Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 18, No. 1, 2025

Towards Accurate and Interpretive Disease Diagnosis for Aging Adults Through Deep Learning Framework

Meiyi Dong

Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

Abstract: Population aging has become a serious problem all around the world. The world's population is aging faster than ever before. What's worse is that as the population ages, a serious problem arises: as the number of elderly people increases, medical resources begin to become scarce. For example, during the COVID-19 pandemic, many elderly people died due to lack of treatment and other medical resources such as masks. Similar to the COVID-19, some diseases are believed to be closely related to age and lead to higher mortality rates among the elderly. It is becoming more important to establish an efficient and accurate diagnostic system for such diseases. In this way, the impact of population aging on society can also be further alleviated. The rapid development of artificial intelligence technology can bring a turning point to this problem. Deep learning has become one of the most current topics in the field of artificial intelligence. With the swelling data of the Internet, deep learning showed emerging power to solve problems from various subjects that traditional machine learning methods could not handle well. Disease diagnosis is exactly one of them. In this research, We are aimed to build an accurate and interpretive disease diagnosis model leveraging the deep learning framework, specifically the Transformer, to diagnose the physical conditions of the elderly and reveal the relationship between specific microbiomes and disease. The key to the problem lies in what the classification model uses as features. Human gut microbiome has been proved to involve in a variety of chemical and biological processes in the human body and have complex connection with our health including specific diseases. In fact, sampling gut microbiome from patients to do disease diagnosis has been expected to be a more efficient and non-invasive method comparing to existed method such as blood test. Furthermore, the rapid development of gene sequencing technology also provides strong hardware support for this idea such as High-throughput sequencing. The core idea of this study is to use the relative abundance of human intestinal microorganisms as features to train a high-precision classification model for specific disease diagnosis. We evaluated five algorithms on six different and independent datasets with three age-related diseases. The results showed that Transformer could show the best performance in most of the cases, although it demands more computing resources to train. Furthermore, we managed to improve the generalization ability with the MK-MMD method. Finally, we did biomarker discovery to all the models with decent performance by SHAP and student t-test.

Keywords: Deep Learning; Machine Learning; Transformer; Disease Diagnosis; Human Gut Microbiome.

1. Introduction

In recent years, with the population aging grows, age-related diseases are gaining more attention of the research community. One important topic is the disease diagnosis. With the development of the sequencing technology, we could get a better view of human body, including the microbiome living in our gut. Metagenomic data is what we get from the sequencing tool applied on the biological samples. With the metagenomic sequencing, we could get the genomic DNA sequence of all microbiomes in the environment. Usually, we can quantify the microbiome with their abundance or relative abundance, which is obtained by the annotation of the metagenomic sequence data. With this process, we could get a better view of the distribution of the microbiome. In fact, many researchers have already found the connection between the gut microbiome and several diseases. In this research, we will focus three of them respectively. The first one is Type 2 Diabetes, the second one Alzheimer's Disease and the final one is Sarcopenia. They have different pathologies and symptoms but share one common feature, they are all related closely to people's age.

Type 2 diabetes is the most common type of diabetes. Because it often occurs in adults, it is also called adult-onset diabetes. Its pathogen could be various, including natural aging, unhealthy diet, the lack of exercises. T2D has been considered as one of the most common diseases for the elder people in the world especially for the developed country. In fact, the prevalence is still increasing rapidly nowadays. Even worse, many people does not realize that they have T2D because the early symptom is not obvious and when it comes to notice the disease is almost impossible to cure. T2D is a chronic disease and usually not fatal, but it's not the case for the elderly. For aging people, T2D becomes a more dangerous disease because of the deterioration of body functions. Furthermore, T2D is gradually taking over a larger part of the aging population. The elderly patients of T2D will soon constitute the majority in many regions [1]. T2D leads to a higher incidence and mortality rate compared to non-diabetic patients [2]. There has been sufficient research proved that gut microbiota has a strong connection with the disease status of the T2D [3].

Alzheimer's disease is a troublesome disease that would causing serious damage to people's cognitive competence, such as memory and thinking ability. AD has been discovered only in adults and especially worse for the elder people who is over 60. According to the research, AD is the one of the most common form neuropathic diseases that the elder people would get and have been a big threat to the elderly welfare. People with AD would have serious problem for their cognitive capability and physical coordination. According to the report, Alzheimer's disease has been ranked as the 6th biggest cause of mortality in the America [4]. How gut microbiome could influence the central nervous system have

been received greatly attention in neurosciences area, which is usually referred to the “gut-brain axis”. The study of gut-brain axis has revealed the connection between gut microbiome and the state of neuropathic diseases [5]. Furthermore, recent research has studied that the composition of human gut microbiome can be powerful indicators of the Alzheimer’s disease [6].

Sarcopenia, also clinically known as “skeletal muscle aging”, proposed by Evans WJ et al. in 1991, refers to the decrease in skeletal muscle mass, muscle strength and body function caused by aging. Skeletal muscles are the main type of muscle in the human body. They are the power source of the human body. Muscle aging and atrophy are important signs of human aging and can easily lead to various physical problems such as fractures. People with Sarcopenia would be challenged with the high rate of physical injury. To be specific, the elderly people with sarcopenia would have problem on basic body motion and more easily to fall because the weak body strength, and get fracture because the insufficient muscle for protecting bones.

These facts have all confirmed that the importance of establishing an accurate diagnostic mechanism for geriatric diseases. The human gut microbiome is considered a powerful indicator of various diseases. Research has shown that the gut microbiome is closely linked to human health and is playing an increasingly important role in disease diagnosis [7]. Therefore, it can be stated with confidence that applying a human gut microbiome-based classification model to do disease diagnosis is feasible.

2. Literature Survey

A large amount of remarkable work has been done regarding disease diagnosis leveraging machine learning methods based on microbiome data. Most of them served two common goals: to build a more precise disease diagnosis model with the human microbiome and to discover how certain microbiomes would connect with diseases, which is also called biomarker discovery. Some common problems have been revealed and emphatically discussed for reaching the research goals. Firstly, the dimension of the microbiome data is usually considerably high, which could reach hundreds of thousands in strain-level marker profile. Comparatively, the bio-sample number is meager, typically limited to about a hundred. This peculiarity of the data would inevitably cause the high sparsity of the training set, leading to the unsatisfactory performance of the prediction model. Secondly, as multiple papers have discussed, the generalization ability of the model to fit the data from different domains is essential for practical usage. For patients from different nations and areas, the composition of the microbiome can be vastly different. These challenges building a generalized model for all kinds of subjects. Finally, the learning process of deep learning is a black box. It is difficult to extract the interpretive features of the data from the model. This makes the biomarker discovery inconvenient to implement.

Many researchers have proposed solutions to address these problems. One relevant work is DeepMicro [8], a model focused on dimension reduction, attempting to address the high-dimension problem. The core methodology of DeepMicro is to combine various autoencoders, trying to extract the latent feature and preserve the internal structure with robustness at the same time. It is worth noting that the dimension reduction trick could conspicuously improve the training efficiency, which would benefit the tuning process of the hyper-parameters. According to their experiments, the autoencoder shows better accuracy than the traditional methods, such as PCA, in most situations. However, dimension reduction means unavoidable information loss, indicating that it is not necessarily the optimal option. GDmicro [9] tried to solve this problem from a different perspective. The main component of the GDmicro is the Graph Convolution Network, which is often applied in semi-supervised learning. By making use of the adjacency matrix of nodes and graph convolution to combine the information of labeled and unlabeled nodes, the model could learn from both the labeled and unlabeled samples. In this way, the shortage of the sample is mitigated. The input of the GCN is an inter-host microbiome similarity graph. The vital information this graph expresses is that a similar microbiome composition comes with similar disease status. Because the edges between nodes are built with the k-means clustering method, the hosts with similar microbiome composition will also be near each other. Similarly, PopPhy-CNN [10] applied the Convolution Neural Network to grab the spatial information of the microbiome data. But instead of using the graph structure, PopPhy-CNN tried to leverage the tree structure to address the problem. A phylogenetic tree is built with abundance data, and the similarity of the taxa is represented by the closeness of the nodes in the tree. The value of the internal nodes is the sum of its children nodes' abundance value. The feature number is reduced by transforming the tree into a 2-D matrix, indicating that fewer training samples are required to train the deep-learning model. In addition, to avoid the possible overfitting caused by the lack of samples, the researchers applied both L1 and L2 regularizations and the dropout method. GDmicro tends to solve the generalization problem by applying the Deep Adaptation Network. The researchers combined the data from multiple domains as the input of the Deep Adaptation Network to learn the domain-transferable latent features. With the latent features as the input, GDmicro could mitigate the influence of the domain discrepancy on the prediction model. The meta-analysis of large metagenomic datasets [11] tried to comprehensively assess the generalization ability of the model using the microbiome as the disease predictor. The methodology leverages the metagenomics-based machine learning model trained with the data collected from eight studies and six microbiome datasets. The researchers built the prediction model with SVM, RF, Lasso, and Enet. Then, cross-validation and cross-study were applied, respectively, to assess the accuracy and generalization of the model. Judging from the experiments, the model showed strong cross-stage generalization ability evaluated with samples from different stages. However, less generalization ability is observed for completely different datasets because of the cohort effects. The paper proceeded with an intriguing test by adding some healthy samples from other independent datasets, and the model's generalization ability was improved. This could be a good practice for improving the generalization of the transformer model. To be extended, adding more patient samples may also impact the generalization ability. GDmicro proposed a simple yet straight method based on the trained model. The main idea is to use the variable control method to test each species’ score of the model. The one with the high score can be a biomarker for the host status. As an extension of this idea, the paper proposed a way to analyze certain species’ contributions by modifying their abundance independently and then comparing the classification results to

test if they have a specific impact on the probability of different classes. To get the interpretive features for biomarker discovery, PopPhy-CNN applied a single convolution layer to do feature extraction for all samples grouped by classes. However, this may miss the crucial non-linear relationship between disease and microbiome, which can only be discovered by multiple layers.

Overall, these studies have certain similarities and principles worth considering, such as combining species-level abundance with strain-level biomarkers to obtain more information, making discriminative feature selection for biomarker discovery and training models, and evaluating their models on famous public datasets. Comparatively, the methods these studies used to accomplish the problem are various, but the basic pipeline of these studies remains analogical, which is to extract the feature of data first, then train the machine learning model with the training sets, and finally evaluate them on the testing sets.

3. Methodology

3.1. Algorithms and Models

Several algorithms are considered in the research including not only the traditional machine learning algorithms such as support vector machines(SVM), random forests(RF), as well as deep learning algorithms like multilayer perceptron(MLP) and Transformers. Among all of them, we are primarily focused on applying Transformer to build the classification models, because as to our knowledge, there are no existing studies that have done so.

Transformer is a deep learning algorithm which was introduced in 2017 by Google [12]. It has the ability to show better efficiency compared to former deep learning models by avoiding the complex recurrence or convolution such as CNN and RNN. Additionally, by applying the self-attention mechanism, Transformer would be able to grab the deeper relationships between the subparts of the sequence data. This is significant for doing natural language processing. Encoder-decoder architecture is the core part of the Transformer. To be more specific, the output of the encoder layer would be the latent feature representation of the input data. The output of the decoder part is the probabilities of the output sequence. The decoder part is important for the generative model. In related fields, there have been many successful examples of applying Transformer to dealing with sequence data, such as ChatGPT. However, there is insufficient number of research studying how to apply the Transformer on the human gut microbiome data. Therefore, we managed to take the Transformer encoder component for feature extraction, and then use it as the input to a fully connected neural network to build a disease diagnosis model.

We trained these models based on the microbiome data we collected and tested their prediction ability on the unseen dataset. We built both binary and multi-class classification models for disease diagnosis to explore the functionality of the Transformer. Then, we evaluated the classification models from multiple perspectives and select those with fair performance. With the models with decent performance, we then analyze their feature importance respectively to get a view of biomarkers for all three diseases.

3.2. Dataset and Feature Extraction

The features considered in the research are the relative abundances of microbiome. Namely it is the percentage of the certain microbiome detected in the biological sample. All the metagenomic data were obtained from the NCBI public SRA database, which is a public database for biological research. It is worth to note that the metagenomic data collected are sequenced with the same technology called 16s rRNA sequencing. 16s rRNA is a special fragment that commonly existed in most of the microbiome. More importantly, the 16s rRNA could be very different and recognizable for different species, which makes it significant for recognizing bacteria. We collected six different and independent datasets along with three diseases. The statistics of each dataset are shown in Table 1.

Table 1. Six datasets collected from NCBI

Cohort	Disease	Accession	Region	# of Sample	Ref
T2D-CN	T2D	PRJEB25715	China (Hebei)	58	[13]
T2D-PK	T2D	PRJNA554535	Pakistan	60	[14]
AD-US	AD	PRJEB51982	United States	50	[15]
AD-GER	AD	PRJEB59009	Germany	146	[16]
SAR-BJ	Sarcopenia	PRJNA691136	China (Beijing)	87	[17]
SAR-GD	Sarcopenia	PRJNA1005560	China (Guangdong)	60	[18]

All metagenomic sequences collected were annotated using KrakenUniq [19], a metagenomic sequence annotation software that based on the k-mers method. Through the annotation of metagenomic sequences, we obtained taxonomic information from superkingdom to species. We were concentrated on seven taxonomic ranks: 1.Clade 2.Phylum 3.Class 4.Order 5.Family 6.Genus 7.Species. Other ranks were not considered because rather they lose too much information during the annotation process, or the taxonomy is too ambiguous. For example, In the T2D-CN dataset, the average abundances at the order, family, genus, and species levels are 96%, 94%, 92%, and 74% respectively among all biological samples. However, the average abundance rapidly drops to 14% at the strain level, which is the lower level of species. For higher level taxonomies like super kingdom, they often have only 1 or 2 features in most cases. We consider the taxonomy to be ambiguous and there is not enough information to train a robust classification model.

The detailed statistics of this feature are shown in Table 2.

Table 2. Feature number for different biological ranks in datasets

Ranks	T2D-CN	T2D-PK	AD-US	AD-GER	SAR-BJ	SAR-GD
Clade	36	13	33	31	35	32
Phylum	44	32	44	38	43	42
Class	81	68	85	76	84	84
Order	180	152	187	169	190	187
Family	377	342	407	375	437	413
Genus	981	888	1149	1118	1440	1302

It is worth to notice that the feature number increase rapidly when the rank get to the genus and species ranks. However, the number of the samples are restricted within 100. This singularity becomes a challenge for building accurate models.

4. Experiments

4.1. Experiment Design

The experiments are held on two kinds of models respectively: 1.traditional machine learning models: SVM, Random Forest, Logistic Regression 2.deep learning models: MLP, Transformer. In order to get an unbiased performance result, we applied the k-folds cross validation method. To be specific, k-folds cross validation is to divide the whole dataset into k batches with equal number of samples. Take one batch as the test set and the other batches as the training set and then do the training and testing. Such process will be held k times with each batch set taken as the test set and get k different models. Finally calculate the average performance of these k models as the unbiased performance. In this experiment we set k as 5 and since which the sample number in training and testing set is divided with the ratio of 4:1.

For data processing part, we applied Z-score standardization and Min-Max normalization method to flatten the discrepancy between different individuals. Z-score standardization is to set the feature value as the original value menus the feature mean and the divided with the standard deviation. Z-score standardization would scale data to zero mean and unit standard deviation. This would help to mitigate the impact of individual differences. Min-Max normalization could scale data into the range between 0 and 1. Similarly, this could also help to mitigate the impact of the data volume, but the difference is that Min-Max normalization has a solid range and less sensitive to anomaly.

For model tuning, multiple rounds of train-test processes were executed to find the best hyper parameters. All random seed is set to 0 to keep the experiment result trackable. The models are firstly built as the binary classification models meaning that the output of the models is the possibility for the two class:1. healthy 2. diseased. The class with higher possibility will be considered as the classification result. Multi-class classification model is also considered in this project. Specifically, there will be four class: 1. healthy 2.T2D 3.AD 4. Sarcopenia. It's worth to mention that multi-class classification model is trained on the dataset with domain discrepancy for datasets are collected on samples from different regions.

4.2. Performance Evaluation

For the evaluation metrics of the classification model, three criteria are used to test the predictive power of the classifiers. The first is prediction accuracy, which is calculated based on the match between the ground truth and the predicted values. Another is the AUC score. AUC is computed using the area under the Receiver Operating Characteristic (ROC) curve [20], which is commonly used to evaluate binary classifiers. The AUC score ranges from 0.5 to 1. More closer it is to 1, more stronger the predictive power of the classifier. On the other hand, the closer it is to 0.5, the more it resembles random guessing. And the final one is the recall rate. Recall rate only consider the positive samples: TP / (TP + FN). In this way, recall rate could focused more on the prediction ability on positive samples, which is an important feature for disease diagnosis because it's reasonable that predicting the true diseased samples is more important than the loss of predicting the false healthy samples. The specific performance of the binary classification model is showed below. It's noted that the best result is picked from the model trained on original data, data with z-score standardization, data with min-max normalization.

The detailed evaluation of all binary classification models on each dataset is shown as followed.

Table 3. Cross validation results on T2D-CN with traditional machine learning model

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.83	0.86	0.91	0.76	0.79	0.65
Phylum	0.90	0.95	0.91	0.90	0.93	0.82
Class	0.88	0.94	0.71	0.90	0.97	0.74
Order	0.91	0.95	0.80	0.88	0.98	0.71
Family	0.91	1.0	0.82	0.91	0.99	0.83
Genus	0.95	0.99	0.86	0.93	0.97	0.85
Species	0.86	0.98	0.62	0.86	0.97	0.66

Table 4. Cross validation results on T2D-CN with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.75	0.91	0.88	0.91	0.98	1.0
Phylum	0.90	0.94	0.86	0.93	1.0	0.91
Class	0.89	0.93	0.87	0.96	0.98	0.91
Order	0.95	0.95	0.91	0.98	0.99	0.95
Family	0.93	0.95	0.95	1.0	1.0	1.0
Genus	0.93	0.98	1.0	1.0	1.0	1.0
Species	0.91	0.96	1.0	1.0	1.0	1.0

Table 5. Cross validation results on T2D-PK with traditional machine learning models

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.90	0.92	1.0	0.95	0.95	1.0
Phylum	0.90	0.90	1.0	0.92	0.99	1.0
Class	0.95	0.98	1.0	0.93	0.99	1.0
Order	0.98	1.0	1.0	0.95	0.99	1.0
Family	0.95	0.98	1.0	0.97	0.98	1.0
Genus	0.97	1.0	1.0	0.95	0.98	1.0
Species	0.97	1.0	1.0	0.95	1.0	1.0

Table 6. Cross validation results on T2D-PK with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.90	0.85	1.0	0.97	0.98	1.0
Phylum	0.92	0.97	1.0	0.97	1.0	1.0
Class	0.96	1.0	1.0	1.0	1.0	1.0
Order	0.98	1.0	1.0	1.0	1.0	1.0
Family	0.98	0.98	0.98	1.0	1.0	1.0
Genus	1.0	1.0	1.0	1.0	1.0	1.0
Species	1.0	1.0	1.0	1.0	1.0	1.0

Table 7. Cross validation results on AD-US with traditional machine learning models

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.70	0.86	0.74	0.78	0.85	0.76
Phylum	0.70	0.83	0.81	0.76	0.81	0.81
Class	0.70	0.84	0.74	0.66	0.79	0.73
Order	0.80	0.93	0.81	0.78	0.88	0.88
Family	0.96	0.99	0.94	0.92	0.96	0.90
Genus	0.76	0.89	0.88	0.94	0.98	1.0
Species	0.74	0.83	0.85	0.96	0.99	0.97

Table 8. Cross validation results on AD-US with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.64	0.80	0.71	0.92	0.96	0.96
Phylum	0.68	0.80	0.78	0.90	0.94	0.90
Class	0.76	0.86	0.77	0.94	0.92	0.97
Order	0.76	0.92	0.75	0.96	0.97	1.0
Family	0.88	1.0	0.91	0.98	1.0	0.97
Genus	0.86	0.89	0.84	0.94	0.97	0.97
Species	0.78	0.79	0.81	0.94	0.98	0.94

Table 9. Cross validation results on AD-GER with traditional machine learning models

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.56	0.58	0.91	0.57	0.54	0.63
Phylum	0.65	0.67	0.81	0.64	0.65	0.71
Class	0.59	0.61	0.88	0.62	0.62	0.68
Order	0.66	0.66	0.83	0.64	0.62	0.74
Family	0.55	0.54	1.0	0.68	0.64	0.74
Genus	0.60	0.62	0.92	0.70	0.69	0.78
Species	0.65	0.60	0.95	0.72	0.70	0.75

Table 10. Cross validation results on AD-GER with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.61	0.63	0.83	0.71	0.76	0.82
Phylum	0.60	0.67	0.93	0.76	0.76	0.83
Class	0.65	0.61	0.76	0.70	0.69	0.80
Order	0.61	0.61	0.69	0.66	0.74	0.80
Family	0.56	0.57	0.59	0.67	0.69	0.73
Genus	0.66	0.66	0.69	0.72	0.71	0.77
Species	0.72	0.70	0.80	0.76	0.76	0.84

Table 11. Cross validation results on SAR-BJ with traditional machine learning models

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.74	0.80	0.21	0.76	0.76	0.45
Phylum	0.73	0.76	0.39	0.78	0.80	0.45
Class	0.73	0.79	0.29	0.85	0.90	0.69
Order	0.75	0.80	0.41	0.82	0.89	0.55
Family	0.83	0.88	0.62	0.86	0.92	0.74
Genus	0.79	0.74	0.51	0.83	0.94	0.56
Species	0.84	0.88	0.61	0.81	0.87	0.45

Table 12. Cross validation results on SAR-BJ with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.70	0.65	0.30	0.87	0.85	0.83
Phylum	0.77	0.73	0.41	0.88	0.86	0.89
Class	0.78	0.81	0.56	0.92	0.85	0.85
Order	0.84	0.83	0.69	0.81	0.87	0.87
Family	0.82	0.85	0.70	0.86	0.90	0.87
Genus	0.78	0.81	0.53	0.79	0.84	0.77
Species	0.83	0.88	0.60	0.90	0.86	0.81

Table 13. Cross validation results on SAR-GD with traditional machine learning models

Ranks	SVM			RF		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.74	0.79	0.39	0.76	0.76	0.45
Phylum	0.77	0.75	0.40	0.78	0.80	0.45
Class	0.77	0.81	0.72	0.85	0.90	0.69
Order	0.78	0.81	0.73	0.82	0.89	0.55
Family	0.79	0.83	0.60	0.86	0.92	0.74
Genus	0.76	0.88	0.53	0.83	0.94	0.56
Species	0.74	0.90	0.45	0.81	0.87	0.45

Table 14. Cross validation results on SAR-GD with deep learning models

Ranks	MLP			Transformer		
	Accuracy	AUC	Recall	Accuracy	AUC	Recall
Clade	0.57	0.70	0.88	0.87	0.88	0.77
Phylum	0.68	0.78	0.77	0.83	0.84	0.86
Class	0.80	0.83	0.67	0.84	0.90	0.84
Order	0.83	0.79	0.73	0.85	0.92	0.86
Family	0.75	0.82	0.66	0.86	0.93	0.85
Genus	0.79	0.88	0.88	0.92	0.96	0.98
Species	0.75	0.87	0.75	0.93	0.94	0.98

5. Domain Discrepancy

It is a fact that for different biological samples, the data feature could be considerably different even for the samples with same disease status. This is especially serious for samples from different regions because there are multiple factors would affect the distribution of the gut microbiome such as diet, race, and climate. This problem is called domain discrepancy. With the inevitably domain discrepancy, the classification model trained on sole dataset would have rather poor generalization ability. For example, model trained on the T2D-CN dataset would have poor performance on T2D-PK.

Table 15. Cross validation results of Transformer trained on T2D-CN and tested on T2D-PK

Ranks	Transformer		
	Accuracy	AUC	Recall
Clade	0.73	0.76	0.75
Phylum	0.71	0.80	0.60
Class	0.75	0.81	0.62
Order	0.73	0.86	0.65
Family	0.68	0.78	0.60
Genus	0.7	0.77	0.57
Species	0.72	0.79	0.6

Similarly, in order to train a multi-class classification model, we need to combine the datasets from different regions, which would also introduce domain discrepancy for most of the samples are from different regions or nations.

Table 16. Cross validation results of Transformer trained on T2D-CN, AD-US, SAR-BJ

Ranks	Transformer		
	Accuracy	AUC	Recall
Clade	0.69	0.58	0.52
Phylum	0.77	0.72	0.65
Class	0.74	0.58	0.58
Order	0.79	0.73	0.70
Family	0.82	0.67	0.72
Genus	0.79	0.80	0.75
Species	0.87	0.94	0.80

Maximum Mean Discrepancy(MMD) [21] is the method that we chose to overcome this problem. MMD is a mathematical measurement of the difference between two different distributions. More specifically, MMD measures the statistical distance between two distributions to quantify how different they are.

$$\text{Dist}(P(X_S), P(X_T)) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \Phi(x_{S_i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \Phi(x_{T_j}) \right\|_{\mathcal{H}}$$

Multi-kernel MMD is an extension of MMD. The difference is that MK-MMD combines multiple kernel functions to measure the distance between the distributions. It provides a more powerful and adaptable way to measure distributional differences.

By calculating the MMD distance and adding it to the training loss, we can mitigate the impact of the domain discrepancy to the generalization ability of the models.

Table 17. Cross validation results of Transformer trained on T2D-CN and tested on T2D-PK with MK-MMD

Ranks	Transformer		
	Accuracy	AUC	Recall
Clade	0.75	0.72	0.75
Phylum	0.77	0.80	0.65
Class	0.82	0.87	0.70
Order	0.85	0.89	0.80
Family	0.86	0.87	0.85
Genus	0.90	0.91	0.90
Species	0.94	0.98	0.92

Table 18. Cross validation results of Transformer trained and tested on T2D-CN, AD-US, SAR-BJ with MK-MMD

Ranks	Transformer		
	Accuracy	AUC	Recall
Clade	0.70	0.65	0.54
Phylum	0.77	0.72	0.65
Class	0.78	0.60	0.62
Order	0.80	0.75	0.72
Family	0.82	0.67	0.72
Genus	0.80	0.78	0.70
Species	0.87	0.94	0.80

6. Biomarker Discovery

For the biomarker discovery, we mainly focused on the contribution of the features in different biological ranks. Those with larger contribution to the classification result would be considered as the potential biomarker for the disease.

There are two methods we applied. One is that the weight of the trained classification models. This is reasonable for the traditional machine learning methods, however it does not work for the neural network-based models Because the weights of these models are combined through complex non-linear functions in a multi-layer network, the volume of the weights does not directly reflect the importance of the features. The alternative method is SHAP [22]. SHAP (Shapley Additive Explanations) is a method that can interpretate any machine learning model with the shapely value in the game theory. If the shapely value of a feature gets positively larger with the feature value, then this feature is considered as the indicator for the positive class, in this condition is diseased. On the contrary, if the shapely value gets negatively larger with the feature value, then the feature is considered as the indicator for negative class, in this condition is healthy. Furthermore, we make a comparison between the median value of microbiome relative abundance of disease group and control group to validate the SHAP result. Furthermore, we also considered applying student t-test, which is a commonly used method for testing the relationship between two distributions, as validation.

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}.$$

T-test of top factors:

Bacteroidetes/Chlorobi group: T-statistics: 3.09 P-value: 0.003

FCB group: T-statistics: 3.09 P-value: 0.003

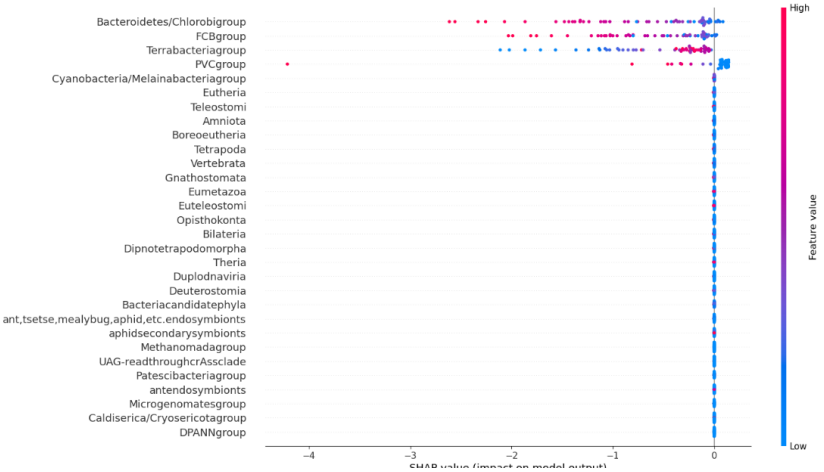

Figure 1: SHAP beeswarm graph of Transformer on T2D-CN dataset in clade rank

Terrabacteria group: T-statistics: -3.15 P-value: 0.002

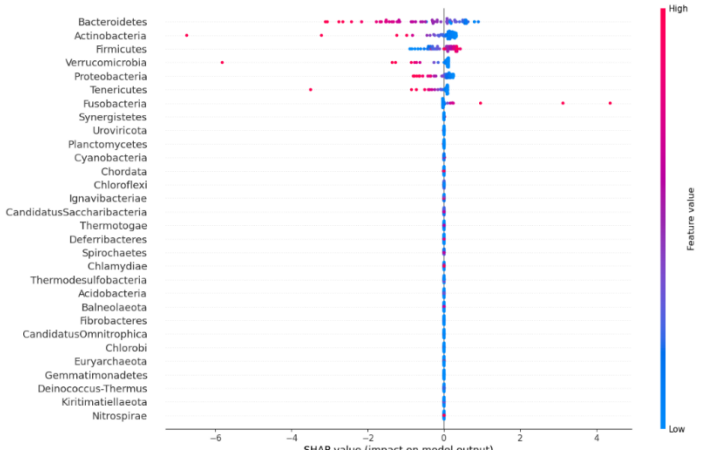

Figure 2: SHAP beeswarm graph of Transformer on T2D dataset in phylum rank

T-test of top factors:
Bacteroidetes: T-statistics: 3.09 P-value: 0.003
Actinobacteria: T-statistics: 1.17 P-value: 0.245
Firmicutes: T-statistics: -3.6 P-value: 0.0006
Proteobacteria: T-statistics: 2.18 P-value: 0.033
Fusobacteria: T-statistics: -2.23 P-value: 0.029

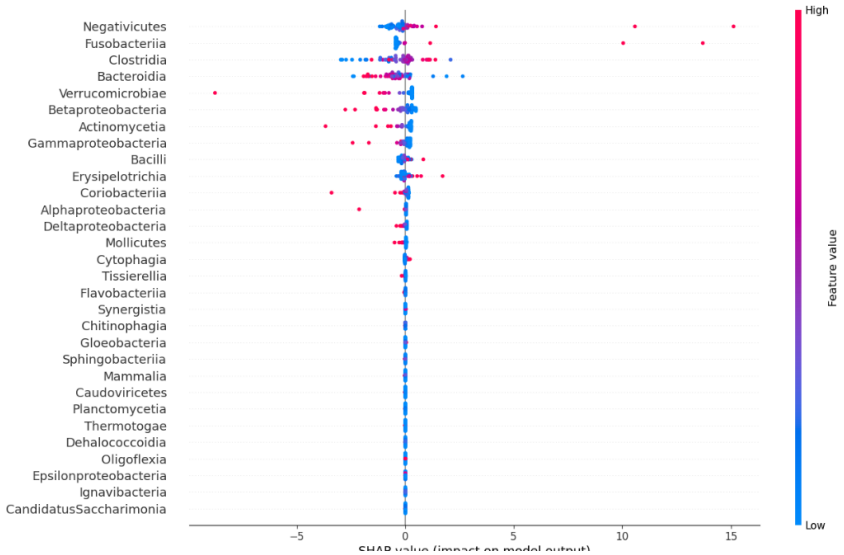

Figure 3: SHAP beeswarm graph of Transformer on T2D dataset in class rank

T-test of top factors:
Negativicutes: T-statistics: -2.69 P-value: 0.009
Fusobacteriia: T-statistics: -2.23 P-value: 0.029
Clostridia: T-statistics: -2.43 P-value: 0.018
Bacteroidia: T-statistics: 3.11 P-value: 0.002
Betaproteobacteria: T-statistics: 2.47 P-value: 0.016

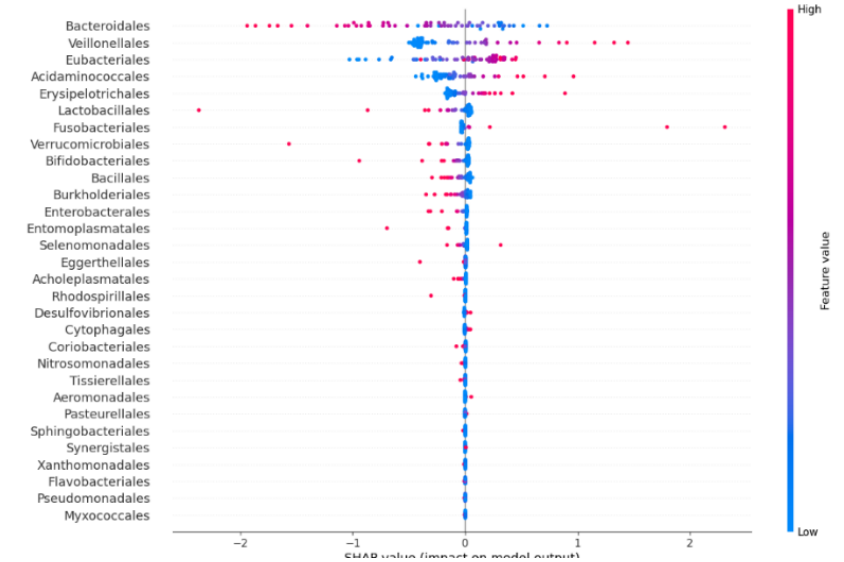

Figure 4: SHAP beeswarm graph of Transformer on T2D dataset in order rank

T-test of top factors:
Bacteroidales: T-statistics: 3.11 P-value: 0.002
Eubacteriales: T-statistics: -2.44 P-value: 0.018
Fusobacteriales: T-statistics: -2.23 P-value: 0.029
Burkholderiales: T-statistics: 2.44 P-value: 0.017

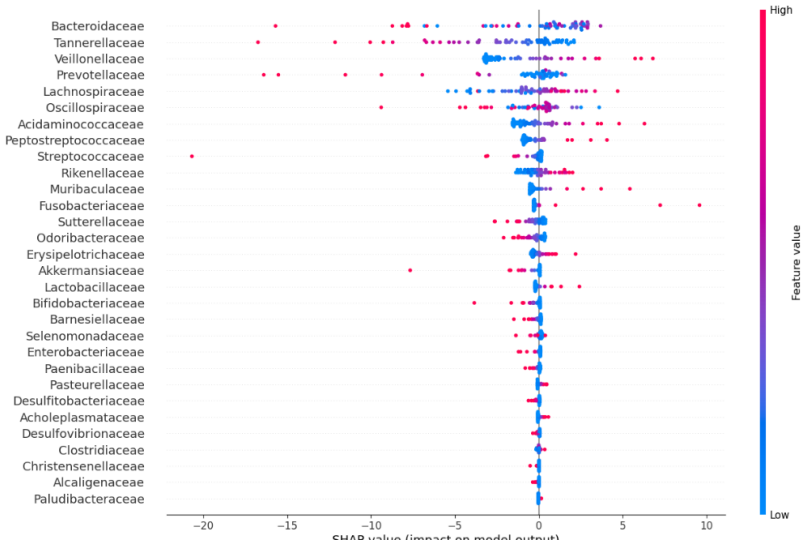

Figure 5: SHAP beeswarm graph of Transformer on T2D dataset in family rank

T-test of top factors:
Tannerellaceae: T-statistics: 3.79 P-value: 0.0003
Lachnospiraceae: T-statistics: -2.67 P-value: 0.009
Peptostreptococcaceae: T-statistics: -4.02 P-value: 0.0001
Fusobacteriaceae: T-statistics: -2.23 P-value: 0.029
Sutterellaceae: T-statistics: 2.35 P-value: 0.022
Odoribacteraceae: T-statistics: 2.35 P-value: 0.022

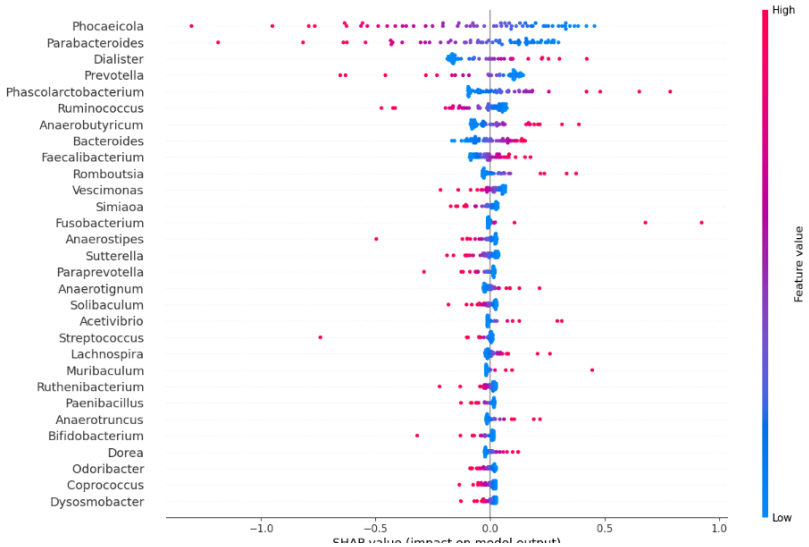

Figure 6: SHAP beeswarm graph of Transformer on T2D dataset in genus rank

T-test of top factors:
Parabacteroides: T-statistics: 3.83 P-value: 0.0003
Romboutsia: T-statistics: -4.12 P-value: 0.0001
Fusobacterium: T-statistics: -2.23 P-value: 0.029
Sutterella: T-statistics: 2.35 P-value: 0.02
Dorea: T-statistics: -2.02 P-value: 0.04

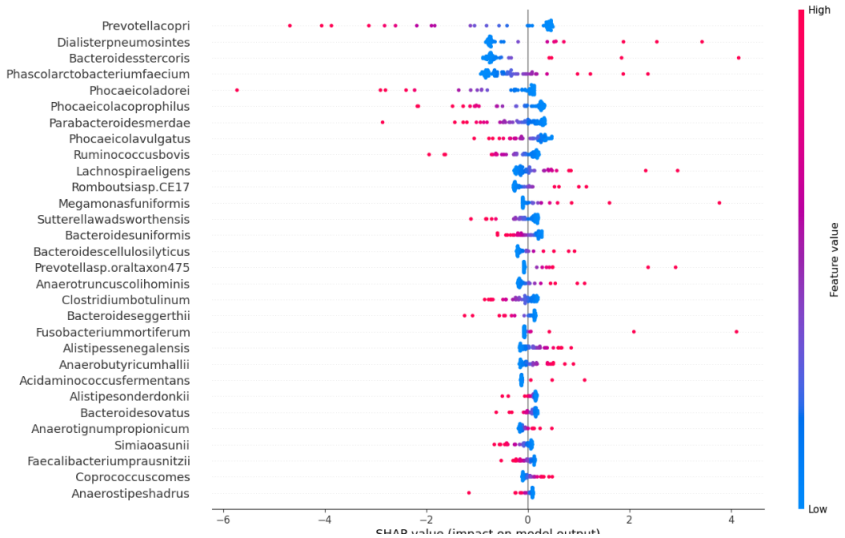

Figure 7: SHAP beeswarm graph of Transformer on T2D dataset in species rank

T-test of top factors:
Dialisterpneumosintes: T-statistics: -2.09 P-value: 0.04
Parabacteroidesmerdae: T-statistics: 3.07 P-value: 0.003
Romboutsiasp.CE17: T-statistics: -4.12 P-value: 0.0001
Sutterellawadsworthensis: T-statistics: 2.43 P-value: 0.018
Prevotellasp.oraltaxon475: T-statistics: -2.1 P-value: 0.039
Anaerotruncuscolihominis: T-statistics: -2.2 P-value: 0.03
Fusobacterium-mortiferum: T-statistics: -2.17 P-value: 0.03

7. Conclusions

It is showed that for traditional machine learning models, RF shows the best performance comparing to SVM and LR. RF could even show the perfect performance in some cases(accuracy:1, auc:1). For deep learning models, Transformer basically shows better performance than the MLP in all three datasets and biological ranks. This result indicates that the encoder part of the Transformer has successfully extracted the latent feature of the data and improved the prediction ability. Furthermore, it is noticed that for the lower ran00000000000ks such as genus and species, deep learning models achieved better results than traditional machine learning models, indicating that neural network do have the better ability to learn the com0plex feature structure with its non-linear transformation. For mitigating the domain discrepancy, MK-MMD achieves a success on the binary classification model but failed to improve the performance of the multi-class classification model.

With the shap score, we could analyze the weight of the feature that on each biological ranks. After combining the shap value and the student t-test, we had the more detailed and validated biomarker deduction as followed:

Pathogenic factors of the T2D diagnosis:

1. Lachnospiraceae;Eubacteriales;Clostridia;Firmicutes;Terrabacteria group.
2. Fusobacterium-mortiferum;Fusobacterium;Fusobacteriaceae;Fusobacteriales;Fusobacteriia;Fusobacteria.
3. Romboutsiasp.CE17;Romboutsia;Peptostreptococcaceae;Eubacteriales;Clostridia;Firmicutes;Terrabacteria group.

Healthy factors of the T2D diagnosis:

1. Bacteroidales;Bacteroidia;Bacteroidetes;Bacteroidetes/Chlorobi group.
2. Parabacteroides-merdae;Parabacteroides;Tannerellaceae;Bacteroidales;Bacteroidia;Bacteroidetes; Bacteroidetes/Chlorobi group.
3. Sutterella-wadsworthensis;Sutterella;Sutterellaceae;Burkholderiales;Betaproteobacteria;Proteobacteria; Bacteroidetes/Chlorobi group.

Pathogenic factors of the AD diagnosis:

1. Phocaeicola-vulgatus;Phocaeicola;Bacteroidaceae;Bacteroidales;Bacteroidia;Bacteroidetes; Bacteroidetes/Chlorobigroup;
2. Massilistercora-timonensis;Massilistercora;Oscillospiraceae;Eubacteriales;Clostridia;Firmicutes; Bacteroidetes/Chlorobigroup
3. Alistipesonderdonkii;Alistipes;Rikenellaceae;Bacteroidales;Bacteroidia;Bacteroidetes;Bacteroidetes/Chlorobigroup

Healthy factors of the AD diagnosis:

1. Faecalibacterium-prausnitzii;Faecalibacterium;Oscillospiraceae;Eubacteriales;Clostridia;Firmicutes; Bacteroidetes/Chlorobigroup
2. Lachnospira-eligens;Lachnospira;Lachnospiraceae;Eubacteriales;Clostridia;Firmicutes;Terrabacteria group;

Pathogenic factors of the Sarcopenia diagnosis:

1. Bifidobacterium-longum;Bifidobacterium;Bifidobacteriaceae;Bifidobacteriales;Actinomycetes; Actinomycetota; Terrabacteria group;

Healthy factors of the Sarcopenia diagnosis:

1. Phocaeicola-dorei;Phocaeicola;Bacteroidaceae;Bacteroidales;Bacteroidia;Bacteroidetes; Bacteroidetes/Chlorobi group;

2. Roseburia-hominis; Roseburia; Lachnospiraceae; Eubacteriales; Clostridia; Firmicutes; Terrabacteria group;

As for the practical application of the trained predictive model, in some cases the models could have the perfect performance for all three diseases. But still there are various hidden factors in the samples that may have a significant impact on the abundance distribution of the microbiome, for example, the history of antibiotic use. The clinical value of this model still requires extensive clinical trials to explore.

Acknowledgements

We are grateful to Prof. Sun Yanni and Dr. Shang Jiayu for giving their fruitful instructions and suggestions.

References

[1] Bradley, David, and Willa Hsueh. “Type 2 Diabetes in the Elderly: Challenges in a Unique Patient Population.” Journal of geriatric medicine and gerontology Vol. 2,2 (2016): 14.

[2] Bethel, M Angelyn et al. “Longitudinal incidence and prevalence of adverse outcomes of diabetes mellitus in elderly patients.” Archives of internal medicine Vol. 167,9 (2007): 921-7.

[3] Zhou, Zheng et al. “Gut Microbiota: An Important Player in Type 2 Diabetes Mellitus.” Frontiers in cellular and infection microbiology Vol. 12 (2022) 834485.

[4] “2023 Alzheimer's disease facts and figures.” Alzheimer's & dementia : the journal of the Alzheimer's Association Vol. 19,4 (2023): 1598-1695.

[5] Fung, Thomas C et al. “Interactions between the microbiota, immune and nervous systems in health and disease.” Nature neuroscience Vol. 20,2 (2017): 145-155.

[6] Ferreiro, Aura L et al. “Gut microbiome composition may be an indicator of preclinical Alzheimer's disease.” Science translational medicine Vol. 15,700 (2023): eabo2984.

[7] Cho, Ilseung, and Martin J Blaser. “The human microbiome: at the interface of health and disease.” Nature reviews. Genetics Vol. 13,4 260-70. 13 Mar. 2012.

[8] Oh, Min, and Liqing Zhang. “DeepMicro: deep representation learning for disease prediction based on microbiome data.” Scientific reports Vol. 10,1 6026. 7 Apr. 2020.

[9] Herui Liao, Jiayu Shang, Yanni Sun, GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data, Bioinformatics, Vol. 39,12 Dec. 2023.

[10] Reiman, Derek et al. “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 Vol. 24,10 (2020): 2993-3001.

[11] Pasolli, Edoardo et al. “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights.” PLoS computational biology Vol. 12,7 e1004977. 11 Jul. 2016.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[13] Li, Qian et al. “Implication of the gut microbiome composition of type 2 diabetic patients from northern China.” Scientific reports Vol. 10,1 5450. 25 Mar. 2020.

[14] Ahmad, Aftab et al. “Analysis of gut microbiota of obese individuals with type 2 diabetes and healthy individuals.” PloS one Vol. 14,12 e0226372. 31 Dec. 2019.

[15] Vogt, Nicholas M et al. “Gut microbiome alterations in Alzheimer's disease.” Scientific reports Vol. 7,1 13537. 19 Oct. 2017.

[16] Troci, Alba et al. “Disease- and stage-specific alterations of the oral and fecal microbiota in Alzheimer's disease.” PNAS nexus Vol. 3,1 pgad427. 11 Dec. 2023.

[17] Kang, Lin et al. “Alterations in intestinal microbiota diversity, composition, and function in patients with sarcopenia.” Scientific reports Vol. 11,1 4628. 25 Feb. 2021.

[18] Zhou, Jing et al. “Characteristics of the gut microbiome and metabolic profile in elderly patients with sarcopenia.” Frontiers in pharmacology Vol. 14 1279448. 3 Nov. 2023.

[19] Breitwieser, F P et al. “KrakenUniq: confident and fast metagenomics classification using unique k-mer counts.” Genome biology Vol. 19,1 198. 16 Nov. 2018.

[20] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Vol. 27(2006) No.8, p.861-874.

[21] Gretton, A & Borgwardt, K. & Rasch, Malte & Schölkopf, Bernhard & Smola, AJ, A Kernel Two-Sample Test, The Journal of Machine Learning Research, Vol. 13(2012) p.723-773.

[22] Scott M. Lundberg and Su-In Lee: A unified approach to interpreting model predictions, International Conference on Neural Information Processing Systems (NY, USA, December, 2017) Vol. 1, p.4768–4777.