A perspective on the use of machine learning on the human microbiome for phenotype prediction and possible adaptations for plant microbiomes

By

Doğukan Bayraktar

Supervisor: Ronnie de Jonge

GSLS Writing Assignment
Bioinformatics and Biocomplexity
Utrecht University

24/07/2024

Contents

List of Figures

List of Tables

1 Plain language summary

We used to view organisms as individuals when studying their health, but now we understand that their health is significantly influenced by the microbes that live inside or on them. These microbes can affect how an organism, like a plant or human, grows, develops, and stays healthy. Through DNA sequencing, we can detect the presence of these microbes. By studying which microbes are present, how many of them are present, and what genes they carry, we gain insight into how they influence their host. Because unravelling the influence of microbes on a host is so complicated, we make use of machine learning.

In machine learning, we teach computers to recognize patterns and make decisions based on information. This can be done by giving examples with correct answers, providing feedback on predictions, and having computers make decisions by answering a series of questions. Classic machine learning tends to use clear, specific instructions which are easily interpretable by humans. Deep learning is a type of machine learning where a computer "learns" by observing a lot of information and "practising" predictions and getting feedback on it. It's difficult to figure out what a computer exactly learns in deep learning. But when there is enough data, it easily and significantly outperforms the classic approaches.

Because of the success of deep learning in related fields, we investigated if it can be used to predict the health of plants based on the microbes that live on and around them. Deep learning use with microbial data is relatively new and has been in use for human health predictions for a longer time. Hence, we explored the methods used in human predictions and if those methods can be used for health predictions in plants. Plants, however, are more diverse than humans and host a greater variety of microbes, influenced by soil and crop types. Applying human prediction methods to plants isn't technically difficult as the structure of the data is the same, but the diversity requires much more data for effective deep learning. Additionally, deep learning's complexity limits our understanding, posing challenges for plant health predictions. The current use of deep learning is mostly focused on reducing microbial data to key information or representing data in different ways by focusing on aspects such as their evolutionary relationships. Its performance currently isn't much better than traditional machine learning, and we don't have enough big plant datasets to really take advantage of it. In the future, deep learning work should aim to be more understandable, and combine types of data from different fields.

2 Abstract

There has been a fundamental shift from viewing organisms as isolated entities to considering them as a holobiont. This change highlights the critical role of microbiomes in influencing host phenotypes in plants and humans. The microbiome of a host organism has a direct relationship to its phenotype, and it has an impact on various physiological traits playing roles in immune response, nutrient acquisition, and protection against pathogens. To explore the potential of deep learning (DL) models for the classification of host phenotypes in plants, we reviewed studies that use classic machine learning (ML) and DL for the prediction of microbiome related host disease phenotypes (Inflammatory Bowel Disease, colorectal cancer, type II diabetes, obesity, arthritis, liver cirrhosis) in humans using microbiome data. As these methods have been more extensively applied for such use cases.

In classic ML applications, the data is processed in bioinformatic pipelines to produce abundance tables and find gene pathways. Features are often selected through an expert-driven process, and Support vector machines and Random Forests are the most successful classifiers.

Studies either transform standard metagenomic outputs such as abundance tables into different forms such as phylogenetic trees or "synthetic images" or perform dimensionality reduction through DL. Features from these methods are then fed to both DL and classic ML classifiers. DL classifiers such as Convolutional Neural Networks (CNNs), and Multilayer perceptrons (MLPs) occasionally outperform traditional ML techniques by small margins. However, current applications fail to leverage significant advantages such as end-to-end prediction and automatic feature selection. Moreover, there is a severe lack of explainability when using DL classifiers, while classic approaches such as Random Forest are very explainable. Which is critical to understand the underlying processes that cause disease phenotypes in the host.

As human and plant microbiome sequencing data shares the same structure, there is no technical blockade for applying the methods reviewed in this study to plant microbiomes. However, plant microbiomes are much more diverse, exhibiting compositional differences among different types of plants. This necessitates larger datasets to appropriately generalize DL models. As is, DL classifiers already face limitations due to feature-sample size imbalance in metagenomics, on top of limited dataset availability. This issue will only be exacerbated in plant applications. Although DL holds promise for future large-scale microbiome analysis, its current performance and the need for explainable models and extensive datasets remain significant hurdles. Future DL classification in this field should focus on interpretability, end-to-end prediction, and multi-omics integration.

3 Introduction

3.1 The microbiome of an organism is highly correlated with its phenotype

There has been a change in our view of biological systems from seeing organisms as separate entities without considering how they interact with their microbiome to the idea of the holobiont. The concept underlines the relationship between a host organism, such as a human or plant, and their microbiome and its influence on the host phenotype [1]. The microbiome consists of microorganisms, such as bacteria, archaea, viruses, and eukaryotes, residing in a specific environment, like the human body or the rhizosphere of a plant, along with their properties and interactions. [2, 3]. The microbiome of a host organism has a direct relationship to its phenotype, and it has an impact on numerous physiological traits [4, 5, 6]. The immune system and disease processes of a host are linked to the microbiome, which is a critical component of both human and plant health. Plant and human microbiomes are similar in their relationship to their host and its impact on them, despite the differences in composition and diversity [1]. In both instances, the microbiome is associated with the acquisition of nutrients, the functioning of the immune system, and protection against pathogens [1, 7]. For example, genes found in the microbiota of the human gut can help the digestive system by providing enzymes and biochemical pathways that the host themselves does not possess [8]. The human microbiome is also important for the function, and training, of the immune system [9]. An example of this is its ability to provide protection by excluding other microorganisms that could potentially be harmful [8]. Similar to humans, plants also depend on their microbiome for particular traits and functions. Such traits can be nutrient acquisition through the roots and resiliency to stress factors [1]. An example of this is in the rhizosphere, which is comparable to the human gut microbiome. It is the section of soil surrounding the plant roots, here microorganisms such as nitrogen-fixing bacteria help the plant to take up nutrients and other necessary chemical compounds [1]. The effect of the microbiome on the host phenotype and the concept of the holobiont highlight the need to not only study hosts individually. It is essential that the micro-

biome of an organism is also considered when studying them. Not just in the case of illness or dysfunction, but also for potential health enhancements through the use of probiotics. We can study the microbiome of an organism and gain insight into its composition, abundance, and diversity through sequencing.

3.2 High-throughput sequencing increased the amount of microbiome data, which opened the way for ML and DL

Through the advent of next-generation sequencing, we are capable of analysing DNA and RNA at the genetic, transcriptomic, and epigenetic layers, which has allowed us to gain insight into the biological functions of organisms. The study of the microbiome is largely done through Illumina next-generation sequencing, which has allowed for significant developments in the field, with many large datasets being generated [10]. Amplicon sequencing and shotgun metagenomics sequencing are the two most common methods used to study microbiomes.

In amplicon sequencing, the 16s rRNA region is targeted for sequencing due to its high conservation among all bacteria and archaea. The strong conservation of this gene makes it an ideal target for amplification through the use of primers [11]. Conveniently, the gene also includes nine hypervariable regions, which can be used to classify bacteria at the genus level [11]. Furthermore, the 16s rRNA gene is absent in eukaryotes, which makes it useful for studying prokaryotic diversity in environments such as the human gut or plant rhizosphere as no host contamination can occur. This method is selected as an alternative to shotgun metagenomic sequencing, often to save costs. It offers limited taxonomic resolution and does not provide insights into the aspects of a microbial community beyond its composition [12].

In shotgun metagenomics, DNA is extracted directly from an environment and no specific genes are targeted, and therefore the produced samples contain a mixture of genomes including the host. The DNA is then fragmented into small pieces (hence "shotgun") and sequenced randomly. Unlike amplicon sequencing, metagenomic shotgun sequencing allows for taxonomic identification at higher resolutions and functional pathway analysis [13]. Processed reads are either checked against a database or assembled into contigs from which genes are then annotated or predicted, allowing functional pathways to be determined [13] (Figure 1). A disadvantage of shotgun metagenomics is host contamination, this refers to the presence of host DNA in the microbiome samples. This occurs because all DNA in the sample is sequenced and not a specific region that is only present in bacteria and archaea, like in amplicon sequencing. Host contamination is a serious problem that can lead to miss identification of microbial DNA and false genetic variability that will be elaborated on in section 4.2 [14, 15]. These sequencing approaches have made it cost-efficient and accessible to sequence large amounts of microbiome data. However, microbiome data is complex and predicting host phenotypes from it is a challenging task. Normalization of the data is required, and classic machine learning and deep learning methods are applied to make predictions. Machine learning techniques have been applied to human microbiomes since the 2010s, and deep learning applications gained more traction in the 2020s with the aim to predict disease phenotypes of the host.

3.3 Machine learning and deep learning for the microbiome

Machine learning has been increasingly applied in microbiome studies to predict the host phenotypes of humans and plants through the analysis of their microbiome [16, 7, 17]. Machine learning methods can be divided into supervised- and unsupervised-learning. In supervised learning, a model gets trained on a labelled dataset. This means that each data point has a label that describes what outcome it belongs to. In the context of microbiome data, this can

look like two sets of human gut samples, one from a cohort with colorectal cancer and one healthy group, each being labelled as such [17]. The purpose of this approach is for the classifier to learn patterns from the two groups that lead to different outcomes. Then the model can use the patterns to discriminate between healthy and disease microbiome samples in other datasets. This application is referred to as classification and can be done for discrete labels and regression [17]. Some examples of classic supervised learning algorithms are Linear Regression (LM), Support Vector Machines (SVM), and Random Forest (RF). Supervised machine learning models are often applied for such phenotype classifications in metagenomics [16, 17]. Unsupervised learning is performed on unlabelled datasets and is applied when labelled data is not available or for the purpose of dimensionality reduction. A frequently used application is clustering, where data points are grouped together based on similarity. In dimensionality reduction, the variables under consideration are reduced to the most important variables that contain the most information. Unsupervised methods are not applicable for phenotype classification, but are instead often used in data preprocessing and feature selection steps that involve microbiome data [17]. Some examples are the clustering of amplicon sequencing data to form Operational Taxonomic Units (OTUs) and the binning of contigs after assembly in shotgun metagenomics.

Deep learning is a category of machine learning that utilizes artificial neural networks to learn complex patterns in large datasets. Recent advancements in deep learning, the most important being the "self attention" mechanism, have resulted in a resurgence in the utilization of neural networks [18]. Deep learning can be done both supervised and unsupervised. Some techniques used for supervised learning are MLPs, Recurrent Neural Networks (RNNs), and CNNs. These architectures can be used to classify microbiome samples when making discrete prediction or to predict gene expression levels when doing regression tasks. [17, 19]. Autoencoders (AE) and Variational Autoencoders (VAEs) can be used for unsupervised approaches, mainly for the purpose of clustering or dimensionality reduction in metagenomic datasets[19]. Typically, deep learning model performance increases with the amount of available data [19]. Deep learning also takes care of feature engineering, models can automatically learn and extract features from raw data while classic machine learning usually requires expert-driven feature selection, especially in biological applications such as metagenomics [20, 19]. In this context, features could be the presence or absence of specific strains of bacteria or genes that are involved in certain metabolic pathways. Another advantage is that deep learning has the capability for end-to-end learning, this means that data can be transformed directly from input to output without the need for intermediate processing steps. For example, a sequencing read from a microbiome dataset can be directly input into a model and classified as belonging to a healthy or disease sample [21] or be classified as belonging to a particular species of bacteria [22, 19]. Lastly, there is also the concept of transfer learning, models pre-trained on certain datasets can be adjusted for specific tasks, for example, a model that predicts phenotype in humans through microbiome interactions could be transfer learned to apply in plant microbiomes.

Machine learning is being used in many parts of various bioinformatics pipelines for the analysis of microbiome data. SVMs are used in the Illumina basecalling algorithms, Oxford Nanopore sequencing employs deep neural networks for base calling and methylation detection. Clustering is used to create OTUs in amplicon sequencing, and Hidden Markov Models (HMM) and neural networks are used in gene predictions.

3.4 Goal and search strategy

In this review, we will focus on the use of machine learning on human microbiome gut data for the classification of microbiome related host disease phenotypes. As, the application of

deep learning for phenotype prediction in plant microbiomes has not been researched as much. We aim to review scientific articles on the use of machine learning and deep learning in humans to offer a perspective on the possibilities to adapt such methods for application in plant microbiomes.

To achieve this, we formulated the following research questions:

- What are the latest machine learning and deep learning techniques used for predicting human disease phenotypes based on microbiome data?
- How do these techniques perform in metagenomics applications, particularly classic machine learning vs deep learning approaches?
- Can these techniques, be adapted to predict host phenotypes in plant microbiome communities effectively?

The search strategy employed for finding relevant literature included keyword searches using terms such as "machine learning in microbiome research", "deep learning for phenotype prediction", "phenotype prediction using microbiome data", "host phenotype machine learning", "microbiome and disease prediction", "metagenomics machine learning". The main databases used were PubMed, and Google Scholar. The initial literature selection was based on titles and abstracts that seemed relevant. An effort was made to select publications that were published within the past five years. Exceptions were made for relevant papers that exclusively employed traditional machine learning methods for classification. After the initial selection, a snowballing method was employed to incorporate additional studies that were referenced in benchmarks of the initial papers.

4 Key Findings

4.1 The need for normalization of microbiome data

Microbiome data is intrinsically compositional, which means that it represents parts of a whole instead of absolute quantities [2]. In metagenomics, each sample only reflects the relative abundance of microbial taxa in the sampled environment [2]. This means that we do not have true absolute abundance counts of the sequenced taxa. This problem occurs because the total reads in a sample are influenced by factors such as sequencing depth, coverage, sampling bias, and DNA extraction efficiency [23]. Because the total number of reads in a sample is fixed, an increase in the relative abundance of one taxon results in a decrease in the abundance of others. This can selectively inflate negative correlations between taxa [24, 25]. Without normalization, compositional data is subject to such biases. This is why data must be normalized, especially when comparisons are made across different samples. Another issue that must be corrected for is sparseness [26]. Sparseness results in zero counts in the abundance tables after processing. This happens by chance when certain microbes are missed when sampling. Lastly, microbiome data, especially in plants, can be heterogeneous [2]. For example, soil samples from around a plant may contain microhabitats with different microbial communities.

Compositional data can be normalized by performing log-ratio transformations [25], by converting raw read counts to proportions or percentages, or by adjusting the data so that the total read counts across the samples are comparable. However, log-transformations cannot be performed on spurious data because zeros cannot be log-transformed. This is why zeros must be handled in some way. This is done through different approaches, such as adding arbitrary pseudo-counts, replacement techniques or imputations [2]. However, the consequences of normalization techniques such as adding arbitrary pseudo-counts are not clear [2].

A review by Busato et al. [2] showed that many plant microbiome papers do not address these problems, which is a cause for concern. When the mentioned factors are unaccounted for, the resulting technical correlations can lead to biased and incorrect results, which ultimately leads to incorrect interpretations and conclusions [26]. Additionally, training machine learning models on non-normalized data will result in worse outcomes, as noise and sparsity conceal biological signals. Some applications might even be inappropriate, as many statistical and machine learning techniques require the input data to adhere to criteria such as normality. Normalization is a key step in microbiome research that leads to more accurate, and interpretable results.

4.2 Bioinformatic workflows for microbiome data and their use with classic machine learning classifiers

Data obtained through amplicon sequencing are either clustered into OTUs or Amplicon sequence variants (ASVs). With OTUs, sequences are clustered based on similarity, where 97% similarity is the most commonly used threshold [27]. This threshold was widely adapted from a 2005 study by Konstantinidis and Tiedje [28]. The study looked at seventy closely related bacterial genomes and found that bacterial genomes with 16s rRNA gene sequence identity of 97% or higher often belonged to the same species. Creating OTUs has computational benefits, and avoids the problem of differentiating biological from technical sequence variations [27]. A 16s amplicon run has millions of reads, by clustering them the data is reduced to a few thousand OTUs, represented by a single sequence. This single sequence then receives a 16s classification, which is then applied to the entire cluster, which reduces the downstream computational costs. Since technical variations such as sequencing errors with Illumina-reads are low, they are not likely to be less than 97% dissimilar and therefore still get assigned to the correct cluster under the assumption that 97% is an appropriate threshold. The use of percentage-based sequence similarity is not without criticisms. The first being that the commonly used 97% threshold is only a rough approximation that was derived by analysing only seventy genomes, as mentioned earlier. It has been shown that species can have multiple copies of 16s rRNA sequences that differentiate by as much as 5% [29] and that differences between copies of varying species can be as little as 99% [30]. Which is why amplicon sequencing is not precise enough to determine species or strains in a sample, and is instead limited to the genus level. Additionally, the percent similarity does not consider evolutionary processes, such as varying substitution rates and biases [27]. These criticisms imply that it is likely that sequences that do not belong to the same taxa get clustered together and then get incorrectly annotated, which then leads to incorrect conclusions regarding the microbial community composition and diversity. ASVs do not get clustered, instead they get "denoised", which is a process where technical variants get removed. Technical variants, as opposed to true biological variation, are errors introduced during sequencing or sample preparation. Technical variants beside sequencing errors can be PCR artefacts such as chimeras and duplicates. After denoising, each unique sequence in the dataset is considered an ASV. ASVs provide higher taxonomic resolution, which allows for comparison at the nucleotide level [31]. As ASVs represent exact biological sequences, they are comparable across different datasets, whereas OTUs are dependent on the used datasets and therefore less reusable [31]. After creating OTUs or ASVs they are quantified, and taxonomic tables are generated [2]. These tables then represent the abundance and composition of the sequenced samples.

Moving on to shotgun metagenomics, the first preprocessing step and perhaps one of the most important procedures is the removal of host contamination. Not removing host contamination has several consequences. For example, host DNA can overwhelm the desired microbial DNA in a sample during sequencing, which makes it more difficult to detect low-abundance mi-

crobes [15]. This issue also has an impact on cost efficiency, as sequencing host DNA consumes resources without providing relevant information. Contamination can also lead to misidentification of microbial species because host sequences can be miss identified when annotating reads through databases or predicting genes on assembled contigs [32]. This can impact the accuracy of the results, leading to an incorrect conclusion being drawn because of false genetic variability [14]. Because of the severity of these consequences, host contamination must be removed [14]. There are two approaches to removing host contamination. The first approach is to make use of sequencing kits that deplete the host DNA or enrich the desired microbial DNA before sequencing. The second approach is to make use of bioinformatics tools such as Kraken that taxonomically classify reads and then remove reads identified as host DNA. Another similar approach is to align the samples against the genome of the host and then only consider reads that did not align to the genome. Once samples are cleaned, reads are either directly compared to curated databases or assembled into contigs. The databases contain annotated reads which are used to do taxonomic and functional profiling. The assembly-based approach bins contigs after which genes are predicted. Through these methods, abundance, taxonomic composition and functional pathways can be determined from a microbiome sample.

The above described processes fall under the classic bioinformatics pipelines for microbiome analyses. An overview of the processes can be seen in Figure 1. To make predictions of disease phenotypes with classifiers from these outputs, features from the taxonomic and functional tables are selected through either expert-driven or data-driven processes. Expert-driven feature selection is the process of selecting features from available data based on medical expertise and established literature that indicate the relevance of certain factors for a disease or condition. Such features can be patient age, gender or medical history. More specific microbiome features could be the presence of certain species of bacteria, combinations of the presence of multiple bacteria, genes that these bacteria carry, and their involvement in certain metabolic pathways. Data-driven processes involve ranking features based on importance scores, which are determined during the training of a classifier. An example for determining such scores in RF classifiers is the use of the gini importance. In data-driven selection, classifiers are first trained on all available features, then the importance of the features are ranked, and the classifiers are retrained on different selections of the top features. Then the number or combination of features that result in the best performance are selected for the final version of the classifier. This is done because certain features or combinations of features might actually lead to worse prediction capabilities. Features selected through these processes are used as input for classic machine learning methods, which are then used to make predictions.

In microbiome studies, SVM and RF have been the most successful implementations, reporting the highest machine learning scores [16]. The study by Pasolli et al. [16] performed analysis on eight public shotgun metagenomics datasets (Table 1) that in total contained 2424 samples of the human gut. Using RF, SVM, Elastic Net Regularization (ENR), and Least absolute shrinkage and selection operator (Lasso) they predicted the host phenotype for liver cirrhosis, colorectal cancer, Inflammatory bowel disease (IBD), obesity and type II diabetes. They then implemented their work into a software framework called MetaAML [16]. The framework they presented produced excellent results, with RF boasting the overall highest performance as measured by the AUC-ROC scores in the liver cirrhosis (0.95), colorectal cancer (0.87), IBD (0.89), obesity (0.66), and two type II diabetes (0.74 and 0.75) datasets (Table 1). The study took advantage of both species abundance and strain markers as features for their machine learning models. The feature selection process was primarily data-driven, employing RF to determine the most relevant features based on importance scores for the final RF and SVM classifiers. The MetAML RF implementation, RF in general, and the eight datasets used by Pasolli et al. are often used in reviews and by other studies presenting new deep learning

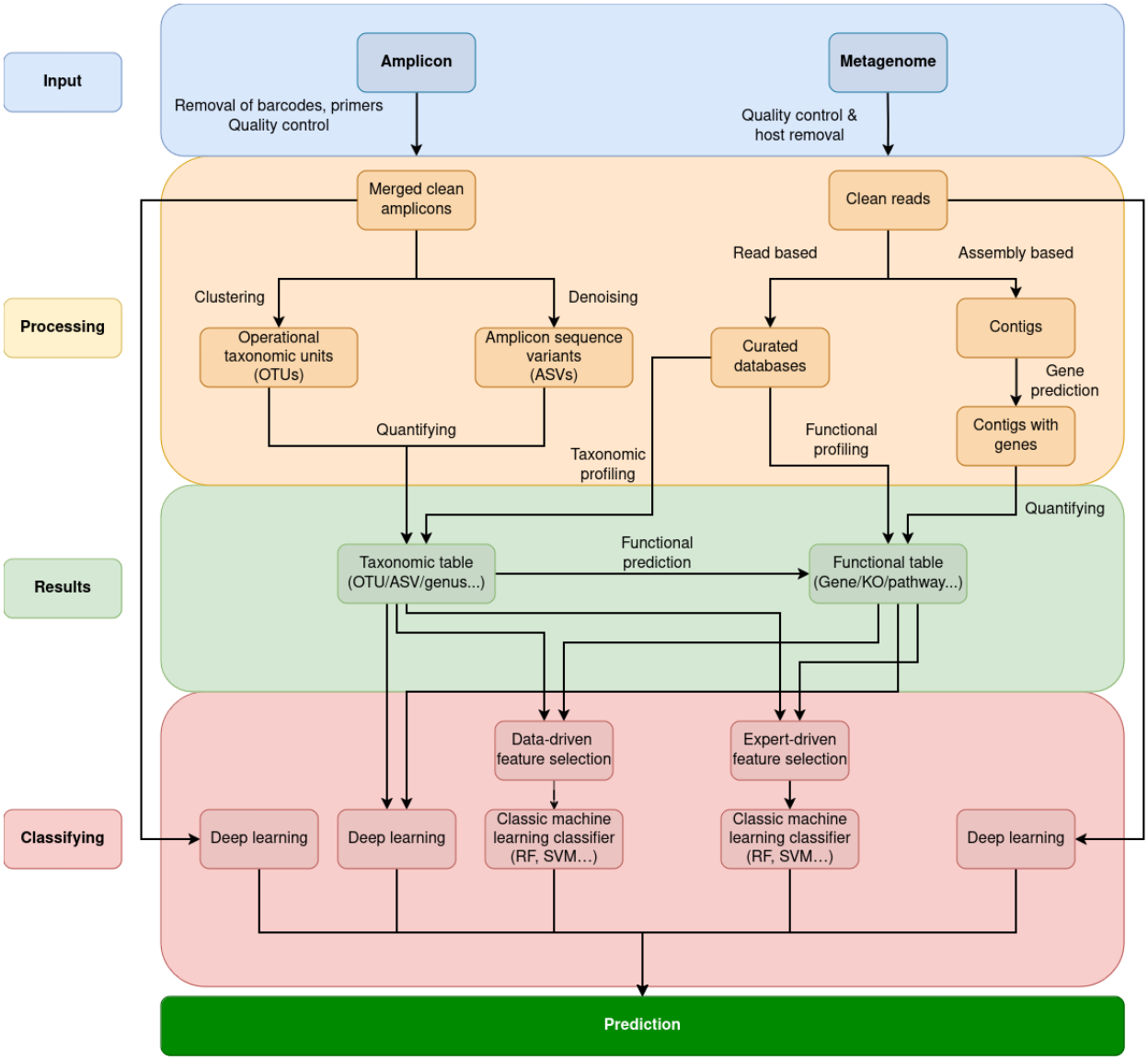

Figure 1: Workflow for amplicon sequencing and shotgun metagenomic sequencing. Figure adapted from Liu et al. [13]. The workflow shows the preprocessing steps from classic machine learning methods and several data flow approaches for deep learning that have been done in reviewed studies.

applications on microbiome data as a benchmark for new phenotype prediction methods [33, 34, 35, 36].

4.3 Datasets used in reviewed research

The key ML and DL microbiome studies (Table 1) predict various disease phenotypes such as liver cirrhosis, colorectal cancer, IBD, obesity, type II diabetes, and arthritis. Most of the reviewed studies use the same public datasets for these predictions, which were originally selected by Pasolli et al. For this reason, we give a short overview of the disease phenotypes and the studies that published the corresponding datasets along with their findings.

The first dataset is for liver cirrhosis, in this disease, the function of the liver is impaired due to the formation of scar tissue as a result of liver damage. As the liver gets more damaged, more scar tissue forms and takes over functioning tissue, further impairing the function of the liver. The disease can be caused by things such as alcohol abuse, various forms of hepatitis, obesity and other types of drug use. The dataset used to study this disease was published by Qin et al. (2014) [6] and consists of 123 samples from healthy patients and 114 samples from patients with liver cirrhosis obtained through shotgun metagenomic sequencing. The study focused on the genes present in the gut microbiota and clustered microbial genes based on abundance profiles across different samples. The authors found that the composition of microbial communities contributed to gene richness, and that gene richness in the disease samples was much lower than in the healthy samples [6]. They concluded that 15 microbial genes can discriminate patients with liver cirrhosis from healthy individuals in this dataset.

The colorectal cancer dataset was published by Zeller et al. [37]. Colorectal cancer is a form of cancer that develops in parts of the large intestine and, like other types of cancer, it occurs when abnormal cells grow uncontrollably, forming tumours. Diagnosis usually occurs through stool blood tests and colonoscopy. The dataset contains 48 case samples and 73 control samples, and explored the potential of gut microbiota for detection of colorectal cancer with shotgun metagenomics. The study suggests that taxonomic markers can be used to distinguish patients from controls using an ENR with accuracy similar to currently used diagnostic tools [37]. Two subspecies of *Fusobacterium nucleatum* (*vincelli* and *animalis*) were found to be major contributors to the model's predictive power in this study.

There are three datasets for the IBD phenotype by Qin et al. (2010) [38], Gevers et al. [39], and Sokol et al. [40]. IBD refers to a group of conditions that affect the colon and small intestines, and is made up of two main types, which are Crohn's disease and ulcerative colitis. The disease causes symptoms such as fatigue, fever, diarrhoea and abdominal pain. Diagnosis can be done through blood tests, stool tests, colonoscopy, and image studies such as CT scans and MRI. The study by Qin et al. (2010) contained 25 samples from both Crohn's disease and ulcerative colitis patients and 85 controls. They found that the gut microbiota in IBD patients had an average of 25% less genes than healthy patients, and were able to clearly separate Crohn's disease patients from healthy patient based on a PCA that used 155 species from the samples [38]. The study by Gevers et al. focused on Crohn's disease and analysed 447 disease samples and 221 control samples using 16s rRNA sequencing [39]. They found positive correlations between the disease and the presence of multiple bacteria such as *Pasteurellaceae* and *Veillonellaceae*. The study by Sokol et al. consists of 235 case samples and 38 control samples using 16s rRNA sequencing and ITS2 sequencing to assess both bacterial and fungal presence.

The obesity dataset was published by Le Chatelier et al., and contained 123 control samples and 169 disease samples. In obesity, body fat has accumulated to a degree that it has a negative impact on one's health. Overconsumption, lack of physical activity, and the use of certain

medications can cause this condition. Obesity is associated with numerous persistent health conditions and health complications, affecting both the quality of life and the life expectancy. In their study, Le Chatelier et al. found two groups across all their samples, one with a high microbial gene count and one with a low microbial gene count [41]. They found that obese patients with low microbial gene counts gained much more weight than obese patients with high microbial gene counts over a period of nine years [41]. The classification scores for obesity tend to be lower than the other disease types across all reviewed studies, as it has proven to be a challenging phenotype to predict from only microbiome data.

There were two datasets for the type II diabetes phenotype by Qin et al. (2012) and Karlsson et al. In type II diabetes, the body becomes resistant to insulin. As the condition progresses, the pancreas fails to produce sufficient amounts of insulin, which leads to high blood sugar levels. If untreated, this imbalance can lead to various metabolic complications. The study by Qin et al. (2012) collected 150 samples, of which 71 did not have the disease. The study identified an imbalance in the gut microbiota of patients with type II diabetes and an increase in opportunistic pathogens. The study by Karlsson et al. had 135 samples with three groups, 53 with type II diabetes, 39 with impaired glucose tolerance and 43 healthy samples. The study determined that there are significant changes in gut microbiota between type II diabetes and healthy individuals, with an increased abundance of *Lactobacillus* species and a decrease in *Clostridium* species in the disease samples.

Table 1: Overview of key reviewed studies and their use of machine learning and deep learning for phenotype prediction. Scores should only be compared within the same studies due to differences in datasets, subsampling and number of cross-validation folds

Study	Study ref	Dataset study	Dataset ref	Dataset type	Dataset origin	Method	Method name	Target Prediction	Normalization	Accuracy	Precision	Recall	F1 Score	AUC	MCC
Pasolli et al., 2016	[16]	Qin et al., 2014	[6]	Shotgun	Human Gut	Random Forest	MetAML-RF	Liver cirrhosis	Yes	-	-	-	-	0.95	-
Pasolli et al., 2016	[16]	Qin et al., 2014	[6]	Shotgun	Human Gut	SVM	MetAML-SVM	Liver cirrhosis	Yes	-	-	-	-	0.92	-
Pasolli et al., 2016	[16]	Qin et al., 2014	[6]	Shotgun	Human Gut	Elastic net	MetAML-ENET	Liver cirrhosis	Yes	-	-	-	-	0.91	-
Pasolli et al., 2016	[16]	Qin et al., 2014	[6]	Shotgun	Human Gut	Lasso	MetAML-LASSO	Liver cirrhosis	Yes	-	-	-	-	0.88	-
Pasolli et al., 2016	[16]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	Random Forest	MetAML-RF	Colorectal cancer	Yes	-	-	-	-	0.87	-
Pasolli et al., 2016	[16]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	SVM	MetAML-SVM	Colorectal cancer	Yes	-	-	-	-	0.81	-
Pasolli et al., 2016	[16]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	Elastic net	MetAML-ENET	Colorectal cancer	Yes	-	-	-	-	0.79	-
Pasolli et al., 2016	[16]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	Lasso	MetAML-LASSO	Colorectal cancer	Yes	-	-	-	-	0.73	-
Pasolli et al., 2016	[16]	Qin et al., 2010	[38]	Shotgun	Human Gut	Random Forest	MetAML-RF	IBD	Yes	-	-	-	-	0.89	-
Pasolli et al., 2016	[16]	Qin et al., 2010	[38]	Shotgun	Human Gut	SVM	MetAML-SVM	IBD	Yes	-	-	-	-	0.86	-
Pasolli et al., 2016	[16]	Qin et al., 2010	[38]	Shotgun	Human Gut	Elastic net	MetAML-ENET	IBD	Yes	-	-	-	-	0.83	-
Pasolli et al., 2016	[16]	Qin et al., 2010	[38]	Shotgun	Human Gut	Lasso	MetAML-LASSO	IBD	Yes	-	-	-	-	0.81	-
Pasolli et al., 2016	[16]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Random Forest	MetAML-RF	Obesity	Yes	-	-	-	-	0.66	-
Pasolli et al., 2016	[16]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	SVM	MetAML-SVM	Obesity	Yes	-	-	-	-	0.65	-
Pasolli et al., 2016	[16]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Elastic net	MetAML-ENET	Obesity	Yes	-	-	-	-	0.64	-
Pasolli et al., 2016	[16]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Lasso	MetAML-LASSO	Obesity	Yes	-	-	-	-	0.60	-
Pasolli et al., 2016	[16]	Qin et al., 2012	[5]	Shotgun	Human Gut	Random Forest	MetAML-RF	Type II diabeters	Yes	-	-	-	-	0.74	-
Pasolli et al., 2016	[16]	Qin et al., 2012	[5]	Shotgun	Human Gut	SVM	MetAML-SVM	Type II diabeters	Yes	-	-	-	-	0.66	-
Pasolli et al., 2016	[16]	Qin et al., 2012	[5]	Shotgun	Human Gut	Elastic net	MetAML-ENET	Type II diabeters	Yes	-	-	-	-	0.70	-
Pasolli et al., 2016	[16]	Qin et al., 2012	[5]	Shotgun	Human Gut	Lasso	MetAML-LASSO	Type II diabeters	Yes	-	-	-	-	0.71	-
Pasolli et al., 2016	[16]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	Random Forest	MetAML-RF	Type II diabeters	Yes	-	-	-	-	0.76	-
Pasolli et al., 2016	[16]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	SVM	MetAML-SVM	Type II diabeters	Yes	-	-	-	-	0.66	-
Pasolli et al., 2016	[16]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	Elastic net	MetAML-ENET	Type II diabeters	Yes	-	-	-	-	0.60	-
Pasolli et al., 2016	[16]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	Lasso	MetAML-LASSO	Type II diabeters	Yes	-	-	-	-	0.54	-
Reiman et al., 2019	[36]	Qin et al., 2014	[6]	Shotgun	Human Gut	CNN	PopPhy-CNN	Liver cirrhosis	Yes	-	-	-	0.857	0.946	-
Reiman et al., 2019	[36]	Qin et al., 2014	[6]	Shotgun	Human Gut	Random Forest	PopPhy-RF	Liver cirrhosis	Unknown	-	-	-	0.876	0.943	-
Reiman et al., 2019	[36]	Qin et al., 2012	[5]	Shotgun	Human Gut	CNN	PopPhy-CNN	Type II diabeters	Unknown	-	-	-	0.620	0.690	-
Reiman et al., 2019	[36]	Qin et al., 2012	[5]	Shotgun	Human Gut	Random Forest	PopPhy-RF	Type II diabeters	Unknown	-	-	-	0.647	0.737	-
Reiman et al., 2019	[36]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	CNN	PopPhy-CNN	Obesity	Unknown	-	-	-	0.621	0.666	-
Reiman et al., 2019	[36]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Random Forest	PopPhy-RF	Obesity	Unknown	-	-	-	0.553	0.683	-
Nguyen et al., 2018	[34]	Qin et al., 2014	[6]	Shotgun	Human Gut	CNN	Met2Img-CNN	Liver cirrhosis	Unknown	0.903	-	-	-	-	-
Nguyen et al., 2018	[34]	Qin et al., 2014	[6]	Shotgun	Human Gut	Random Forest	Met2Img-RF	Liver cirrhosis	Unknown	0.877	-	-	-	-	-
Nguyen et al., 2018	[34]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	CNN	Met2Img-CNN	Colorectal cancer	Unknown	0.798	-	-	-	-	-
Nguyen et al., 2018	[34]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	Random Forest	Met2Img-RF	Colorectal cancer	Unknown	0.812	-	-	-	-	-
Nguyen et al., 2018	[34]	Qin et al., 2010	[38]	Shotgun	Human Gut	CNN	Met2Img-CNN	IBD	Unknown	0.863	-	-	-	-	-
Nguyen et al., 2018	[34]	Qin et al., 2010	[38]	Shotgun	Human Gut	Random Forest	Met2Img-RF	IBD	Unknown	0.808	-	-	-	-	-
Nguyen et al., 2018	[34]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	CNN	Met2Img-CNN	Obesity	Unknown	0.681	-	-	-	-	-
Nguyen et al., 2018	[34]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Random Forest	Met2Img-RF	Obesity	Unknown	0.645	-	-	-	-	-
Nguyen et al., 2018	[34]	Qin et al., 2012	[5]	Shotgun	Human Gut	CNN	Met2Img-CNN	Type II diabeters	Unknown	0.649	-	-	-	-	-
Nguyen et al., 2018	[34]	Qin et al., 2012	[5]	Shotgun	Human Gut	Random Forest	Met2Img-RF	Type II diabeters	Unknown	0.672	-	-	-	-	-
Nguyen et al., 2018	[34]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	CNN	Met2Img-CNN	Type II diabeters	Unknown	0.713	-	-	-	-	-
Nguyen et al., 2018	[34]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	Random Forest	Met2Img-RF	Type II diabeters	Unknown	0.703	-	-	-	-	-
Rahman and Rangwala	[43]	Qin et al., 2014	[6]	Shotgun	Human Gut	NN into RF	ReMIL-RF	Liver cirrhosis	Yes	0.9281	-	-	-	0.9272	-
Rahman and Rangwala	[43]	Qin et al., 2010	[38]	Shotgun	Human Gut	NN into RF	ReMIL-RF	IBD	Yes	0.8466	-	-	-	0.8442	-
Wang et al., 2021	[44]	Scher et al.,2013	[45]	Amplicon	Human Gut	CNN	Mdeep	Arthritis	Yes	07857	0.8245	0.8407	0.8188	-	0.5642
Wang et al., 2021	[44]	Scher et al.,2013	[45]	Amplicon	Human Gut	Random Forest	Mdeep-RF	Arthritis	Yes	0.7629	0.7741	0.8831	0.8089	-	0.5245
Asgari et al., 2018	[46]	Gevers et al., 2014	[39]	Amplicon	Human Gut & body sites	MLP	MicroPheno-MLP	Crohn's disease	Yes	-	0.7	0.7	0.7	-	-
Asgari et al., 2018	[46]	Gevers et al., 2014	[39]	Amplicon	Human Gut & body sites	Random Forest	MicroPheno-RF	Crohn's disease	Yes	-	0.76	0.76	0.76	-	-
Fioravanti et al., 2018	[47]	Sokhol et al., 2017	[40]	Amplicon	Human Gut	CNN	Ph-CNN	IBD	Unknown	-	-	-	-	-	0.751
Fioravanti et al., 2018	[47]	Sokhol et al., 2017	[40]	Amplicon	Human Gut	Random Forest	Ph-CNN	IBD	Unknown	-	-	-	-	-	0.67
Zhao et al., 2021	[21]	Gevers et al., 2014	[39]	Amplicon	Human Gut & body sites	CNN, RNN, attention	Read2Pheno	Crohn's disease	Unknown	0.833	-	-	-	-	-
Li et al., 2021	[48]	Qin et al., 2014	[6]	Shotgun	Human Gut	CNN	TopoPhy-CNN	Liver cirrhosis	Yes	-	85.17	83.24	82.95	90.73	68.38
Li et al., 2021	[48]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	CNN	TopoPhy-CNN	Obesity	Yes	-	62.45	59.29	59.53	61.22	16.65
Li et al., 2021	[48]	Qin et al., 2012	[5]	Shotgun	Human Gut	CNN	TopoPhy-CNN	Type II diabeters	Yes	-	62.14	61.57	61.18	66.67	23.75
Oh and Zhang., 2020	[35]	Qin et al., 2010	[38]	Shotgun	Human Gut	SVM	DeepMicro-SVM	IBD	Unknown	-	-	-	-	0.955	-
Oh and Zhang., 2020	[35]	Karlsson et al., 2013	[42]	Shotgun	Human Gut	Random Forest	DeepMicro-RF	Type II diabeters	Unknown	-	-	-	-	0.899	-
Oh and Zhang., 2020	[35]	Qin et al., 2012	[5]	Shotgun	Human Gut	MLP	DeepMicro-MLP	Type II diabeters	Unknown	-	-	-	-	0.763	-
Oh and Zhang., 2020	[35]	Le chatelier et al., 2013	[41]	Shotgun	Human Gut	Random Forest	DeepMicro-RF	Obesity	Unknown	-	-	-	-	0.659	-
Oh and Zhang., 2020	[35]	Qin et al., 2014	[6]	Shotgun	Human Gut	SVM	DeepMicro-SVM	Liver cirrhosis	Unknown	-	-	-	-	0.940	-
Oh and Zhang., 2020	[35]	Zeller et al. ,2014	[37]	Shotgun	Human Gut	MLP	DeepMicro-MLP	Colorectal cancer	Unknown	-	-	-	-	0.803	-

Table 2: Overview of deep learning phenotype classifiers

Name	Reference	Data	Main objective	Input
PopPhy-CNN	[36]	Shotgun	Phenotype classification	Abundance table + taxon annotation into phylogenetic tree
Met2Img-CNN	[34]	Shotgun	Phenotype classification	Abundance table into "synthetic images"
Mdeep	[44]	Amplicon	Phenotype classification	Abundance table + taxon annotation
MicroPheno-MLP	[46]	Amplicon	Phenotype classification	K-mer distribution
Ph-CNN	[47]	Amplicon	Phenotype classification	Abundance table + taxon annotation into phylogenetic tree
Read2Pheno	[21]	Amplicon	Phenotype classification	Raw reads (amplicon)
TopoPhy-CNN	[48]	Shotgun	Phenotype classification	Abundance table + taxon annotation into phylogenetic tree
DeepMicro-MLP	[35]	Shotgun	Phenotype classification	Abundance table and Gene annotations

4.4 Deep learning approaches for the microbiome

Implementing deep learning in metagenomics is challenging because of the nature of the data [34]. Microbiome data has numerous features (taxa, genes) and the total number of features per sample is much larger than the total sample size. Due to the curse of dimensionality, the feature space will become sparse. As a result, classifiers will require more data to find meaningful patterns among the features. A lack of data in these cases will lead to poor generalization and thus overfitting. Deep learning typically outperforms classic machine learning techniques in datasets where the number of samples is larger than the number of features, but struggles in low-sample size settings like in metagenomics.

Deep learning is currently being applied in metagenomics in multiple ways. One of which is deep learning as a classifier for the prediction of the host phenotype. Multiple approaches are possible regarding the classifier input (Figure 1). For example, amplicon or shotgun sequencing reads can be directly fed to a neural network to perform end-to-end prediction. Some studies apply classic bioinformatic pipelines to obtain taxonomic tables, abundance counts and functional pathways and then use these as input for the deep learning classifier [44, 46]. There are also several studies that take the taxonomic tables and create phylogenetic trees from them, and then use a representation of the tree as input to a CNN (Table 2) [36, 48, 47]. Some studies use deep learning as a substitute for expert-driven feature selection, in this use case, features are selected, transformed, or dimensionally reduced [49]. Afterwards, the features are either still used with deep learning classifiers [35] or with classic machine learning classifiers such as RF [43]. Then there are a multitude of recent studies that use deep learning on microbiome data for the purpose of unsupervised binning [50, 51], viral classification [52], sequence identification [53], taxonomic classification [22], metagenomic profiling [54], predicting genes and ORFs. Such use cases could be used to optimize data preprocessing before classification, but they are out of the scope of this review and will not be discussed further.

4.4.1 Representing microbiome data as phylogenetic trees or synthetic images for use with CNNS

PH-CNN, PopPhy-CNN, and TopoPhy-CNN are tools that make use of deep learning to predict disease phenotypes from human gut microbiome samples. They are distinguished from other approaches by the fact that they represent microbial data as phylogenetic trees. They process metagenomics data using classic bioinformatics pipelines to create these phylogenetic trees, which are then used as input to a CNN classifier. CNNs have been the most successful deep learning method for image classification. They work by using multiple filters to combine different parts of an image. A digital image's matrix structure is the basis for the operations of CNNs. To use the same design for non-image data, you must have a way to measure how close features are. PH-CNN, PopPhy-CNN, and TopoPhy-CNN take advantage of CNNs by embedding the phylogenetic trees in 2D matrixes. This approach is possible because metagenomic data has hierarchical structures when defined as a phylogenetic tree. TopoPhy-CNN provides

additional information to its CNN, such as node distances and the number of child nodes. PH-CNN was used to predict IBD types, and in their study they tested both a CNN classifier and an RF classifier as a benchmark. The CNN outperformed the RF, with an average AUC score of 0.751 vs 0.67 over several IBD subtypes (Table 1). PopPhy-CNN was used to predict liver cirrhosis, type II diabetes, obesity, and IBD. The CNN classifier was outperformed by the RF in three of the four datasets (Table 1). TopoPhy-CNN was used to predict the phenotypes for obesity, liver cirrhosis and type II diabetes. They benchmarked their implementation against PopPhy-CNN and outperformed their scores slightly in AUC in the Cirrhosis, type II diabetes and obesity datasets (Table 1). Mdeep by Wang et al. [44] also uses a phylogenetic tree approach, however instead of directly embedding a phylogenetic tree into a 2D matrix, they group OTUs in vectors based on the distance of the OTUs in the tree. Using this approach, they created a CNN classifier and an RF classifier. The RF slightly outperformed the CNN in the F1-score, with 0.82 vs 0.81 (Table 1).

Nguyen et al. [34] chose a different approach of feeding information into CNNs and developed Met2Img. Instead of capturing human gut microbiome data in a tree, Met2Img represented the microbiome abundance in the form of "synthetic images". The authors create these images through the use of t-SNE embeddings or Fill-up. The Fill-up method preserved the evolutionary relationship between taxa in the samples, as they are arranged in phylogenetic order during processing. Met2Img was trained and tested for the classification of liver cirrhosis, colorectal cancer, IBD, and obesity. In their data preprocessing, they tested several binning techniques, colour scales and gradients for the images. One of the problems they encountered was overlapping data points in the t-SNE representations, which was, for this reason, often outperformed by the Fill-up images. The authors also used features from the synthetic images to test an RF implementation and compared both their RF and CNN classifiers with MetAML-RF. Nguyen et al. only present scores in the form of accuracy and the Met2Img RF classifier outperformed the MetAML-RF classifier in all six datasets that were used (Liver cirrhosis, Colorectal cancer, IBD, Obesity, and two type II diabetes datasets). The Met2Img CNN classifier outperformed their own RF implementation in the liver, IBD, and one of the two diabetes datasets with minimal differences (Table 1).

4.4.2 K-mer distributions from amplicon data

Asgari et al. developed MicroPheno and in their paper investigated the use of k-mer distributions from 16s rRNA sequencing for host phenotype prediction in humans [46]. They tested MicroPheno on predicting the origin of human body site microbes and used human gut samples for the prediction of Crohn's disease. Their primary objective was to compare the use of k-mer distributions versus the use of OTUs in classification tasks. The main problem with their method is that short k-mers make it exceedingly difficult to trace taxa to the phenotypes of interest [46]. The authors employed a bootstrapping framework to select the optimal k-mer sizes and then tested the data with RF, SVM, and Deep Neural Network (DNN) classifiers. The SVM had the worst performance, and the MicroPheno DNN classifier lost from the RF classifier for both the Crohn's disease and body-site prediction datasets. With an F1-score difference of 0.7 vs 0.76 for the Crohn's disease predictions (Table 1).

4.4.3 Dimension reduction focused classification frameworks

In their paper, Oh and Zhang [35] present DeepMicro which is a framework that uses deep learning to reduce dimensionality of metagenomics data. They argue that there is a lack of tools for the representation of microbiome profiles, and that deep learning implementations can reduce dimensionality and accurately represent profiles. The framework makes use of

MetaPhlAn2 to create marker and abundance profiles and then uses a multitude of autoencoders (SAE, DAE, VAE, CAE) to transform data into low-dimensional representation. The authors then trained RF, SVM, and MLP classification models based on the representations from the framework for predicting human host disease phenotypes such as obesity, colorectal cancer, liver cirosis and type II diabetes. They benchmarked their work against MetAML for predicting sample phenotypes from obesity, liver cirrhosis, colorectal cancer and two datasets for type II diabetes. While the DeepMicro "framework" as a whole outperformed MetAML in five out of six datasets, it was often the RF or SVM classifiers that actually outperformed MetAML. The actual deep learning MLP classifier only produced better results in one of the two type II diabetes datasets. The score difference was minimal, with an AUC of 0.76 for DeepMicro-MLP and an AUC of 0.74 for MetAML-RF (Table 1).

Then there is the paper by Rahman and Rangwala [43] who created RegMIL for clincal phenotype prediction from human microbiome data. RegMil does not make use of a deep learning classifier, but rather uses a two-layer neural network in combination with an RF to classify sample disease states. The authors make use of a k-mer bagging approach. After bagging, the bags get positivity scores assigned to them by a neural network. Then the scored bags are used as input to an RF to perform final classification. In their benchmarks, RegMIL did not outperform MetAML for the same datasets (Table 1).

4.4.4 End-to-end prediction from amplicons

Read2Pheno by Zhao et al. [21] is the only reviewed paper that makes use of the end-to-end prediction capabilities of deep learning. The model was developed as a microbial DNA classifier with the aim to classify environments from reads and samples. Interestingly, they do this with 16s rRNA sequencing instead of shotgun sequencing. The authors trained a 16s rRNA read level classifier using an attention based neural network. The classifier predicts the phenotype of individual reads, and could therefore not classify entire samples directly. To still be able to classify samples as a whole, the authors tested various approaches that made use of the single read classifications. The method that produced the best results extracted read embeddings from the neural network and clustered them to form what they call "pseudo OTUs". Unfortunately, they did not benchmark their tool as a whole. Instead, the study focused on how to classify samples as from single read predictions and benchmarked against more traditional methods such as the use of OTU- and ASV-tables within their framework. They did not compare their results to external classifiers.

5 Discussion

5.1 Current deep learning applications do not significantly outperform classic machine learning in metagenomics

Deep learning has revolutionized several fields, becoming the de facto model for applications such as image recognition, natural language processing, and speech recognition. In medical applications, it is capable of classifying ovarian tumours from ultrasound images [55] and detecting brain tumours from CT scans and MRIs [56], significantly outperforming models using classic approaches and even rivalling assessment by experts [55]. There have also been successful application in proteomics, such as DeepDigest [57] for protein proteolytic digestion prediction and AlphaPeptDeep [58] for predicting peptide properties. In metagenomics, however, there has been no significant improvement in phenotype predictions. While in some cases, the deep learning classifiers outperformed more classic approaches, score differences were often minimal.

Unfortunately, due to different data processing and validation steps such as subsampling, cross-validation, and score metrics used, we were unable to compare model performances cross study. All score comparisons that were done or mentioned were performed by authors themselves in the benchmark sections of their papers.

A disadvantage of classic machine learning over deep learning is the need for an elaborate preprocessing pipeline to obtain the taxa and functional pathways present in a microbiome sample. However, most of the studies in this review still use these classic bioinformatics pipelines. Deep learning approaches that take direct advantage of amplicons and shotgun metagenomic reads have the potential to predict host phenotype end-to-end. But most studies in this review do not take advantage of this capability, except for Read2Pheno by Zhao et al. [21] which did not benchmark against external classifiers.

One significant advantage of deep learning is that it can eliminate the requirement for expert-driven feature selection. Both in end-to-end and classic bioinformatic approaches. But most of the studies that are implementing deep learning as classifiers often have elaborate preprocessing and dimension reduction steps that implement deep learning, then they test deep learning and classic machine learning classifiers on the features produced. This somewhat negates this advantage, and increases the compute time.

Currently, the potential of deep learning approaches is not being fully taken advantage of. The studies in this review use relatively small sample sizes from about ten datasets, which might contribute to the lacklustre performance. Additionally, it is still important to perform data normalization and standard filtering and quality checks when using deep learning approaches. Barcodes and primers should still be removed, abundance tables should still be normalized (when they are used), and host contamination should be handled in shotgun metagenomics before passing on amplicons or reads to deep learning approaches that take direct advantage of them. This is to reduce variables and ensure that the deep learning methods actually learn patterns from the bacterial samples rather than host DNA in the case of shotgun metagenomics. Some mentioned studies do not explicitly state whether they performed this preprocessing and normalization, which could be affecting performance.

To make deep learning classifiers robust and perform optimally, the training data must meet certain requirements, such as low error rates. Additionally, the training- and test- set must be diverse to capture the wide range of biological variability to improve the generalization of the classifiers. However, there is a lack of diverse and large amounts of sequencing data with ground truths available. Public databases such as the NCBI and ENA have large amounts of data available, however finding samples that match exact requirements is incredibly laborious because the databases do not uphold strict FAIR principles [59]. Then there is the additional problem of mislabelled microbiome data on these databases, which can severely impact performance [59]. Some solution might be subsampling, perturbation or synthetic data generation. However, this comes with additional dangers. Reiman et al. [36] generated various multiclass synthetic datasets with SparseDOSSA. They then evaluated the performance of PopPhy-CNN on three real multiclass datasets for obesity (three subclasses), IBD (seven subclasses), and a multi-disease set (ten subclasses) and the synthetic datasets. The four synthetic datasets were multiclass with three, five, seven, and nine classes each. The performance on the real datasets was much worse than the performance on the synthetic dataset for both PopPhy-CNN and the RF classifier. For example, obesity scored 0.159 (MCC) compared to the synthetic Syn3 dataset's 0.884 MCC score.

5.2 Lack of multi-omics approaches for phenotype classification

While many of these papers perform dimensionality reduction and data transformation through deep learning applications such as autoencoders, they do not attempt to integrate any multi-omics approaches through use of co-expression networks or use of KEGG pathway mapping. Which is one of the major benefits of deep learning. Approaching diseases through a single -omics field would only be the start of finding treatments. After having determined that the presence of certain bacteria or combinations of them contribute to a disease, it must be further investigated why they contribute. Only metagenomics does not further reveal why a given bacterium causes or contributes to a disease. Multi-omics provides a more comprehensive view of the biological processes in the underlying phenotypes. Transcriptomic data obtained through RNA-sequencing methods could provide insight into the expression of the genes carried by the microbiota in a sample. It can help in determining why a certain bacterium is important for predicting the disease phenotype if a certain gene in the bacteria has very high expression. Proteomic data obtained through mass spectrometry can provide information about the function of proteins present in the microbiome and host, useful for understanding protein-mediated interactions, where the protein interactions could be contributing to inflammation. metabolomic molecule profiles obtained through mass spectrometry or nuclear magnetic resonance spectroscopy would provide a snapshot of the metabolic products and pathways that are active in the microbiome and host. The snapshots could provide insight into the increase or decrease in specific metabolites that are expected to be present or absent. Through such a multifaceted approach, we would gain much more understanding of why a certain bacterium is relevant for a disease. Instead of just knowing that its presence contributes to the disease, we can gain understanding of the specific genes that are active, which interactions caused by the proteins produced cause the symptoms and what metabolites are present as a result of that. Such a multifaceted approach would make it much easier to create treatment plans, as the underlying factors of the disease are much better understood.

5.3 Interpretability of deep learning methods

Interpretability of deep learning classifiers is a problem. In classic machine learning approaches, data points that contribute the most to accurate classifications can be traced back. This means that these data points can be traced to the presence of certain microbes, their genes and associated pathways which can then be studied further, leading to an enrichment of our knowledge. Deep learning approaches are often difficult to explain or even unexplainable. We cannot track the data points that are the most important for the classifier to make accurate predictions. Being able to determine these data points are critical to gaining understanding of the biological processes and influences of the microbiome on host phenotype. Classic machine learning approaches such as RF and SVM, on the other hand, are very explainable. RF provides measures of feature importance, indicating how important a feature is to the models' accuracy. This helps in understanding which variables are most influential. SVMs explicitly identify support vectors, which are the most important data points of the training set that influence where the decision boundary gets placed.

There was an attempt by Reiman et al. [36] (PopPhy-CNN) to make their CNN classifier more interpretable by focusing on the first layer of their neural network classifier. This layer created the feature maps, which captured the spatial relationships of the 2D matrix from the phylogenetic tree. They used these maps to find areas with the most activity in the neural network layer. Following this, they traced these positions to their locations in the 2D matrix. Using this technique, they could assess the importance of each OTU. The algorithm used to trace back the features can be found in their paper [36]. While this is a good attempt

and an important first step at improving the interpretability, it is limited to only the first convolutional layer, which does not accurately represent the full prediction process of their classifier. Additionally, they had to reduce the total number of layers in the network to make it work, which led to overall less performance.

5.4 Adaption to plant microbiomes

Human and plant microbiome sequencing data share the same structure. The preprocessing approaches and the workflows for creating the taxonomic- and- functional tables are the same. There is no technical blockade for applying phylogenetic tree, k-mer, strain-profile and abundance-profile approaches. Optimally, you would want to make use of tools that offer deep learning neural network architectures and then let users prepare their data to train the model. Pre-trained human-specific models that are aimed at predicting exact disease types would be difficult to apply in plants due to biological differences such as genetic structures, disease mechanisms, and pathogen variation. Bacteria present in the human gut that could be important classification features for the human disease phenotype are likely to be irrelevant for a plant disease phenotype.

Roadblocks for applying deep learning in plants are more likely to be related to the availability of large plant microbiome datasets. As, plant microbiomes are more diverse and the composition is very different from human microbiomes. Larger amounts of sequencing data will be required to properly generalize the classifiers. Especially when considering differences between plant species and their environments. The rhizosphere of plants in sandy soil vs clay soil or different types of crops will be significantly different. Another example is that legume crops often have relationships with nitrogen-fixing bacteria which are vital for their growth and development, whereas grass crops do not. It could also be possible that different models will have to be trained in such cases.

Synthetic plant microbiomes might help alleviate some data problems. As these microbiomes are put together based on known interactions and phenotypes, researchers could use them to test the accuracy and robustness of models under different conditions. The synthetic communities could also be used to fine-tune models for underrepresented conditions or rare events.

6 Conclusion

In this review, we gave an overview of the latest machine- and deep-learning methods used for phenotype prediction from microbiome data in humans. The goal was to see whether these methods could be easily adopted for similar use in plant metagenomics. Overall, this is possible and the future of large scale microbiome analysis lies in deep learning, especially regarding plant microbiomes. But currently, deep learning methods do not significantly outperform classic machine learning methods and the lack of explainability and unavailability of larger datasets is a considerable roadblock for widespread adaptation. Through this review, we hope to contribute to the future development and improvement of deep learning classification in metagenomics. It seems that most studies are overly focused on already existing approaches that are used in classic machine learning classification. Future deep learning classification in this field should focus on interpretability, end-to-end prediction, and multi-omics integration.

References

[1] Mendes R, Raaijmakers JM. Cross-Kingdom Similarities in Microbiome Functions. The ISME Journal. 2015 Sep;9(9):1905-7.

[2] Busato S, Gordon M, Chaudhari M, Jensen I, Akyol T, Andersen S, et al. Compositionality, Sparsity, Spurious Heterogeneity, and Other Data-Driven Challenges for Machine Learning Algorithms within Plant Microbiome Studies. Current Opinion in Plant Biology. 2023 Feb;71:102326.

[3] Berg G, Rybakova D, Fischer D, Cernava T, Vergès MCC, Charles T, et al. Microbiome Definition Re-Visited: Old Concepts and New Challenges. Microbiome. 2020 Dec;8(1):1-22.

[4] Simon JC, Marchesi JR, Mougel C, Selosse MA. Host-Microbiota Interactions: From Holobiont Theory to Analysis. Microbiome. 2019 Dec;7(1):1-5.

[5] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes. Nature. 2012 Oct;490(7418):55-60.

[6] Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the Human Gut Microbiome in Liver Cirrhosis. Nature. 2014;513(7516):59-64. Available from: https://www.nature.com/articles/nature13568.

[7] Emmenegger B, Massoni J, Pestalozzi CM, Bortfeld-Miller M, Maier BA, Vorholt JA. Identifying Microbiota Community Patterns Important for Plant Protection Using Synthetic Communities and Machine Learning. Nature Communications. 2023 Dec;14(1):7983.

[8] Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic Analysis of the Human Distal Gut Microbiome. Science (New York, NY). 2006 Jun;312(5778):1355-9.

[9] Cash HL, Whitham CV, Behrendt CL, Hooper LV. Symbiotic Bacteria Direct Expression of an Intestinal Bactericidal Lectin. Science (New York, NY). 2006 Aug;313(5790):1126-30.

[10] Navas-Molina JA, Hyde ER, Sanders JG, Knight R. The microbiome and big data. Current Opinion in Systems Biology. 2017;4:92-6. Available from: https://www.sciencedirect.com/science/article/pii/S2452310017301270.

[11] Patel JB. 16S rRNA Gene Sequencing for Bacterial Pathogen Identification in the Clinical Laboratory. Molecular Diagnosis. 2001 Dec;6(4):313-21.

[12] Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis. Nature Communications. 2019 Nov;10(1):5029.

[13] Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, et al. A Practical Guide to Amplicon and Metagenomic Analysis of Microbiome Data. Protein & Cell. 2021 May;12(5):315-30.

[14] Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability. BMC Biology. 2020 Mar;18:24.

[15] Yap M, Feehily C, Walsh CJ, Fenelon M, Murphy EF, McAuliffe FM, et al. Evaluation of Methods for the Reduction of Contaminating Host Reads When Performing Shotgun Metagenomic Sequencing of the Milk Microbiome. Scientific Reports. 2020 Dec;10(1):21665.

[16] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology. 2016;12(7):e1004977. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977.

[17] Krause T, Wassan JT, Mc Kevitt P, Wang H, Zheng H, Hemmje M. Analyzing Large Microbiome Datasets Using Machine Learning and Big Data. BioMedInformatics. 2021 Nov;1(3):138-65.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need. arXiv; 2017. Available from: https://arxiv.org/abs/1706.03762.

[19] Roy G, Prifti E, Belda E, Zucker JD. Deep Learning Methods in Metagenomics: A Review. Microbial Genomics. 2024;10(4):001231.

[20] Borgman J, Stark K, Carson J, Hauser L. Deep Learning Encoding for Rapid Sequence Identification on Microbiome Data. Frontiers in Bioinformatics. 2022 Jun;2:871256.

[21] Zhao Z, Woloszynek S, Agbavor F, Mell JC, Sokhansanj BA, Rosen GL. Learning, Visualizing and Exploring 16S rRNA Structure Using an Attention-Based Deep Neural Network. PLOS Computational Biology. 2021 Sep;17(9):e1009345.

[22] Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics. 2020 02;2(1):lqaa009. Available from: https://doi.org/10.1093/nargab/lqaa009.

[23] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017 Nov;8.

[24] Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological). 2018 12;44(2):139-60. Available from: https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

[25] Pawlowsky-Glahn V, Egozcue JJ. Compositional Data and Their Analysis: An Introduction. Geological Society, London, Special Publications. 2006 Jan;264(1):1-10.

[26] Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. Frontiers in Genetics. 2019 Nov;10.

[27] Nguyen NP, Warnow T, Pop M, White B. A Perspective on 16S rRNA Operational Taxonomic Unit Clustering Using Sequence Similarity. npj Biofilms and Microbiomes. 2016 Apr;2(1):1-8.

[28] Konstantinidis KT, Tiedje JM. Genomic Insights That Advance the Species Definition for Prokaryotes. Proceedings of the National Academy of Sciences. 2005 Feb;102(7):2567-72.

[29] Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: Differentiating between Closely Related Microbial Taxa Using 16S rRNA Gene Data. Methods in Ecology and Evolution. 2013;4(12):1111-9.

[30] Fox GE, Wisotzkey JD, Jurtshuk P. How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient to Guarantee Species Identity. International Journal of Systematic Bacteriology. 1992 Jan;42(1):166-70.

[31] Callahan BJ, McMurdie PJ, Holmes SP. Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis. The ISME Journal. 2017 Dec;11(12):2639-43.

[32] Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn LJ, et al. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. Frontiers in Microbiology. 2019 Jun;10.

[33] LaPierre N, Ju CJT, Zhou G, Wang W. MetaPheno: A Critical Evaluation of Deep Learning and Machine Learning in Metagenome-Based Disease Prediction. Methods. 2019 Aug;166:74-82.

[34] Nguyen TH, Prifti E, Chevaleyre Y, Sokolovska N, Zucker JD. Disease Classification in Metagenomics with 2D Embeddings and Deep Learning. arXiv. 2018 Jun. Available from: http://arxiv.org/abs/1806.09046.

[35] Oh M, Zhang L. DeepMicro: Deep Representation Learning for Disease Prediction Based on Microbiome Data. Scientific Reports. 2020 Apr;10(1):6026.

[36] Reiman D, Metwally AA, Sun J, Dai Y. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data. IEEE Journal of Biomedical and Health Informatics. 2020 Oct;24(10):2993-3001.

[37] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology. 2014;10(11):766. Available from: https://www.embopress.org/doi/abs/10.15252/msb.20145645.

[38] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. Nature. 2010 Mar;464(7285):59-65.

[39] Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe. 2014 Mar;15(3):382-92.

[40] Sokol H, Leducq V, Aschard H, Pham HP, Jegou S, Landman C, et al. Fungal Microbiota Dysbiosis in IBD. Gut. 2017 Jun;66(6):1039-48.

[41] Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of Human Gut Microbiome Correlates with Metabolic Markers. Nature. 2013 Aug;500(7464):541-6.

[42] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control. Nature. 2013 Jun;498(7452):99-103.

[43] Rahman MA, Rangwala H. RegMIL: Phenotype Classification from Metagenomic Data. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. BCB '18. Association for Computing Machinery; 2018. p. 145-54. Available from: https://doi.org/10.1145/3233547.3233585.

[44] Wang Y, Bhattacharya T, Jiang Y, Qin X, Wang Y, Liu Y, et al. A Novel Deep Learning Method for Predictive Modeling of Microbiome Data. Briefings in Bioinformatics. 2021 May;22(3):bbaa073.

[45] Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of Intestinal Prevotella Copri Correlates with Enhanced Susceptibility to Arthritis. eLife. 2013 Nov;2:e01202.

[46] Asgari E, Garakani K, McHardy AC, Mofrad MRK. MicroPheno: Predicting Environments and Host Phenotypes from 16S rRNA Gene Sequencing Using a k-Mer Based Representation of Shallow Sub-Samples. Bioinformatics. 2018 Jul;34(13):i32-42.

[47] Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, et al. Phylogenetic Convolutional Neural Networks in Metagenomics. BMC Bioinformatics. 2018 Mar;19(2):1-13.

[48] Li B, Zhong D, Jiang X, He T. TopoPhy-CNN: Integrating Topological Information of Phylogenetic Tree for Host Phenotype Prediction From Metagenomic Data. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021. p. 456–461. Available from: https://ieeexplore.ieee.org/abstract/document/9669509.

[49] Wickramaratne D, Wijesinghe R, Weerasinghe R. Human Gut Microbiome Data Analysis for Disease Likelihood Prediction Using Autoencoders. In: 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter); 2021. p. 49-54.

[50] Bao HQ, Vinh LV, Van Hoai T. A Deep Embedded Clustering Algorithm for the Binning of Metagenomic Sequences. IEEE Access. 2022;10:54348-57.

[51] Tran VT, Quach HD, Van PVD, Tran VH. A Novel Metagenomic Binning Framework Using NLP Techniques in Feature Extraction. IPSJ Transactions on Bioinformatics. 2022;15:1-8.

[52] Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. VirNet: Deep attention model for viral reads identification. In: 2018 13th International Conference on Computer Engineering and Systems (ICCES); 2018. p. 623-6.

[53] Busia A, Dahl GE, Fannjiang C, Alexander DH, Dorfman E, Poplin R, et al. A Deep Learning Approach to Pattern Recognition for Short DNA Sequences. bioRxiv : the preprint server for biology. 2019.

[54] Arango-argoty GA, Heath LS, Pruden A, Vikesland PJ, Zhang L. MetaMLP: A Fast Word Embedding Based Classifier to Profile Target Gene Databases in Metagenomic Samples. Journal of Computational Biology. 2021;28(11):1063-74. Available from: https://doi.org/10.1089/cmb.2021.0273.

[55] Giourga M, Petropoulos I, Stavros S, Potiris A, Gerede A, Sapantzoglou I, et al. Enhancing Ovarian Tumor Diagnosis: Performance of Convolutional Neural Networks in Classifying Ovarian Masses Using Ultrasound Images. Journal of Clinical Medicine. 2024 Jan;13(14):4123.

[56] Nagaraju G. A Comparative Analysis of Advanced Machine Learning Techniques for Enhancing Brain Tumor Detection. Journal of Electrical Systems. 2024 Apr;20(2s):901-9.

[57] Yang J, Gao Z, Ren X, Sheng J, Xu P, Chang C, et al. DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning. Analytical Chemistry. 2021 Apr;93(15):6094-103.

[58] Zeng WF, Zhou XX, Willems S, Ammar C, Wahle M, Bludau I, et al. AlphaPeptDeep: A Modular Deep Learning Framework to Predict Peptide Properties for Proteomics. Nature Communications. 2022 Nov;13(1):7238.

[59] Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Scientific Data. 2019 Feb;6(1). Available from: http://dx.doi.org/10.1038/sdata.2019.21.