical Profiles	Fecal	RFE
[102]	Random Forest, SVM, and Generalized Linear Models	Predicting cervical cancer metastasis	Taxonomical and Gene Expression Profiles	Biopsy	Statistical tests
[245]	Logistic Regression, Random Forest, MLP, Gradient Boosted Trees, and SVM	Distinguish poorly differentiated from moderately differentiated colorectal cancer	Taxonomical Profiles	Fecal	None
[246]	Logistic Regression	Predicting the response to immune checkpoint inhibitor treatment	Taxonomical and Functional Profiles	Fecal	None
[247]	Random Forest, Logistic Regression, XGBoost, Gradient Boosted Trees, and SVM	Identifying colorectal adenomas and carcinomas	Taxonomical Profiles	Fecal	None
[103]	Random Forest	Identifying non-small cell lung carcinoma	Taxonomical Profiles	Fecal and sputum	Statistical tests and RFE
[248]	Random Forest	Identifying pre and post-treatment patients with colorectal carcinoma and adenoma	Taxonomical Profiles	Fecal	None

Publications found in the literature survey (continued).

Publication	ML Model	Learning Task	Input Features	Sample Type	Feature Selection Method
[249]	All models available in the Weka 3.6.13 package for Java	Identifying colorectal cancer	Taxonomical Profiles	Fecal	None
[109]	Random Forest, Naïve Bayes, Gradient Boosted Trees	Predicting recurrence and metastasis of lung cancer	Taxonomical Profiles and Host Gene Expression	Biopsy	LEfSe and Statistical Tests
[250]	Random Forest, SVM, MLP	Predicting response to treatment of melanoma patients	Taxonomical Profiles	Fecal	Autoencoder
[251]	Logistic Regression, Random Forest, and SVM	Identifying colorectal cancer and Crohn's disease, and predicting immunotherapy response	Taxonomical Profiles	Fecal	Boruta
[179]	AggMapNet (CNN)	Identifying cirrhosis, obesity, type 2 diabetes, inflammatory bowel disease, and colorectal cancer	Taxonomical Profiles	Fecal	MEGMA
[252]	Logistic Regression, Random Forest, MLP, Gradient Boosted Trees, and SVM	Identifying colorectal cancer lymph node metastasis	Taxonomical Profiles	Fecal	None

References

- [1] Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, et al. Current understanding of the human microbiome. *Nature Medicine*, 24(4):392–400, April 2018. doi:[10.1038/nm.4517](https://doi.org/10.1038/nm.4517).
- [2] Sam Behjati and Patrick S Tarpey. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.*, 98(6):236–238, December 2013.
- [3] Duo Jiang, Courtney R Armour, Chenxiao Hu, et al. Microbiome multi-omics network analysis: Statistical considerations, limitations, and opportunities. *Front. Genet.*, 10:995, November 2019.
- [4] Juan Jovel, Jordan Patterson, Weiwei Wang, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.*, 7:459, April 2016.
- [5] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.
- [6] Nihal Hasan and Hongyi Yang. Factors affecting the composition of the gut microbiota, and its modulation. *PeerJ*, 7:e7502, August 2019. doi:[10.7717/peerj.7502](https://doi.org/10.7717/peerj.7502).
- [7] Kerri L Glassner, Bincy P Abraham, and Eamonn M M Quigley. The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.*, 145(1):16–27, January 2020.
- [8] Bing-Nan Liu, Xiao-Tong Liu, Zi-Han Liang, and Ji-Hui Wang. Gut microbiota in obesity. *World J. Gastroenterol.*, 27(25):3837–3850, July 2021.
- [9] Weiguang Chen, Fanlong Liu, Zongxin Ling, et al. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One*, 7(6):e39743, June 2012.
- [10] Marilia Carabotti, Annunziata Scirocco, Maria Antonietta Maselli, and Carola Severi. The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems. *Ann. Gastroenterol. Hepatol.*, 28(2):203–209, 2015.
- [11] R Teles and C-Y Wang. Mechanisms involved in the association between periodontal diseases and cardiovascular disease: Periodontal diseases and cardiovascular disease. *Oral Diseases*, 17(5):450–461, July 2011. doi:[10.1111/j.1601-0825.2010.01784.x](https://doi.org/10.1111/j.1601-0825.2010.01784.x).
- [12] R. C. V. Casarin, A. Barbagallo, T. Meulman, et al. Subgingival biodiversity in subjects with uncontrolled type-2 diabetes and chronic periodontitis: *Subgingival biodiversity in diabetes mellitus*. *Journal of Periodontal Research*, 48(1):30–36, February 2013. doi:[10.1111/j.1600-0765.2012.01498.x](https://doi.org/10.1111/j.1600-0765.2012.01498.x).
- [13] Guruduth Banavar, Oyetunji Ogundijo, Ryan Toma, et al. The salivary metatranscriptome as an accurate diagnostic indicator of oral cancer. *npj Genomic Medicine*, 6(1):105, December 2021. doi:[10.1038/s41525-021-00257-x](https://doi.org/10.1038/s41525-021-00257-x).
- [14] Xiaozhou Fan, Alexander V Alekseyenko, Jing Wu, et al. Human oral microbiome and prospective risk for pancreatic cancer: A population-based nested case-control study. *Gut*, 67(1):120–127, January 2018. doi:[10.1136/gutjnl-2016-312580](https://doi.org/10.1136/gutjnl-2016-312580).
- [15] Burkhardt Flemer, Ryan D Warren, Maurice P Barrett, et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, 67(8):1454–1463, August 2018. doi:[10.1136/gutjnl-2017-314814](https://doi.org/10.1136/gutjnl-2017-314814).
- [16] Ikuko Kato, Adrian A. Vasquez, Gregory Moyerbrailean, et al. Oral microbiome and history of smoking and colorectal cancer. *Journal of Epidemiological Research*, 2(2):p92, April 2016. doi:[10.5430/jer.v2n2p92](https://doi.org/10.5430/jer.v2n2p92).
- [17] Eun Ju Cho, Sangseob Leem, Sun Ah Kim, et al. Circulating Microbiota-Based Metagenomic Signature for Detection of Hepatocellular Carcinoma. *Sci. Rep.*, 9(1):7536, May 2019.
- [18] Glenn Hogan, Julia Eckenberger, Neegam Narayanan, et al. Biopsy bacterial signature can predict patient tissue malignancy. *Sci. Rep.*, 11(1):18535, September 2021.

- [19] Victoria Pascal, Marta Pozuelo, Natalia Borruel, et al. A microbial signature for Crohn's disease. *Gut*, 66(5):813, May 2017. doi:10.1136/gut.jnl-2016-313235.
- [20] Congmin Xu, Man Zhou, Zhongjie Xie, et al. LightCUD: A program for diagnosing IBD based on human gut microbiome data. *BioData Mining*, 14(1):2, December 2021. doi:10.1186/s13040-021-00241-2.
- [21] Jonas Halfvarson, Colin J. Brislaw, Regina Lamendella, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2(5):17004, February 2017. doi:10.1038/nmicrobiol.2017.4.
- [22] Yvonne Yijuan Lim, Yung Seng Lee, and Delicia Shu Qin Ooi. Engineering the Gut Microbiome for Treatment of Obesity: A Review of Current Understanding and Progress. *Biotechnology Journal*, 15(10):2000013, October 2020. doi:10.1002/biot.202000013.
- [23] J Ferlay, M Colombet, I Soerjomataram, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer*, 144(8):1941–1953, April 2019.
- [24] *WHO Methods and Data Sources for Country-Level Causes of Death: 2000-2019*. World Health Organization, 2020.
- [25] Shuwen Han, Jing Zhuang, Yuefen Pan, et al. Different Characteristics in Gut Microbiome between Advanced Adenoma Patients and Colorectal Cancer Patients by Metagenomic Analysis. *Microbiology Spectrum*, 10(6):e01593–22, December 2022. doi:10.1128/spectrum.01593-22.
- [26] Joseph P. Zackular, Nielson T. Baxter, Grace Y. Chen, and Patrick D. Schloss. Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere*, 1(1):e00001–15, February 2016. doi:10.1128/mSphere.00001-15.
- [27] Yueqiong Ni, Zoltan Lohinai, Yoshitaro Heshiki, et al. Distinct composition and metabolic functions of human gut microbiota are associated with cachexia in lung cancer patients. *ISME J.*, 15(11):3207–3220, November 2021.
- [28] Yajuan Zheng, Zhaoyuan Fang, Yun Xue, et al. Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes*, 11(4):1030–1042, July 2020. doi:10.1080/19490976.2020.1737487.
- [29] Gi-Ung Kang, Da-Ryung Jung, Yoon Hee Lee, et al. Dynamics of Fecal Microbiota with and without Invasive Cervical Cancer and Its Application in Early Diagnosis. *Cancers*, 12(12):3800, December 2020. doi:10.3390/cancers12123800.
- [30] Na Wang, Jun Yang, Wenjie Han, et al. Identifying distinctive tissue and fecal microbial signatures and the tumor-promoting effects of deoxycholic acid on breast cancer. *Frontiers in Cellular and Infection Microbiology*, 12:1029905, December 2022. doi:10.3389/fcimb.2022.1029905.
- [31] Marco Rossi, Salvatore M Aspromonte, Frederick J Kohlhapp, et al. Gut Microbial Shifts Indicate Melanoma Presence and Bacterial Interactions in a Murine Model. *Diagnostics*, 12(4):958, 2022.
- [32] Leore T. Geller, Michal Barzily-Rokni, Tal Danino, et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, 357(6356):1156–1160, September 2017. doi:10.1126/science.aah5043.
- [33] Vyara Matson, Jessica Fessler, Riyue Bao, et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science*, 359(6371):104–108, January 2018. doi:10.1126/science.aao3290.
- [34] Bertrand Routy, Emmanuelle Le Chatelier, Lisa Derosa, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*, 359(6371):91–97, January 2018. doi:10.1126/science.aan3706.
- [35] Yusuke Konishi, Shintaro Okumura, Tomonori Matsumoto, et al. Development and evaluation of a colorectal cancer screening method using machine learning-based gut microbiota analysis. *Cancer Medicine*, 11(16):3194–3206, 2022.
- [36] Ryza Rynazal, Kota Fujisawa, Hirotsugu Shiroma, et al. Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biology*, 24(1):21, February 2023. doi:10.1186/s13059-023-02858-4.
- [37] Manasi S Shah, Todd Z DeSantis, Thomas Weinmaier, et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut*, 67(5):882–891, May 2018.
- [38] Zhongkun Zhou, Shiqiang Ge, Yang Li, et al. Human Gut Microbiome-Based Knowledgebase as a Biomarker Screening Tool to Improve the Predicted Probability for Colorectal Cancer. *Front. Microbiol.*, 11:596027, November 2020.

- [39] Xin Li, Xiaoqi Wang, Ruihao Huang, et al. The Machine-Learning-Mediated Interface of Microbiome and Genetic Risk Stratification in Neuroblastoma Reveals Molecular Pathways Related to Patient Survival. *Cancers*, 14(12), June 2022.
- [40] Hai Liang, Jay-Hyun Jo, Zhiwei Zhang, et al. Predicting cancer immunotherapy response from gut microbiomes using machine learning models. *Oncotarget*, 13(1):876–889, 2022. doi:10.18632/oncotarget.28252.
- [41] Yi Ma, Mantang Qiu, Shaodong Wang, et al. Distinct tumor bacterial microbiome in lung adenocarcinomas manifested as radiological subsolid nodules. *Transl. Oncol.*, 14(6):101050, June 2021.
- [42] Xuan-Yu Mao, Jesus Perez-Losada, Mar Abad, et al. iCEMIGE: Integration of Cell-morphometrics, Microbiome, and GENE biomarker signatures for risk stratification in breast cancers. *World Journal of Clinical Oncology*, 13(7):616–629, July 2022. doi:10.5306/wjco.v13.i7.616.
- [43] Emmanuel Montassier, Gabriel A. Al-Ghalith, Tonya Ward, et al. Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Medicine*, 8(1):49, December 2016. doi:10.1186/s13073-016-0301-4.
- [44] Gregory D. Poore, Evguenia Kopylova, Qiyun Zhu, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, 579(7800):567–574, March 2020. doi:10.1038/s41586-020-2095-1.
- [45] The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. doi:10.1038/ng.2764.
- [46] Rebecca M. Rodriguez, Brenda Y. Hernandez, Mark Menor, et al. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing. *Computational and Structural Biotechnology Journal*, 18:631–641, 2020. doi:10.1016/j.csbj.2020.03.003.
- [47] Lei Chen, Zhandong Li, Tao Zeng, et al. Identifying Robust Microbiota Signatures and Interpretable Rules to Distinguish Cancer Subtypes. *Front Mol Biosci*, 7:604794, November 2020.
- [48] Jianhua Zhou, Lili Wang, Rongtao Yuan, et al. Signatures of Mucosal Microbiome in Oral Squamous Cell Carcinoma Identified Using a Random Forest Model. *Cancer Management and Research*, Volume 12:5353–5363, July 2020. doi:10.2147/CMAR.S251021.
- [49] Xinxuan Zhou, Yu Hao, Xian Peng, et al. The Clinical Potential of Oral Microbiota as a Screening Tool for Oral Squamous Cell Carcinomas. *Front. Cell. Infect. Microbiol.*, 11:728933, August 2021.
- [50] Yan Chen, Fa-Hong Wu, Peng-Qiang Wu, et al. The Role of The Tumor Microbiome in Tumor Development and Its Treatment. *Frontiers in Immunology*, 13:935846, July 2022. doi:10.3389/fimmu.2022.935846.
- [51] Joanne Cummins and Mark Tangney. Bacteria and tumours: Causative agents or opportunistic inhabitants? *Infectious Agents and Cancer*, 8(1):11, December 2013. doi:10.1186/1750-9378-8-11.
- [52] Tamara Glyn and Rachel Purcell. Circulating Bacterial DNA: A New Paradigm for Cancer Diagnostics. *Frontiers in Medicine*, 9:831096, April 2022. doi:10.3389/fmed.2022.831096.
- [53] Hong Sheng Cheng, Sin Pei Tan, David Meng Kit Wong, et al. The Blood Microbiome and Health: Current Evidence, Controversies, and Challenges. *International Journal of Molecular Sciences*, 24(6):5633, March 2023. doi:10.3390/ijms24065633.
- [54] Graeme Lawrence, Ingvild Midtervoll, Sven Ove Samuelsen, et al. The blood microbiome and its association to cardiovascular disease mortality: Case-cohort study. *BMC Cardiovascular Disorders*, 22(1):344, December 2022. doi:10.1186/s12872-022-02791-7.
- [55] Cedric C. S. Tan, Karrie K. K. Ko, Hui Chen, et al. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. *Nature Microbiology*, 8(5):973–985, March 2023. doi:10.1038/s41564-023-01350-w.
- [56] Georg Zeller, Julien Tap, Anita Y Voigt, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.*, 10(11):766, November 2014.
- [57] James A. Nichols, Hsien W. Herbert Chan, and Matthew A. B. Baker. Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, 11(1):111–118, February 2019. doi:10.1007/s12551-018-0449-9.
- [58] Ali Kishk, Asmaa Elzizy, Dina Galal, et al. 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*, 2018.
- [59] Andrea McDowell, Juwon Kang, Jinho Yang, et al. Machine-learning algorithms for asthma, COPD, and lung cancer risk assessment using circulating microbial extracellular vesicle data and their application to assess dietary effects. *Experimental & Molecular Medicine*, 54(9):1586–1595, September 2022. doi:10.1038/s12276-022-00846-5.

- [60] Matthew Uzelac, Yuxiang Li, Jaideep Chakladar, et al. Archaea Microbiome Dysregulated Genes and Pathways as Molecular Targets for Lung Adenocarcinoma and Squamous Cell Carcinoma. *International Journal of Molecular Sciences*, 23(19):11566, September 2022. doi:10.3390/ijms231911566.
- [61] Katriina Whitaker. Earlier diagnosis: The importance of cancer symptoms. *Lancet Oncol.*, 21(1):6–8, January 2020.
- [62] Bernard Levin, Durado Brooks, Robert A Smith, and Amy Stone. Emerging technologies in screening for colorectal cancer: CT colonography, immunochemical fecal occult blood tests, and stool screening using molecular markers. *CA Cancer J. Clin.*, 53(1):44–55, 2003.
- [63] Heidi D Nelson, Amy Cantor, Linda Humphrey, et al. *Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation*. Agency for Healthcare Research and Quality (US), Rockville (MD).
- [64] Katalin Balázs, Lilla Antal, Géza Sáfrány, and Katalin Lumniczky. Blood-derived biomarkers of diagnosis, prognosis and therapy response in prostate cancer patients. *J. Pers. Med.*, 11(4):296, April 2021.
- [65] Adriana F Neves, Jaqueline D Does Dias-Oliveira, Thaise G Araújo, et al. Prostate cancer antigen 3 (PCA3) RNA detection in blood and tissue samples for prostate cancer diagnosis. *Clin. Chem. Lab. Med.*, 51(4):881–887, April 2013.
- [66] Minetta C Liu. Transforming the landscape of early cancer detection using blood tests-Commentary on current methodologies and future prospects. *Br. J. Cancer*, 124(9):1475–1477, April 2021.
- [67] Jeongshin An, Jinho Yang, Hyungju Kwon, et al. Prediction of breast cancer using blood microbiome and identification of foods for breast cancer prevention. *Scientific Reports*, 13(1):5110, March 2023. doi:10.1038/s41598-023-32227-x.
- [68] Jakob Woerner, Yidi Huang, Stephan Hutter, et al. Circulating microbial content in myeloid malignancy patients is associated with disease subtypes and patient outcomes. *Nat. Commun.*, 13(1):1038, February 2022.
- [69] Jinho Yang, Hyo Eun Moon, Hyung Woo Park, et al. Brain tumor diagnostic model and dietary effect based on extracellular vesicle microbiome data in serum. *Exp. Mol. Med.*, 52(9):1602–1613, September 2020.
- [70] Ezzat Dadkhah, Masoumeh Sikaroodi, Louis Korman, et al. Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterol*, 6(1):e000297, May 2019.
- [71] Laura Bukavina, Ilaha Isali, Rashida Ginwala, et al. Global Meta-analysis of Urine Microbiome: Colonization of Polycyclic Aromatic Hydrocarbon-degrading Bacteria Among Bladder Cancer Patients. *European Urology Oncology*, 6(2):190–203, April 2023. doi:10.1016/j.euo.2023.02.004.
- [72] Yu Cao, Séamus Fanning, Sinéad Proos, et al. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. *Front. Microbiol.*, 8:1829, September 2017.
- [73] Cecilia Noecker, Colin P McNally, Alexander Eng, and Elhanan Borenstein. High-resolution characterization of the human microbiome. *Transl. Res.*, 179:7–23, January 2017.
- [74] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014. doi:10.1186/gb-2014-15-2-r29.
- [75] Ruizhong Miao, Taylor C Badger, Kathleen Groesch, et al. Assessment of peritoneal microbial features and tumor marker levels as potential diagnostic tools for ovarian cancer. *PLoS One*, 15(1):e0227707, January 2020.
- [76] Yan He, J Gregory Caporaso, Xiao-Tao Jiang, et al. Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. *Microbiome*, 3:20, May 2015.
- [77] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.*, 20(4):1125–1136, July 2019.
- [78] Roni Rasnic, Nadav Brandes, Or Zuk, and Michal Linial. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer*, 19(1):783, August 2019.
- [79] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, September 2020. doi:10.1093/nargab/lqaa078.
- [80] Seung Jae Lee and Mina Rho. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Sci. Rep.*, 12(1):824, January 2022.
- [81] Dandan Zhao, Hong Liu, Yuanjie Zheng, et al. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med. Biol. Eng. Comput.*, 57(4):901–912, April 2019.

- [82] Wodan Ling, Youran Qi, Xing Hua, and Michael C Wu. Deep ensemble learning over the microbial phylogenetic tree (DeepEn-Phy). In *2021 IEEE International Conference on Bioinformatics and Biomedicine*, 2021.
- [83] Carlos S Casimiro-Soriguer, Carlos Loucera, María Peña-Chilet, and Joaquin Dopazo. Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer. *Sci. Rep.*, 12(1):450, January 2022.
- [84] Shuang Jiang, Guanghua Xiao, Andrew Y Koh, et al. HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity. *Front. Genet.*, 11:445, June 2020.
- [85] Noemi Corsini and Cinzia Viroli. Dealing with overdispersion in multivariate count data. *Computational Statistics & Data Analysis*, 170:107447, June 2022. doi:10.1016/j.csda.2022.107447.
- [86] Abazar Arabameri, Davud Asemani, and Pegah Teymourpour. Detection of Colorectal Carcinoma Based on Microbiota Analysis Using Generalized Regression Neural Networks and Nonlinear Feature Selection. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 17(2):547–557, 2020.
- [87] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.*, 97(105524):105524, December 2020.
- [88] Angelina De Martin, Mechthild Lütge, Yves Stanossek, et al. Distinct microbial communities colonize tonsillar squamous cell carcinoma. *Oncoimmunology*, 10(1):1945202, July 2021.
- [89] Mwenge Mulenga, Sameem Abdul Kareem, Aznul Qalid Md Sabri, and Manjeevan Seera. Stacking and Chaining of Normalization Methods in Deep Learning-Based Classification of Colorectal Cancer Using Gut Microbiome Data. *IEEE Access*, 9:97296–97319, 2021. doi:10.1109/ACCESS.2021.3094529.
- [90] Mwenge Mulenga, Sameem Abdul Kareem, Aznul Qalid Md Sabri, et al. Feature Extension of Gut Microbiome Data for Deep Neural Network-Based Colorectal Cancer Classification. *IEEE Access*, 9:23565–23578, 2021.
- [91] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.
- [92] A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015. doi:10.1109/MIPRO.2015.7160458.
- [93] Jianyu Miao and Lingfeng Niu. A survey on feature selection. *Procedia Comput. Sci.*, 91:919–926, 2016.
- [94] Biao Yuan, Bin Ma, Jing Yu, et al. Fecal Bacteria as Non-Invasive Biomarkers for Colorectal Adenocarcinoma. *Front. Oncol.*, 11:664321, August 2021.
- [95] Amhar Jabeer, Aysegul Kocak, Huseyin Akkas, et al. Identifying Taxonomic Biomarkers of Colorectal Cancer in Human Intestinal Microbiota Using Multiple Feature Selection Methods. In *2022 Innovations in Intelligent Systems and Applications Conference*, 2022.
- [96] Leandro C Hermida, E Michael Gertz, and Eytan Ruppin. Predicting cancer prognosis and drug response from the tumor microbiome. *Nat. Commun.*, 13(1):2896, May 2022.
- [97] Edoardo Pasolli, Duy Tin Truong, Faizan Malik, et al. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.*, 12(7):e1004977, July 2016.
- [98] Ian T Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.*, 374(2065):20150202, April 2016.
- [99] H B Mann and D R Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.*, 18(1):50–60, March 1947.
- [100] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, 47(260):583–621, December 1952.
- [101] Tae Kyun Kim. Understanding one-way ANOVA using conceptual figures. *Korean Journal of Anesthesiology*, 70(1):22, 2017. doi:10.4097/kjae.2017.70.1.22.
- [102] Lu Jiang, Baofeng Duan, Peng Jia, et al. The Role of Intratumor Microbiomes in Cervical Cancer Metastasis. *Cancers*, 15(2):509, January 2023. doi:10.3390/cancers15020509.
- [103] Hui Lu, Na L. Gao, Fan Tong, et al. Alterations of the Human Lung and Gut Microbiomes in Non-Small Cell Lung Carcinomas and Distant Metastasis. *Microbiology Spectrum*, 9(3):e00802–21, December 2021. doi:10.1128/Spectrum.00802–21.
- [104] J Jaeger, R Sengupta, and W L Ruzzo. Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing 2003*, 2003.

- [105] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, 3(2):185–205, April 2005.
- [106] Nicola Segata, Jacques Izard, Levi Waldron, et al. Metagenomic biomarker discovery and explanation. *Genome Biol.*, 12(6):R60, June 2011.
- [107] Frank Wilcoxon. Individual comparisons by ranking methods. *Biom. bull.*, 1(6):80, December 1945.
- [108] M.Y. Chen, J.W. Chen, L.W. Wu, et al. Carcinogenesis of Male Oral Submucous Fibrosis Alters Salivary Microbiomes. *Journal of Dental Research*, 100(4):397–405, April 2021. doi:10.1177/0022034520968750.
- [109] Xiangfeng Zhou, Lei Ji, Yanyu Ma, et al. Intratumoral Microbiota-Host Interactions Shape the Variability of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma in Recurrence and Metastasis. *Microbiology Spectrum*, pages e03738–22, April 2023. doi:10.1128/spectrum.03738-22.
- [110] Jian-Hang Huang, Jie Wang, Xiao-Qiang Chai, et al. The Intratumoral Bacterial Metataxonomic Signature of Hepatocellular Carcinoma. *Microbiology Spectrum*, 10(5):e00983–22, October 2022. doi:10.1128/spectrum.00983-22.
- [111] Bjoern H Menze, B Michael Kelm, Ralf Masuch, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213, July 2009.
- [112] Christoph K. Stein-Thoeringer, Neeraj Y. Saini, Eli Zamir, et al. A non-antibiotic-disrupted gut microbiome is associated with clinical responses to CD19-CAR-T cell cancer immunotherapy. *Nature Medicine*, 29(4):906–916, April 2023. doi:10.1038/s41591-023-02234-6.
- [113] B Venkatesh and J Anuradha. A review of Feature Selection and its methods. *Cybern. Inf. Technol.*, 19(1):3–26, March 2019.
- [114] Fang Chen, Shenghui Li, Ruochun Guo, et al. Meta-analysis of fecal viromes demonstrates high diagnostic potential of the gut viral signatures for colorectal cancer and adenoma risk assessment. *Journal of Advanced Research*, page S2090123222002144, October 2022. doi:10.1016/j.jare.2022.09.012.
- [115] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, April 2015.
- [116] Jung-Wen Chen, Jer-Horng Wu, Wei-Fan Chiang, et al. Taxonomic and Functional Dysregulation in Salivary Microbiomes During Oral Carcinogenesis. *Frontiers in Cellular and Infection Microbiology*, 11:663068, September 2021. doi:10.3389/fcimb.2021.663068.
- [117] Yoon Hee Lee, Gi-Ung Kang, Se Young Jeon, et al. Vaginal Microbiome-Based Bacterial Signatures for Predicting the Severity of Cervical Intraepithelial Neoplasia. *Diagnostics*, 10(12):1013, November 2020. doi:10.3390/diagnostics10121013.
- [118] Sohyun Bang, DongAhn Yoo, Soo-Jin Kim, et al. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Scientific Reports*, 9(1):10189, July 2019. doi:10.1038/s41598-019-46249-x.
- [119] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. *Mach. Learn.*, 46(1/3):389–422, 2002.
- [120] Shi-Long Zhang, Li-Sha Cheng, Zheng-Yan Zhang, et al. Untangling determinants of gut microbiota and tumor immunologic status through a multi-omics approach in colorectal cancer. *Pharmacological Research*, 188:106633, February 2023. doi:10.1016/j.phrs.2022.106633.
- [121] Baris Senliol, Gokhan Gulgezen, Lei Yu, and Zehra Cataltepe. Fast Correlation Based Filter (FCBF) with a different search strategy. In *2008 23rd International Symposium on Computer and Information Sciences*, 2008.
- [122] Faisal Bin Ashraf, Md Shafiur Raihan Shafi, and Md Rayhan Kabir. Host trait prediction from human microbiome data for Colorectal Cancer. In *2020 23rd International Conference on Computer and Information Technology*, 2020.
- [123] Begüm D Topçuoğlu, Nicholas A Lesniak, Mack T Ruffin, 4th, et al. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *MBio*, 11(3), June 2020.
- [124] Katherine M. Watson, Ivy H. Gardner, Sudarshan Anand, et al. Colonic Microbial Abundances Predict Adenoma Formers. *Annals of Surgery*, 277(4):e817–e824, April 2023. doi:10.1097/SLA.0000000000005261.
- [125] Puzi Jiang, Senying Lai, Sicheng Wu, et al. Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics*, 21(1):348, December 2020. doi:10.1186/s12864-020-6749-z.
- [126] Min Yang, Huandong Yang, Lei Ji, et al. A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.*, 146:105516, July 2022.

- [127] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, August 2006.
- [128] Kaiyang Qu, Feng Gao, Fei Guo, and Quan Zou. Taxonomy dimension reduction for colorectal cancer prediction. *Computational Biology and Chemistry*, 83:107160, 2019. doi:10.1016/j.compbiolchem.2019.107160.
- [129] Julia D. Warnke-Sommer and Hesham H. Ali. Evaluation of the oral microbiome as a biomarker for early detection of human oral carcinomas. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2069–2076, 2017. doi:10.1109/BIBM.2017.8217979.
- [130] Rachel L. Shrode, Jessica E. Knobbe, Nicole Cady, et al. Breast cancer patients from the Midwest region of the United States have reduced levels of short-chain fatty acid-producing gut bacteria. *Scientific Reports*, 13(1):526, January 2023. doi:10.1038/s41598-023-27436-3.
- [131] Matthew Uzelac, Yuxiang Li, Jaideep Chakladar, et al. Archaea Microbiome Dysregulated Genes and Pathways as Molecular Targets for Lung Adenocarcinoma and Squamous Cell Carcinoma. *International Journal of Molecular Sciences*, 23(19):11566, September 2022. doi:10.3390/ijms231911566.
- [132] Qi Su, Qin Liu, Raphaela Iris Lau, et al. Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nature Communications*, 13(1):6818, November 2022. doi:10.1038/s41467-022-34405-3.
- [133] Dinithi Wickramaratne, Rupika Wijesinghe, and Ruwan Weerasinghe. Human Gut Microbiome Data Analysis for Disease Likelihood Prediction Using Autoencoders. In *2021 21st International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 49–54, 2021. doi:10.1109/ICTer53630.2021.9774811.
- [134] Puzi Jiang, Sicheng Wu, Qibin Luo, et al. Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. *mSystems*, 6(3):e00112–21, June 2021. doi:10.1128/mSystems.00112-21.
- [135] Yoshitaro Heshiki, Ruben Vazquez-Urbe, Jin Li, et al. Predictable modulation of cancer treatment outcomes by the gut microbiota. *Microbiome*, 8(1):28, March 2020. doi:10.1186/s40168-020-00811-2.
- [136] Ben Liu, Justin Chau, Qun Dai, et al. Exploring Gut Microbiome in Predicting the Efficacy of Immunotherapy in Non-Small Cell Lung Cancer. *Cancers*, 14(21):5401, November 2022. doi:10.3390/cancers14215401.
- [137] Jill C Stoltzfus. Logistic regression: A brief primer. *Acad. Emerg. Med.*, 18(10):1099–1104, October 2011.
- [138] J Ranstam and J A Cook. LASSO regression. *Br. J. Surg.*, 105(10):1348, September 2018.
- [139] Andrew Y Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Twenty-First International Conference on Machine Learning - ICML '04*, New York, New York, USA, 2004. ACM Press.
- [140] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 2010.
- [141] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.
- [142] William S Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.
- [143] C Schuldt, I Laptev, and B Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. IEEE, 2004.
- [144] G Camps-Valls and L Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1351–1362, 2005.
- [145] Nick Guenther and Matthias Schonlau. Support vector machines. *Stata J.*, 16(4):917–937, December 2016.
- [146] G Camps-Valls, L Gomez-Chova, J Munoz-Mari, et al. Composite kernels for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.*, 3(1):93–97, January 2006.
- [147] Sasan Karamizadeh, Shahidan M Abdullah, Mehram Halimi, et al. Advantage and drawback of support vector machine functionality. In *2014 International Conference on Computer, Communications, and Control Technology*, 2014.
- [148] S B Kotsiantis, I D Zaharakis, and P E Pintelas. Machine learning: A review of classification and combining techniques, 2006.
- [149] J R Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.
- [150] J Ross Quinlan. *C4.5: Programs for Machine Learning*. Elsevier, June 2014.
- [151] S B Kotsiantis. Decision trees: A recent overview. *Artificial Intelligence Review*, 39(4):261–283, 2013.
- [152] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nat. Biotechnol.*, 26(9):1011–1013, September 2008.
- [153] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

- [154] Artur J Ferreira and Mário A T Figueiredo. Boosting algorithms: A review of methods, theory, and applications. In *Ensemble Machine Learning*, pages 35–85. Springer US, Boston, MA, 2012.
- [155] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pages 23–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995.
- [156] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 2016.
- [157] Guolin Ke, Qi Meng, Thomas Finley, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pages 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [158] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, et al. CatBoost: Unbiased boosting with categorical features. 2017.
- [159] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: An overview and their use in medicine. *J. Med. Syst.*, 26(5):445–463, October 2002.
- [160] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.
- [161] Trevor Hastie and Robert Tibshirani. Generalized Additive Models; Some Applications. *Generalized Linear Models*, pages 66–81, 1985.
- [162] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.
- [163] Aaron E Maxwell, Maneesh Sharma, and Kurt A Donaldson. Explainable Boosting Machines for Slope Failure Spatial Predictive Modeling. *Remote Sensing*, 13(24):4991, 2021.
- [164] Rich Caruana, Yin Lou, Johannes Gehrke, et al. Intelligible Models for HealthCare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.
- [165] Yoav Goldberg. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.
- [166] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU). *arXiv*, 2018. doi:10.48550/ARXIV.1803.08375.
- [167] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016.
- [168] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, August 2018.
- [169] Mufti Mahmud, Mohammed Shamim Kaiser, Amir Hussain, and Stefano Vassanelli. Applications of Deep Learning and Reinforcement Learning to Biological Data. *IEEE Trans Neural Netw Learn Syst*, 29(6):2063–2079, June 2018.
- [170] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J Big Data*, 8(1):53, March 2021.
- [171] Frank Emmert-Streib, Zhen Yang, Han Feng, et al. An Introductory Review of Deep Learning for Prediction Models With Big Data. *Front Artif Intell*, 3:4, February 2020.
- [172] Derek Reiman, Ahmed A Metwally, Jun Sun, and Yang Dai. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data. *IEEE J Biomed Health Inform*, 24(10):2993–3001, October 2020.
- [173] Beatriz García-Jiménez, Jorge Muñoz, Sara Cabello, et al. Predicting microbiomes through a deep latent space. *Bioinformatics*, 37(10):1444–1451, June 2021.
- [174] Min Oh and Liqing Zhang. DeepMicro: Deep representation learning for disease prediction based on microbiome data. *Sci. Rep.*, 10(1):1–9, April 2020.
- [175] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6(1), December 2019.
- [176] D F Specht. A general regression neural network. *IEEE Trans. Neural Netw.*, 2(6):568–576, 1991.
- [177] Shaikh Abdul Hannan, R R Manza, and R J Ramteke. Generalized regression neural network and radial basis function for heart disease diagnosis. *Int. J. Comput. Appl.*, 7(13):7–13, October 2010.

- [178] Ahmad Jobran Al-Mahasneh, Sreenatha G Anavatti, and Matthew A Garratt. Review of applications of Generalized Regression Neural Networks in identification and control of dynamic systems. 2018.
- [179] Wan Xiang Shen, Shu Ran Liang, Yu Yang Jiang, and Yu Zong Chen. Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations. *Patterns*, 4(1):100658, January 2023. doi:10.1016/j.patter.2022.100658.
- [180] K. Syama, J. Angel Arul Jothi, and Namita Khanna. Automatic disease prediction from human gut metagenomic data using boosting GraphSAGE. *BMC Bioinformatics*, 24(1):126, March 2023. doi:10.1186/s12859-023-05251-x.
- [181] Zonghan Wu, Shirui Pan, Fengwen Chen, et al. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, January 2021. doi:10.1109/TNNLS.2020.2978386.
- [182] Bruno Cavadas, Rui Camacho, Joana C. Ferreira, et al. Gastric Microbiome Diversities in Gastric Cancer Patients from Europe and Asia Mimic the Human Population Structure and Are Partly Driven by Microbiome Quantitative Trait Loci. *Microorganisms*, 8(8):1196, August 2020. doi:10.3390/microorganisms8081196.
- [183] Martin Lauss, Ilhami Visne, Albert Kriegner, et al. Monitoring of technical variation in quantitative high-throughput datasets. *Cancer Inform.*, 12:193–201, September 2013.
- [184] Yiwen Wang and Kim-Anh LêCao. Managing batch effects in microbiome data. *Brief. Bioinform.*, 21(6):1954–1970, December 2020.
- [185] Susanna C Larsson, Nicola Orsini, and Alicja Wolk. Diabetes mellitus and risk of colorectal cancer: A meta-analysis. *J. Natl. Cancer Inst.*, 97(22):1679–1687, November 2005.
- [186] K. K. Tsilidis, J. C. Kasimis, D. S. Lopez, et al. Type 2 diabetes and cancer: Umbrella review of meta-analyses of observational studies. *BMJ*, 350(jan02 1):g7607–g7607, January 2015. doi:10.1136/bmj.g7607.
- [187] Wei-Zheng Li, Kyle Stirling, Jun-Jie Yang, and Lei Zhang. Gut microbiota and diabetes: From correlation to causality and mechanism. *World J. Diabetes*, 11(7):293–308, July 2020.
- [188] Zheng Zhou, Bao Sun, Dongsheng Yu, and Chunsheng Zhu. Gut Microbiota: An important player in type 2 diabetes mellitus. *Front. Cell. Infect. Microbiol.*, 12:834485, February 2022.
- [189] Nathalie Japkowicz. Assessment metrics for imbalanced learning. In *Imbalanced Learning*, pages 187–206. John Wiley & Sons, Inc., Hoboken, NJ, USA, June 2013.
- [190] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. 2017.
- [191] Chenzhao Feng, Tianyu Xiang, Zixuan Yi, et al. A Deep-Learning Model With the Attention Mechanism Could Rigorously Predict Survivals in Neuroblastoma. *Frontiers in Oncology*, 11, 2021.
- [192] Lionel P Joseph, Erica A Joseph, and Ramendra Prasad. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. *Comput. Biol. Med.*, 151(Pt A):106178, December 2022.
- [193] Mei Lin, Vanya Jaitly, Iris Wang, et al. Application of Deep Learning on predicting prognosis of acute myeloid leukemia with cytogenetics, age, and mutations. *arXiv*, 2018.
- [194] Yu-Feng Huang, Yen-Ju Chen, Tan-Chi Fan, et al. Analysis of microbial sequences in plasma cell-free DNA for early-onset breast cancer patients and healthy females. *BMC Medical Genomics*, 11(S1):16, February 2018. doi:10.1186/s12920-018-0329-y.
- [195] Thomas N.Y. Kwong, Xiansong Wang, Geicho Nakatsu, et al. Association Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer. *Gastroenterology*, 155(2):383–390.e8, August 2018. doi:10.1053/j.gastro.2018.04.028.
- [196] Abraham Gihawi, Colin S. Cooper, and Daniel S. Brewer. Caution Regarding the Specificities of Pan-Cancer Microbial Structure. Preprint, Bioinformatics, January 2023. doi:10.1101/2023.01.16.523562.
- [197] Derrick E Wood and Steven L Salzberg. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, March 2014. doi:10.1186/gb-2014-15-3-r46.
- [198] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010. doi:10.1186/gb-2010-11-3-r25.
- [199] Brigham H. Meacham, Peter S. Nelson, and John D. Storey. Supervised normalization of microarrays. *Bioinformatics*, 26(10):1308–1315, May 2010. doi:10.1093/bioinformatics/btq118.
- [200] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. 2012.
- [201] Léon Bottou. Stochastic Gradient Descent Tricks. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-35289-8_25.

- [202] Ben Carterette. Precision and Recall. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 2126–2127. Springer US, Boston, MA, 2009. doi:10.1007/978-0-387-39940-9_5050.
- [203] Cyril Goutte and Eric Gaussier. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In David Hutchison, Takeo Kanade, Josef Kittler, et al., editors, *Advances in Information Retrieval*, volume 3408, pages 345–359. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. doi:10.1007/978-3-540-31865-1_25.
- [204] Helen R. Sofaer, Jennifer A. Hoeting, and Catherine S. Jarnevich. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577, April 2019. doi:10.1111/2041-210X.13140.
- [205] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. doi:10.1016/S0031-3203(96)00142-2.
- [206] André M. Carrington, Paul W. Fieguth, Hammad Qazi, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, 20(1):4, December 2020. doi:10.1186/s12911-019-1014-6.
- [207] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. doi:10.1613/jair.953.
- [208] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
- [209] Hiroshi Fukui. Increased Intestinal Permeability and Decreased Barrier Function: Does It Really Influence the Risk of Inflammation? *Inflammatory Intestinal Diseases*, 1(3):135–145, 2016. doi:10.1159/000447252.
- [210] Greta M. De Waal, Willem J. S. De Villiers, Timothy Forgan, et al. Colorectal cancer is associated with increased circulating lipopolysaccharide, inflammation and hypercoagulability. *Scientific Reports*, 10(1):8777, May 2020. doi:10.1038/s41598-020-65324-2.
- [211] Yi-Hui Zhou and George Sun. Improve the Colorectal Cancer Diagnosis Using Gut Microbiome Data. *Frontiers in Molecular Biosciences*, 9:921945, August 2022. doi:10.3389/fmolb.2022.921945.
- [212] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6, April 2018. doi:10.1016/j.cell.2018.03.022.
- [213] Kendra K. Maass, Paulina S. Schad, Agnes M. E. Finster, et al. From Sampling to Sequencing: A Liquid Biopsy Pre-Analytic Workflow to Maximize Multi-Layer Genomic Information from a Single Tube. *Cancers*, 13(12):3002, June 2021. doi:10.3390/cancers13123002.
- [214] Patrizia D’Aquila, Robertina Giacconi, Marco Malavolta, et al. Microbiome in Blood Samples From the General Population Recruited in the MARK-AGE Project: A Pilot Study. *Frontiers in Microbiology*, 12:707515, July 2021. doi:10.3389/fmicb.2021.707515.
- [215] Sandrine Païssé, Carine Valle, Florence Servant, et al. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing: BLOOD MICROBIOME 16S METAGENOMIC SEQUENCING. *Transfusion*, 56(5):1138–1147, May 2016. doi:10.1111/trf.13477.
- [216] Yifan Xie, Feng Xie, Xiaoxue Zhou, et al. Microbiota in Tumors: From Understanding to Application. *Advanced Science*, 9(21):2200470, July 2022. doi:10.1002/advs.202200470.
- [217] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, December 2013. doi:10.1186/1471-2105-14-91.
- [218] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, March 2012. doi:10.1093/bioinformatics/bts034.
- [219] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
- [220] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426.
- [221] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [222] David Sims, Ian Sudbery, Nicholas E. Ilott, et al. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, February 2014. doi:10.1038/nrg3642.
- [223] Georgy Meshcheryakov. Conorm, October 2021.

- [224] Guy E. Thwaites and Vanya Gant. Are bloodstream leukocytes Trojan Horses for the metastasis of *Staphylococcus aureus*? *Nature Reviews Microbiology*, 9(3):215–222, March 2011. doi:10.1038/nrmicro2508.
- [225] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. doi:10.48550/ARXIV.1502.03167.
- [226] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 2015. doi:10.48550/ARXIV.1512.03385.
- [227] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.
- [228] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. 2013. doi:10.48550/ARXIV.1308.3432.
- [229] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7045–7054. PMLR, July 2020.
- [230] Mehmet E. Aktas, Esra Akbas, and Ahmed El Fatmaoui. Persistence homology of networks: Methods and applications. *Applied Network Science*, 4(1):61, December 2019. doi:10.1007/s41109-019-0179-3.
- [231] Bastian Rieck. Pytorch-topological.
- [232] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [233] Richard Liaw, Eric Liang, Robert Nishihara, et al. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018. arXiv:1807.05118.
- [234] Pearson’s Correlation Coefficient. In Wilhelm Kirch, editor, *Encyclopedia of Public Health*, pages 1090–1091. Springer Netherlands, Dordrecht, 2008. doi:10.1007/978-1-4020-5614-7_2569.
- [235] Shufeng Kong, Junwen Bai, Jae Hee Lee, et al. Deep hurdle networks for zero-inflated multi-target regression: Application to multiple species abundance estimation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*, Yokohama, Yokohama, Japan, 2021.
- [236] Yao Lu and Yong Liao. STS: A novel deep learning method for zero-inflated crime prediction. In *Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, RICAI ’22*, pages 1097–1103, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3584376.3584571.
- [237] Mingqiu Wei, Rongjie Liu, Yue Julia Wang, and Chao Huang. Deep zero-inflated negative binomial model and its application in scRNA-seq data integration. In *SoutheastCon 2023*, pages 901–905, Orlando, FL, USA, April 2023. IEEE. doi:10.1109/SoutheastCon51012.2023.10115099.
- [238] Takeshi Osawa, Hiromune Mitsushashi, Yuta Uematsu, and Atushi Ushimaru. Bagging GLM: Improved generalized linear model for the analysis of zero-inflated data. *Ecological Informatics*, 6(5):270–275, September 2011. doi:10.1016/j.ecoinf.2011.05.003.
- [239] Thanh Hai Nguyen and Jean-Daniel Zucker. Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches. In *2019 11th International Conference on Knowledge and Systems Engineering*, 2019.
- [240] Xinhua Ren, Jun Xu, Yuanyuan Zhang, et al. Bacterial Alterations in Post-Cholecystectomy Patients Are Associated With Colorectal Cancer. *Frontiers in Oncology*, 10:1418, August 2020. doi:10.3389/fonc.2020.01418.
- [241] Yue Liu, Shu-Lin Wang, Jun-Feng Zhang, et al. DMFMDA: Prediction of Microbe-Disease Associations Based on Deep Matrix Factorization Using Bayesian Personalized Ranking. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5):1763–1772, September 2021. doi:10.1109/TCBB.2020.3018138.
- [242] Edi Prifti, Yann Chevalere, Blaise Hanczar, et al. Interpretable and accurate prediction models for metagenomics data. *GigaScience*, 9(3):giaa010, March 2020. doi:10.1093/gigascience/giaa010.
- [243] D.N. Wickramaratne, C.R. Wijesinghe, and A.R. Weerasinghe. A Deep Learning Approach to Predict Health Status Using Microbiome Profiling. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 075–079, Colombo, Sri Lanka, November 2022. IEEE. doi:10.1109/ICTer58063.2022.10024102.

- [244] Changzhi Jiang, Yousi Fu, Shuting Jin, et al. A multi-task learning method for analyzing microbiota as cancer immunotherapy signal. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 437–441, Seoul, Korea (South), December 2020. IEEE. doi:10.1109/BIBM49941.2020.9313509.
- [245] Zhang Qi, Zuo Zhibo, Zhuang Jing, et al. Prediction model of poorly differentiated colorectal cancer (CRC) based on gut bacteria. *BMC Microbiology*, 22(1):312, December 2022. doi:10.1186/s12866-022-02712-w.
- [246] Karla A. Lee, Andrew Maltez Thomas, Laura A. Bolte, et al. Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nature Medicine*, 28(3):535–544, March 2022. doi:10.1038/s41591-022-01695-5.
- [247] Courtney R. Armour, Begüm D. Topçuoğlu, Andrea Garretto, and Patrick D. Schloss. A Goldilocks Principle for the Gut Microbiome: Taxonomic Resolution Matters for Microbiome-Based Classification of Colorectal Cancer. *mBio*, 13(1):e03161–21, February 2022. doi:10.1128/mbio.03161-21.
- [248] Marc A. Sze, Nielson T. Baxter, Mack T. Ruffin, et al. Normalization of the microbiota in patients after treatment for colonic lesions. *Microbiome*, 5(1):150, December 2017. doi:10.1186/s40168-017-0366-3.
- [249] Luoyan Ai, Haiying Tian, Zhaofer Chen, et al. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget*, 8(6):9546–9556, February 2017. doi:10.18632/oncotarget.14488.
- [250] Min Oh and Liqing Zhang. DeepGeni: Deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy. *Scientific Reports*, 13(1):4599, March 2023. doi:10.1038/s41598-023-31210-w.
- [251] Kuncheng Song and Yi-Hui Zhou. Leveraging Scheme for Cross-Study Microbiome Machine Learning Prediction and Feature Evaluations. *Bioengineering*, 10(2):231, February 2023. doi:10.3390/bioengineering10020231.
- [252] Wu Yinhang, Zhuang Jing, Zhou Jie, et al. Prediction model of colorectal cancer (CRC) lymph node metastasis based on intestinal bacteria. *Clinical and Translational Oncology*, 25(6):1661–1672, January 2023. doi:10.1007/s12094-022-03061-w.