EVOLUTION IN ELECTRICAL AND ELECTRONIC ENGINEERING
e-ISSN: 2756-8458

EEEE

Vol. 5 No. 2 (2024) 214-219
https://publisher.uthm.edu.my/periodicals/index.php/eeee

Diet Prediction and Feature Importance of Gut Microbiome using Machine Learning

Vaibhav Godase[1]*, Jyoti Godase[2]

[1] *Faculty of Electronics & Telecommunication Engineering, SKN Singhad College of Engineering,Pandharpur , 413304, India*
[2] *Faculty of Electronics & Telecommunication Engineering PVPIT , Budhgaon,413314, India*

*Corresponding Author: vaibbhav.godse@gmail.com
DOI: https://doi.org/10.30880/eeee.2024.05.02.026

Article Info

Received: 28 April 2024
Accepted: 28 October 2024
Available online: 30 October 2024

Keywords

Gut Microbiome, Machine Learning, Diet Prediction, OTUs, ANN

Abstract

The microbiome is the genetic material of all the microbes, bacteria, fungi, protozoa and viruses that live on and inside the human body. The most exciting venture of present-day medicine is determining the microscopic dwellers of every organ and restoring them in balanced proportion when it gets disrupted. It is challenging to conclude whether disease or disrupted microbiota occurred first based on observation of health problems. By collecting the microbiome data and analysing it, precise medicines can be given to the patients. Precision Medicine depends on microbiome composition inside a patient. There are various medicines used to cure the specific disease. Medicines reacts differently on every human being, one of the reason is the difference in the microbiome composition of the person. If one is able to analyse the composition of microbiome, the medicine apt can even be predicted to cure that disease. Diet has abundant influence on once gut microbiome population. This work analyses human gut microbiome data to find the microbiomes which has more influence on human health from different diet pattern scrutinized on humanized gnotobiotic mice. Machine Learning (ML) techniques like Logistic Regression (LR), Random Forest (RF) and Artificial Neural Network (ANN) have been applied on an accessible data set from MINE (Maximal Information-based Non-parametric exploration). By using Machine Learning, the type of diet can be predicted from microbiome composition with great accuracies as well as diet sensitive microbiomes are found which are very less compared to a huge number of microbiomes in data set.

1. Introduction

During incubation, humans are disinfected and it becomes host to massive variations of microbes, bacteria, archaea, fungus and virus that develops on various body parts during and after birth. In maintenance of immune system and digestion of our food, microbiomes plays an important role and dysfunction of microbiomes leads to various disease [1]. If human gut microbiome is scrutinized, the potential diseases can be predicted by using machine learning techniques before they actually affect the person. Early predictions of the disease will help people to take preventive measures and to start the treatment accordingly and will prevent risk of life. In near future, healthy microbiomes with microbiota might be able to be restored from the donor, which will reduce the risk of antibiotic resistance and increase the cure rate.

Evolution in Electrical and Electronic Engineering Vol. 5 No. 2 (2024) p. 214-219

However, the most interesting and challenging task for the doctors is the diagnosis of the disease. Machine Learning (ML) techniques can be used to help them to predict the disease with high accuracy. Relationship between microbiomes and the disease and their patterns can be identified by using machine learning techniques. It is also a remarkable technology in bio-medical research [2] with several applications. By using various ML techniques and algorithms, one can obtain acceptable results by examining the biological samples.

The microbiome data set implies the Influence of Diet on Human Gut Microbiome on Humanized Gnotobiotic Mice [3]. Mice are populated at particular life phases with diverse microbial populations which is referred to as gnotobiotic animals and offer admirable system for controlling different conditions such as host genotype, microbial community composition, diet, and housing conditions. Before colonizing, mice were upraised in germ free environment without being in contact with microbes.

Culture-independent comparisons shows that the distal gut microbiota of mice and humans harbors the same bacterial phyla, most bacterial genera and species found in mice are not seen in humans. Diet induced obesity is used as a model, and it shows how these humanized animals can be used to conduct the control. Young and adult male mice of 5 to 7 week old were colonized using the microbial community [4] present in a freshly evacuated faecal samples from a healthy adult human. The samples collected were instantly stored in an anaerobic environment. The recipient mice were conserved on ordinary low-fat plant polysaccharide–rich (LF/PP) diet in distinct cages in a gnotobiotic isolators. Mice were swapped to high-fat, high-sugar Western diet after the faecal samples were collected in 1 day, 1 week, and 1 month. For further 2 months mice were maintained on the same diet, faecal samples were collected weekly, and at this point they were killed.

Samples of Microbiomes are collected based on different combination of diet, gender, and donor and collection method. Total collected samples are 676. And total microbiomes (OTUs) considered for every sample is 6,696. In this paper, analysis is done on the diet of the mice and the microbiome samples injected to them based on the different experimental condition.

Section II refers to literature review, section III describe the objective of the work, section IV describes proposed Method, section V explains results and discussion and section VI refers to conclusion.

2. Literature Review

The research by Laura V. Blanton, et al. emphasis on the microbiome composition collected from the children of Malawi in East Africa. Malnourished offspring show impaired development of their gut microbiome. Mice were colonized on Malawian diet and in germ free environment as they were transplanted with the microbiome samples from 6 and 18 months old healthy or malnourished donors from Malawian [4] and it was observed that immature microbiota from malnourished offspring and children conveyed impaired growth phenotypes. Standardized offspring gut microbiota was developed by developing RandomForest model. Relation between age and growth discriminatory taxa and femoral bone phenotype was created. Their results show that microbiota of 18 months old donor couldn't make remarkable effects as compared to the microbiota from 6-months old donor in freshly weaned mice.

Yet another study by Pia S. Pannaraj, et al. related to the microbiome is, explains, and varied population of bacteria that are imagined seeding newborn's gut by suckling. In healthy mother-infant pair, they estimated the relation between maternal bacteria and infant's stool. The samples were collected from the areolar skin swabs and breast milk from mother and stool from infants used diapers during their clinical survey and even home study visit [5]. Difference in the diversity was compared by using various comparison methods based on feeding characteristics. Here, authors thought of considering various factors such as age of infant, geographic area, type of delivery, demographic and suckling characteristics to calculate the source of infant stool microbiome variation. RandomForest model was developed to predict the newborn's stool bacteria based on suckling methods.

The latest research on microbiome is by Firas S. Midani, et al. The authors considered the cases which were positive to V. cholerae based on their stool test. After recognizing the patients, they collected all the samples from their households within 6 hours of diagnosis and found that patients didn't have enough access to the uncontaminated water and sanitation. For the research purpose only, the patients between the ages of 2 to 60 years were selected from Dhaka, Bangladesh. Data was collected with blood group and rectal swab specimen for V. cholerae with 16SrRNA [6]. Authors managed to measure the parameters such as vibriocidal titters and symptom histories for the period of 30 days during survey. Further they tried to develop a relationship between human gut bacterial communities and V. cholerae which may help to understand steps for disease prevention. Models were built using sklearn in Python. The train test split strategy was used with a training set of 48 instances and the testing set of 28 instances. SVM was used to analyze relative abundance of OTUs and classified infected and uninfected human.

3. Proposed Methodology

Based on microbiome composition, this work is to predict the diet of the humanized mice and also to find the diet induced important microbiome from the microbiome data set.

Diet has been selected as target variable for analyzing microbiomes. Diet is classified into five classes such aLFPP FATR (Full fat ingredient), Suckling diet and Human diet. Features are 6,696 OTUs of microbiomes. Diet is predicted by using Logistic Regression, Random Forest, and Artificial Neural Network (ANN).

Correlation between dependent variables and one or more independent variable can be estimated by using Logistic Regression. Random Forest also can be used for classification problems, mapping several decision tree and combining them to get exact and firm prediction. ANN model is influenced from the animal's central nervous system, it can be applied for both classification and regression with continuous target attributes [7]. It comprises of three types of layers such as Input layer, Hidden layer, and Output layer. ANN required a huge amount of data for good performance.

Performance of Logistic Regression, Random Forest and ANN are analyzed using accuracy and Cohen's kappa metrics. In Machine Learning, accuracy and Cohen's kappa are the most used metrics to determine which model is best at identifying relationships and patterns between input features (independent variables) and output variable (dependent or target variable), in case of classification problems. Accuracy is a metric which represents the percentage of matches between predicted labels and actual labels. Multiclass and imbalanced class problems can be handled effectively by using Cohen's Kappa metric. Kappa helps us to know the agreement between two annotators. Due to inadequate data set, the microbiome composition was not able to be predicted based on the diet.

4. Result

The models have been trained with a total of 6,696 OTUs as input features and diet as target variable. Used data set has total 667 samples for 5 diet classes out of which only two classes are selected, LF/PP with 389 samples and Western Diet with 289 samples. A total of 658 samples were selected for this work, remaining classes are neglected due to fewer samples.

Logistic Regression and Random Forest are applied with 10-fold stratified cross-validation and their results are compared in Fig. 1 and Fig. 2. Accuracy and Cohen's kappa score for Logistic Regression is 96.86% and 0.93, respectively. Similarly, accuracy and Cohen's kappa for Random Forest is 98.43% and 0.96, respectively shown in Fig. 1 and Fig. 2.

For ANN, the data into is split into 90% and 10% by using stratified train-test split function of scikit-learn package of Python. 90% of the dataset is used for training and the remaining 10% is used for testing purposes. ANN is implemented using Python's TensorFlow library, with two hidden layers, learning rate of 0.003, batch size of 40, AdamOptimizer as optimization function and trained for 15 epochs. For ANN training and testing, accuracy observed is 98.47% and 92.42% as shown in Fig. 3 respectively. Cohen's kappa score is 0.9686 and 0.8476.

Table 1 shows the comparison of Logistic Regression, Random Forest and ANN performance with respective accuracy and Cohen Kappa score. It is expected that ANN will perform better than Logistic Regression and Random Forest if using adequate data.

Out of 6,696 OTUs (i.e., microbiomes), it is observed that only 338 are diet sensitive microbiomes. The top five most important observed microbiomes are OTU453, OTU5370, OTU5948, OTU5417, and OTU1462 shown in Fig. 4. OTUs of the first 100 important microbiomes are shown in Fig. 5.

Table 1 *Logistic Regression, Random Forest and ANN (testing) with respective Accuracy and Cohen Kappa*

Modes	Accuracy	Cohen Kappa
Logistic Regression	96.86	0.93
Random Forest	98.43	0.96
ANN	92.42	0.84

Evolution in Electrical and Electronic Engineering Vol. 5 No. 2 (2024) p. 214-219

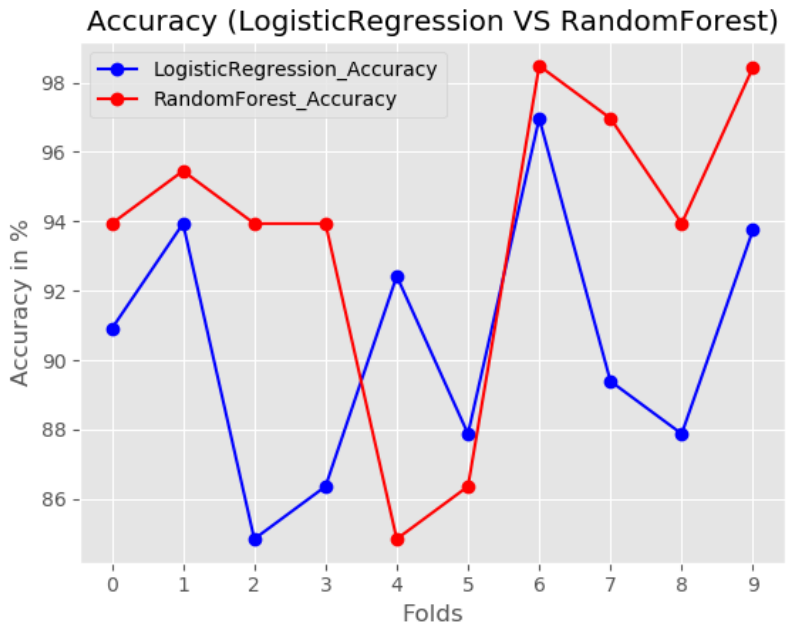

Fig. 1 *Plot of accuracy for LogisticRegression and RandomForest for 10 folds.*

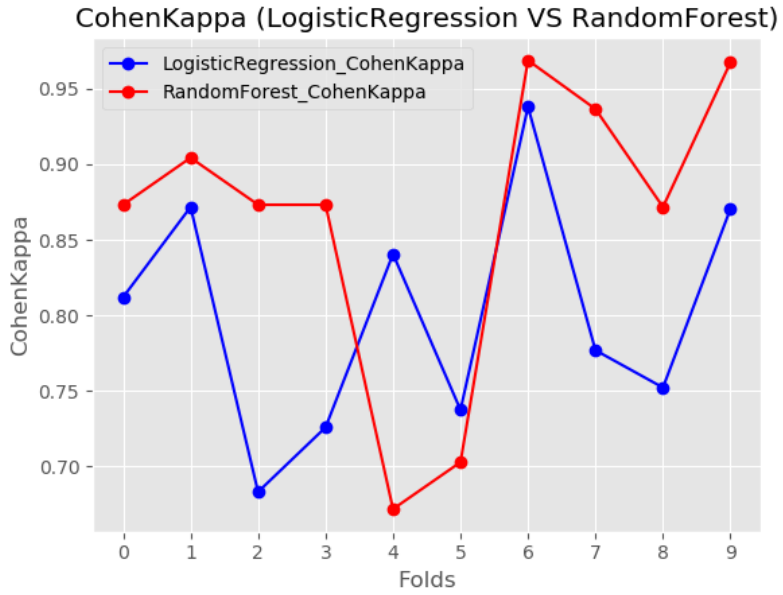

Fig. 2 *Plot of CohenKappa for LogisticRegression and RandomForest for 10 folds cross validation*

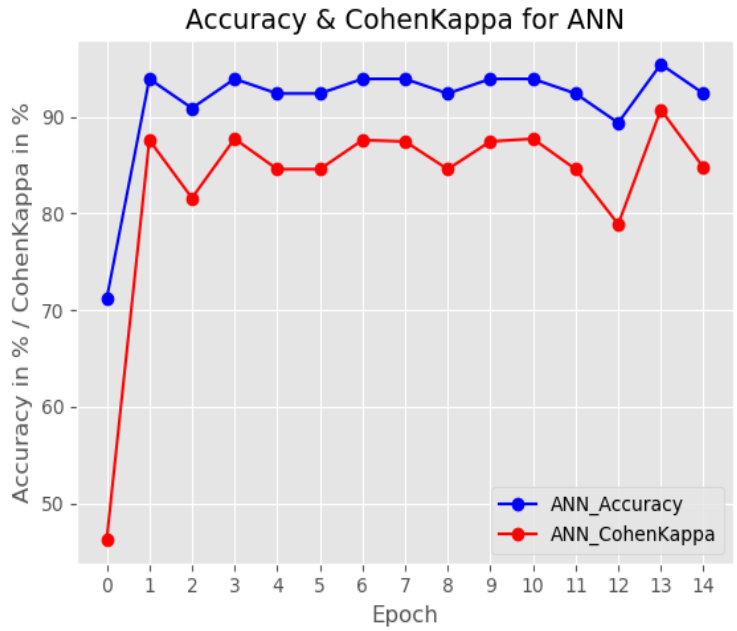

Fig. 3 *Plot shows the Accuracy and CohenKappa for ANN with 15 epochs.*

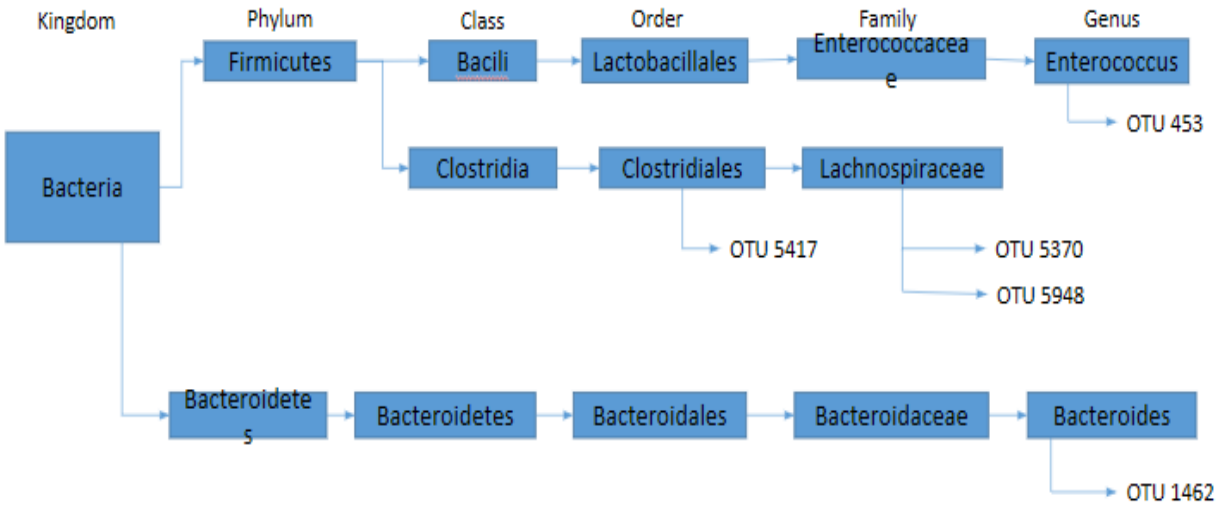

Fig. 4 *Phylogenetic Tree for top 5 OTUs from diet sensitive microbiomes out of 338*

OTU453	OTU2970	OTU1193	OTU546	OTU3845	OTU6314	OTU501	OTU6678	OTU2809	OTU5987
OTU5370	OTU4083	OTU1288	OTU1780	OTU4365	OTU678	OTU2957	OTU2358	OTU3046	OTU3696
OTU5948	OTU434	OTU5429	OTU5625	OTU4832	OTU1044	OTU2454	OTU1530	OTU3725	OTU903
OTU5417	OTU5691	OTU6470	OTU4368	OTU949	OTU3933	OTU3490	OTU5420	OTU3386	OTU3110
OTU1462	OTU3406	OTU5139	OTU4020	OTU3165	OTU6231	OTU1618	OTU2786	OTU6675	OTU2763
OTU155	OTU5937	OTU3818	OTU2490	OTU4380	OTU2418	OTU336	OTU4003	OTU3858	OTU483
OTU4695	OTU2036	OTU3400	OTU5741	OTU3964	OTU3010	OTU1274	OTU2187	OTU5490	OTU6524
OTU1629	OTU3855	OTU2340	OTU4154	OTU5236	OTU1171	OTU675	OTU2930	OTU368	OTU2473
OTU6224	OTU3837	OTU4931	OTU2705	OTU3612	OTU4511	OTU2489	OTU1827	OTU4264	OTU4257
OTU453	OTU1333	OTU2998	OTU4810	OTU5167	OTU3213	OTU5348	OTU2413	OTU730	OTU101

Fig. 5 *for first 100 diet sensitive OTUs from the total of diet sensitive 338 OTUs from the complete dataset.*

5. Conclusion

This is clear from the work done here that the common Machine Learning techniques can figure out the relation between gut microbiomes and diet. With such a smaller number of samples, the results are more than 95% accuracy by all the three ML techniques (Logistic Regression, Random Forest and ANN) applied here. Increase in dataset will help to achieve close to 100% accuracy. Along with this, ML techniques also provide us the diet sensitive microbiomes. With the help of a medical expert or microbiologist, diet sensitive microbiomes can be further analyzed to find the effect of diet on human health and also figure out ways to maintain gut microbiome composition.

Acknowledgement

The data set used has been downloaded from: http://www.exploredata.net/Downloads/Microbiome-Data-Set. And the significance of the data set has been understood from "The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice" by Peter J. Turnbaugh, Vanessa K. Ridaura, et al.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Vaibhav Godase, Jyoti Godase;* ***data collection:*** *Vaibhav Godase, Jyoti Godase;* ***analysis and interpretation of results****: Vaibhav Godase;* ***draft manuscript preparation:*** *Vaibhav Godase,jyoti Godase. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Xochitl C. Morgan, Curtis Huttenhower. "Human Microbiome Analysis PLOS". PLOS , December 27, 2012
[2] Niknejad A, Petrovic D. "Introduction to computational intelligence techniques and areas of their applications in medicine". Med Appl Artif Intell 2013; Pg.no.51.
[3] Peter J. Turnbaugh, Vanessa K. Ridaura, et al. "The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice" ScienceTranslationMeg.org Published on 11 November 2009; Volume 1 Issue 6 6ra14, Pg.no.1
[4] Laura V. Blanton, Mark R. Charbonneau, et al. "Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children". sciencemag.org 19 February 2016, Vol 351 Issue 6275, Pg.no. 830
[5] Pia S. Pannaraj, Fan Li,Chiara Cerini, et al. "Association between Breast Milk Bacterial Communities and Establishment andDevelopment of the Infant Gut Microbiome". JAMA Pediatr, Published online May 8, 2017.Pg no. E1
[6] Firas S. Midani, Ana A. Weil, et al. "Human Gut Microbiota Predicts Susceptibility to Vibrio Cholerae Infection". The Journal of Infectious Diseases, 10 April 2018, Pg. no. 1-9.
[7] ANN in Machine Learning https://data-flair.training/blogs/artificial-neural-network/ . Last accessed 25 September 2018