Phy-PMRFI : Phylogeny-aware Prediction of Metagenomic Functions using Random Forest Feature Importance

Jyotsna Talreja Wassan, Haiying Wang, Fiona Browne, *Member, IEEE,* Huiru Zheng*, *Senior Member, IEEE*

***Abstract*—High-throughput sequencing techniques have accelerated functional metagen*omics* studies through the generation of large volumes of '*omics*' data. The integration of these data using computational approaches is potentially useful for predicting metagenomic functions. Machine learning models can be trained using microbial features (e.g. taxonomical units in human microbiome) which are then used to classify microbial data into different functional classes (e.g. healthy versus diseased states). For analyzing the *omics* data, features (i.e. the microbial taxas) as well as taxonomical relations between the features are important. The relationships are potentially uncoverable from the phylogenetic tree of microbial taxas. In this paper, we propose a novel integrative framework, namely Phy-PMRFI, driven by phylogeny-based modelling of *omics* data to predict metagenomic functions by using important features selected by a Random Forest Importance (RFI) strategy. The proposed framework integrates the underlying phylogenetic tree information with abundance measures of microbial species (features) by creating a novel phylogeny and abundance aware matrix structure (PAAM). Phy-PMRFI progresses by ranking the columns of the obtained matrix (i.e. the microbial features) by using the RFI measure, which are further used as input for the microbiome classification. The resultant feature set enhances the performance of the most popular state-of-art methods such as Support Vector Machines. Our proposed integrative framework also outperforms the state-of-the-art pipeline of Phylogenetic Isometric Log-Ratio Transform (PhILR) and MetaPhyl (e.g. obtaining 90 % accurate predictions with Phy-PMRFI over human throat microbiome in comparison to other approaches of PhILR with 53% and MetaPhyl with 71% Accuracy).**

***Index Terms*—Metagenomics, Phylogeny, Classification, Machine Learning (ML), Operational Taxonomic Units (OTUs), Random Forest Importance (RFI)**

I. Introduction

Metagenomics provides a non-cultured approach for the analysis of genomic content of microbial communities [1]. A metagenomic sample typically consists of the quantative abundance of microbial taxas at different taxonomic levels of taxonomy (*Kingdom, Phylum, Class, Order, Family, Genus*, and *Species*), which are represented as nodes on a phylogenetic tree [2].

The article is submitted in March 2019. This research is supported by Research Strategy Fund of Ulster University, United Kingdom. Huiru Zheng (h.zheng@ulster.ac.uk) is the corresponding author of this research.

Jyotsna Talreja Wassan is a researcher with the School of Computing, Ulster University, Co. Antrim, BT37 0QB, U.K. E-mail: wassan-jt@ulster.ac.uk .Haiying Wang is a Reader with the School of Computing, Ulster University, Co. Antrim, BT37 0QB, U.K. E-mail: hy.wang@ulster.ac.uk.Fiona Browne is a Lecturer with the School of Computing, Ulster University, Co. Antrim, BT37 0QB. U.K. E-mail: f.browne@ulster.ac.uk.Huiru Zheng* is a Professor of Computer Science with the School of Computing, Ulster University, Co. Antrim, BT37 0QB, U.K. E-mail: h.zheng@ulster.ac.uk

Apart from the tree topology, evolutionary distance of each node also serves as an important basis to evaluate microbial communities [3]. The distance annotated on a branch separating two metagenomic sequences in a phylogenetic tree, represents an estimate of their evolutionary divergence [3,4]. The evolutionary distances between sequences connected through multiple branches is the sum of the evolutionary distances represented by each branch [4]. In this paper, we introduce a method that integrates node-by-node information from the phylogenetic tree by incorporating evolutionary distances and abundance of taxas to improve predictive models over microbial taxa at different taxonomic levels. The goal of these predictive models is to associate several taxa of varying phylogenetic depth with the environmental phenotype (for example, to associate human microbiome taxa with different body sites or with human disease states) [5]. Since each node on the phylogenetic tree share a certain degree of evolutionary similarity; integrating such information in metagenomics analysis is useful as phylogenetically close microbial taxas tend to have similar effects on the functional phenotype [6,7].

We here present a novel framework that aids in determining which taxa at different taxonomic levels matter in order to associate a metagenomic sample with environmental phenotypes. The framework progresses by implementing a novel phylogeny and abundance-aware integrative approach, creating a matrix structure combining phylogeny with the abundance counts of microbial species (i.e. phylogeny and abundance aware matrix named PAAM). The feature columns of PAAM comprise of quantative profiles of both leaf level and intermediate nodes of the phylogenetic tree. We used Random Forest (RF) to identify important microbial features (i.e. the columns of PAAM), that are useful in classifying between different phenotypic groups to improve the metagenomic predictions. RF [8] has been widely used in omics data analysis providing good predictive accuracy and information on variable importance which is useful for classification tasks [9]. The informative microbial taxas obtained from applying RF, was inputted to three commonly used machine learning (ML) classifiers: (1) Support Vector Machines (SVMs) [10], (2) Logistic Regression (LR) [11], or (3) Naive Bayes (NB) [12].

In this paper we outline our proposed framework Phy-PMRFI (Phylogeny-aware modelling for prediction of metagenomic functions using RF Feature Importance), for metagenomics functional classification which integrates quantitative profiles of taxas and biological information derived from phylogeny of the microbial taxas. By integrating metagenomes structure and function we aim to answer an important research question "Is phylogenetic relatedness a good predictor of functional similarity (similar niche, states, environmental factors) in metagenomic studies?" In order to do this, we used three microbiome datasets as Use Cases to demonstrate the utility of Phy-PMRFI framework in predicting functions of metagenomic data. The inclusion of tree structure could help in determining the metagenomic functions according to the natural properties of a microbial community.

The rest of the paper is organized as follows. Section II highlights the related work. Materials and methods used in the current study are listed in Section III. Experimental results and discussions are enlisted in Section IV. Section V provides a summary and future research directions.

II. Related Work

Metagenomics [1] supports the investigation of complete microbial communities' present in an environment and their relationship with environmental metadata. Uncovering the function of the microbial genome from its structure forms an important problem in the ML domain [13-15]. An extensive review of ML modelling in metagenomics can be found in [13] in addition to the tools highlighted in [14,15]. Research in [16,17] has suggested the use of supervised classification ML techniques to effectively categorize quantative information of abundances of microbial genes into environmental functions (phenotype). Such analysis addresses the question of associating the structure of microbiome communities present in an environment with their functional potential. For example, the Human Microbiome project (HMP) [18], is serving as a catalyst to understand the relationship between the human microbiome and health; and to study how microbial composition varies between distinct body site niches [5,18].The basic unit for microbiome analysis is formed by grouping genes (such as the marker 16S rRNA gene sequences) of microorganisms at a threshold percentage sequence similarity, this measure is termed Operational Taxonomic Units (i.e. OTUs) [19]. However, abundance counts of OTUs alone fail to include the relatedness and distribution of a species in a microbial sample. Therefore, appropriate use of hierarchical ancestral structures in metagenomics could lead to a more informative analysis. Phylogeny is important as it incorporates the evolutionary history and diversity of the microbial taxas [6]. The common ancestry (i.e. from *Phylum* to *Species* taxonomic levels) is usually structured in the form of a phylogenetic tree (also known as *tree of life*) [2], which could be incorporated into the ML analyses of microbiome datasets in addition to the abundance counts of species. Studies [16,17] have suggested that it has mainly been the abundance counts of microbial organisms that have been considered when applying ML to metagenomic data for determining the biological functional roles; instead of their phylogenetic relatedness. Although, there have been some recent efforts to develop phylogenetically-aware computational methods for predicting metagenomic functions which we review as follows.

Ning and Beiko [20] recently performed an analysis classifying oral microbiome samples from HMP [18], using a ML framework involving weighted and unweighted UniFrac phylogeny-driven distances [21] to customize the kernel for SVMs [10,20,21]. The workflow of PhyloRelief as proposed by Albanese et al. [22], is driven by the relief strategy of selecting OTU features based on phylogenetic weights annotated on tree branches. The Phylogenetic Isometric Log-Ratio Transform (PhILR) approach described in [23] primarily focuses on the compositional nature of the microbiome where compositional parts are transformed using the "Isometric Log-Ratio transform (ILR)", utilizing the reference weights obtained from a phylogenetic tree. PhILR provides an approach to overcome the challenges associated with the compositional nature of OTU data with evolutionary analysis. MetaPhyl, proposed by Tanaseichuk et al. [24], is based on tuning the ML model of multinomial LR based on natural grouping and relatedness of species, as encoded in a phylogenetic tree. The method aims to optimize the coefficients of LR model by regularizing the tree-guided penalty function which is based on a hierarchical grouping of leaf-level OTU features. The aMiSPU test (adaptive Microbiome based Sum of Power statistical test) [25], which was designed with the aim that differential weightings of OTUs in accordance to their importance, can potentially improve the association of microbiome into functional roles. Reiman et al. [26] recently proposed a novel deep learning approach based on Convolution Neural Networks (CNNs) using phylogenetic distance, for classifying metagenomes.

However, challenges in this domain pertains to the areas of data handling, data integration and data analysis [27]. The high-dimensionality of metagenomic data (i.e. greater number of microbial features than the number of samples), along with the sparse nature of such data (due to the absence of some microorganisms) and probable existence of high variability in species of a microbial community makes classical ML challenging.

To address the challenges of the metagenomic analysis we developed Phy-PMRFI, a novel framework that: (1) integrates evolutionary distance measures (annotated on phylogenic tree branches) and abundances of microbial species into a mathematical matrix structure; (2) performs feature engineering using RFI to handle high-dimensional metagenomic data; (3) performs classification of microbial samples using a supervised ML. The framework is an extension of preliminary study conducted in [28]. The framework is inspired from the observations in [7, 28], that have indicated that the subset of intermediate nodes of the phylogenetic tree could play an important role in metagenomic classification models for complex microbial communities rather solely considering the leaf nodes (OTUs) of trees. We have modelled the quantative profile of internal nodes by integrating evolutionary distances and abundance count of the children nodes. We have then benchmarked Phy-PMRFI framework with 1) ML methods using only raw abundance counts, 2) other feature importance measuring strategies and 3) the other phylogenetic measures of PhILR [23] and MetaPhyl [24].

III. Material and Methods

In this section, we provide a brief description of materials and methodology used in the current study.

A. Materials

We applied the methods described below to 16S rRNA publicly available data sets. These datasets have been obtained from three different sources and form the Use Cases in this paper.

- Throat Dataset (Use Case 1): The dataset is obtained from a study by Charlson et al. [29], investigating the effect of cigarette smoking on the bacterial communities present in human's respiratory tract. It contains measurements obtained from 28 smoking and 32 non-smoking individuals. The dataset is comprised of 856 OTUs, and 60 samples. The source files from R package are also available as part of MiSPU package (*https://cran.rproject .org/web/ packages/MiSPU/)*.
- HTS-SIP Experiment Dataset (Use Case 2): This dataset relates to the DNA-based stable isotope probing (SIP) experiments incubating aliquots of soil with either 13C-glucose/Cellulose (treatment) or 12C the unlabelled control) [30]. It aids in determining microorganisms in soil that help in incorporating isotopically labeled substrate into biomass. The dataset is available as part of the HTSSIP package in R (*https://cran.rstudio.com/web/packages/HTSSIP/vignettes/HTSSIP_intro.html)*, developed for analysing high throughput sequence (HTS) data from DNA- or RNA-based SIP experiments [30].
- Human Microbiome Dataset (Use Case 3): This dataset consists of 16S rRNA sequences from Human Microbiome Project (HMP) [18], funded as an initiative of the NIH Roadmap for Biomedical Research. The dataset is comprised of 3285 Samples and 5830 OTUs to be classified into four body sites of Oral (1818), Skin (966), Vaginal (291) and Stool (210). The source is available at a GitHub repository (*https://github.com/twbattaglia/MicrobeDS*) for large-scale microbiome datasets [18].

B. Methods

In this subsection, we describe the proposed framework Phy-PMRFI and its major components (as shown in Algorithm 1). The proposed

approach integrates biological relatedness from the phylogenetic tree (structure) along with the abundance counts of OTUs to classify microbiome into phenotypes (functions). We base our framework on the observation that the natural properties of a microbial community as driven by a phylogenetic measure, could aid in determining metagenomic functions in an improved way [6,7,24,28]. A 2D matrix data structure PAAM (i.e. Phylogeny and abundance aware matrix) is introduced. To maximize the performance of our experimental design with Phy-PMRFI we followed an integrated workflow focusing on the following steps.

a. Inputs. The metagenomic predictions are derived from a reference microbial population with the quantitative metagenomic profiles as inputs and functional phenotypes as the outcome of interest. The quantitative metagenomic profiles include OTU abundances and phylogenetic distances. The OTUs in the metagenomic samples are clustered based on their DNA sequence similarity at a certain threshold to further create an abundance count matrix X (as shown in Eq. (1)), with 'm' metagenomic samples and 'n' OTUs.

$$X(m,n) = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots\ x_{i,j} & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \quad (1)$$

A phylogenetic tree is defined as a connected acyclic undirected graph with N nodes and B branches and $|B| = |N| - 1$. The leaves/tips in the tree represent OTUs. An important piece of information that is extracted from the phylogeny is the evolutionary phylogenetic distance (PD) embarked on the branches, which are required to span a given set of taxas on the phylogenetic tree [4].

b. Pre-processing. We transformed the microbial profile of metagenomic samples into a phylogeny and abundance aware matrix (PAAM), which combines abundance counts from the OTU abundance table and the PD annotated on branches of a phylogenetic tree, following a hierarchal manner (as indicated in Step 2. Algorithm 1). PAAM of size $i\ X\ j$, was constructed using the proposed approach, where i represents the number of samples and j represents the total number of nodes in the tree. The nodes of the tree consist of OTUs (i.e. leaf level nodes) as well as the ancestral nodes (i.e. internal nodes). The abundance weights of leaf level OTUs remained the same in PAAM. However, the entries for ancestral nodes in the PAAM matrix were computed by combining PD and abundances of the constituting OTUs. The abundance of each OTU was weighted by the PD to span ancestral nodes at each level of a tree, forming a hierarchal topology. The procedure for calculating weighted abundances of ancestral nodes is detailed in Step 2 of Algorithm 1 used for constructing the PAAM feature space.

c. Feature Engineering. In our proposed approach, the feature-set included both OTUs (leaf-level nodes) as well as the combination of OTUs based on abundances and phylogeny in the analysis (i.e. ancestral nodes in the topology). Feature engineering over such high dimensional feature space is expected to yield an advantage in the classification of metagenomes to reduce the complexity.

Feature engineering aids in selecting feature variables that are useful in predicting the response phenotypes when building supervised predictive models. We primarily attained feature importance by using the ML algorithm RF [8,31], suitable for high-dimensional metagenomic data [31]. Breiman [8] proposed the RF technique which works by generating several decision subtrees by bootstrapping over the learning data and suggested the Gini coefficient index [31] to calculate node impurity, as one of the key contributions of RF, to decide upon the important features (i.e. the features that are useful as splitting nodes in the RF subtrees). In RFI, the importance of features is calculated by a total decrease in tree-node impurities from splitting on the predictor feature variable and is averaged over all sub-trees in the RF strategy [31]. We found that the ancestral nodes modelled in PAAM contributed as important features by RFI, apart from the leaf-level taxas. In order to benchmark this approach, we compared it with other featuring engineering techniques listed below:

i) Variable Importance by ML using optimized distributed gradient boosting (i.e. XgBoost): This method calculates an importance score for each feature based on its participation in making key decisions with boosted decision trees as suggested in [32].

ii) RreliefF Measure: This method [33], calculates the weights of features to determine how well their value is distinguished between different samples, based on finding the nearest hits (i.e. a feature is observed close to a neighbor having the same class) and misses (i.e. a feature is observed with a neighbor having a different class), of instanced samples [33].

iii) glmnet: The method tends to fit a generalized regression model via penalized maximum likelihood, tuning the parameter settings of alpha, with lasso (alpha ~ 1), elastic-net (alpha ~ 0.5), or ridge (alpha ~ 0) penalty [34]. The regularization method tends to introduce penalty constraints to the coefficients of feature variables using the predictive model of LR.

Our proposed framework was further compared with the phylogeny-aware computational method PhILR [23]. Implementation of PhILR required three additional steps of pre-filtering i) removing OTUs that were not seen with more than 3 counts in at least 20% of samples; ii) with a coefficient of variation ≤ 3; iii) a pseudo counts of 1 was added to the remaining OTUs to avoid calculating log-ratios involving zeros [23].

d. ML Classification. To develop the most suitable ML model for classification, different classifiers were evaluated for their suitability in the prediction task using the selected features by RFI. These include: i) kernel-based modelling with SVMs with *Poly kernel* [10], ii) regression-based modelling with LR [11], and iii) the probabilistic-based approach with NB [12]; along our proposed framework. Leave-one-out cross validation (LOOCV) [35], was used as a validation strategy to fit these ML models used over Use Cases 1 and 2, allowing a single sample as the validation data, and the remaining samples as the training data. Each observation in the sample was used once as the validation data. However, we partitioned the input data set to train and test sets for ten-folds cross-validation for the Use Case 3.

In case of small-sized datasets, a ML model tends to be more sensitive towards any noise or sampling artefacts. In such cases, ten- folds cross-validation would lead to a very small sized training set; and may face high variance and bias issues [35]. It is better to use LOOCV in such examples. Hence, we used LOOCV for Use Case 1 and 2, but 10-folds cross validation for Use Case 3. We also characterized an experimental set-up with i) ML modelling over the PhILR [23] and ii)) MetaPhyl [24] for metagenomic analysis of OTUs with phylogenetic annotations.

e. Performance Evaluation. The Accuracy [36] and *K*appa [37] performance assessment metrics were used for

evaluating the classification models in our study. The Accuracy is defined as the fraction of correctly classified samples trained with LOOCV in Use Cases 1 and 2; and with 10-folds cross validation in Use Case 3 (as shown in Eq.2.).

$$Accuracy = \frac{NP_c}{NP_t} \quad (2)$$

where NP_c are total number of correct predictions and NP_t represents the total number of predictions.

Similarly, *K*appa [37] is used to evaluate the inter-rater agreement between classifications on ordinal or nominal scales and is considered a more robust measure than simple percent agreement calculation since it considers the agreement occurring by chance (as indicated in Eq.3.) [37].

$$Kappa = \frac{P_a - P_e}{1 - P_e} \quad (3)$$

where P_a is the actual probability of occurrence (agreement between raters), and P_e is the expected probability of occurrence (chance agreement), with the class labels.

*K*appa is more robust than scalar metrics of *A*ccuracy as it considers the marginal distribution of the response variable well [36,37].

IV. Results and Discussions

It is important to understand the performance of the proposed framework when compared to alternative state-of-the-art prediction methods (as enlisted in Section II). We performed a comprehensive set of experiments focused on evaluating the performance of the predictive models for benchmarking.

A. Performance of Prediction Models

In this work, we investigated the combination of feature subsets obtained from RFI ranking applied over PAAM and the classifier models. We conducted this study to devise an efficient framework for downstream metagenomic analysis and subsequently to evaluate the efficiency of functional predictions, with and without the inclusion of phylogeny for benchmarking. The experimental environment used was an Intel(r) Core (TM) i7 -8650U CPU @ 1.90 GHz/2112 GHz, RAM 16.0 GB, x64 Windows OS. For, preprocessing a *Perl* script was created to generate the PAAM structure. We used R (*https://www.rstudio.com/products/rpackages/*) platform for conducting the experiments for measuring the features' importance using RFI implemented as part of the randomForest package [31]. We used the caret package (*http://CRAN.R-project.org/package=caret)* [38] in R for the other supervised learners (SVM, NB, LR, glmnet) to implement models using cross validation. Implementation of RReliefF [33] was undertaken using functions from the FSelector package (*http://CRAN.R-project.org/package=FSelector*). The results of the ML models were obtained by resampling across the tuning parameters' settings of the models implemented in the caret package [38], to achieve the best performance. A summary of these results is presented below.

1) Distinguishing Microbiome samples using Phylogenetic Information

In the first experiment, we obtained the performance of state-of-the-art classifiers RF [8], SVMs [9] and LR [10] using the raw abundance count table of the OTUs in all three Use Case data sets (Section 2). Furthermore, we benchmarked the ML models when built using PAAM (i.e. phylogenetic measure), against the results obtained using only the raw OTU abundances (i.e. non-phylogenetic measure). Here, we observed that the phylogenetic measure did not yield an improved performance relative to the non-phylogenetic measure in Use Case 1 and 2 except in the case of LR applied over PAAM in Use Case 2

Algorithm 1: Workflow of Phy-PMRFI

1. **Input**: A phylogenetic tree 'T_n' with '*n*' OTUs and '*n-1*' ancestral nodes; taxa abundance Matrix *X (m, n)* with 'm' as number of samples & '*n*' as number of OTU features. Also predefined functional (phenotype) class categories for supervised learning.
2. **Pre-processing (Construction of PAAM Feature Space):** - A new phylogeny and taxa abundance aware matrix *X (m,2n-1)* is constructed with 'm' samples and '*2*n-1*' features containing *n-1* ancestral nodes as the new features as per following procedure.

   Procedure:

   $i \leftarrow 0$
   $j \leftarrow 0$
   For each sample row *'i'* in Matrix *X (m, n)* do
   $j \leftarrow n+1$ //indexing for newly constructed feature in PAAM
   For each ancestral node '*v*' in T_n do
   X (i, j) ← 0
   For each OTU '*u*' in T_n and *X (m, n)* do
   If OTU '*u*' in sample *i,* is descendent of node '*v*' in the T_n, then
   $PD_{u,v}$ ← phylogenetic distance of OTU '*u*' from node '*v*'
   $A_{u,i}$ ← abundance count of OTU '*u*' in sample *'i'*
   Weighted abundance of ancestral node '*v*' i.e. $WA_v = \frac{A_{u,i}}{PD_{u,v}}$
   $X(i,j) \leftarrow X(i,j) + WA_v$
   End
   End
   $j \leftarrow j + 1$
   End
   End.
3. **Feature Engineering: Apply Feature Engineering with Random Forest Importance over constructed *X (m, 2n-1)***

   It uses Gini impurity index to evaluate each feature-column in *X (m,2n-1)*. RF are an ensemble of decision trees. Gini Impurity of each node of a RF tree is formulated as follows:

   $$\sum_{i=1}^{L} -f_i(1-f_i),$$

   where $\mathbf{f_i}$ denotes the frequency of class label *i* at a node in RF tree and *L* are the total number of class labels

   Mean Decrease in Gini Index is calculated as a weighted average of the total decrease in the Gini Impurity metric weighted by the proportion of samples reaching a given node in a RF tree. The index is averaged for all the constituent trees of RF. The higher the mean decrease in Gini, the more important the feature is considered. The top *N*% features were modeled (*N*=5/10/20/40/60).
4. **Classification: Apply state-of-the-art ML Supervised Classifier** This uses classification functional model (SVM, LR or NB) over selected features in Step (3), by evaluating its performance with the measures of: - i) Accuracy and ii) *K*appa
5. **Output:** Microbial taxa classified into Functional Phenotypes (Classes)

TABLE 1
Performance of ML Models with LOOCV over Use Case 1

ML Model	Non-Phylogenetic (over raw abundances)		Phylogenetic (over PAAM)	
(LOOCV)	Accuracy	*K*appa	Accuracy	*K*appa
RF	0.700	0.386	0.683	0.359
SVM	0.716	0.429	0.700	0.394
LR	0.733	0.461	0.716	0.429
NB	0.466	0.095	0.566	0.129

(Tables 1 and 2). However, in Use Case 3, RF, SVM and LR applied over PAAM produced better performance results than when applied over the raw OTU abundances only (Table 3). PAAM has almost twice as many features than the OTU abundance table and hence has higher

dimensionality. The increased dimensionality may influence building an accurate model. Therefore, applying feature selection as detailed in the next steps of analysis has the potential to improve ML modelling over the phylogenetic measures of PAAM.

TABLE 2
PERFORMANCE OF ML MODELS WITH LOOCV OVER USE CASE 2

ML Model	Non-Phylogenetic (over raw abundances)		Phylogenetic (over PAAM)	
(LOOCV)	Accuracy	*K*appa	Accuracy	*K*appa
RF	0.942	0.913	0.913	0.870
SVM	**0.964**	**0.946**	0.920	0.881
LR	0.835	0.751	**0.885**	**0.823**
NB	0.570	0.360	0.580	0.360

TABLE 3
PERFORMANCE OF ML MODELS WITH 10-FOLDS CV OVER USE CASE 2

ML Model	Non-Phylogenetic (over raw abundances)		Phylogenetic (over PAAM)	
(LOOCV)	Accuracy	*K*appa	Accuracy	*K*appa
RF	0.967	0.944	**0.975**	**0.959**
SVM	0.931	0.884	0.939	0.898
LR	0.923	0.889	0.943	0.909
NB	0.911	0.854	0.841	0.745

2) Classification of the Microbiome using RF- based selected Important Features

Since the inclusion of the phylogenetic measure did not improve the overall classification Accuracy in every Use Case (as discussed in sub-part (1) above), we further investigated the use of RFI [31] technique (as described in Section III-B) for selecting important microbial taxas from the PAAM (termed as Phy-PMRFI i.e. phylogeny aware predictive modelling with RFI). The study in [8,9,31], suggested the application of the RF methodology to obtain variable importance measures. We benchmarked the performance of our approach by applying RFI over the OTU table with abundance count information only. The goal of this task is to analyze whether considering relationships among OTUs might lead to better prediction performance. The integration of phylogeny with OTU abundance could result in the improvement of microbiome sample classification by using a smaller number of RF-ranked features over the three Use Cases (Tables 4,5, and 6). Hence, modeling the phylogenetic measure between OTUs allows the RFI method to exploit the biological relationships and produce improved predictions.

When comparing the classification methods over the top 5, 10, 20, 40 or 60 % of the feature set obtained from RFI applied over PAAM as part of Phy-PMRFI framework, we observed that various resultant ML models indicated an improvement over original OTU abundance table in all three Use Cases. In Use Case 1, where we used the throat microbiome data, an ensemble of Phy-PMRFI with SVM produced better results (e.g. Accuracy: 0.900, *K*appa: 0.798 over top 20% features) than SVM applied over top 20 % of raw OTU abundance features (Accuracy: 0.883, *K*appa: 0.765) (Table 4). Overall performance of LR and NB was also improved, when applied over the RFI selected feature-set obtained from PAAM in comparison to the set obtained from original OTU abundances (Table 4) in Use Case 1. Also, the modelling with RFI (Table 4) indicated a significant improvement over the models applied over raw OTU abundance in Use Case 1 (Table1). For, example, Phy-PMRFI improved the classification Accuracy of NB from 0.466 (over the raw abundances of taxas; Table 1) to 0.767 (using over 20 % of the RFI selected features from PAAM; Table 4). In Use Case 2, we observed improvement in the performance of SVM over the Top 5 % and Top 10 % of the feature-set obtained from PAAM over the features obtained from original OTU abundances (Table 5). However, in Use Case 2 particularly, modelling with LR and NB attained similar results with PAAM features and OTU abundance counts (Table 5). Only SVM along Phy-PMRFI (Table 5) attained better performance than SVM over raw OTU abundances (Table 2). Phy-PMRFI based models provided better performance in Use Case 3 also, in comparison to ML applied over original abundance counts (Table 6); establishing again the evidence that integration of phylogeny could deliver more robust performances across the different classifiers of SVM, LR and NB.

To summarize, RFI based modelling improved the Accuracy and interpretability of SVMs over all the Use Cases. RFI yields variable importance measures for each candidate predictor by using ensembles of trees and reduces the computational complexity of subsequently applied supervised classifiers in learning, and more importantly helps in making a better prediction (as indicated in results above). Also, the results over the Use Case 3 (Table 6) indicate that they are not overly sensitive to the numbers of features selected by RFI approach. It appears that explicitly modeling over the phylogeny between OTUs allows the RFI method to exploit the relationships between different microbial taxas and, hence also capturing the biological diversity. Furthermore, feature selection could be used to retrain the ML classifiers (SVM, LR, and NB) to make them more robust (Table 6).

TABLE 4
PERFORMANCE OF ML MODELS WITH LOOCV OVER USE CASE 1 (TOP N =5, 10, 20, 40, 60 % FEATURES) RELATED TO PAAM CONSTRUCTED FOR HUMAN THROAT MICROBIOME

Feature Selection	ML Model	Phy-PMRFI		Raw OTU Abundances	
RFI ranked Top N % Features where N=5/10/20/40/60	LOOCV	Accuracy	*K*appa	Accuracy	*K*appa
Top 5 %	SVM	0.867	0.730	0.750	0.497
Top 10 %	SVM	0.900	0.797	0.867	0.730
Top 20 %	SVM	**0.900**	**0.798**	0.883	0.765
Top 40 %	SVM	0.817	0.631	0.750	0.494
Top 60 %	SVM	0.767	0.529	0.716	0.716
Top 5 %	LR	0.867	0.729	0.750	0.496
Top 10 %	LR	0.900	0.797	0.767	0.529
Top 20 %	LR	**0.900**	**0.798**	0.767	0.531
Top 40 %	LR	0.883	0.766	0.750	0.496
Top 60 %	LR	0.783	0.563	0.750	0.494
Top 5 %	NB	0.717	0.426	0.567	0.125
Top 10 %	NB	0.767	0.531	0.667	0.336
Top 20 %	NB	0.700	0.397	0.617	0.231
Top 40 %	NB	0.700	0.400	0.567	0.125
Top 60 %	NB	0.500	0.193	0.500	0.001

3) An Empirical Study of the impact of Feature Selection Strategies

We conducted a further empirical study using different ML models along with sets of features obtained using the three feature engineering methods of i) XgBoost ranked features [32], ii) relief-measure based raked features [33] and iii) glmnet [34].In the scope of current paper, we evaluated the models over the feature-subset obtained over the Use Case 1. We used the variable importance evaluation functions (i.e. varImp()) available in caret package in R [38], and the model information function of XgBoost [32,39] with default settings and a cut-off threshold provided by the top 20 features. The results of ML applied over the feature subset selected from XgBoost are recorded in Table 7. The ensemble of XgBoost as feature selector and SVM as ML model over PAAM, provided good performance in this Use Case

(Table 7). In the next experiment, rank importance and weights of the predictors were obtained using the RReliefF strategy [33].

TABLE 5
PERFORMANCE OF ML MODELS WITH LOOCV OVER USE CASE 2 (TOP N =5, 10, 20, 40, 60 % FEATURES) RELATED TO PAAM CONSTRUCTED FOR SOIL MICROBIOME

Feature Selection	ML Model	Phy-PMRFI		Raw OTU Abundances	
RFI ranked Top N % Features where N=5/10/20/40/60	LOOCV	Accuracy	*K*appa	Accuracy	*K*appa
Top 5 %	SVM	**0.992**	**0.989**	0.985	0.978
Top 10 %	SVM	0.992	0.989	0.985	0.978
Top 20 %	SVM	**0.992**	**0.989**	0.985	0.978
Top 40 %	SVM	0.985	0.978	0.985	0.978
Top 60 %	SVM	0.985	0.978	0.985	0.978
Top 5 %	LR	1.000	1.000	1.000	1.000
Top 10 %	LR	1.000	1.000	1.000	1.000
Top 20 %	LR	1.000	1.000	1.000	1.000
Top 40 %	LR	1.000	1.000	1.000	1.000
Top 60 %	LR	1.000	1.000	1.000	1.000
Top 5 %	NB	0.791	0.687	0.791	0.687
Top 10 %	NB	0.791	0.687	0.726	0.590
Top 20 %	NB	0.683	0.525	0.661	0.493
Top 40 %	NB	0.647	0.572	0.647	0.572
Top 60 %	NB	0.647	0.572	0.647	0.572

TABLE 6
PERFORMANCE OF ML MODELS WITH 10-FOLDS CV OVER USE CASE 3 (TOP N =5, 10, 20, 40, 60 % FEATURES) RELATED TO PAAM CONSTRUCTED FOR HUMAN MICROBIOME

Feature Engineering with RFI	ML Model	Phy-PMRFI		RAW OTU ABUNDNACES	
Top N % where N=5/10/20/40/60	10-folds CV	Accuracy	*K*appa	Accuracy	*K*appa
Top 5 %	SVM	0.937	0.902	0.929	0.891
Top 10 %	SVM	0.941	0.901	0.933	0.898
Top 20 %	**SVM**	**0.943**	**0.910**	**0.934**	**0.899**
Top 40 %	SVM	0.940	0.901	0.932	0.897
Top 60 %	SVM	0.940	0.901	0.930	0.896
Top 5 %	**LR**	**0.946**	**0.918**	**0.945**	**0.910**
Top 10 %	LR	0.913	0.886	0.886	0.819
Top 20 %	LR	0.934	0.903	0.929	0.889
Top 40 %	LR	0.943	0.909	0.923	0.884
Top 60 %	LR	0.940	0.902	0.929	0.889
Top 5 %	NB	0.934	0.903	0.923	0.889
Top 10 %	NB	0.934	0.904	0.923	0.888
Top 20 %	NB	0.937	0.908	0.910	0.875
Top 40 %	**NB**	**0.945**	**0.922**	**0.943**	**0.904**
Top 60 %	NB	0.944	0.909	0.940	0.902

The FSelector package in R, was used to obtain the relief measure-based ranking with settings defaulting to *k* nearest neighbours (i.e. *neighbours.count* = 5 and *sample.size* = 10. The results are summarized in Table 8. The results indicate the features (i.e. the internal nodes and leaf nodes of taxonomical tree) selected from PAAM play an important role in determining the functional roles and attain higher predictive performance (e.g. SVM applied on relief-based feature set attained from the PAAM constructed in Use Case 1 achieved highest Accuracy of 0.850 and *K*appa of 0.697) (Table 8). glmnet applies a regularization method that penalizes the linear or logistic models with a proportion to the weights of the coefficients in regression modelling [34]. This results in reducing the coefficients of certain unwanted features to zero and removing the unwanted variables. It was implemented using the glmnet function in the caret R package [32,34]. The cross-validated Accuracy and *K*appa of the ML models trained with glmnet over PAAM was noted to have obtained a higher predictive performance than the model obtained using only OTU abundance with different tuning parameters of alpha and lambda [32,34] (Table 9).

The performance Accuracy obtained from all three types of feature selection as listed above was in range of 0.700–0.850 (Tables 7, 8 and 9) and the Phy-PMRFI feature permutation-based approach yielded a maximum accuracy score of 0.900 with only 20% of the phylogeny integrated features (Table 4). The results presented in this sub-section provide a supporting evidence for selecting RFI-based feature selection in our proposed framework. From the right of Tables 4 and 8, we observed that ML does not need many features in the selected subspace to achieve the best prediction performance.

TABLE 7
PERFORMANCE OF ML MODELS WITH LOOCV OVER THE XGBOOST OBTAINED FEATURE SUB-SET IN USE CASE 1

ML Model	Non-Phylogenetic (over raw abundances)		Phylogenetic (over PAAM)	
(LOOCV)	Accuracy	*K*appa	Accuracy	*K*appa
RF	0.716	0.429	**0.783**	**0.563**
SVM	0.783	0.559	**0.833**	**0.665**
LR	0.800	0.596	0.800	0.596
NB	0.667	0.334	0.650	0.289

TABLE 8
PERFORMANCE OF ML MODELS WITH LOOCV OVER THE RRELIEFF OBTAINED FEATURE SET IN USE CASE 1

Feature Engineering with RReliefF	ML Model	Phy-PMRFI		RAW OTU ABUNDNACES	
Top N % where N=10/20/40/60	LOOCV	Accuracy	*K*appa	Accuracy	*K*appa
Top 10 %	SVM	0.800	0.596	0.766	0.522
Top 10 %	RF	0.717	0.424	0.700	0.394
Top 10 %	LR	0.800	0.600	0.800	0.598
Top 10 %	NB	0.650	0.310	0.633	0.279
Top 20 %	**SVM**	**0.850**	**0.697**	**0.783**	**0.565**
Top 20 %	RF	0.700	0.386	0.700	0.386
Top 20 %	LR	0.816	0.629	0.816	0.634
Top 20 %	NB	0.633	0.273	0.600	0.181
Top 40 %	SVM	0.816	0.629	0.750	0.494
Top 40 %	RF	0.683	0.356	0.733	0.452
Top 40 %	LR	0.833	0.662	0.733	0.457
Top 40 %	NB	0.600	0.196	0.516	0.013
Top 60 %	SVM	0.766	0.524	0.766	0.522
Top 60 %	RF	0.717	0.426	0.700	0.375
Top 60 %	LR	0.816	0.632	0.783	0.565
Top 60 %	NB	0.566	0.129	0.466	-0.018

The RFI method has potential to provide highly informative features for classifying human throat microbiome. Experimental results have demonstrated the improvements in increasing of predictive performance for metagenomics classification for the 16S rRNA dataset, with RFI in comparison with other feature selection models, including modelling with XgBoost [32], RreliefF [33], and glmnet [34]. Also, the improvement in performance over the Use Case 1 suggests that phylogeny can provide useful information in the prediction of metagenomic functions, along with Phy-PMRFI. The RFI method has potential to provide highly informative features for classifying human throat microbiome. Experimental results have

demonstrated the improvements in increasing of predictive performance for metagenomics classification for the 16S rRNA dataset, with RFI in comparison with other feature selection models, including modelling with XgBoost [32], RreliefF [33], and glmnet [34]. Also, the improvement in performance over the Use Case 1 suggests that phylogeny can provide useful information in the prediction of metagenomic functions, along with Phy-PMRFI.

TABLE 9
PERFORMANCE OF EMBEDDED ML MODEL OF GLMNET WITH LOOCV IN USE CASE 1

ML Model (glmnet function)		Non-Phylogenetic (over raw abundances)		Phylogenetic (over PAAM)		
alpha	lambda	Accuracy	*K*appa	lambda	Accuracy	*K*appa
0.10	0.012	0.633	0.269	0.014	0.650	0.295
0.10	0.039	0.633	0.269	0.045	0.650	0.295
0.10	0.125	**0.700**	**0.402**	0.141	0.633	0.260
0.55	0.012	0.683	0.353	0.014	0.667	0.327
0.55	0.039	0.650	0.282	0.044	0.650	0.288
0.55	0.125	0.616	0.210	0.141	**0.716**	**0.421**
1.00	0.012	0.600	0.178	0.014	0.700	0.386
1.00	0.039	0.633	0.246	0.044	**0.716**	**0.421**
1.00	0.125	0.600	0.174	0.141	0.667	0.318

4) Comparison with other Phylogeny-Aware Supervised Learning Methods

Unlike many popular classification methods [14-17], which consider features (OTUs) as independent, phylogeny-aware methods [20-28] take advantage of the similarities between OTUs derived from the phylogenetic tree. We further evaluated the proposed approach by following a systematic comparison with other phylogeny-aware models of: - i) PhILR [23] and ii) MetaPhyl [24] from the state-of-the-art, over all the Use Cases. The results are indicated in Table 10. Our proposed framework Phy-PMRFI provided significantly better predictive performance than the other phylogenetic methods of PhILR ($p < 0.01$) and MetaPhyl as depicted in Table 10, except in Use Case 3 for which MetaPhyl provided best performance.

TABLE 10
COMPARING PHYLOGENY-AWARE MODELS OF PHILR, PHY-PMRFI, AND METAPHYL OVER USE CASE 1, USE CASE 2 AND USE CASE3

	Use case 1		Use Case 2		Use Case 3	
	Accuracy	*K*appa	*Accuracy*	*K*appa	Accuracy	*K*appa
PhILR with						
SVM	0.533	0.000	0.611	0.415	0.874	0.837
LR	0.433	0.164	0.618	0.427	0.882	0.847
NB	0.450	0.087	0.503	0.254	0.818	0.762
Phy-PMRFI with						
SVM	**0.900**	**0.797**	**0.992**	**0.989**	0.940	0.907
LR	**0.900**	**0.797**	**1.000**	**1.000**	0.943	0.909
NB	0.767	0.531	0.791	0.687	**0.945**	**0.922**
MetaPhyl						
	0.717	0.472	0.755	0.666	**0.984**	**0.976**

MetaPhyl [24] method takes advantage of the similarities between only leaf level OTUs, as encoded by the phylogenetic tree. On the other hand, our method progressed by considering internal nodes as well in the metagenomics downstream analysis, rather than only leaf-level OTUs. PhILR [23] also demonstrated one-to-one correspondence between the features and the internal nodes on the phylogenetic tree, incorporating evolutionary distance into the PhILR transform of feature space [23]. However, Phy-PMRFI proved to be a powerful classification framework attaining high predictive performance and considering several lineages of varying phylogenetic depth in the microbial analysis. Table 10 shows the performance of Phy-PMRFI that attained the highest Accuracy [36] and *K*appa [37] values with different numbers of engineered features.

5) Biological Relevance of the Features Selected

It was observed that more than half of the important features as ranked by RFI were not original OTUs, indicating the importance of hierarchical combinations formed by combining OTUs based on phylogeny (i.e. the internal nodes) in metagenomic applications. The percentages of internal nodes proved to be dominant in Use Cases 1 and 3, when considering the top 5, 10, 20, 40 and 60 % of the feature-set. (Table 11). Some important groups of microbial species that played an important role in classification were identified along with the proposed framework. In Use Case 3, the human microbiome (HMP) body sites, *Phylum: Firmicutes with Genus: Lactobacillus and Weissella*; were noted as top ranked features to differentiate different body site niches. This observation is supported by the study in [39], which highlighted the potential of *Weissella* in the human microbiome. In Use Case 2, which relates to the soil microbiome, *Proteobacteria* with genus *Pseudomonas* served as top ranked features. The other predominant *Phylum* of *Actinobacteria* served as an important role in classifying metagenomes into sugar treated substrates. In Use Case 1, taxonomical mapping is not available at the data source (*https://cran.rproject.org/web/packages/MiSPU/);* hence we noted internal nodes as combination of important OTUs (i.e. their IDs) for any reference (APPENDIX A).

TABLE 11
PERCENTAGES OF INTERNAL NODES IN THE FEATURES MODELLED IN THE PHY-PMRFI FRAMEWORK

Feature Set	Top 5%	Top 10 %	Top 20 %	Top 40 %	Top 60 %
Use Case 1	**82%**	**81 %**	80%	72%	70%
Use Case 2	58%	58%	58%	60%	58%
Use Case 3	**90%**	**85%**	80%	76%	67%

V. SUMMARY

In this paper, we presented a novel phylogeny-aware modelling framework, Phy-PMRFI, for predicting functions of taxas in microbiome sequencing data sets. The framework integrates abundance counts of OTUs and the relationships between various OTUs at different taxonomic levels for metagenomics downstream analysis. Phylogenetically close microbial taxas tend to have similar effects on the functional phenotypes in metagenomic studies [7,24,28]. Hence, inclusion of the phylogenetic measure potentially maximizes the opportunity of classifying microbiome functions according to naturally inherent properties of taxas.

In this work, we implemented a novel 2D matrix PAAM which combines level-by-level evolutionary information obtained from a phylogenetic tree with OTU abundance counts. However, PAAM has almost twice as many features than the OTU abundance count table and hence the data has higher dimensionality. Therefore, applying feature selection approaches across the columns of PAAM we obtained improved ML modelling through determining predictive functions.

Along with our proposed framework Phy-PMRFI, the ML approach RF was implemented in order to selected informative features from the PAAM, as suggested by [7,24,28]. The output of RFI as a rank ordering of important predictors was worthy of further investigation. The different classifiers of SVM, LR and NB were applied over the engineered features. RFI proved attractive for providing insight regarding the discriminative ability of individual predictor variables. The proposed methodology provided better or competitive Accuracy in comparison to state-of-the-art ML models applied over the raw abundances by selecting important features. For example, the classifier of SVM on a feature's subset obtained by RFI over PAAM performed

better than RFI applied over the raw OTU abundances. Also, for the Use Case 1, modelling over RFI selected features attained best results when benchmarked with other feature selection strategies. The approach overall indicates that phylogeny could play an important role in differentiating samples obtained from metagenomic environments into functional phenotypes. We found feature-sets derived from columns of PAAM are better predictors in characterizing metagenomic functions. This indicates that a subset of intermediate nodes of the phylogenetic tree, rather than defining features as OTUs at the tree leaves could lead to better classification of microbiome. The method provided significantly better performance Accuracy in Use Case 1 and Use Case 3; competitive Accuracy in Use Case 2; when compared with state-of-the-art methods over the raw OTU abundances. However, it outperformed PhILR [23] in all Use Cases. Also, the approach improved over the MetaPhyl [24] method in Use Case 1 and 2. Our proposed approach facilitates the extraction of a ranked microbial taxonomic set for the interpretation of the learned predictive model to determine functional roles in metagenomic classification. Thus, we hope that the characterized framework in this study would inform future microbiome studies.

As microbial taxas substantially outnumber the number of microbial samples, our study could be extended to benchmark with other regularization algorithms that promote the learning of important features. Another potential future application of this work would be to analyse the minimum redundancy maximum relevance of the feature sets obtained from engineering.

As metagenomics has accelerated the understanding of microbiome, we would like to extend on our analysis on phylogenetic advancements with advances in ML such as deep forest approach of gcForest [40]. There are other possibilities to work towards the development of a novel classification method for 16S rRNA sequence taking advantage of the natural structuring of microbiome as encoded by a phylogenetic tree in the ML classifier itself; in comparison to current work of incorporating phylogeny at pre-processing level.

ACKNOWLEDGEMNET

This work is supported by the grant from the MetaPlat project, (www.metaplat.eu), under H2020-MSCA-RISE-2015-19 and additionally by the Research Challenge Strategy Fund of Ulster University, U.K.

REFERENCES

1. C. S. Riesenfeld, P. D. Schloss, and J. Handelsman, “Metagenomics: genomic analysis of microbial communities,” *Annu. Rev. Genet*., vol. 38, pp. 525–552, 2004.
2. D. R. Maddison, K.-S. Schulz, and W. P. Maddison, “The tree of life web project,” *Zootaxa*, vol. 1668, no. 1, pp. 19–40, 2007.
3. S. Whelan, P. Lio, and N. Goldman, “Molecular phylogenetics: state-of-the-art methods for looking into the past,” *TRENDS in Genetics*, vol. 17, no. 5, pp. 262–272, 2001.
4. L. Brocchieri, “Phenotypic and Evolutionary Distances in Phylogenetic Tree Reconstruction,” *Journal of Phylogenetics & Evolutionary Biology*, vol. 01, no. 04, 2013.
5. D. McDonald, A. Birmingham, and R. Knight, “Context and the human microbiome.,” *Microbiome*, vol. 3, no. 1, p. 52, 2015.
6. S. M. Scheiner, E. Kosman, S. J. Presley, and M. R. Willig, “The components of biodiversity, with a particular focus on phylogenetic information,” *Ecology and Evolution*, vol. 7, pp. 6444–6454, 2017.
7. J. B. H. Martiny, S. E. Jones, J. T. Lennon, and A. C. Martiny, “Microbiomes in light of traits: A phylogenetic perspective,” *Science*, vol. 350, no. 6261, 2015.
8. L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
9. Y. Qi, “Random Forest for Bioinformatics,” *Ensemble Machine Learning*, pp. 307–323, 2012.
10. Joachims, Thorsten. "Training linear SVMs in linear time." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 217-226. ACM, 2006.
11. J.Hosmer, D.W., S.Lemeshow, R.Sturdivant, "Applied logistic regression",vol. 398,. John Wiley & Sons,2013.
12. Murphy, Kevin P. "Naive bayes classifiers." University of British Columbia 18 ,2006.
13. H. Soueidan and M. Nikolski, "Machine learning for metagenomics: methods and tools*", Metagenomics, vol. 1, no. 1, 2017.*
14. D. R Mingle, "Machine Learning Techniques on Microbiome -Based Diagnostics", *Advances in Biotechnology & Microbiology*, vol. 6, no. 4, 2017. Available: 10.19080/aibm.2017.06.555695.
15. E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, “Machine learning meta-analysis of large metagenomic datasets: tools and biological insights,” *PLoS computational biology*, vol. 12, 2016.
16. D. Knights, E. Costello and R. Knight, "Supervised classification of human microbiota", *FEMS Microbiology Reviews*, vol. 35, no. 2, pp. 343-359, 2011.
17. A. Statnikov et al., “A comprehensive evaluation of multicategory classification methods for microbiomic data.,” *Microbiome*, 2013.
18. The NIH HMP Working Group, “The NIH Human Microbiome Project,” Genome Res., vol. 19, no. 12, pp. 2317–2323, 2009.
19. W. Chen, C. Zhang, Y. Cheng, S. Zhang and H. Zhao, "A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs", *PLoS ONE*, vol. 8, no. 8, p. e70837, 2013.
20. J. Ning and R. G. Beiko, “Phylogenetic approaches to microbial community classification.,” *Microbiome*, vol. 3, no. 1, p. 47, 2015.
21. C. Lozupone, M. Lladser, D. Knights, J. Stombaugh and R. Knight, "UniFrac: an effective distance metric for microbial community comparison", *The ISME Journal*, vol. 5, no. 2, pp. 169-172, 2010
22. D. Albanese, C. De Filippo, D. Cavalieri, and C. Donati, “Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting,” *PLoS Comput. Biol*., vol. 11, pp. 1–18, 2015.
23. J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, “A phylogenetic transform enhances analysis of compositional microbiota data,” *Elife*, vol. 6, pp. 1–20, 2017.
24. O.Tanaseichuk, J. Borneman and T. Jiang, "Phylogeny-based classification of microbial communities", *Bioinformatics*, vol. 30, no. 4, pp. 449-456, 2013
25. C. Wu, J. Chen, J. Kim, and W. Pan, “An adaptive association test for microbiome data,” *Genome Med.,* vol. 8, no. 1, pp. 1–12, 2016.
26. D. Reiman, A. A. Metwally, and Y. Dai, “PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data,” bioRxiv, p. 257931, 2018.
27. H. Li, “Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis,”*Annual Rev. Stat. Its Appl.*, vol. 2, 2015.
28. J.T. Wassan, et al. "PAAM-ML: A novel Phylogeny and Abundance aware Machine Learning Modelling Approach for Microbiome Classification." IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018.
29. E. Charlson, J. Chen, R. Custers-Allen, K. Bittinger, H. Li, R. Sinha, J. Hwang, F. Bushman and R. Collman, "Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers", *PLoS ONE*,2010.
30. N. D. Youngblut, S. E. Barnett, and D. H. Buckley, “SIPSim: A Modeling Toolkit to Predict Accuracy and Aid Design of DNA-SIP Experiments,” *Frontiers in Microbiology*, vol. 9, 2018.
31. K.Archer and R. Kimes, "Empirical characterization of random forest variable importance measures*", Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249-2260, 2008.
32. T. Chen and C. Guestrin, “XGBoost : A Scalable Tree Boosting System,” pp. 785–794, 2016.
33. I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF", European conference on machine learning. Springer, Berlin, Heidelberg, 1994.
34. J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, vol. 33, no. 1, 2010.
35. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", *Pattern Recognition*, vol. 48, no. 9, pp. 2839-2846, 2015.
36. P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: an overview,” *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.
37. A. Bendavid, “Comparison of classification accuracy using Cohen’s Weighted Kappa,” *Expert Systems with Applications*, vol. 34, 2008.
38. M. Kuhn, “Building predictive models in r using the caret package,” *Journal of Statistical Software*, Articles, vol. 28, no. 5, pp. 1–26, 2008.
39. V. Fusco, G. M. Quero, et al., “The genus Weissella: taxonomy, ecology and biotechnological potential,” *Frontiers in Microbiology*, vol. 6, 2015.

40. Z. Zhou and J. Feng, "Deep Forest: Towards an Alternative to Deep Neural Networks", arXiv.org, 2019. [Online].