Developing a new Phylogeny-driven Random Forest Model for Functional Metagenomics

Developing a new Phylogeny-driven Random Forest Model for Functional Metagenomics

Jyotsna Talreja Wassan[1], Haiying Wang [2], and Huiru Jane Zheng [2]*, Sr. *Member, IEEE*
1 University of Delhi, India
2 University of Ulster, United Kingdom

***Abstract*—Metagenomics is an unobtrusive science linking microbial genes to biological functions or environmental states. Classifying microbial genes into their functional repertoire is an important task in the downstream analysis of Metagenomic studies. The task involves Machine Learning (ML) based supervised methods to achieve good classification performance. Random Forest (RF) has been applied rigorously to microbial gene abundance profiles, mapping them to functional phenotypes. The current research targets tuning RF by the evolutionary ancestry of microbial phylogeny, developing a Phylogeny-RF model for functional classification of metagenomes. This method facilitates capturing the effects of phylogenetic relatedness in an ML classifier itself rather than just applying a supervised classifier over the raw abundances. The idea is rooted in the fact that closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits. Such microbes behave similarly; and hence tend to be selected together, or one of these could be dropped from the analysis, to improve the ML process. The proposed Phylogeny-RF algorithm has been compared with state-of-the-art classification methods including RF and the phylogeny-aware method of MetaPhyl, using two real-world 16S rRNA metagenomic datasets. The proposed method was observed to perform better than the other phylogeny-driven benchmarks. For example, Phylogeny-RF attained a high AUC of 0.949 over soil microbiomes in comparison to other benchmarks.**

Index Terms— Metagenomics, Phylogeny, Random Forest, Operational Taxonomic Units (OTUs), Classification, Clustering

I. Introduction

THE area of functional metagenomics [1] has been expanding, seeking to predict functional roles associated with microbial genes i.e. Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs) of high-dimensional, heterogeneous, and complex 16S rRNA genes present in metagenomic datasets[2]. Mapping of microbial samples consisting of OTUs/ASVs as microbial features to functional phenotypes (categorical class); is the most commonly used phenomenon (known as functional metagenomics) and forms a classical problem of supervised classification in Machine Learning (ML)[3]. The progress of using Random Forest classifier for classifying genotypes into phenotypes has been observed [4]–[11]. The application of RF in these studies considered OTUs as independent features, but naturally ,OTUs are related by their evolutionary and taxonomic hierarchal relationships, also known as phylogeny[12],[13]. The construction of quantitative profiles of OTUs in 16S metagenomics pipelines such as QIIME[14] ignores the fine structure present in their phylogeny. A current challenge in functional metagenomics is to model the phylogenetic effects with high-dimensional abundance profiles of microbial genes in ML models. Very limited research in this area has been conducted[15]–[19]. Some regularization methods have been recommended as classifiers in literature [20],**Error! Reference source not found.**[21] for functional metagenomics to jointly considering the phylogenetic effect. However, research in the literature lacks in exploring and modeling phylogeny in decision tree-based models.

The possible ways in which phylogeny could be integrated into functional microbiome analysis are - i) integrating phylogeny in modeling feature (data) space[17],[22],[23], and ii) integrating phylogeny in the ML models itself [15],[20],[24]. In this work, for the first time, it has been demonstrated how the phylogenetic information could be used to guide and model the RF approach (and is termed as Phylogeny-RF), accounting for functional classification of metagenomes based on the biological domain knowledge. Importantly, the proposed approach addresses the modelling of phylogeny aware OTU features or taxa within the RF model, which could prove useful in its application to metagenomic datasets. This sets up the stage to identify biologically relevant OTUs that enable this good prediction via RF and; contain more diverse biological information.

The traditional RF model of classification progresses by using bagging strategy over multiple decision trees[25]. Each decision tree in RF chooses m number of predictors to make decisions; and $m = \sqrt{Number\ of\ \ OTU\ features}$ [25],[26].

Phylogeny-RF particularly regularizes m to be guided by a phylogenetic measure assuming that phylogenetic similar OTU features tend to share the same functional responses[27]. The phylogenetic refinement learns the decision nodes of all trees under a global objective function so that microbiome information within multiple trees of RF is biologically diverse. The proposed new approach progresses by further clustering (grouping) original OTUs/ASVs based on their phylogenetic similarity into m groups and, thereafter, choosing an OTU/ASV element randomly from each group as a potential

feature to be chosen as a decision node in a constituting decision tree of RF model. This approach is inspired from the principle that closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits [13], [27]–[29].

The performance of an RF classifier is highly dependent on the accuracy of each component decision tree. In RF, randomization could cause the occurrence of redundant and correlated trees as this could include phylogenetically correlated features. This may lead to an inefficient ensemble classification decision. Better RF modelling could be achieved through the random selection of uncorrelated and diverse microbial features. The purpose of this research is to model RF over the phylogenetically diverse features to improve the quality of functional phenotypic predictions. As when two OTU features have high phylogenetic similarity, they will behave similarly; and henceforth, one of the two features could be chosen to minimize the redundancy and make learning over the more biologically relevant signatures in RF modelling. The key idea behind the proposed classification approach is to model RF over the OTU features with minimum redundancy and maximal biological relevance; to enhance its application in metagenomic studies.

The proposed Phylogeny-RF over a different number of trees (ranging from 1-300) showed a significant improvement (p-value < 0.05) with regards to predictive metrics (AUC and Kappa values) over the two datasets under study, in comparison to the traditional RF method proposed by Breiman[25].The comprehensive analysis results also showed less variation in AUC using Phylogeny-RF than using traditional RF over different runs of 5-fold cross-validation on soil and throat microbiome. The current research primarily focused to optimize over the high-dimensional feature set fed to RF in functional metagenomic studies by the selection of only biologically diverse and uncorrelated features to be modelled in constituting decision trees with attaining good classification performance.

The rest of the paper is organized as follows. The data sources used in the study are presented in Section II. Section III describes the methodology and implementation involved in the proposed approach. Experimental results are presented in Section IV followed by discussions in Section V. Section VI provides a conclusion and future research directions.

II. Material

- High throughput sequence (HTS) data from DNA- or RNA-based stable isotope probing (SIP) experiments[30]:- The data were sampled from SIP experiments by incubating microorganisms in aliquots of soil receiving both cellulose and glucose treatments. It contains 139 samples with 1072 OTUs. The three classes associated with this data are 13C-Cel (Cellulose treated) (46), 13C-Glu (Glucose treated) (47) and 12C-Con (Control) (46). The source is available at https://cran.rproject.org/web/packages/HTSSIP/index.html.
- Human Throat Microbiome (HTM) [31]:- Charlson et al. [31]investigated the effect of cigarette smoking on the microbial communities present in the human respiratory tract. The dataset is comprised of 28 cigarette smokers and 32 non-smokers individuals. It has 60 samples and 856 OTUs. It is available as part of the MiSPU package in R [32] at https://cran.r project.org/web/packages/MiSPU/).
- Costello et al. Body Habitats (CBH) [33]:- The microbial communities present in different human body sites are responsible for determining healthy or disease states. This data set included a microbial OTU table with 622 samples and 2683 OTUs present in six major body niches: External Auditory Canal (EAC) (44), Gut (45), Hair (14), Nostril (46), Oral cavity (54), and Skin (419) showcasing their variability across the human body. However, this example data set is a case of an imbalanced number of classes. This dataset could be availed from https://www.knightslab.org/data.

The summary of the datasets is shown in Table 1.

TABLE 1
SUMMARY OF DATASETS USED IN THE CURRENT RESEARCH FOR FUNCTIONAL ANALYSES.

Data Source	Samples	OTUs	Phenotype
HTS	139	1072(absolute abundances)	Substrates treatment of glucose (47) or cellulose (46) or controls (46)
HTM	60	856 (absolute abundances)	Smoking (28) vs. Non-smoking individuals (32)
CBH	622	2683 (relative abundances)	Body habitats of External Auditory Canal (EAC) (44), Gut (45), Hair (14), Nostril (46), Oral cavity (54), and Skin (419)

III. Methodology

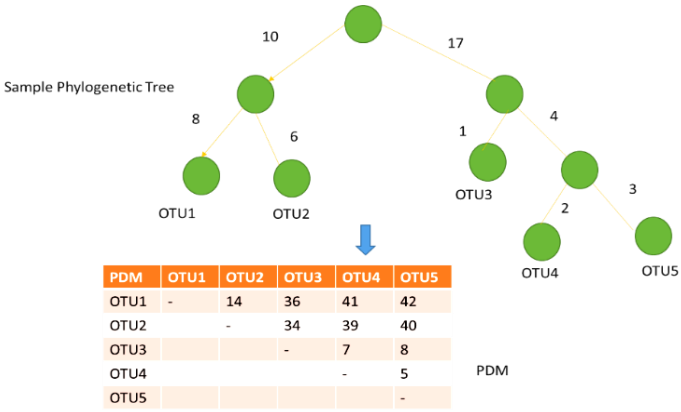
This section indicates a step-wise procedure for defining the novel application of tuning RF model with phylogeny for functional metagenomic predictions. The inclusion of phylogeny would help in classifying microbes into phenotypic functions, according to the natural structure and properties of a microbial community.

A. Inputs

A source phylogeny tree that is in the parenthetical Newick format [34] was read and further processed using the read.tree() function in the ape R package[35]. An abundance count table of OTUs/ASVs and meta-data are other inputs to the pipeline. Metagenomic pipelines such as QIIME[14], aids in generating these inputs from the raw metagenomic sequences.

B. Generating Phylogenetic Distance Matrix from Phylogeny

A matrix depicting the phylogenetic distances between each pair of leaf nodes of a phylogenetic tree is obtained by using the cophenetic () function in the *ape* package in R[35] . It is referred to as a phylogenetic distance matrix (PDM), and is serving as an important background information for incorporating phylogenetic diversity in the current approach [36]. The values in the cells of PDM are the sum of the branch lengths separating each pair of OTUs. Thus, if two OTUs are close relatives, their intersection cell contains a smaller value in comparison to OTUs that are far apart on the phylogeny tree. A toy example of converting phylogenetic tree to PDM is shown in Fig. 1.



PDM	OTU1	OTU2	OTU3	OTU4	OTU5
OTU1	-	14	36	41	42
OTU2		-	34	39	40
OTU3			-	7	8
OTU4				-	5
OTU5					-

Fig.1. A toy example illustration for tree to PDM

C. Clustering

Clustering determines interesting patterns in metagenomic data based on the similarity of the microbial genes [37]. In our approach, clustering has been applied to the PDM obtained in step (B), to group microbial OTUs/ASVs based on their shared phylogeny. Intuitively, the optimal choice of the number of clusters in the proposed approach is chosen equivalent to the number of features to be chosen by RF modelling to strike a balance of selecting a microbial feature (OTU) from each cluster. A well-known *k*-means [38], [39] clustering approach was chosen along the workflow of Phylogeny-RF as a precursor step to supervised learning. The parameter commonly referred as '*k*' in *k*-means specifies the number of clusters to be generated; however no such explicit choice of a number of clusters exists in other commonly used strategy of hierarchal clustering over distance-based matrices [40].

The points in dimension coordinates (along with the two principal directions) are derived from the similarity distance matrix (PDM) to be input to the k-means clustering, as the k-means algorithm primarily deals with calculating Euclidean distances between data points in a cartesian coordinate space for finding groups of similar data features. This also aided in reducing the dimensions while dealing with the high-dimensional metagenomes.

To obtain points-in-dimensions, coordinate data from the similarity distance matrix of PDM was obtained by calculating eigenvectors from an eigen decomposition following a path of- PDM (input) → Covariance Matrix → Eigen-Decomposition → 2D Coordinate System [41],[42],[43]. Centering of PDM would make it a covariance matrix as indicated in Algorithm 1.

Algorithm 1 Perform " Centering" of PDM

Input. Phylogenetic Distance Matrix (PDM) of size (Number of data features X Number of data features)

Method.
n = nrow (PDM) where nrow () calculates the number of rows of PDM
P = diag(n) – 1/n where diag () extracts the diagonal of the matrix or constructs a diagonal matrix
$PDM_c = -0.5 * P * PDM * P$

where * represents the multiplication operator

Output.
PDM_c

The eigen values and corresponding eigen vectors are calculated from centered PDM (i.e., PDM_c) (as indicated in Algorithm 2) and serves similar to fitting straight principal-component lines in a 2D coordinate system in accordance with the variance in data. Eigenvalues are the coefficients associated with eigenvectors and; provide the measure of the data's covariance. The two-component axes are determined by ranking eigenvectors in order of their eigenvalues (highest to lowest) (Algorithm 2). The reduced feature space is referred to as 2DCS-OTU.

k-means clustering is applied to reduced feature space (2DCS-OUT) as shown in Algorithm 3. The algorithmic procedure follows an iterative process (Algorithm 3). The function of *kmeans()* in the vegan R package, has been used to obtain clustering[44].

Obtained clusters from *k*-means are inputted to RF implementation in subsequent steps of the proposed method.

Algorithm 2 Construction of Cartesian Coordinates

Input. Centred Phylogenetic Distance Matrix (PDM_c)

Method.

Eigen = Eigen Values of PDM_c that are calculated by solving determinant ($PDM_c - \lambda I$) = 0; and finding the value of λ where 'I' is an identity matrix. λ serves as eigen values.

twoDEValue= Eigen [1:2] # a set of first two eigen values is taken

twoDEValue is a matrix of size 2X2.

Eigen Vectors = Two Eigen Vectors are calculated for corresponding twoDEValue by following ($PDM_c - \lambda I$) *x = 0; where x values are considered as corresponding eigen vectors of λ eigen values.

2DCS-OTU = Eigen Vectors * diag (twoDEValue) where diag () constructs a diagonal matrix; and here * represents matrix multiplication (Eigen Vectors is a matrix of size Number of data features X 2).

Output.

2DCS-OTU (i.e., a reduced two-dimensional coordinate system)

D. Supervised Learning with Phylogeny-RF

A traditional RF model constructs a constituting decision tree by selecting a random subset of the predictors from the entire feature set to determine the best split markers based on a Gini Index splitting criterion [45]. However, we propose an idea of random selection of subset of microbial features from phylogeny-based created clusters, also equivalent to $\sqrt{number\ of\ input\ features}$. Randomly choosing features from each cluster to model a decision tree intends to cover maximal phylogenetic diversity with minimum redundancy (Algorithm 4). The idea gains insights from the assumption that the similarity patterns are present between the microbial features based on their phylogeny and similar features tend to behave similarly to functions. This would reduce redundancy between microbial genes while constructing decision trees in RF. Hence, one of the OTUs/ASVs could be chosen randomly from its cluster, as the rest of the OTUs/ASVs in that cluster would behave and function similarly. This is repeatable for each cluster created. Thus, the proposed phylogeny-guided RF model (Algorithm 4) has the potential to cover maximal phylogenetic diversity and minimize redundancy between microbial genes.

Algorithm 3. k -Means Clustering applied over the obtained Coordinate System in Algorithm 2

Input. The number of clusters as *k* and *2DCS-OTU* is containing *n* OTU data points. Let X = $\{x_1, x_2, .. x_n\}$ denotes the set of data points.

Method. *k* points are chosen as initial cluster centroids (c_1, ..., c_k);

The algorithmic procedure follows an iterative process, as indicated below.

Repeat

1. Each data point x is assigned to its nearest centroid, based on an objective of minimizing distance(c_i, x)^2, where distance (·) is the standard Euclidean distance between cluster centroid (c_i) and the data point (x) and $i \in 1,2, \ldots . k$
2. Recompute/Update the cluster centroid, i.e., calculating the new centroid value by calculating the mean of the data points for each cluster created in step 1;

i.e. $c_i = \frac{1}{N_i} \sum_{j=1 to N_i} x_j$, where N_i represents the number of data points in an ith cluster.

Until no change (arbitrarily for *n* times or till no point changes its cluster and convergence is attained)

Output. A set of *k* clusters that meets the convergence criteria.

E. Performance Evaluation

The performance has been evaluated in this chapter using the average performance of popular assessment techniques of Accuracy [7], Kappa[46], and AUC-ROC[47] over different phenotypic classes in the datasets mentioned in Table 1. Kappa and AUC-ROC serve as a logical extension to the interpretation of classification performance by Accuracy, especially in the case of imbalanced classes[48].

The framework based on the above steps (A-E) is summarized in Fig.2. Phylogeny-RF is implemented as a modification of a Python script to code RF functionality[49]. The settings of RF parameters with maximum depth = 6, number of folds = 5 (chosen 5 folds to uniformly evaluate all data sources with cross-validation settings), and the number of trees is set to 1, 20, 64, 100, 128, 164, 200, 225, and 300 respectively for experiments in the study.

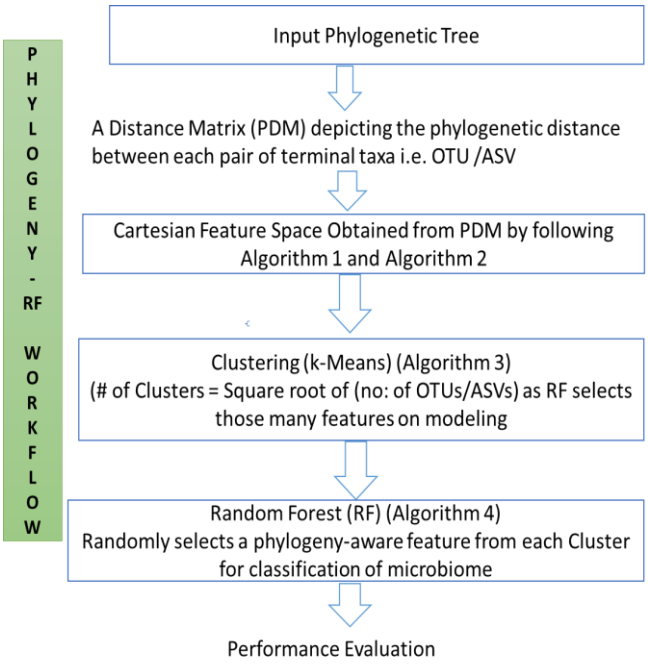


Fig.2. Proposed Phylogeny-RF Approach considering Microbial gene Similarity for 16S rRNA metagenomic classification

Based on the predictive values, the significance of the differences between the multiple ML methods (in a benchmark analysis) has been performed using Analysis of Variance (ANOVA)[50].

Algorithm 4. Phylogeny-RF

Input. The input set consists of *N* metagenomic samples; the dimension of the feature space is equal to the total number of OTU features (***F***)

Method.

1. OTU features (*F*) obtained from the input are clustered into $m = \sqrt{F}$ a number of groups using the *k*-Means algorithm (shown in Algorithm 3) [38] based on their phylogenetic similarity (PDM).
2. A random sub-sample is generated from *N* that serves as the training set (particularly in this research *kf*-folds with *kf* = 5 is implemented keeping *4*-folds over *N* for training, and *1*-fold of data serves as the test set.
3. Amongst the m groups of clustered OTU features (obtained in step 1), the algorithm randomly chooses a predicting node from each clustered group to construct a decision tree. Hence, the number of predicting nodes in a decision tree would be equal to the number of clusters. It serves an alternative way to choose $\sqrt{F}$ predicting nodes in a decision tree based on the phylogenetic clusters. Chosen predictors intend to cover maximal phylogenetic diversity in this case of RF. A node is chosen from the predictors as a split point in a decision using the traditional split criteria of the Gini Impurity criterion [25].
4. Build forest by repeating steps 2 to 3 for "*n*" number times to create "*n*" number of trees.; with the application of bagging strategy [51].

The built forest is used to predict output class for the test set by calculating the votes for each predicted class by each constituting tree. The highly voted class is considered as the final prediction obtained from the RF algorithm

Output. Phylogeny-based Phenotypic class predictions for test set samples

IV. EXPERIMENTS AND RESULTS

Experimental results demonstrate a promising performance with the preliminary investigation of Phylogeny-RF into the prediction of the microbiome associated with different functions in human and soil samples (over three data sources as indicated in Table 1). We have estimated the performance of the classification algorithms using the three trial runs of the 5-fold cross-validation procedure.

A. Performance Comparison of Phylogeny-RF and traditional RF

Tables 2-4 record predictive Accuracy, Kappa and AUC-ROC results of the performance across Phylogeny-RF and classical RF over three microbiome data sources of HTS[52], HTM and CBH[52], respectively. Results indicate that the phylogeny distance-driven RF model can successfully be applied to the problem area of functional metagenomics and has the potential to outperform state-of-the-art RF.

The best performance with AUC-ROC of 0.949 was achieved by Phylogeny-RF (at Number of trees = 164) applied over HTS dataset to classify microbial genes into the phenotypes of substrates treatment of glucose (47) or cellulose (46) or controls (46).

TABLE 2 SUMMARY OF AVERAGE PERFORMANCES OBTAINED BY APPLICATION OF PHYLOGENY-RF AND RF WITH 5-FOLDS (3 TRIAL RUNS) CROSS-VALIDATION DEALING WITH THE DIFFERENT NUMBER OF TREES IN RF OVER SOIL MICROBIOME (HTS) (BEST VALUES ATTAINED IN EACH COLUMN ARE HIGHLIGHTED IN BOLD IN EACH CATEGORY) [52].

	Phylogeny-RF			**RF**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.768	0.666	0.831	0.600	0.395	0.691
10	0.787	0.684	0.849	0.674	0.505	0.765
20	0.871	0.812	0.904	0.731	0.595	0.806
40	0.895	0.841	0.924	0.750	0.626	0.816
64	0.906	0.861	0.933	0.788	0.68	0.848
100	0.916	0.872	0.938	0.812	0.718	0.861
128	0.919	0.874	0.944	0.815	0.722	0.868
164	**0.929**	**0.891**	**0.949**	0.807	0.709	0.864
200	0.901	0.851	0.943	0.820	0.725	0.870
225	0.921	0.88	0.946	**0.827**	**0.74**	**0.875**
300	0.910	0.883	0.949	0.815	0.724	0.868
500	0.906	0.868	0.935	0.810	0.713	0.867

TABLE 3. SUMMARY OF AVERAGE PERFORMANCES OBTAINED BY APPLICATION OF PHYLOGENY-RF AND RF WITH 5-FOLDS (3 TRIAL RUNS) CROSS-VALIDATION DEALING WITH THE DIFFERENT NUMBER OF TREES IN RF OVER HUMAN THROAT MICROBIOME (HTM) (BEST VALUES ATTAINED IN EACH COLUMN ARE HIGHLIGHTED IN BOLD IN EACH CATEGORY).

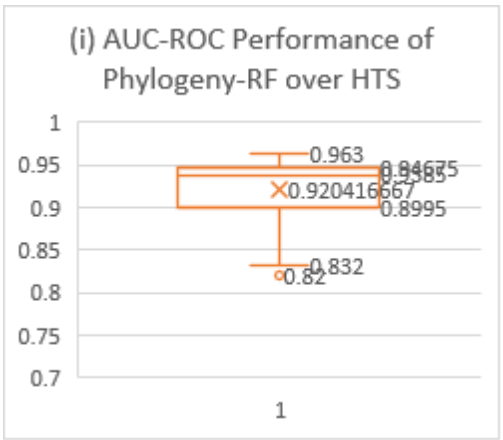
	Phylogeny-RF			**RF**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.559	0.140	0.570	0.498	0.080	0.499
10	0.594	0.239	0.635	0.545	0.119	0.555
20	0.628	0.364	0.647	0.589	0.176	0.591
40	0.689	0.381	0.696	0.583	0.223	0.632
64	**0.700**	**0.394**	**0.712**	0.616	0.252	0.637
100	0.616	0.232	0.622	0.600	0.233	0.620
128	0.639	0.323	0.674	0.639	0.315	0.669
164	0.672	0.365	0.689	**0.667**	**0.317**	**0.678**
200	0.672	0.396	0.711	0.589	0.243	0.643
225	0.661	0.344	0.694	0.633	0.268	0.638
300	0.694	0.370	0.696	0.628	0.302	0.672
500	0.655	0.326	0.671	0.561	0.229	0.627

. Over HTM, the highest performance of Phylogeny-RF employed with number of trees = 64, noticed an Accuracy of 0.700 and Kappa of 0.390 which is better than RF (Accuracy of 0.667 and Kappa of 0.317 with number of trees = 164). Over the CBH dataset, the best performance was achieved with Phylogeny-RF (AUC-ROC of 0.609) method. This indicates performance of Phylogeny-RF with respect to functional classification in terms of Accuracy, Kappa, and AUC-ROC, has achieved an overall improvement in comparison to the traditional RF[25] for identifying phenotypes (Tables 2-4). Phylogeny-RF over a different number of trees (ranging from 1-500) reported this improvement as significant (p-value < 0.05) with regards to predictive metrics over the three data sources in comparison to the traditional RF method proposed by Breiman [25]. It has also been observed that the performance value entries almost intend to follow a U-shaped curve, attaining maximum performance in the somewhere middle range of a number of trees. However, in the case of Phylogeny-RF, better performance is being achieved in a smaller number of trees in comparison to the traditional RF. For example, in the case of HTS data, Phylogeny-RF attained the highest performance at 164 trees whilst 225 trees in traditional RF.

TABLE 4 SUMMARY OF AVERAGE PERFORMANCES OBTAINED BY APPLICATION OF PHYLOGENY-RF AND RF WITH 5-FOLDS (3 TRIAL RUNS) CROSS-VALIDATION DEALING WITH THE DIFFERENT NUMBER OF TREES IN RF OVER COSTELLO BODY SITES (CBH) (BEST VALUES ATTAINED IN EACH COLUMN ARE HIGHLIGHTED IN BOLD IN EACH CATEGORY)[52].

	Phylogeny-RF			**RF**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.745	0.352	0.613	0.729	0.287	0.576
10	0.737	0.287	0.591	0.729	0.252	0.577
20	0.745	0.312	0.603	0.729	0.246	0.575
40	0.738	0.295	0.593	0.718	0.213	0.564
64	0.743	0.307	0.595	0.721	0.217	0.564
100	0.727	0.250	0.575	0.716	0.210	0.567
128	**0.753**	**0.348**	**0.609**	0.726	0.242	0.574
164	0.740	0.301	0.589	0.724	0.239	0.573
200	0.735	0.279	0.584	0.727	0.257	0.577
225	0.748	0.337	0.602	**0.735**	**0.280**	**0.585**
300	0.737	0.283	0.585	0.734	0.259	0.577
500	0.735	0.274	0.581	0.732	0.268	0.579

Furthermore, the variation in AUC was analyzed in the three trials of 5-folds cross-validation employed over the three data sets used in current study. The results also indicate comparatively less variation in the results of Phylogeny-RF than traditional RF **Error! Reference source not found.** over the datasets in the comprehensive analysis and hence indicated the effectiveness of Phylogeny-RF.



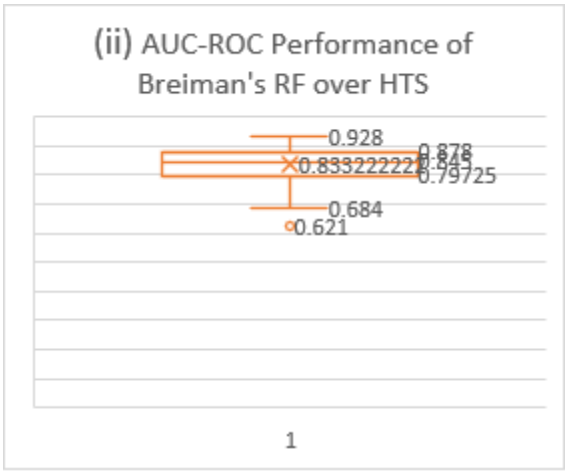
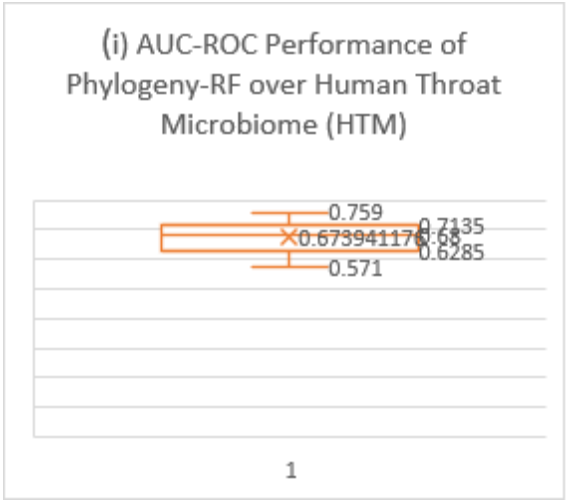


Fig.3a. A graphical representation of Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-folds cross-validation with (i) Phylogeny-RF over HTS (ii) Traditional RF over HTS.

For the use cases of HTS and HTM, the related boxplot graphs indicating variation in AUC-ROC are presented in Fig.3. The results demonstrated a comparatively less variation in the results of Phylogeny-RF than traditional RF[25] over Soil(HTS) and Human Throat Microbiome (HTM) in the comprehensive analysis and hence indicated the effectiveness of Phylogeny-RF.



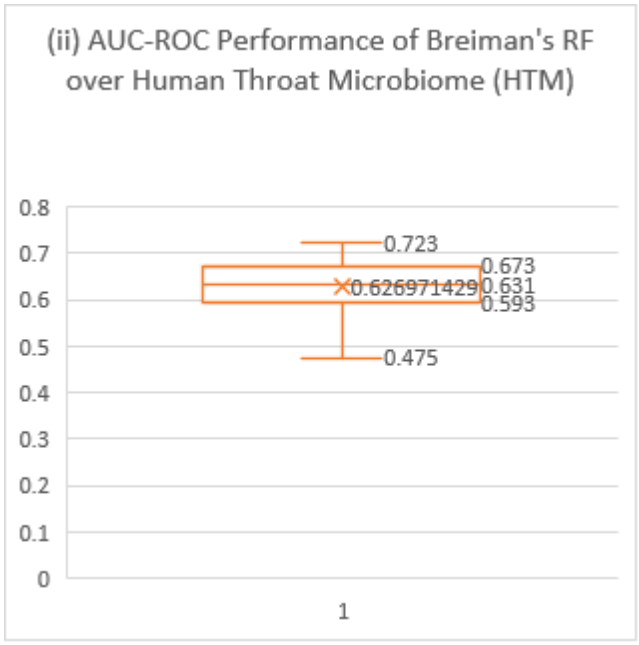


Fig.3b. A graphical representation of Box Plots indicating variation in AUC-ROC obtained over 3 trial runs of 5-folds cross-validation with (i) Phylogeny-RF over HTM (ii) Traditional RF over HTM.

B. Performance Comparison of Phylogeny-RF and state-of-the-art Phylogenetic Models

Phylogeny-RF method is further compared to two popular state-of-the-art methods of MetaPhyl [20] and PhILR[53]. MetaPhyl is a supervised classifier that involves regularization of Logistic Regression by taking advantage of the natural characteristics as encoded in the phylogenetic tree. Analysis in current research shows benchmarking of developed Phylogeny-RF method (attaining best results while varying number of trees in 1-500) with the phylogeny-aware classifier of MetaPhyl (with its default settings recommended in [20]). The newly developed classifier of Phylogeny-RF performed relatively better than MetaPhyl with respect to classification performance metrics over HTS and CBH (Table 5). However, comparing Phylogeny-RF and MetaPhyl over all three data sources, reported marginal difference in the significance of results with (p-value = 0.042 < 0.05).

TABLE 5. SUMMARY OF PERFORMANCE RESULTS OBTAINED BY APPLICATION OF PHYLOGENY-RF WITH 5-FOLDS CROSS-VALIDATION AND METAPHYL WITH DEFAULT SETTINGS OVER DS1, DS3, DS4.

	Phylogeny-RF				**MetaPhyl**		
Data Source	Number of Trees at which best performance is attained	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
HTM	64	0.700	0.394	0.712	**0.733**	**0.471**	**0.739**
DS3	128	**0.753**	**0.348**	**0.609**	0.553	0.132	0.557
DS4	164	**0.929**	**0.891**	**0.949**	0.755	0.633	0.816

The current research further compares the traditional RF model applied to PhILR[53] transformed microbiota data (a phylogeny-based transform) for a further benchmark analysis. The newly constructed Phylogeny-RF was applied to abundance datasets and compared with RF applied over the PhILR-transformed data. The results (averaged over all 3 trial runs over all the 5-folds) are reported in Table 6-8 for sources HTS, HTM and CBH, respectively.

TABLE 6. A COMPARISON OF AVERAGE PERFORMANCE RESULTS OBTAINED BY APPLICATION OF PHYLOGENY-RF AND CLASSICAL RF APPLIED OVER PHILR TRANSFORMED DATA WITH 5-FOLDS CROSS-VALIDATION (3 TRIALS) ON HTS (SOIL MICROBIOME) (BEST VALUE IN EACH COLUMN IS BOLD-FACED).

	Phylogeny-RF			**RF applied over PhILR Transformed Data**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.768	0.666	0.831	0.418	0.113	0.561
10	0.787	0.684	0.849	0.555	0.337	0.665
20	0.871	0.812	0.904	0.567	0.356	0.673
40	0.895	0.841	0.924	0.562	0.349	0.67
64	0.906	0.861	0.933	**0.612**	**0.399**	**0.703**
100	0.919	0.874	0.944	0.567	0.355	0.678
128	0.916	0.872	0.938	0.574	0.363	0.678
164	**0.929**	**0.891**	**0.949**	0.550	0.333	0.667
200	0.901	0.851	0.943	0.581	0.375	0.682
225	0.921	0.88	0.946	0.579	0.375	0.687
300	0.906	0.862	0.935	0.572	0.357	0.654
500	0.923	0.883	0.949	0.576	0.372	0.686

TABLE 7. A COMPARISON OF AVERAGE PERFORMANCE RESULTS OBTAINED BY APPLICATION OF PHYLOGENY-RF AND CLASSICAL RF APPLIED OVER PHILR TRANSFORMED DATA WITH 5-FOLDS CROSS-VALIDATION (3 TRIALS) ON HTM (HUMAN THROAT MICROBIOME) (BEST VALUE IN EACH COLUMN IS BOLD-FACED).

	Phylogeny-RF			**RF applied over PhILR Transformed Data**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.559	0.14	0.57	0.407	-0.189	0.397
10	0.594	0.239	0.635	0.461	0.008	0.472
20	0.628	0.364	0.647	0.455	-0.012	0.482
40	0.689	0.381	0.696	0.439	-0.099	0.462
64	**0.700**	**0.394**	**0.712**	**0.505**	**0.055**	**0.53**
100	0.616	0.232	0.622	0.500	0.059	0.517
128	0.639	0.323	0.674	0.428	-0.087	0.471
164	0.672	0.365	0.689	0.442	-0.054	0.443
200	0.672	0.396	0.711	0.422	-0.072	0.469
225	0.661	0.344	0.694	0.322	-0.142	0.361
300	0.694	0.37	0.696	0.41	0.045	0.457
500	0.655	0.326	0.671	0.422	-0.154	0.427

TABLE 8. A COMPARISON OF AVERAGE PERFORMANCE RESULTS OBTAINED BY APPLICATION OF PHYLOGENY-RF AND CLASSICAL RF APPLIED OVER PHILR TRANSFORMED DATA WITH 5-FOLDS CROSS-VALIDATION (3 TRIALS) ON CBH (COSTELLO BODY HABITATS) (BEST VALUE IN EACH COLUMN IS BOLD-FACED).

	Phylogeny-RF			**RF applied over PhILR Transformed Data**		
Number of Trees	Accuracy	Kappa	AUC-ROC	Accuracy	Kappa	AUC-ROC
1 (Decision Tree)	0.745	0.352	0.613	0.756	0.462	0.676
10	0.717	0.207	0.561	0.790	0.469	0.655
20	0.745	0.312	0.603	0.802	0.512	0.666
40	0.738	0.295	.593	0.781	0.439	0.641
64	0.743	0.307	0.595	**0.815**	**0.547**	**0.683**
100	0.727	0.250	0.575	0.813	0.540	0.680
128	**0.753**	**0.348**	**0.609**	0.800	0.505	0.682
164	0.740	0.301	0.589	0.800	0.505	0.662
200	0.735	0.279	0.584	0.808	0.526	0.680
225	0.753	0.337	0.602	0.800	0.507	0.669
300	0.737	0.283	0.585	0.810	0.534	0.676
500	0.735	0.274	0.581	0.797	0.505	0.671

Phylogeny-RF significantly improved classification performance (p-value < 0.01) in 2 of the 3 benchmark datasets (i.e., HTS, HTM) relative to the PhILR transform averaged over a different number of trees (1 – 500 trees) (Table 7-8).

C. Use of k-means in the Scope of Current Study

k-means as a precursor to RF is proposed in our approach as it helps in achieving good clustering quality and subsequently aids RF in attaining good quality predictions using diversified features. To investigate and validate the cluster quality obtained by k-means more objectively, a benchmark was carried out by generating clusters over phylogenetic diversity matrix with two other popular clustering techniques of Partition Around Medoids (PAM) [54] and hierarchical clustering (hclust)[55] and by using validation index of silhouette coefficient [56]. The value of the silhouette coefficient varies from -1 to 1, and; higher value represents better clustering quality. Experiments performed on data sets in the current study verified the effectiveness of k-means in the proposed computational method. The results of cluster validation indicated a higher value of silhouette coefficient (> 0.60) with regards to k-means in the scope of the current study, in comparison to the other methods of PAM and hclust **Error! Reference source not found.**.

V. DISCUSSIONS

There exists an association between phylogenetic understanding of microbial evolution and the host of microbial communities. The current research continued constructing a computational model for linking microbiome evolution with a host by designing a tree-based ML model that uses phylogenetic distances to tune internal feature modelling within the popular RF classifier enabling a novel evolutionary supervised analysis. In this paper, the construction of an evolutionary-driven model over human and soil microbiomes with different phenotypes is undertaken. The preceding section has shown that the proposed Phylogeny-RF performs better than traditional RF over the three metagenomic data sources. The accuracy of traditional RF is improved by integrating phylogenetic knowledge by minimizing the correlation between microbial features of trees in the RF and maximizing the predictive ability. A decision tree-based modelling of Phylogeny-RF seemed to perform better (in terms of AUC-ROC, Kappa) than MetaPhyl, which regularized the LR model with a phylogenetic measure driven penalty to present an optimization [20], over two of the data sources. This indicates the potential of integrating phylogeny in decision tree-based methods in addition to regularization methods, as suggested in [14].

Nonetheless, what distinguishes the Phylogeny-RF from classical RF applied over the abundances, is the interpretability of microbial features (attributes) corresponding to phylogenetic similarity and reducing the redundancy between microbial features while modelling, which can be a source for biological insight.

This exploratory study found phylogenetic diversity of metagenomes to be useful in each functional classification task capturing the relevant functional adaptations of microbial communities. For this reason, phylogenetic relationships of the input 16S rRNA microbial genes was used to regularize the popular state-of-the-art classifier RF. To meet the needs of the proposed new classifier in the current study, OTUs (microbial features) were clustered based on their phylogenetic similarity as observed in PDM which reports the pair-wise phylogenetic distances between OTUs. PDM serves as a simplified similarity model in which OTUs that are phylogenetically close have been grouped into a cluster. RF model selecting microbial features (i.e., OTUs) from different clusters provided the phylogenetic diverse features.

Specifically, some limitations in current research exist relating to the choice of a phylogenetic distance, the clustering scheme used, and the modelling parameters used for Phylogeny-RF. However, these could be viewed as preliminary heuristics in the current research. Additionally, if it is important that the Phylogeny-RF has meaningful features participating in making class decisions, that should not complicate the interpretation of the traditional RF model. In terms of clustering scheme utilized as a precursor to Phylogeny-RF, current research further worked towards calculating the clustering quality. The paper proposed the use of k-means as a precursor to RF as it helps in achieving good clustering quality and subsequently aids RF in attaining good quality predictions.

VI. CONCLUSION AND FUTURE WORK

We developed a novel and improved approach Phylogeny-RF for classifying 16S rRNA genes in metagenomic studies by incorporating biological domain knowledge into ML. Results indicate that phylogeny could play an important role in shaping metagenomes to determine their functional repertoire. The current metagenomic approaches consider OTUs as independent features for metagenomic analysis with RF; however, OTUs are linked by phylogeny in real-world [13]. Hence, we characterized an approach involving the tuning of the ML model with phylogenetic similarity (derived from a tree), and this resulted in better outcomes while performing metagenomic classification tasks. We constructed Phylogeny-RF, as a predictive model for characterizing metagenomic functions. Phylogeny-RF is inspired from a principle to included maximal phylogenetic diversity while classifying 16SrRNA genes into their functional roles. Similar genes will behave similarly and will be correlated. One of the aims is to remove correlated genes to reduce the redundancy of the data set. Choosing decision nodes of the constituting tree in RF-based on the phylogenetic diversity, supported better metagenomic classification. We report that the contribution of the clustering stage in propose approach is important as it improves the quality of predictions in the subsequent supervised learning stage. The proposed approach attained significantly better performance than state-of-the-art RF over the raw abundance counts with ($p < 0.05$), considering a different number of trees ranging from 1-500, dealing with the 3 metagenomic environments. The proposed method also outperformed the state-of-the-art of MetaPhyl [20] and PhILR [17] in 2 of the benchmarks, indicating potential direction for the future research. The key highlights are summarized in Table 5.

TABLE 5. THE KEY HIGHLIGHTS OF THE PROPOSED RESEARCH

S.No.	Highlights
1	A new classifier Phylogeny-RF (Fig.1) is proposed which is driven by the observation from the literature that closely related microbes by phylogeny are highly correlated and tend to have similar genetic and phenotypic traits. The proposed approach considers phylogenetic dependencies (similarity) between microbial OTUs. It was discovered that the functional analysis of metagenomes with RF benefitted from including the phylogenetic similarity.
2	RF proposed by Breiman selects randomly the subset of features ($Number\ of\ features$) from the original feature space for classification. However, the newly proposed approach in this study (i.e. Phylogeny-RF) also selects randomly the $Number\ of\ features$ but with an intention to select phylogenetically diverse features as closely related microbes by phylogeny tend to have similar genetic and phenotypic traits
3	The similarity between microbial features (OTUs/ASVs) is calculated from a phylogenetic tree in the form of a matrix (PDM) depicting the phylogenetic distances between each pair of leaf nodes (OTUsAaSVs) which is clustered to find similar microbial features by phylogeny and selecting only one amongst these similar features (supporting Minimum redundancy and Maximum relevance principle of involved microbial features). However, an interesting point to observe here is phylogeny has been used to regularize RF model which is applied to the abundance count table which still supports variety in trees as per the quantitative profiles of microbial features.
4	Phylogeny-RF achieves better performance in lesser number of constituent trees in comparison to the RF

The current approach poses limitations as it is time intensive for a large number of samples. It is useful to extend upon the analysis by using the phylogeny-aware context in a more scalable manner or designing a scalable method. Furthermore, this initial analysis could be extended to the utilization of other different distance metrics [36] to explore the area of embedding phylogeny at ML model level further. Also, in current approach similarity between OTU-OTU is considered, however in future we would like to incorporate similarity between internal nodes of a phylogenetic trees as well extending and combining with our previous work published in [23].

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] L. Chistoserdova, "Biotechnology and Genetic Engineering Reviews Functional Metagenomics: Recent Advances and Future Challenges," *Biotechnol. Genet. Eng. Rev.*, vol. 26, no. 1, pp. 335–352, doi: 10.5661/bger-26-335.

[2] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Computational Biology*. 2010, doi: 10.1371/journal.pcbi.1000667.

[3] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

[4] H. Rangwala, A. Charuvaka, and Z. Rasheed, "Machine Learning Approaches for Metagenomics," *Mach. Learn. Knowl. Discov. Databases*, pp. 512–515, 2014.

[5] D. Knights, E. K. Costello, and R. Knight, "Supervised classification of human microbiota," *FEMS Microbiology Reviews*. 2011, doi: 10.1111/j.1574-6976.2010.00251.x.

[6] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning: Methods and ApplicatiOns*, 2012.

[7] A. Statnikov *et al.*, "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, 2013, doi: 10.1186/2049-2618-1-11.

[8] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nat. Commun.*, 2017, doi: 10.1038/s41467-017-01973-8.

[9] 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 Comput. Biol.*, vol. 12, no. 7, pp. 1–26, 2016, doi: 10.1371/journal.pcbi.1004977.

[10] D. Gevers *et al.*, "The treatment-naive microbiome in new-onset Crohn's disease," *Cell Host Microbe*, 2014, doi: 10.1016/j.chom.2014.02.005.

[11] E. Asgari, K. Garakani, A. C. McHardy, and M. R. K. Mofrad, "MicroPheno: Predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples," *Bioinformatics*, vol. 34, no. 13, pp. i32–i42, 2018, doi: 10.1093/bioinformatics/bty296.

[12] S. Rakoff-Nahoum, K. R. Foster, and L. E. Comstock, "The evolution of cooperation within the gut microbiota," *Nature*, 2016, doi: 10.1038/nature17626.

[13] M. Berlanga and M. Berlanga Herranz, "Molecular phylogeny of microorganisms," *Int. Microbiol. Off. J. Spanish Soc. Microbiol.*, 2010, doi: 10.2436/im.v13i4.53878.

[14] J. Kuczynski, J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight, "Using QIIME to analyze 16s rRNA gene sequences from microbial communities," *Curr. Protoc. Microbiol.*, 2012, doi: 10.1002/9780471729259.mc01e05s27.

[15] M. G. I. Langille *et al.*, "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nat. Biotechnol.*, vol. 31, no. 9, pp. 814–821, 2013, doi: 10.1038/nbt.2676.

[16] A. D. Washburne *et al.*, "Phylogenetic factorization of compositional data," *bioRxiv*, pp. 1–30, 2016, doi: 10.1101/074112.

[17] 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, doi: 10.7554/eLife.21887.

[18] Q. Zhu, Q. Zhu, M. Pan, X. Jiang, X. Hu, and T. He, "The Phylogenetic Tree based Deep Forest for Metagenomic Data Classification," 2019, doi: 10.1109/BIBM.2018.8621463.

[19] J. T. Wassan *et al.*, "An Integrative Framework for Functional Analysis of Cattle Rumen Microbiomes," 2019, doi: 10.1109/BIBM.2018.8621104.

[20] O. Tanaseichuk, J. Borneman, and T. Jiang, "Phylogeny-based classification of microbial communities," *Bioinformatics*, vol. 30, no. 4, pp. 449–456, 2014, doi: 10.1093/bioinformatics/btt700.

[21] T. Wang and H. Zhao, "A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms," *Biometrics*, vol. 73, no. 3, pp. 792–801, 2017, doi: 10.1111/biom.12654.

[22] Q. Zhu, Q. Zhu, M. Pan, X. Jiang, X. Hu, and T. He, "The Phylogenetic Tree based Deep Forest for Metagenomic Data Classification," *Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018*, pp. 279–282, Jan. 2019, doi: 10.1109/BIBM.2018.8621463.

[23] J. T. Wassan, H. Wang, F. Browne, and H. Zheng, "PAAM-ML: A novel Phylogeny and Abundance aware Machine Learning Modelling Approach for Microbiome Classification," 2019, doi: 10.1109/BIBM.2018.8621382.

[24] J. Ning and R. G. Beiko, "Phylogenetic approaches to microbial community classification," *Microbiome*, vol. 3, no. August, p. 47, 2015, doi: 10.1186/s40168-015-0114-5.

[25] L. Breiman, "Random Forrest," *Mach. Learn.*, 2001, doi: 10.1023/A:1010933404324.

[26] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," 2012, doi: 10.1007/978-3-642-31537-4_13.

[27] J. B. H. Martiny, S. E. Jones, J. T. Lennon, and A. C. Martiny, "Microbiomes in light of traits: A phylogenetic perspective," *Science (80-.).*, vol. 350, no. 6261, 2015, doi: 10.1126/science.aac9323.

[28] T. Wang and H. Zhao, "Constructing Predictive Microbial Signatures at Multiple Taxonomic Levels," *J. Am. Stat. Assoc.*, vol. 112, no. 519, pp. 1022–1031, 2017, doi: 10.1080/01621459.2016.1270213.

[29] T. H. Nguyen, Y. Chevaleyre, E. Prifti, N. Sokolovska, and J.-D. Zucker, "Deep Learning for Metagenomic Data: using 2D Embeddings and Convolutional Neural Networks," no. Nips, 2017, [Online]. Available: http://arxiv.org/abs/1712.00244.

[30] N. D. Youngblut, S. E. Barnett, and D. H. Buckley, "HTSSIP: An r package for analysis of high throughput sequencing data from nucleic acid stable isotope probing (sip) experiments," *PLoS One*, 2018, doi: 10.1371/journal.pone.0189616.

[31] E. S. Charlson *et al.*, "Disordered microbial communities in the upper respiratory tract of cigarette smokers," *PLoS One*, 2010, doi: 10.1371/journal.pone.0015216.

[32] 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, doi: 10.1186/s13073-016-0302-3.

[33] E. E. K. Costello, C. C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight, "Bacterial community variation in human body habitats across space and time," *Science (80-.).*, vol. 326, no. 5960, pp. 1694–7, 2009, doi: 10.1126/science.1177486.Bacterial.

[34] J. Felsenstein, "The Newick tree format," *http://evolution.genetics.washington.edu/phylip/newicktree.html*, 1857.

[35] E. Paradis, J. Claude, and K. Strimmer, "APE: Analyses of phylogenetics and evolution in R language," *Bioinformatics*, 2004, doi: 10.1093/bioinformatics/btg412.

[36] C. M. Tucker *et al.*, "A guide to phylogenetic metrics for conservation, community ecology and macroecology," *Biol. Rev.*, 2017, doi: 10.1111/brv.12252.

[37] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," *J. Comput. Biol.*, 1999, doi: 10.1089/106652799318274.

[38] A. Ng, K. Soo, A. Ng, and K. Soo, "k-Means-Clustering," in *Data Science – was ist das eigentlich?!*, 2018.

[39] L. Czech and A. Stamatakis, "Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples," *PLoS One*, vol. 14, no. 5, pp. 1–72, 2019, doi: 10.1371/journal.pone.0217050.

[40] P. Kumar and S. K. Wasan, "Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets," *Proc. CSIT 2009 Int. Symp. Comput. Commun. Control (ISCCC 2009)*, 2011.

[41] N. Hansen and A. Ostermeier, "Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation," 1996, doi: 10.1109/icec.1996.542381.

[42] G. T. Shi, G. Gao, Y. M. Jiang, and G. Y. Kuang, "Revision of the similarity metrics and validity proof of the second eigenvalue," *Tien Tzu Hsueh Pao/Acta Electron. Sin.*, 2009.

[43] D. L. Elliott, "Matrix Algebra," in *Applied Mathematical Sciences (Switzerland)*, 2009.

[44] J. Oksanen *et al.*, "Package vegan: Community Ecology Package," 2013. doi: 10.4135/9781412971874.n145.

[45] R. W. Kolb, "Gini Index," in *The SAGE Encyclopedia of Business Ethics and Society*, 2018.

[46] S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of

statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability," *Soft Comput.*, 2009, doi: 10.1007/s00500-008-0392-y.
[47] D. Brzezinski and J. Stefanowski, "Prequential AUC: properties of the area under the ROC curve for data streams with concept drift," *Knowl. Inf. Syst.*, 2017, doi: 10.1007/s10115-017-1022-8.
[48] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.
[49] "RF CODE at DATA QUEST BLog." https://www.dataquest.io/blog/learning-curves-machine-learning/.
[50] Stevens *et al.*, "Post-hoc Tests in ANOVA," *J. Supercrit. Fluids*, 2017, doi: doi:10.1201/9781420036909.ch4.
[51] L. Breiman, "Bagging predictors," *Mach. Learn.*, 2004, doi: 10.1007/bf00058655.
[52] J. T. Wassan, H. Wang, and H. Zheng, "A New Phylogeny-Driven Random Forest-Based Classification Approach for Functional Metagenomics," *Proc. - 2022 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2022*, pp. 32–37, 2022, doi: 10.1109/BIBM55620.2022.9995554.
[53] J. Silverman, "Phylogenetic partitioning based ILR transform for metagenomics data." https://bioconductor.org/packages/release/bioc/html/philr.html.
[54] J. Chen, "Statistical Methods for Human Microbiome Data Analysis Statistical Methods for Human Microbiome Data Analysis," 2012.
[55] P. Langfelder and S. Horvath, "Fast R functions for robust correlations and hierarchical clustering," *J. Stat. Softw.*, 2012, doi: 10.18637/jss.v046.i11.
[56] D. B. Knights and D. Knights, "Predictive Modeling of Metagenomes," 2012.

Author Details

Dr Jyotsna Talreja Wassan has obtained PhD degree from the School of Computing, Ulster University, United Kingdom, 2020 and her current research area is Integrative Data Analytics in Metagenomics. She has been serving as an Assistant Professor in the Department of Computer Science, Maitreyi College, University of Delhi, India, since 2010. She has published papers in international journals and conference proceedings, e-lessons, and book chapters. She worked as a software engineer in her early career at ST. Microelectronics Pvt. Ltd., India, and received Silver Recognition for the FALCON project.

Dr. Haiying Wang received the PhD degree in artificial intelligence in biomedicine, in 2004, he is currently a reader in the School of Computing at Ulster University, UK. His research area includes artificial intelligence, complex network analysis, computational biology and bioinformatics. He has a particular research interest and expertise in network-based approaches to the field of systems biology and meta-genomics. Since 2004, he has published more than 150 peer-reviewed research papers in international journals and conference proceedings.

Prof. Huiru Zheng (Senior Member, IEEE) received the PhD degree in bioinformatics, in 2003 and a postgraduate certificate in teaching in Higher Education, in 2005 from Ulster University, Coleraine, Northern Ireland. She is a professor of computer science with the School of Computing, Ulster University, U.K.; and a fellow of the U.K. Higher Education Academy. She is an active researcher in bioinformatics and healthcare informatics. Within her broad interests in data mining, data integration, machine learning and healthcare decision support, she has a particular research interest and expertise in integrative data analytics in the field of systems biology, and intelligent data analysis and assistive technology to support healthcare and independent living. She has published more than 300 peer reviewed scientific research papers.