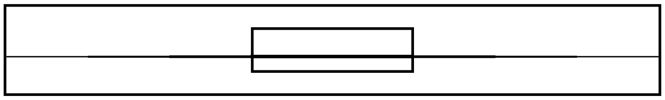


Model Agnostic Meta Learning For Disease Prediction From Metagenomic Data

Esha Manchanda[1]

[1] *Department of Computer Science, Ashoka University*

I. INTRODUCTION

The use of metagenomic data for disease prediction is a rapidly growing field, and it involves identifying patterns in the types and relative abundances of different microorganisms that are associated with different disease states.

Alterations in the gut microbiome, often referred to as dysbiosis, have been associated with a variety of diseases. It's thought that dysbiosis may contribute to disease through several mechanisms:
1. Inflammation: Dysbiosis can lead to an overactive immune response, causing chronic inflammation. This has been implicated in diseases like inflammatory bowel disease (IBD), obesity, and cardiovascular disease.
2. Loss of Barrier Function: Dysbiosis can impair the gut barrier, allowing bacteria and bacterial products to cross into the bloodstream, a condition known as "leaky gut". This can trigger inflammation and has been implicated in diseases like IBD and celiac disease.
3. Metabolic Disruption: Changes in the gut microbiome can disrupt normal metabolic functions, potentially leading to conditions like obesity, metabolic syndrome, and type 2 diabetes.
4. Neurological Effects: Changes in the gut microbiome may impact the gut-brain axis and contribute to neurological and mental health conditions, such as depression, anxiety, autism, and possibly even neurodegenerative diseases like Parkinson's and Alzheimer's.
5. Pathogen Expansion: Dysbiosis can allow harmful pathogens to proliferate, which can lead to infectious diseases or contribute to chronic diseases like IBD.

The relationships observed between the human gut microbiome and overall health highlight the potential of machine learning in predicting diseases from metagenomic data. The complex nature of these associations offers a fitting application for machine learning techniques, which can navigate high-dimensional data to detect patterns and inform predictive models.

II. LITERATURE REVIEW

Early research focused on the use of traditional statistical methods and ML models like logistic regression and decision trees to identify associations between microbial populations and diseases. For instance, Qin et al. [1] successfully used a logistic regression model to distinguish patients with liver cirrhosis from healthy controls using metagenomic data. Another notable study by Pasolli et al.[2] employed decision trees to identify specific microbial markers associated with colorectal cancer.

With advancements in technology and computational power, more sophisticated ML models such as support vector machines (SVMs), random forests, and neural networks have been applied to this domain. These models are capable of capturing complex, non-linear relationships within high-dimensional metagenomic data. A study by Zeller et al. [3] demonstrated the power of SVMs in accurately predicting colorectal cancer from gut microbiome profiles. On the other hand, Knights et al. [4] showcased the effectiveness of random forests in classifying various diseases, including inflammatory bowel disease and obesity, based on microbial community data.

More recently, the advent of deep learning has opened up new opportunities for disease prediction from metagenomic data. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in handling the complexity and high-dimensionality of metagenomic data. For example, a study by Fiannaca et al. [5] reported that a CNN-based approach outperformed traditional ML models in predicting type 2 diabetes from gut microbiota data.

Despite these advances, the application of ML in metagenomics is still in its infancy and is confronted with several challenges. These include issues related to data sparsity, overfitting, interpretability, and the need for large, well-curated datasets for model training. Additionally, most existing studies have adopted a 'one-size-fits-all' approach, designing a single model to predict a particular disease from metagenomic data.

In light of these challenges, novel approaches such as few-shot learning and meta-learning have emerged. One such approach, Model-Agnostic Meta-Learning, has been proposed as a promising avenue for disease prediction from metagenomic data.

III. PROJECT DESCRIPTION AND GOALS

In this work, I have implemented the Model-Agnostic Meta-Learning framework by Finn et al. [6] coupled with few-shot learning to assess their efficacy in the classification of diseases. I primarily endeavored to address two questions. The first one pertained to the comparative effectiveness of Model-Agnostic Meta-Learning versus a traditional neural network in the context of a specific task. Given the inherent complexity and variability of metagenomic data, I hypothesised that the adaptive capacity of MAML would offer a superior performance. The second research objective was to explore the impact of task variability on the model's performance. Specifically, I evaluated how altering the number of classes and samples within each task influenced the predictive outcomes.

IV. THE DATASET

A. Data Extraction

The dataset used in this project is processed using eight publicly available studies containing 2424 shotgun metagenomic samples and six different diseases. Shotgun metagenomic analysis is a method for studying the genetic content of microbial communities in various environmental samples. It involves randomly sequencing DNA fragments from the entire genomic content of a mixed population of microorganisms, without first isolating and culturing individual organisms. During the process, DNA is extracted from a sample of interest, such as feces, and then fragmented into small pieces. These fragments are then sequenced using high-throughput sequencing technologies. The resulting data can then be analyzed using bioinformatics tools to identify the different microbial taxa present in the sample, infer their functional capabilities, and investigate their interactions within the community. Shotgun metagenomic analysis allows the study of genetic diversity and functional potential of entire microbial communities in a single experiment, providing insights into the complex relationships between microorganisms and their environment.

The tool MetaPhlAn2 (Metagenomic Phylogenetic Analysis) was used to process these metagenomic datasets. MetaPhlAn2 is a computational tool designed to profile the composition of microbial communities such as bacteria, archaea, viruses, and eukaryotes from metagenomic sequencing data. It works by mapping the metagenomic sequencing data against a database of known clade-specific marker genes. These marker genes are unique to particular microbial strains or species and can be used to identify the presence and relative abundance of that taxon in the metagenomic sample.

B. Data Specifications

The data file contains metadata about the samples, including the disease state of the sample, the sample_id field corresponding to the sample identifier and the abundance data of various species found in the metagenomics samples. The data is presented as a large matrix, with each row corresponding to a sample and each column corresponding to a different species. The values in the matrix represent the relative abundance of each species in each sample.

C. Cleaning and Preprocessing

For the purpose of this project, I consolidated the disease labels to streamline the dataset. The labels 'n' and 'leaness' were merged under the label 'healthy'. The two subtypes of Inflammatory Bowel Disease (IBD), namely Ulcerative Colitis and Crohn's Disease, were merged under the 'ibd' label. The label 'impaired_glucose_tolerance' was reclassified as 't2d', and the two categories of 'adenoma' were merged into a single 'adenoma' label. Consequently, the refined dataset retained only the data corresponding to the following disease labels: 'ibd', 'cancer', 'cirrhosis', 'obesity', 't2d', 'adenoma', and 'healthy'. All other data was excluded from the analysis.

In the process of feature selection, species were chosen based on the variability of their abundance values across all samples. Specifically, only those species were considered whose standard deviation of abundance values exceeded a threshold of 1. This criterion ensures the inclusion of species that exhibit substantial variability, thereby allowing us to focus on species that exhibited significant differential abundance in relation to different diseases.

D. Dataset Limitations

The complexity of metagenomics data, encapsulating a vast number of microbial species, creates a high-dimensional feature space that may lead to model overfitting and difficulties in training. Additionally, the human microbiome showcases considerable inter-individual and intra-individual variability, influenced by factors such as diet, genetics, and environmental factors, which can introduce significant noise and undermine model generalization. Also, the heterogeneity inherent to many diseases, such as inflammatory bowel disease (IBD) or cancer, with various subtypes and stages, may not be entirely encapsulated by the dataset labels, potentially affecting model performance.

V. METHODOLOGY

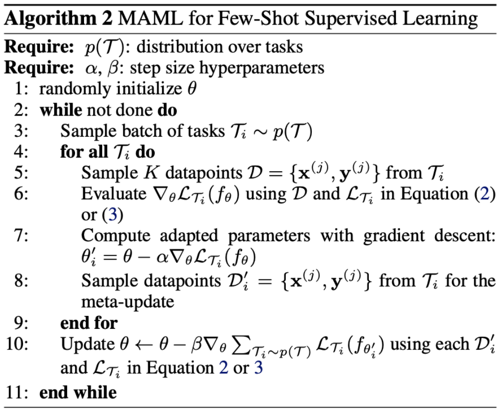
A. Few Shot Learning

For disease prediction based on gut microbiome data, we might have a large number of different diseases we're interested in predicting, but only a small number of examples of each disease. This problem can be solved using few shot learning, where the aim is to design machine learning models that can learn useful information from a small number of examples - typically 1-10 training examples. The motivation behind few-shot learning is to mimic the human ability to learn quickly and effectively from a small number of examples
The model in this work learns to classify the species abundance data into k unseen disease classes, given only n labelled examples, i.e. the k-shot n-way classification task.

B. Model Agnostic Meta Learning

The key idea behind meta-learning is that, during meta-training, the model learns to recognize common patterns and structures that are shared across various tasks sampled from the task distribution. By learning these shared patterns, the model becomes better at adapting to new tasks, even if it has never seen those specific tasks before. In a traditional machine learning setting, the algorithm is trained on a fixed dataset and then tested on a new, unseen dataset. In contrast, meta-learning is concerned with training a model to learn how to learn from a few examples of a new task, rather than relying on a fixed dataset.
The key idea of MAML is to find a model initialization that is not optimal for any single task, but instead is able to adapt quickly to any task from a distribution of tasks. The model is then fine-tuned on each individual task for a few gradient steps to achieve good performance. This procedure allows the model to learn task-invariant features and makes it capable of generalising from a small number of samples.
The MAML algorithm has the following phases:

- Task Definition: Each task might be the classifying different diseases based on the metagenomic features. For example, one task could be predicting between ”Type 2 Diabetes” and ”Inflammatory Bowel Disease” from the abundance of different bacterial species, and so on.
- Meta-Training: During training, a batch of tasks is sampled, and for each task, a copy of the model is created and trained for a few steps on the training samples of that task. The loss calculated on a test set from the same task is then used to update the original model by minimising a meta-objective (aggregate test loss) function.
- Meta-Testing: During testing, the model is quickly adapted to each new task (a new disease) using a few gradient steps. The adapted model is then used for prediction on the test set of that task.



Algorithm 2 MAML for Few-Shot Supervised Learning

```
Require: $p(\mathcal{T})$: distribution over tasks
Require: $\alpha$, $\beta$: step size hyperparameters
randomly initialize $\theta$
while not done do
    Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
    for all $\mathcal{T}_i$ do
        Sample $K$ datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$
        Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation (2)
        or (3)
        Compute adapted parameters with gradient descent:
        $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
        Sample datapoints $\mathcal{D}'_i = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$ for the
        meta-update
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$ using each $\mathcal{D}'_i$
    and $\mathcal{L}_{\mathcal{T}_i}$ in Equation 2 or 3
end while
```

FIG. 1. MAML Algorithm[6]

C. Implementation Details

A n-shot, k-way task sampler was implemented to generate batches of tasks for each meta-iteration. For each task, I randomly selected k diseases and for each disease, n support samples were randomly chosen. An equal number of query samples (q_samples) were also randomly selected for each disease, making sure they were not already included in the support set. This process was repeated for a specified number of tasks (num_tasks) to form a batch.

The model architecture chosen is a simple feed-forward neural network created using the PyTorch library, having a single input, hidden and output layer.

The MAML algorithm is used for training. For each task in a batch, the model was deep-copied and trained using the task's support set for a predefined number of inner loop steps (inner_loop_steps) with an Adam optimizer. The loss function used was the Cross-Entropy Loss. After inner loop training, the model was evaluated on the task's query set and the loss was computed. The meta-loss, i.e., the average of these task-specific losses, was then calculated. The gradient of the meta-loss was computed with respect to the global model parameters, and a meta-optimizer was used to perform the meta-update. This procedure was performed for a predefined number of meta-iterations (meta_iters).

After the completion of the meta-training phase, I conducted a meta-testing phase to evaluate the performance of the trained model on unseen tasks. I defined a number of test tasks (num_test_tasks), for which a new set of n-shot, k-way

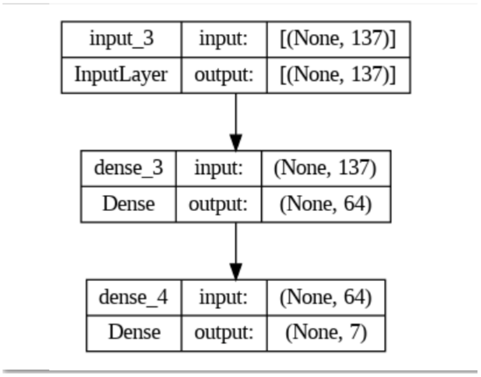


FIG. 2. Model Layout for a 7 Class Task

tasks using the task sampler was created. For each test task, I carried out a similar procedure to the inner loop of the meta-training phase. I sampled a task-specific dataset from each task, which included both training and test data. The training data was used to fine-tune the model (copied from the final global model after meta-training), and the Adam optimizer was used for this inner-loop training. The fine-tuned model was then evaluated on the test data to calculate the task-specific test loss using the Cross-Entropy Loss. I collected these test losses for each task and computed the average test loss across all the test tasks. This average test loss provided a metric to measure the overall performance of the model on unseen tasks.The trained model parameters were directly utilized in this phase, showcasing the model's ability to generalize and quickly adapt to new tasks with limited data.
Code implementation can be found here.

VI. EXPERIMENTAL RESULTS & INFERENCES

Cross-Entropy Loss is employed as the choice of loss function which is defined as:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad \text{for } i \in \{1, \ldots, C\}$$

where y_i is the true label (1 if the true class is i, 0 otherwise), $\hat{y}_i$ is the predicted probability of class i, and C is the number of classes.

Sample Set	MAML Model	Conventional Neural Net
Train	0.55629	0.89595
Test	1.1921	1.4145

Table 1. Cross Entropy Loss Values for MAML and Traditional Model

The performance of the models in disease classification tasks was also quantitatively assessed using accuracy scores. The accuracy score provides a measure of how well the models are able to correctly predict disease labels in relation to the total number of instances evaluated.

Task	MAML Model	Conventional Neural Net
Train: Cancer Vs Type 2 Diabetes	0.91	0.96
Test: Adenoma Vs Healthy	0.79	0.62
Train: Adenoma Vs Type 2 Diabetes	0.89	0.93
Test: IBD Vs Cancer	0.64	0.43
Train: Adenoma Vs IBD	0.93	0.91
Test: Healthy Vs Cancer	0.74	0.69

Table 3. Accuracy Scores of Different Tasks

Model-Agnostic Meta-Learning algorithm shows superior performance over the conventional neural network because of the inherent adaptability of MAML, which is designed to accommodate new tasks effectively, a property that traditional neural networks lack.

Number of diseases (k)	Train Accuracy	Test Accuracy
2	0.97	0.77
4	0.76	0.45
7	0.72	0.34

Table 3. Accuracy Score for Varying Classes

The effectiveness of MAML decreased with an increase in the number of classes within each task due to the increased complexity.

Additionally, the complexity of the meta-testing task impacts how well the model performs. When the meta-training tasks chosen were simple (e.g. binary classification), and the meta-testing tasks were complex (e.g. multi-class classification), the model struggled and showed poor accuracy. As with all machine learning models, the more similar the training and testing tasks are, the better the model is likely to perform.

VII. CONCLUSIONS

This work demonstrates the potential applicability of Model-Agnostic Meta-Learning and few-shot learning in the domain of disease prediction from metagenomic data. Employing a meta-learning strategy has shown promising results in handling the microbiome data, which is often characterized by high dimensionality, heterogeneity, and sparsity. Moreover, this study's findings highlights the importance of microbiome composition as a potential diagnostic tool.

The use of few-shot learning, in particular, has addressed the issue of scarcity of labeled samples, a common limitation in medical datasets. The model's ability to generalize from a small number of examples in the meta-training phase to complex tasks in the meta-testing phase, especially compared to a traditional neural network model underscores its robustness and flexibility.

However, it's important to note that while many studies such as this one have found associations between dysbiosis and disease, it's not yet clear whether changes in the microbiome are a cause or a consequence of disease.

VIII. FUTURE DIRECTIONS

One potential direction is to incorporate temporal information into our model by considering longitudinal microbiome data, which could offer insights into the dynamic changes in microbiome composition over time and their association with disease progression or treatment response.

Additionally, the human microbiome's susceptibility to various external influences, such as lifestyle, diet, and medication use, which are often not accounted for in the dataset, can significantly impact the disease state. Thus, integrating other types of data, could provide a more holistic view of the host-microbiome-disease relationship and potentially enhance prediction performance.

IX. REFERENCES

[1] Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., ... Li, L. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature, 513(7516), 59-64.

[2] Pasolli, E., Truong, D. T., Malik, F., Waldron, L., Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology, 12(7), e1004977.

[3] Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., ... Bork, P. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology, 10(11), 766.

[4] Knights, D., Costello, E. K., Knight, R. (2011). Supervised classification of human microbiota. FEMS Microbiology Reviews, 35(2), 343-359.

[5] Fiannaca, A., La Rosa, M., La Paglia, L., Renda, G., Rizzo, R. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinformatics, 19(Suppl 7), 198.

[6] Finn, C., Abbeel, P., Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML) (Vol. 70, pp. 1126-1135).