ISSN2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol.10, No.2, March 2022, pp.138-146

Clinical decision making for prediction of otitis using machine learning approach

Fatima Abdel Monem Trtak[1], Kanita Karađuzović-Hadžiabdić[1]
[1]Department of Engineering, International University of Sarajevo, Bosnia

ABSTRACT

This study investigates the relationship between autoimmune disease otitis and gut microbial community abundance by using machine learning as an aid in the medical decision-making process. Stool samples of healthy and otitis diseased infants were obtained from the *curatedMetagenomicData* package. Class imbalance present in the dataset was handled by oversampling a minority class. Afterwards, we built several machine learning models (support vector machine, k-nearest neighbour, artificial neural networks, random forest and gradient boosting) to predict otitis from gut microbial samples. The best overall accuracy was obtained by the random forest classifier, 0.99, followed by support vector machine and gradient boosting algorithms, both achieving 0.96 overall accuracy. We also obtained the most informative predictors as potential microbial biomarkers for the otitis disease. The obtained results showed better accuracy in prediction of otitis from microbial metagenome than previously proposed methods found in literature.

Keywords: microbiome, artificial intelligence, machine learning, biomarker identification imbalanced data analysis, outlier detection

Corresponding Author:

Kanita Karađuzović-Hadžiabdić
Department of Engineering
International University of Sarajevo
Hrasnička cesta 15, 71210 Sarajevo, Bosnia
E-mail: kanita@ius.edu.ba

1. Introduction

Human beings possess a microbial community (also known as microbiome) composed of different bacteria, archaea, fungi, protists and viruses. 500-1000 species of bacteria are estimated to exist in the human body, each with thousands of genes indicating genetic diversity. Furthermore, every one of us has a unique microbiome that varies in taxonomic composition [1]. There are also differences in microbial diversity and community composition across different body sites. The most studied site in the human body is the gut microbiome, as it contains large and diverse microbial communities. The change in the composition of the intestinal microbiome community and its interaction with the immune and nervous system has been found to correlate with different illnesses. This condition of imbalance in the gut has been found to be associated with various diseases such as cancer, inflammatory bowel disease, allergy, schizophrenia, asthma, hypertension, etc. [2]. The fact that 50% of the human being cell population is composed of bacteria attracted researchers' interest for deeper investigation [3]. Recent advances in high throughput sequencing have led to increased availability of microbiome data resulting in a large amount of microbiome-related research in disease diagnostic, prediction and therapeutics. In addition, due to the rise in computer processing power and storage capacity, machine learning methods are increasingly used in microbiome analysis. Many studies have

explored the relationship between the taxonomical abundance of bacteria and the presence of specific diseases.

For example, Beck et al. [4] used logistic regression and random forest classifiers to analyze the relationship between the microbial community and bacterial vaginosis (BV). By using random subsets of features the authors identified features linked to BV that are in line with other performed studies.

Zhu et al. [5] used MicroPto, a computational tool to perform metagenomic data analysis considering reads of known and unknown microbial organisms to investigate the association between viruses and complex diseases. The results of the performed research show that including reads from unknown organisms increases the disease prediction accuracy based on metagenomic data. The authors also showed that some viruses have an important role in developing colorectal cancer and liver cirrhosis, but not in type-2 diabetes.

Casimiro-Soriguer et al. [6] used 1042 fecal metagenomic samples from seven publicly available studies to perform meta-analysis using machine learning. They used functional metagenomic profiles instead of taxonomic profiles to predict colorectal cancer (CRC) and distinguish CRC from adenoma. They observed that functional profiles reach superior accuracy in predicting CRC and adenoma conditions than taxonomic profiles.

Vatanen et al. [7] showed that Bacteroides lipopolysaccharide (LPS) is structurally distinct from E. coli l LPS and inhibits innate immune signaling and endotoxin tolerance by following the gut microbiota development of 222 infants in Northern Europe (Russian, Finnish, and Estonian children) from their birth until the age of three. The authors also trained a set of random forest classifiers using genus-level data from samples collected between 170 and 260 days of age. They were able to predict the country from which the samples belong with high accuracy of 0.94 for Finns versus Russians but low accuracy of 0.55 between Finns and Estonians.

Common challenges in microbiome research often include dealing with high dimensional data low sample size, data heterogeneity and scarcity, imbalanced class size, etc. In [8], the authors handled extreme class imbalance by computing the size of the largest class (healthy) and randomly resampling from every class until each class had the same number of samples. Sayyari et al. [9] addressed low sample size by introducing the TADA algorithm. The algorithm uses available data and a statistical generative model taking into account phylogenetic relationships between microbial species, to create new samples augmenting existing ones.

In a recent study, Khan et al. [8] developed a multiclass microbiome disease classifier using *curatedMetagenomicData*. To perform the classification, the authors used random forests (RF), deep neural networks (DNN), and graph convolutional neural networks (GCN) machine learning models. They observed that in general, GCN performed similar or better than DNN. Implemented classifiers were able to distinguish between 18 different diseases and healthy controls achieving greater than 70% accuracy on a dataset of over 7000 samples, 92% average area under the receiver operating characteristic curve (AUC) and 50% average area under precision recall (AUPR). In the performed multiclass classification, overall accuracy obtained for the otitis disease was 12% by GCN, 13% by DNN and 0% by the RF classifier. Periodontitis was the most accurately classified disease with 93% overall accuracy achieved by all three models. However, the RF model achieved an excellent accuracy (99%) in classification of healthy vs. non-healthy samples.

In this study we analyzed the relationship between taxonomical abundances of the gut microbial community extracted from the fecal samples and otitis disease using machine learning approach using the *curatedMetagenomicData*. We also identified the most important microbial community related to otitis. Otitis is a group of inflammatory diseases of the middle ear. It mostly appears in young children as a result of pulling at the ear, increased crying, and poor sleep. Other causes can be decreased eating and a fever. Otitis was also found to be associated with hearing loss [10].

2. Materials and Methods

2.1. Dataset

The otitis data selected for this study was obtained from the *curatedMetagenomicData* [11]. With the large availability of shotgun metagenomic data, *curatedMetagenomicData* was generated as an initiative by Pasolli et al. [11] to provide a standardized microbiome data that will allow the study of taxonomic composition and functional potential of microbiome to the research community. It contains uniformly processed taxonomic and metabolic functional profiles for more than 5,500 whole metagenome shotgun sequencing samples from 26 publicly available studies, providing a standardized, microbiome data for novel analysis. More specifically, the data includes species-level taxonomic profiles expressed as relative abundance from kingdom to strain level, presence of unique clade-specific markers, the abundance of unique clade-specific markers, abundance of gene families, metabolic pathway coverage, and metabolic pathway abundance collected from different body sites. This study uses species-level relative abundance data. Microbial taxonomic abundances for each sample were generated by MetaPhlAn2, and metabolic functional potential was computed with HUMAnN2. Standardized metagenomic data and its manually curated metadata are integrated and documented as ExpressionSet objects distributed through the Bioconductor ExperimentHub.

For otitis disease prediction, we used the available Operational Taxonomic Units (OTU) table. OTU tables are an essential part of the microbial data study. They are made by clustering DNA sequences based on their similarity (usually 97%) to represent the abundance of a particular bacterial taxon [12]. After clustering, a table that records corresponding abundances per sample is generated.

Obtained otitis data contains 785 samples from infants aged between 33 days and 3 years collected from the subject stool, and the available OTU table consists of 1584 taxa. The proportion of samples belonging to healthy controls is 78%, 14% to otitis diseased infants, and remaining 8% belong to infants with other autoimmune diseases.

2.2. Preprocessing and feature selection

As a first preprocessing step, we removed all the samples annotated as infant autoimmune disease other than otitis and all phylum-level taxa with NA (not available) values, leaving a total of 721 otitis and healthy samples in the dataset. Then, to ensure comparability of data across samples we performed min-max normalization using the following formula:

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where the x_i is the i^{th}data instance, $min(x)$ and $max(x)$ are the minimum and maximum values respectively, y_i is the obtained minimization result. Afterwards, we analyzed 1584 taxa for possible outliers. Taxonomy filtering was done to only keep the effective taxa. We explored the relationship of prevalence and total abundance at the phylum level (Fig. 1). Prevalence indicates the number of samples in which a genus was positively detected, while total abundance indicates the average fractional representation of a single genus only when present. Some bacteria such as *Acidobacteria, Candidatus_Saccharibacteria, Chlorobi, Apicomplexa, Deinococcus_Thermus, Synergistetes, Tenericutes, Verrucomicrobia, and Euryarchaeota* appeared to have low prevalence at the phylum level, meaning they do not appear in many samples. Filtering was thus done to remove the taxa present in less than 0.01% across all samples. This reduced the amount of false-positive bacteria. It also considerably lowered the dimensionality of the dataset from 1584 to 178 taxa, reducing the number of OTUs (i.e. the initial features) to be used as inputs into the machine learning methods.

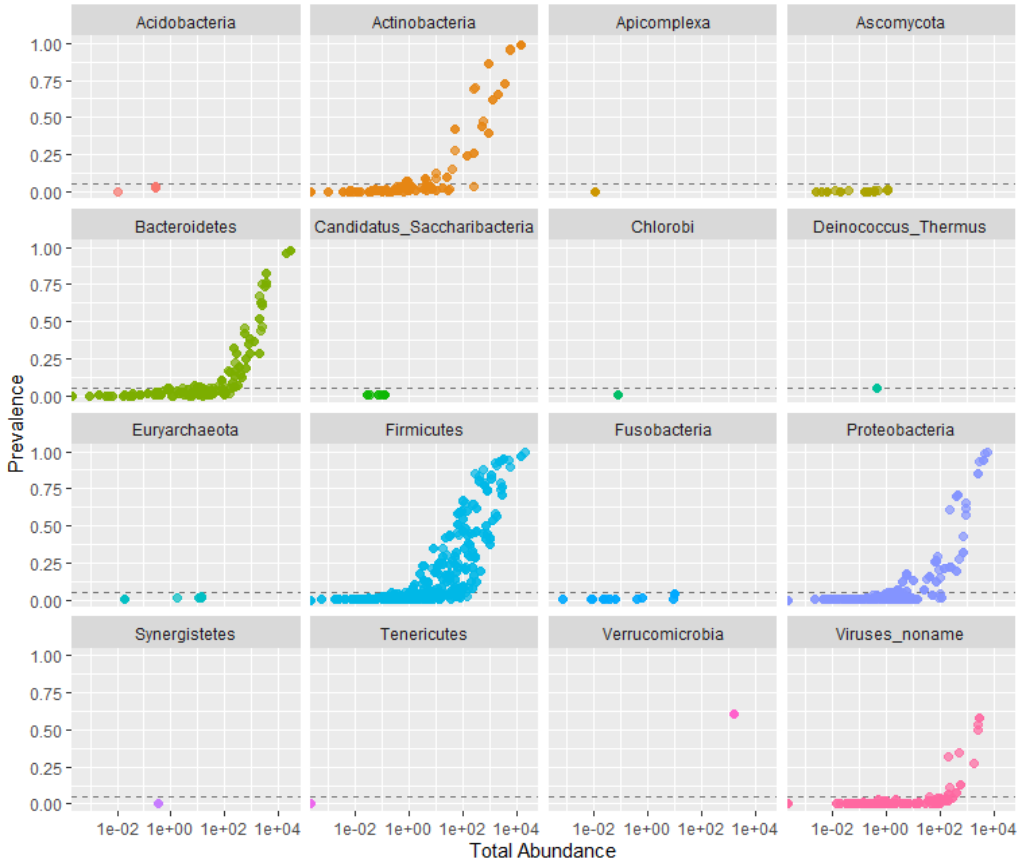

Figure 1. Taxa prevalence and total counts at phylum level before filtering unique taxa

One of the challenges in this study was dealing with large class imbalance present in the dataset. Class imbalance occurs when the dataset has a considerable disproportion of samples belonging to different classes. In general, imbalanced data results in reduced model accuracy where the model is biased towards the majority class. This is a serious concern, since we are typically most interested in the correct classification of the minority class. By decreasing the class imbalance, the overall accuracy of the learning models increases. In particular, prediction of the minority class is improved [13], [14]. After data preprocessing was performed, 85% samples belonged to the healthy class and 15% belonged to otitis. The class imbalance problem was handled by simply generating new sample data by randomly oversampling the minority class (i.e. otitis). This resampling method was selected due to its simplicity, which has also shown to yield successful results. In addtion, many resampling methods have been observed to perform similarly [15]. Otitis samples were generated to be equal in numbers to the healthy samples. Afterwards, to optimize the hyperparameters we retrained the classifiers models with different configurations.

For comparison purposes, classification was performed using both imbalanced and balanced data. Finally, to identify the most informative microbiome taxa for the prediction of otitis, feature importance was performed.

2.3. Methods

We built five models for otitis classification: k-nearest neighbors (K-NN) [16], support vector machines (SVM) [17], artificial neural networks (ANN) [18], gradient boosting (GB) [19], and random forest (RF) [20]. These methods are some of the most common methods used in microbiome host trait prediction [21], [22]. For each method, hyperparameter optimization was performed. A brief summary of each method is as follows:

K-NN algorithm is one of the simplest machine learning methods that assignes the test sample to a label based on the nearest *k* training samples in the feature space. Nearest neighbours are computed by using a distance metric in a multidimensional feature space.

SVMs are a popular machine learning algorithm that apply a kernel which is a mathematical function that maps the inputs into a multidimensional space. The model aims to maximize the margin between the samples while minimizing the error when separating the data. Some common SVM kernels include linear, polynomial, sigmoid and radial basis function kernels.

ANN algorithm is another frequently used algorithm applied to medical problems. ANN is made up of interconnected neurons to form a network. The general architecture consists of an input layer, hidden layer(s), and an output layer. The input neurons are task specific (e.g. OTUs for microbiome data), however the number of hidden layers need to be determined. Neurons in each layer are connected to the neurons in the next layer by a weighted connection. During the training phase, ANN aims to minimize the error between the desired output and the predicted output by adjusting the weights. Weight adjustment is done using the backpropagation algorithm that uses the gradient descent method to minimize the error.

Gradient boosting is an ensemble method that applies boosting to iteratively update the weights for each weak learner (in general decision trees) in order to improve the predictions of misclassified samples. The algorithm uses the gradient descent method on a loss function to update the predictions.

RF method is another ensemble method that builds a "forest" of decision trees to make a prediction. For each decision tree, random forest applies bootstrap aggregation, where subsets of randomly selected samples are used for the training set. In addition, each decision tree is trained with a random subset of features. The final decision is made by combing the results of many decision trees by computing the majority vote for classification or by averaging decision tree results for regression tasks.

3. Results and discussion

To estimate the model performance and mitigate overfitting, stratified k-fold (k=10) cross-validation was used. Briefly, k-fold cross-validation randomly splits the data into k folds (i.e. subsets) approximately of the same size, where the data in k-1 folds is used for the training set and the data in the remaining fold is used for the test set. Due to stratification, each fold approximately contains equal proportion of each class as in the complete dataset. The model is then trained on the training set and validated on the test set. The procedure is repeated k times, where each time, different fold is used for validation and remaining folds for training. Performance results are obtained by averaging the validation results.

The most common performance measures applied for diagnostic predictors are measured in terms of the overall accuracy, sensitivity, and specificity defined as follows:

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \quad (4)$$

where true positives indicate correctly classified otitis samples; true negatives indicate correctly classified healthy samples; false positives indicate incorrectly classified otitis samples, and false negatives indicate incorrectly classified healthy samples. Accuracy refers to the correctly classified samples. Sensitivity refers to the ability of a test to correctly classify an individual as *diseased* (also called true positive rate or TPR). Specificity is the ability of a test to correctly classify an individual as *disease-free* (also called true negative rate or TNR).

Performance results using class imbalanced data are displayed in Table 1. Overall accuracy of 0.85 was achieved by all tested classifiers. Furthermore, the results show that classifiers performed well when classifying the majority class, i.e. healthy cases. Obtained specificity for all methods was 0.93 and above.
However, a poor classification was observed across all classifiers to correctly classify otitis samples (i.e. the rare occurrences in our dataset that we are most interested in predicting): sensitivity results range from 0.0 for K-NN method, to 0.16 achieved by GB.

Table 1. Binary classification results using class imbalanced data

Algorithm	Accuracy	Sensitivity	Specificity
K-NN	0.85	0.0	0.97
SVM	0.85	0.01	0.93
ANN	0.85	0.0	100
RF	0.85	0.06	0.98
GB	0.85	0.16	0.96

Table 2. shows the results of each classifier on balanced class data. The table also shows optimized hyperparameter values. With a balanced class size data, all classifiers achieved high overall accuracy: RF 0.99, SVM and GB 0.96, K-NN 0.93, ANN 0.89. All methods except GB (with very low specificity) also achieved high sensitivity and specificity results. Sensitivity: SVM and K-NN 0.99, RF 0.97, GB 0.92 and ANN 0.81. Specificity: GB 0.1, RF 0.99, ANN 0.98, SVM 0.93, KNN 0.87. The results show that the sensitivity score for all classifiers was considerably increased when compared with the sensitivity score obtained with class imbalanced data.

Table 2. Binary classification results using class balanced data

Algorithm	Hyperparameter Configuration	Accuracy	Sensitivity	Specificity
K-NN	K = 1	0.93	0.99	0.87
SVM	kernel: RBF, cost:64, gamma: 0.00471	0.96	0.99	0.93
ANN	hidden layers: 1 (5 hidden nodes), learning rate 0.01	0.89	0.81	0.98
RF	no. of features: 2, no. of trees: 500	0.99	0.97	0.99
GB	no. of trees: 250, interaction depth: 5, shrinkage (learning rate): 0.1, minobsinnode no. (min. no. of observations in trees' terminal nodes) : 10	0.96	0.92	0.1

From the tested methods (Table 2.), random forest achieved the highest overall accuracy (0.99) as well as high scores for both sensitivity (0.97) and specificity (0.99). In general, random forest is one of the most successful machine learning algorithms. Main reasons for its successful performance is that it is an ensemble-based method that successfully combines bootstrap aggregation (i.e. each model performs bootstrap sampling of the original training data) with random feature selection to provide the essential diversity and low correlation between the decision trees that tends to reduce both overfitting and error due to bias. Feature importance was thus performed using the random forest classifier to establish microbial taxa that have the most predictive power in the classification of otitis. Fig. 2, shows 20 most important features (OTUs). Identified OTUs (representing species-level abundance of particular bacterial taxon) were then mapped to the genus level using the taxonomy table and are as follows: *Villanelle, Streptococcus, Siphoviridae, Clostridium, Sutterella,*

Parasutterella, Neisseria, Granulicatella, Eubacterium, Subdoligranulum, Bacteroides, Escherichia, Blautia, and Bifidobacterium. From the 14 identified bacteria, the highest impact to the classification accuracy was found to be the abundance of *Villanelle, Clostridium* and *Bifidobacterium* bacteria. The identified bacteria can further be explored by the microbial community researchers as potential microbial biomarkers for otitis.

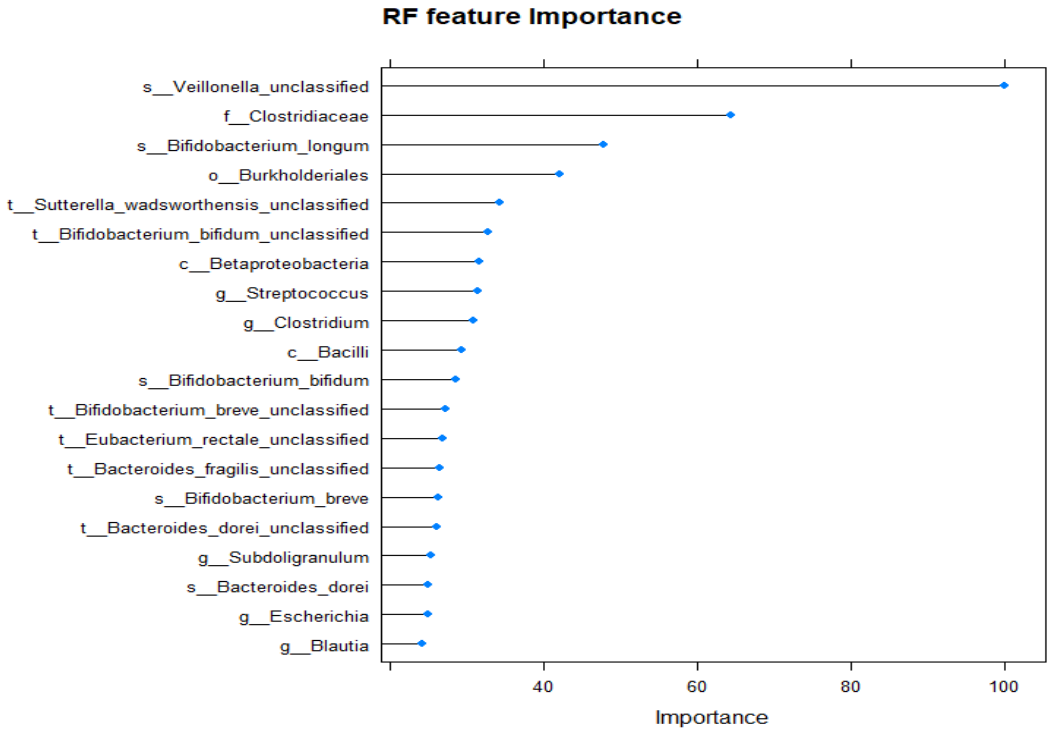

Figure 2. Random Forest feature importance plot

4. Conclusion

In this work we performed classification of otitis in infants aged 33 days to 3 year using otitis microbiome data obtained from the *curatedMetagenomicData*. After the initial exploratory analysis, data preprocessing was performed that included missing values removal, data normalization and outlier removal by removing low prevalence bacteria at the phylum level, (using a threshold of 0.01%). We then built five machine learning models for otitis classification: support vector machine, k-nn, artificial neural networks, random forest and gradient boosting. Even though achieved overall accuracy for all tested models was 0.85, very low sensitivity was observed across all models (between 0.0 – 0.16). Due to the highly class-imbalanced data, oversampling the minority class was done using random sampling to handle the class imbalance. Afterwards, hyperparameter tuning was performed for each model. This resulted in an increased overall accuracy with random forest method outperforming other methods (0.99). More importantly sensitivity was substantially increased for all models. For the RF model, sensitivity increased from 0.06 for imbalanced class size data to 0.97 for balanced class size data. The results of the specificity were also slightly improved (from 0.98 to 0.99). High results obtained by the random forest method are in line with the multiclass disease classification results obtained by Khan et al. [8] on the same dataset (*curatedMetagenomicData*). High results were achieved when the authors performed binary classification of non-healthy vs. healthy samples (0.99 overall accuracy achieved by RF). However, authors achieved low accuracy for otitis disease in multiclass classification (highest accuracy of 0.13 was obtained by deep neural network model).

Finally, we identified 14 genus bacteria that have the most impact on the target class. From the identified 14 bacteria, *Villanelle, Clostridium and Bifidobacterium* bacteria were the three highest ranking bacteria found to be most predictive of the target class.

The identified bacteria may further be used by the researchers towards personalized medicine as a potentially modifiable novel otopathogens for bacterial therapeutics to treat otitis in infants. Future work can also include other demographic-related data such as age, gender, and country to further analyze the relationship between the gut microbiome and otitis media disease. This analysis would help to answer to what extent does demographic-related data affect the development of microbiome and hence the health status of an individual.

Declaration of competing interest

The authors declare that they have no any known financial or non-financial competing interests in any material discussed in this paper.

References

[1] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch and R. Knight, "Current understanding of the human microbiome," *Nature Medicine*, vol. 24, no. 4, pp. 392-400, 2018.

[2] T.H. Nguyen, E. Prifti, N. Sokolovska and J.-D. Zucker, "Disease Prediction Using Synthetic Image Representations of Metagenomic Data and Convolutional Neural Networks," in *IEEE-RIVF International Conference on Computing and Communication Technologies* RIVF, pp. 1-6, 2019.

[3] R. Sender, S. Fuchs and R. Milo, "Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans," *Cell*, vol. 164, no. 3, pp. 337-340, 2016.

[4] Beck, D., Foster, J.A. "Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis," *BioData Mining***,** vol. 8, pp. 23, 2015. doi: 10.1186/s13040-015-0055-3

[5] Z. Zhu, J. Ren, S. Michail and F. Sun, "MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations," *Genome Biology*, vol. 20, no. 1, pp. 154, 2019.

[6] C. S. Casimiro-Soriguer, C. Loucera, M. Peña-Chilet and J. Dopazo, "Interpretable machine learning analysis of functional metagenomic profiles improves colorectal cancer prediction and reveals basic molecular mechanisms," *Scientific reports*, vol. 12, no. 1. pp. 450, 2022.

[7] T. Vatanen, A. D. Kostic, E. d'Hennezel, H. Siljander, E. A. Franzosa, M. Yassour, R. Kolde, H. Vlamakis, T. D. Arthur, A. P. A. Hamalainen, V. Tillmann, R. Uibo, S. Mokurov and et. al, "Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans," *Cell*, vol. 165, no. 4, pp. 842-853, 2016.

[8] S. Khan and L. Kelly, "Multiclass Disease Classification from Microbial Whole-Community Metagenomes," *Pacific Symposium on Biocomputing*, vol. 25, pp. 55-66, 2020.

[9] E. Sayyari, B. Kawas and S. Mirarab, "TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification," *Bioinformatixs (Oxford, England)*, vol. 35, no. 14, pp. 31-40, 2019.

[10] G. K. John and G. E. Mullin, "The Gut Microbiome and Obesity," *Currentn oncology reports*, vol. 18, no. 7, pp. 45, 2016.

[11] E. Pasolli, L. Schiffer, P. Manghi, et al. "Accessible, curated metagenomic data through ExperimentHub," *Nature Methods*, vol. 14, no. 11, pp. 1023-1024, 2017. doi:10.1038/nmeth.4468

[12] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, E. Abebe, "Defining operational taxonomic units using DNA barcode data," *Philosophical transactions of the Royal Society of London, Series B, Biological Sciences*, vol. 360, no. 1462, 2005. doi: 10.1098/rstb.2005.1725

[13] G. Menardi and N. Torelli, "Training and assessing classification rules with unbalanced data," *Data Mining and Knowledge Discovery*, vol. 2014, no. 1, pp. 92-122, 2014. doi: 10.1007/s10618-012-0295-5

[14] J.M. Johnson, T.M. Khoshgoftaar, "Survey on deep learning with class imbalance." *Journal of Big Data*, vol. 6, no. 1, pp. 27, 2019. doi.org/10.1186/s40537-019-0192-5

[15] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "*A study of the behavior of several methods for balancing machine learning training data,*" ACM SIGKDD Explorations Newsletter vol. 6, no. 1, pp. 20-29, 2004.

[16] T. Cover, P. Hart, "Nearest neighbor pattern classification", IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27. 1967. DOI: 10.1109/TIT.1967.1053964

[17] S. Haykin, Neural Networks and Learning Machines, Pearson Education, Inc., New Jeresey, 2009. ISBN-10: 0131471392

[18] S. Dreiseitl and L. Ohno-Machado "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review," *Journal of Biomedical Informatics*, vol. 35, no. 5, pp. 352-359, 2002. doi: 10.1016/S1532-0464(03)00034-0.

[19] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367-378, 2002. doi: 10.1016/S0167-9473(01)00065-2

[20] G. Biau and E. Scornet. "A random forest guided tour", *TEST*, vol. 25, no. 2, pp. 197-227, 2016. doi:10.1007/s11749-016-0481-7

[21] Y.-H. Zhou and P. Gallins, "A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction, *Frontiers in Genetics*, vol. 10, 2019

[22] L. J. Zambrano, K. Karaduzovic-Hadziabdic, T. Loncar Turukalo Tatjana, et al., "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment," *Frontiers in Microbiology*, vol. 12, 2021. doi: 10.3389/fmicb.2021.634511