RESEARCH

Randomized Feature Selection Based Semi-Supervised Latent Dirichlet Allocation for Microbiome Analysis

Namitha V Pais[1*], Nalini Ravishanker[1], Sanguthevar Rajasekaran[2], George Weinstock[3] and Dong-Binh Tran[3]

[1]Department of Statistics, University of Connecticut, Storrs, CT, USA
[2]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
[3]Jackson Laboratory for Genomic Medicine, Farmington,CT, USA

*Correspondence:
namitha.pais@uconn.edu
[1]Department of Statistics,University of Connecticut, Storrs, CT, USA
Full list of author information is available at the end of the article

Abstract

Background: Health and disease are fundamentally influenced by microbial communities and their genes (the microbiome). An in-depth analysis of microbiome structure that enables the classification of individuals based on their health can be crucial in enhancing diagnostics and treatment strategies to improve the overall well-being of an individual.

In this paper, we present a novel semi-supervised methodology known as Randomized Feature Selection based Latent Dirichlet Allocation (RFSLDA) to study the impact of the gut microbiome on a subject's health status. Since the data in our study consists of fuzzy health labels, which are self-reported, traditional supervised learning approaches may not be suitable. As a first step, based on the similarity between documents in text analysis and gut-microbiome data, we employ Latent Dirichlet Allocation (LDA), a topic modeling approach which uses microbiome counts as features to group subjects into relatively homogeneous clusters, without invoking any knowledge of observed health status (labels) of subjects. We then leverage information from the observed health status of subjects to associate these clusters with the most similar health status making it a semi-supervised approach. Finally, a feature selection technique is incorporated into the model to improve the overall classification performance.

Results: The proposed method provides a semi-supervised topic modelling approach that can help handle the high dimensionality of the microbiome data in association studies. Our experiments reveal that our semi-supervised classification algorithm is effective and efficient in terms of high classification accuracy compared to popular supervised learning approaches like SVM and multinomial logistic model.

Conclusion: The RFSLDA framework is attractive because it (i) enhances clustering accuracy by identifying key bacteria types as indicators of health status, (ii) identifies key bacteria types *within each group* based on estimates of the proportion of bacteria types within the groups, and (iii) computes a measure of within-group similarity to identify highly similar subjects in terms of their health status.

Keywords: classification; feature selection; health status; microbiome; topic modeling

1 Introduction

Humans coexist with trillions of single-cell organisms living within their bodies, labeled as human microbiota or microbiome. Most of these organisms reside in the human gut, where most are bacteria, although viruses and fungi are also part of the microbiome. These microbes can be symbiotic or pathogenic and coexist without conflict in a healthy body. However, changes in diet, antibiotics, etc., can disturb this balance, making the body more susceptible to diseases. Recent advances in human microbiome research show evidence of the impact of microbiomes on the host's well-being [2, 6]. The objectives of the research are to characterize the composition of a normal microbiome in healthy individuals, and to investigate similarities and differences between individuals by characterizing microbiomes based on their core functions, ecological characteristics, or temporal dynamics.

With recent developments in machine learning techniques, researchers have adopted several computational approaches to diagnose and understand the microbiome data and its implications on human health. [13] discussed modeling techniques for microbiome profiling to obtain a biologically-interpretable mathematical formula for predicting the likelihood of disease. [4, 8] discussed and compared standard machine learning approaches in terms of predictive accuracy and interpretability. Significant advances in microbiome research are being made in order to address characteristics of the human microbiome structure such as high diversity and presence of rare bacteria types, as well as to handle insufficient samples of individuals at risk of several diseases.

In this paper, we discuss a novel semi-supervised topic modeling approach that analyzes patterns in the microbiome data to identify relatively homogeneous groups and compares their association with observed health status (fuzzy labels), to classify subjects based on health status. We begin by exploring an unsupervised topic modeling approach that provides a powerful tool for discovering and exploiting the hidden structure in the microbiome data. Given the microbiome counts in the subject's gut, we are interested in checking whether these counts can be used as features to group (cluster) subjects without any prior information about their health status. Subsequently, we can classify the subjects into different health status levels by assessing the similarity between the observed health status and the clusters. We develop a methodology called "Randomized Feature Selection based Latent Dirichlet Allocation" (RFSLDA) that identifies important bacteria types to distinguish between different levels of health status and classify subjects based on their counts. Our method (i) provides a semi-supervised topic modeling approach to classify subjects into different health status based on their gut-microbiome composition and, (ii) provides a feature selection technique to identify important bacteria types from the high dimensional microbiome data to improve model performance substantially. Experimental results indicate that our algorithm is very accurate.

2 Data Description

We analyze data provided by the Jackson Laboratory in Farmington, CT to examine how microbiomes affect human health. Medical professionals collected blood from $M = 89$ subjects for host molecular omics profiling. They also collected two types of samples (stool and nasal swabs). Microbiome profiling then recorded counts on each

$B = 109$ bacteria. Suppose $Y_{i,\ell}$ for $i = 1, 2, \ldots, M$, and $\ell = 1, 2, \ldots, B$ represents a count of the ℓth type of bacterium of the ith subject. Initial exploration shows that the observed read counts $Y_{i,\ell}$ exhibit a wide range of values. Following expert opinion that it may be counter-productive to include all bacterial types into the data analysis, we identify *top* bacterial types based on two measures, i.e., *abundance* and *prevalence*. While abundance captures the composition of bacteria types in the microbiome, prevalence captures the presence/absence of bacteria types above a given detection threshold.

Let $\mathbb{P} = \{p_{i,\ell}\}$ be an $M \times B$ matrix of proportions of the ℓ^{th} bacterium in subject i, where,

$$p_{i,\ell} = \frac{Y_{i\ell}}{\sum_{\ell=1}^{B} Y_{i\ell}}. \tag{1}$$

For bacterial type $\ell = 1, 2, \ldots, B$, abundance and prevalence are defined by

$$\mathcal{A}_\ell = \sum_{i=1}^{M} p_{i,\ell}, \text{ and} \tag{2}$$

$$\mathcal{P}_\ell = \sum_{i=1}^{M} I_{i\ell}, \text{ where } I_{i,\ell} = \mathbf{1}[p_{i,\ell} \geq \omega]. \tag{3}$$

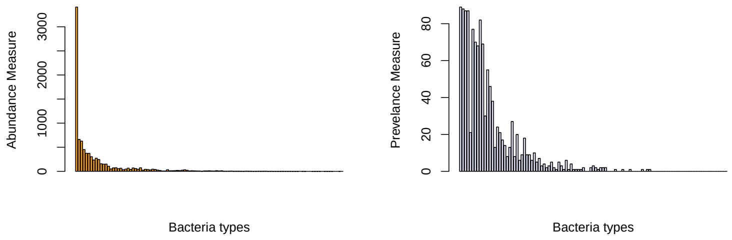

Figure 1: Empirical distributions of bacterial abundance and prevalence measures.

Figure 1 shows the abundance (left) and prevalence (right) for the $B = 109$ bacteria types. We see that some types are highly abundant, while some (possibly other) types are highly prevalent.

We use the tau-path method [14, 15] for identifying the top $K = 50$ bacteria types determined by the high concordance between their abundance and prevalence measures (details are shown in the Appendix). We aggregate the remaining bacteria types into a single category labeled "Others". Let $B_0 = K + 1 = 51$. Let $Z_{i,l}$, $i = 1, 2, \ldots, M$ and $\ell = 1, 2, \ldots, B_0$ denote counts on the top bacteria types on M subjects. The left plot in Figure 2 displays the distribution of all bacteria based on the average read counts across all subjects, while the right plot shows that the the top bacteria types have reasonably sufficient read counts.

In the following sections, we provide details on our RFSLDA algorithm which uses microbiome read counts on $B_0 = 51$ types to group subjects based on their

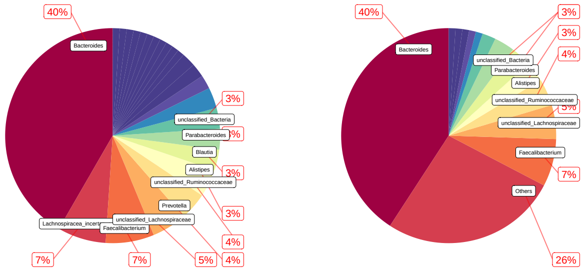

Figure 2: Pie charts represent total percentages of average read counts taken over all the subjects for (i) all the 109 bacteria types and (ii) the top 50 bacteria types obtained using the Tau Path Method.

health status. For convenience, Table 1 presents the notations used in the rest of the article.

3 Randomized Feature Selection Based Latent Dirichlet Allocation

This section describes our algorithm for classifying subjects into different groups which characterize the relationship between their gut micriobiome and health status. First, we build a latent Dirichlet allocation (LDA) model which uses only the microbiome counts as features to determine relatively homogeneous clusters in an *unsupervised way*, yielding latent topic labels for the subjects (Section 3.1.) Second, we do *semi-supervised LDA* to match information on observed health-status labels of subjects with their latent topic labels from Section 3.1 in order to classify them into different levels (see Section 3.2). A final *feature selection* step in Section 3.3 helps us to identify bacteria types which optimally drive the classification of subjects into suitable health status groups.

3.1 Unsupervised latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) [1] is an unsupervised, mixed-membership model mainly used in document analysis. LDA assumes T unobserved topics (clusters) associated with a collection of subjects where each subject exhibits these topics in different proportions. This model uses the observed microbiome counts as features to infer the hidden topic structure of each subject. Suppose we have a corpus $\mathcal{C}$ consisting of M subjects where the observed microbiome counts on each of the subjects are represented as $\mathcal{D} = (b_1, b_2, \ldots, b_{B_0})$. If $N = \sum_{l=1}^{B_0} b_l$ is the total microbiome count, then one can also represent $\mathcal{D}$ as $\mathcal{D} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)$ where $\boldsymbol{w}_n$ corresponds to the n^{th} bacterium present in the subject's gut represented by a $B_0 \times 1$ vector corresponding to the v^{th} bacteria type such that $w^v = 1$ and $w^u = 0$ for $u \neq v$. LDA assumes the following generative process for each subject $\mathcal{D}_d$ with total microbiome count N_d, for $d \in \{1, 2, \ldots M\}$ present in the corpus $\mathcal{C}$:

1 Choose $N_d \sim Pois(\lambda)$

2 Choose $\boldsymbol{\theta}_d \mid \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$ where,
$\boldsymbol{\theta}_d = (\theta_1, \theta_2, \ldots, \theta_J)$ and $Dir(.)$ is a symmetric Dirichlet distribution.

3 For each of the N_d bacterium, $\boldsymbol{w}_{d,n}$

Notation	**Description**
M	Number of subjects.
B	Number of bacteria types.
$Y_{i,\ell}$	Overall read count of ℓ^{th} type of bacterium on the i^{th} subject.
C_i	Health status of the i^{th} subject.
T	Number health status levels/ Number of latent topics identified using LDA.
$p_{i,l}$	Proportion of ℓ^{th} bacterium level in subject i's microbiome .
$\mathbb{P}$	$M \times B$ proportion matrix .
ω	Detection threshold.
$I_{i,l}$	Indicates the presence or absence of ℓ^{th} bacterium level in subject i's microbiome based on detection threshold ω.
$\mathbb{I}$	$M \times B$ incidence matrix .
$\mathcal{A}$	Abundance of ℓ^{th} bacterium type.
$\mathcal{P}$	Prevalence of ℓ^{th} bacterium type.
K	top bacteria types obtained using the tau-path method.
B_0	$K+1$.
$Z_{i,\ell}$	Overall read count of ℓ^{th} top bacterium type on the i^{th} subject.
b_l	observed count of ℓ^{th} top bacterium type.
$\boldsymbol{w}_n$	$B_o \times 1$ vector indicates the bacteria type present in the n^{th} position of $\mathcal{D}$.
$\pi^{'}_{t,i}$	Estimated topic proportions on subject i for topic t.
π_i	Topic label corresponding to subject i.
w_t	Weights assigned to class level t.
λ	proportion threshold to identify within-group similarity.

Table 1: Description of Notations

(a) Choose a topic $\boldsymbol{Z_{d,n}} \mid \boldsymbol{\theta}_d \sim Mult(1, \boldsymbol{\theta_d})$

(b) Choose a bacterium $\boldsymbol{w}_{d,n} \mid \{\boldsymbol{Z}_{d,n}, \boldsymbol{\beta}\} \sim Mult(1, \boldsymbol{\beta}_{\boldsymbol{Z}_{d,n}})$, a multinomial probability distribution conditioned on the topic $\boldsymbol{Z}_{d,n}$.

We use the LDA model to obtain $T = 3$ hidden topics (clusters) using the read counts from the $B_0 = 51$ bacteria types on $M = 89$ subjects. Once the structure of the model is defined, the goal is to estimate the model parameters and compute the posterior distribution for inference. The LDA model estimation is done using the R package *topicmodels* [3], which uses variational expectation maximization (VEM) algorithm to estimate the model parameters and variational inference (VI) algorithm to approximate the posterior distribution. Details are shown in the Appendix.

The latent topical structure is represented by the estimated topic proportions $\pi'_{t,i}$ on each subject, for $i = 1, 2, \ldots, M$ and $t = 1, 2, \ldots T$. Estimating the per-subject topic proportion using LDA allows us to associate subjects with multiple topics, unlike many clustering algorithms that assign one topic per subject. In addition, the topic with the highest proportion is assigned as the topic label $\pi_i = argmax\{\pi'_{t,i}\}$ for a given subject. While π_i enables us to group the subjects into T topics (clusters), $\pi'_{t,i}$ helps us compare the degree of similarity of subjects within clusters.

Using the LDA model, we group 57 subjects into cluster 1, 15 subjects into cluster 2 and 17 subjects into cluster 3. These clusters via their compositional proportions $\pi'_{t,i}$ are shown in Figure 3.

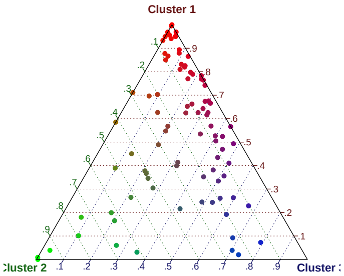

Figure 3: Clustering $M = 89$ subjects into $T = 3$ clusters using the LDA model

3.2 Semi-supervised latent Dirichlet allocation

In addition to the bacteria counts, we obtain information on observed health status C_i for each subject. These were obtained by the medical professional through a battery of molecular and clinical laboratory tests, complemented by self-reported online surveys which documented changes in medication, physical activity, diet preference, and perceived stress level. Being self-reported, the health status levels can be considered *fuzzy* labels that provide only ballpark information about an subjects's health. We have retained the health status levels *Healthy* and *Infection*, but grouped a few levels with insufficient data that indicated medical stress (immunization, antibiotics, travel, fiber, colonoscopy, surgery, weight gain, weight loss, stress, and allergies) into a single level, *Stress*. The distribution of observed health status is shown in Figure 4. Despite the rare occurrence of the level *infection*, it is an extremely important level to identify.

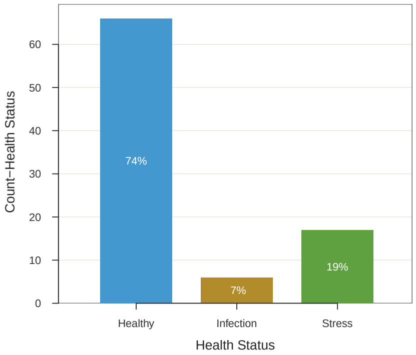

Figure 4: Barchart: Observed health status

We bring in the information from the observed health labels to provide a qualitative interpretation of the clusters learned from the topic model approach. We predict the health status levels of subjects by associating the *topic labels* π_i to the *observed* health status C_i, for $i = 1, 2, \ldots, M$. This is achieved by considering all the $T!$ matches between the topic labels and observed health status, and calculating a classification metric for each scenario defined as,

$$A_w = \sum_{t=1}^{T} w_t \times TPR_t, \tag{4}$$

where w_1, w_2, w_3 are weights such that $\sum_{i=1}^{T} w_t = 1$, and TPR_t corresponds to the true positive rate (TPR) indicating the proportion of correct predictions for class t where, $t = 1, 2, \ldots, T$. Since we have unbalanced classes (see Figure 4), we use weighted accuracy A_w as a metric. The subjects are then classified into different health status levels based on the optimal match identified by the metric. As a result, our framework becomes semi-supervised.

We conduct a grid search and set $w_1 = 0.6, w_2 = 0.15$ and $w_3 = 0.25$. Considering the subject-specific microbiome data with $B_0 = 51$ bacterial types, we are able to only achieve a weighted accuracy of 46.67% with just one infected subject correctly classified. The corresponding confusion matrix is shown in Table 2.

3.3 RFSLDA Algorithm

We incorporate a feature selection technique [12] into the semi-supervised LDA model in Section 3.2 in order to identify an optimal subset of bacteria types and improve model performance. That is, we identify a subset of the B_0 bacteria types that are most important in improving the LDA based classification for determining a subject's health status, by eliminating bacteria types which reduce the predictive power of the classifier. We develop a *Randomized Feature Selection based Latent Dirichlet Allocation*(RFSLDA) algorithm; see Algorithm 1.

		Actual		
		H	I	S
Predicted	H	44	4	9
	I	10	1	4
	S	12	1	4

Table 2: Confusion matrix obtained from semi-supervised LDA using data from $M = 89$ subjects

Algorithm 1 Randomized Feature Selection based Latent Dirichlet Allocation

```
Input: The set $\mathcal{F}$ of all possible features and LDA algorithm $\mathcal{L}$
Output: A near optimal subset $\mathcal{F}'$ of features.
Parameters: Total number of features: $n$, Initial number of features: $p$, Threshold: $t_0$ and constant: $c$.
Randomly sample a subset $\mathcal{F}'$ of features from $\mathcal{F}$ of length $p$.
Run the LDA algorithm $\mathcal{L}$ using the features in $\mathcal{F}'$.
Compute the weighted accuracy $A$ after matching the latent topics of LDA to the observed health status.
repeat
    Flip an unbiased three-sided coin with sides 1, 2, and 3.
    if (the outcome of the coin flip is 1) then
        Choose a random feature $f$ from $\mathcal{F} - \mathcal{F}'$ and add it to $\mathcal{F}'$.
        Remove a random feature $f'$ from $\mathcal{F}'$ to get $\mathcal{F}''$.
    else if (the outcome of the coin flip is 2) then
        Choose a random feature $f$ from $\mathcal{F} - \mathcal{F}'$ and add it to $\mathcal{F}'$ to get $\mathcal{F}''$.
    else
        Remove a random feature $f$ from $\mathcal{F}'$ to get $\mathcal{F}''$.
    end if
    Run the LDA algorithm $\mathcal{L}$ using the features in $\mathcal{F}''$.
    Compute the weighted accuracy $A'$ after matching the latent topics of LDA to the observed health status.
    if $(A' > A)$ then
        set $\mathcal{F}' := \mathcal{F}''$ and $A := A'$; Perform the search from $\mathcal{F}'$.
    else
        with probability $u$ perform the search from $\mathcal{F}'$ and
        with probability $1 - u$ perform the search from $\mathcal{F}''$ with $A := A'$.
        A relevant choice for $u$ is $\exp\{-c(A - A')\}$ for some constant c.
    end if
until no significant improvement in the weighted accuracy can be obtained, i.e., $|A - A'| < t_0$.
Output $\mathcal{F}'$
```

We use a grid search to obtain tuning parameter values of $n = B_0$, $p = \frac{n}{2}, t_0 = 0.0001$ and $c = 1$. We then repeat Algorithm 1 $R = 50$ times to ensure that the optimal subset of bacteria types chosen is not sensitive to the choice of the initial subset. At each iteration, we record optimal set of bacteria types and the corresponding accuracy. We identify the iteration with the highest weighted accuracy and select the corresponding subset of bacteria types as the optimal subset. This step improves the overall weighted accuracy of grouping to 72.12%, which is significantly higher than the weighted accuracy of 46.67% (shown in section 3.2) we obtained before feature selection. The corresponding confusion matrix is shown in Table 3.

		Actual		
		H	I	S
Predicted	H	50	1	5
	I	8	4	3
	S	8	1	9

Table 3: Confusion matrix obtained from RFSLDA using data from $M = 89$ subjects

The RFSLDA algorithm correctly classifies four out of six infected subjects (which is a crucial in medical practice) and improves the overall model performance. It may be less problematic that a few healthy subjects are misclassified as infected or stressed.

4 Interpreting results from the RFSLDA algorithm

We present the interpretation of results from our RFSLDA algorithm.

4.1 Optimal Bacteria Types by health status levels

The feature selection step of the RFSLDA algorithm enables us to select a subset of the bacteria types that are the most important in determining the health status. These are: *Clostridium.XlVa, unclassified_Bacteroidales, Paraprevotella, Dialister, Coprococcus, Enterobacter, unclassified_Fusobacteriaceae, Gemella, unclassified_Proteobacteria and, Comamonas.* The composition of these bacteria types on each of the subjects are shown in Figure 5.

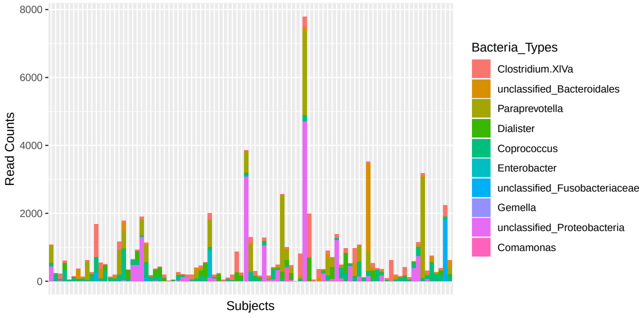

Figure 5: Composition of optimal bacteria types across subjects based on their read counts

From the RFSLDA algorithm, we obtain estimates on the distribution of the different bacteria types by health status levels. Figure 6 shows the selected bacteria types across each health status level, along with their estimated proportions. We see that unclassified_Proteobacteria is most commonly found in subjects classified under infection, with an estimated proportion of 0.9, while Paraprevotella is more commonly found in subjects classified as stress, with an estimated proportion of 0.67. Among healthy subjects, the bacteria types Clostridium.XlVa and Coprococcus are predominant, with estimated proportions 0.35 and 0.33, respectively.

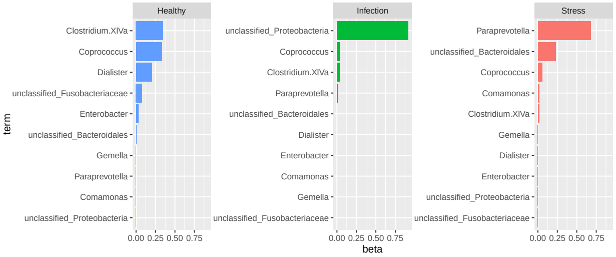

Figure 6: Top bacteria across different health status with their estimated proportions

4.2 Within group similarity of subjects by health status

We evaluate the within-class similarity using information from the estimated topic structure $\pi'_{t,i}$. Let $\pi_{i,[1]}$ and $\pi_{i,[2]}$ denote the highest and second highest topic proportions for each subject and let $\lambda = 0.7$ denote a user-specified threshold.

Suppose we consider the subjects classified as healthy.

1 Subjects with maximum estimated topic proportion $\pi_{i,[1]} \geq \lambda$ are grouped as similar, and called *Predominantly Healthy*.

2 Subjects with $\pi_{i,[1]} < \lambda$, and $\pi_{i,[2]}$ associated to health status level *infection* (or,*stress*) are grouped similar and called *Healthy-Infection* (or, *Healthy-Stress*).

As an example, suppose we have four subjects with estimated topic proportions associated to stress, healthy, and infection as $S_1 = (0.2, 0.75, 0.05)$, $S_2 = (0.15, 0.5, 0.35)$, $S_3 = (0.001, 0.9, 0.09)$, and $S_4 = (0.2, 0.4, 0.4)$. Each subject is classified as healthy according to the maximum estimated topic proportion. However, within the healthy group, we are able to identify that Subjects 1 and 3 are more similar to each other, and Subjects 2 and 4 are more similar to each other.

To visualize three-dimensional topic proportions, we employ the t-distributed stochastic neighbor embedding (t-SNE) method [7], to represent similar objects by nearby points and dissimilar objects by distant points. The t-SNE plot is shown in Figure 7.

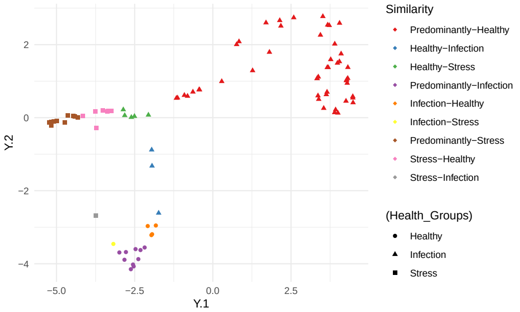

Figure 7: Visualizing the within-group similarity across each health status level using the t-SNE method

5 Comparison to Other Supervised Learning Methods

The semi-supervised RFSLDA method outperforms popular supervised learning methods like the multinomial logistic model [9] and support vector machines (SVM) [5]. To see this, we run all three methods using an 80-20 train-test split of the data. We fit the three methods to the training data with $M = 69$ subjects, and evaluate the fits on the test data with $M = 20$ subjects. The multinomial logit model is implemented using the R package *nnet* [11]. The SVM (with a linear kernel) is implemented using the R package *e1071* [10].

Tables 4a, 5a and 6a show the confusion matrices obtained from the three methods on the train data. The multinomial logistic regression and support vector machines (SVM) perform better than RFSLDA by perfectly classifying subjects into the health status levels. However, the usefulness of an approach and generalizability is best assessed by its performance on a test data. Tables 4b, 5b and 6b shows the confusion matrices obtained on the test data. By detecting one out of two subjects with level infection and two out of four subjects with level stress, RFSLDA performs substantially better than the supervised methods and avoids over-fitting. The RFSLDA model achieves a weighted accuracy of 63%, which surpasses the results obtained by the SVM model (60%) and the multinomial logit model (21%).

	H	I	S
H	52	0	0
I	0	4	0
S	0	0	13

(a) Confusion matrix on train data

	H	I	S
H	5	1	2
I	4	0	2
S	5	1	0

(b) Confusion matrix on test data

Table 4: Confusion matrices based on Multinomial logistic model

	H	I	S
H	52	0	0
I	0	4	0
S	0	0	13

(a) Confusion matrix on train data

	H	I	S
H	14	2	4
I	0	0	0
S	0	0	0

(b) Confusion matrix on test data

Table 5: Confusion matrices based on SVM

	H	I	S
H	41	1	3
I	5	3	3
S	6	0	7

(a) Confusion matrix on train data

	H	I	S
H	10	0	1
I	3	1	0
S	1	1	2

(b) Confusion matrix on test data

Table 6: Confusion matrices based on RFSLDA

6 Discussion and Conclusions

The current study using subject-specific data reveals that gut microbes play a fundamental role in human health. As an initial data preprocessing, we determine the top bacteria levels from the zero-inflated microbiome data based on abundance and prevalence to filter out potentially uninformative bacteria types using the tau-path method. The RFSLDA method then uses a semi-supervised LDA model to classify subjects based on the health status. Through a feature selection incorporated in our framework, we identify significant bacteria types that can accurately distinguish between different health statuses and enhance the accuracy of classification. We identify the top bacteria types in each class. Unclassified_Proteobacteria, for instance, is predominant in subjects classified as infected, but rare in subjects classified as healthy and stress. Further, using the estimated topic structure we obtain a within group similarity of subjects by the health status levels.

The results of a comparative study demonstrate that our RFSLDA method, despite being a semi-supervised approach, outperforms traditional supervised methods, such as SVM and Multinomial logit in identifying rare yet crucial health levels (infected, stress) in the test data. This is an important finding. Moreover, our results enable practitioners to identify the specific types of bacteria associated with each health status, providing valuable insight into the intricate connections between gut microbiome and human health.

The present study has some limitations. First, despite showing a significant relationship between gut microbiome profiles and health status, our model misclassifies 25% of healthy people as infected or stressed. Although we acknowledge this shortcoming, since the model is able to detect 4 out of 6 infected subjects correctly, we are willing to accommodate these false negatives. Moreover, since fuzzy labels are considered to represent the true observed health status of subjects, it could be possible for them to self-report themselves as healthy when they are not. Thus, it is imperative to investigate these misclassifications more closely in order to gain a deeper understanding of their health condition. Future research could further investigate the causal relationship between the gut microbiome and human health using a supervised topic modeling approach. Our RFSLDA approach can also be extended to a longitudinal data setup, where subjects visit the facility at different times.

Declarations

- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: Raw data included in this study provided by the Jackson Laboratory in Farmington, CT, are hosted on the NIH Human Microbiome 2 project site https://portal.hmpdacc.org.
- Competing interests : No competing interest is declared.
- Funding: Not applicable
- Authors' contributions: All authors have contributed equally in analyzing the data and reviewing the manuscript
- Acknowledgements: Not applicable

Author details
[1]Department of Statistics,University of Connecticut, Storrs, CT, USA. [2]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA. [3]Jackson Laboratory for Genomic Medicine, Farmington,CT, USA.

References

1. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

2. Cho Ilseung BMJ (2012) The Human Microbiome: at the interface of health and disease. Nature Reviews Genetics, vol 13
3. Grün B, Hornik K (2011) topicmodels: An R package for fitting topic models. Journal of Statistical Software 40(13):1–30,
4. Gupta Vinod K BU Kim Minsuk, et al (2020) A Predictive Index for Health Status using Species-level Gut Microbiome Profiling. Nature Communications, vol 11
5. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer
6. Lloyd-Price Jason HC Abu-Ali Galeb (2016) The Healthy Human Microbiome. Genome Medicine, vol 8
7. van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research 9(86):2579–2605, URL http://jmlr.org/papers/v9/vandermaaten08a.html
8. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, Berland M, Gruca A, Hasic J, Hron K, Klammsteiner T, Kolev M, Lahti L, Lopes MB, Moreno V, Naskinova I, Org E, Paciência I, Papoutsoglou G, Shigdel R, Stres B, Vilne B, Yousef M, Zdravevski E, Tsamardinos I, Carrillo de Santa Pau E, Claesson MJ, Moreno-Indias I, Truu J (2021) Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Frontiers in Microbiology 12,
9. McCullagh P (2019) Generalized linear models. Routledge
10. Meyer D, Wien F (2001) Support vector machines. R News 1(3):23–26
11. Ripley B, Venables W, Ripley MB (2016) Package 'nnet'. R package version 7(3-12):700
12. Saha S, Rajasekaran S, Ramprasad R (2015) Novel randomized feature selection algorithms. International Journal of Foundations of Computer Science 26(03):321–341
13. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD, Blaser MJ (2020) A Framework for Effective Application of Machine learning to Microbiome-Based Classification Problems. mBio 11(3):e00,434–20,
14. Yu L, Verducci JS, Blower PE (2011) The Tau-Path test for monotone association in an unspecified subpopulation: Application to chemogenomic data mining. Statistical Methodology 8(1):97–111, , URL https://www.sciencedirect.com/science/article/pii/S1572312710000079, advances in Data Mining and Statistical Learning
15. Zhang Y, Ravishanker N, Ivan J, Mamun S (2018) An Application of the Tau-Path Method in Highway Safety. Journal of the Indian Society for Probability and Statistics 20