Detect feature edges for diagnosis of bacterial vaginosis

Jie Li [1], **yaotang Li** [Corresp. 1]

[1] School of mathematics and statistics, Yunnan University, kunming, yunnan, China

Corresponding Author: yaotang Li
Email address: liyaotang@ynu.edu.cn

One of the most common diseases among women of reproductive age is bacterial vaginosis (BV). However, the etiology of BV is still unknown. In this paper, we use a network to model a temporal sample of the vaginal microbiome and study the relationship between the edges of the network and BV. We use the machine learning algorithms decision tree and ReliefF to select the network feature edges that are related to BV and then validate those features using logistic regression and support vector machine. We discover that a few features are required to achieve high BV classification accuracy; logistic regression and support vector machine performs nearly identically under the same feature edges; decision tree feature edges outperform ReliefF feature edges in classification performance, and the feature edges selected by those two algorithms are very different. The feature edges may serve as indicators for personalized diagnosis of BV and may aid in the clarification of a more mechanistic interpretation of its etiology.

Detect feature edges for diagnosis of bacterial vaginosis

Jie Li[1], YaoTang Li[1]
[1] School of mathematics and statistics, Yunnan University, Kunming, Yunan, China

Corresponding Author: YaoTang Li
YaoTang Li [1]
No. 2, Cuihu North Road, Wuhua District, Kunming, Yunan, 650091, China
Email address: liyaotang@ynu.edu.cn

Abstract

One of the most common diseases among women of reproductive age is bacterial vaginosis (BV). However, the etiology of BV is still unknown. In this paper, we use a network to model a temporal sample of the vaginal microbiome and study the relationship between the edges of the network and BV. We use the machine learning algorithms decision tree and ReliefF to select the network feature edges that are related to BV and then validate those features using logistic regression and support vector machine. We discover that a few features are required to achieve high BV classification accuracy; logistic regression and support vector machine performs nearly identically under the same feature edges; decision tree feature edges outperform ReliefF feature edges in classification performance, and the feature edges selected by those two algorithms are very different. The feature edges may serve as indicators for personalized diagnosis of BV and may aid in the clarification of a more mechanistic interpretation of its etiology.

Keywords: Bacterial vaginosis; machine learning; network; feature edges;

Introduction

Bacterial vaginosis (BV) has been identified to be an independent risk factor for women's health (Koumans et al. 2001), including preterm delivery and low infant birth weight, the development of pelvic inflammatory disease increased susceptibility to HIV infection, and other chronic health issues (Hay et al., 1994; Ness et al., 2005; Sha et al., 2005a; Atashili et al., 2008; van de Wijgert et al., 2008; Ma et al. 2012). BV is frequently characterized by changes in the vaginal microbiome; however, the causes of these changes are unknown (Redelinghuys et al.,2020). Historically, BV has been diagnosed using the Nugent score and/or Amsel's clinical criteria (Nugent et al., 1991; Amsel et al., 1983). The Nugent score is based on the presence or absence of lactobacilli on the Gram stain and generates a score ranging from 1 to 10. A score of seven or greater indicates a positive BV diagnosis. Three of the following four Amsel's criteria yield a positive diagnosis: 1) the presence of a fishy-like odor, 2) the presence of a white discharge, 3) a vaginal pH of >4.5, and 4) a minimum of 20% "clue cells" detection. The "gold standard" for BV diagnosis is Amsel's criteria and the Nugent scoring system. These methods have the disadvantages of being difficult to standardize and subject to interobserver variability because the assessment of the diagnostic criteria is dependent on the observer's skill and experience (Klebanof et al., 2004; Modak et al., 2011).

The recent advancement of molecular and high-throughput sequencing technologies allows for the detection of a large number of unculturable microorganisms from clinical samples (Adzitey et al., 2013). As a result, high-throughput biomolecular data are used to track the history of BV or to identify the pathogens of BV (Srinivasan et al. 2010; Ravel et al. 2011, 2013, White et al. 2011, Gajer et al. 2012, Hickey et al. 2012, Ma et al. 2012, Romero et al. 2014; Doyle et al. 2018). Ravel et al. (2013), for example, report on the temporal dynamics of 25 vaginal communities over 10 weeks using daily samples collected from women who were diagnosed with symptomatic BV, asymptomatic BV, and healthy. Srinivasan et al. (2010) conducted deep sequencing of the 16S rRNA gene in an attempt to investigate the variety and composition of vaginal bacteria in BV-positive women.

In the meantime, machine learning techniques have been used in this field. Baker et al. (2014) used genetic programming, random forests, and logistic regression machine learning methods on two BV datasets in the hopes of discovering BV-related microbial relationships. Later, Beck and Foster (2015) used random forests and logistic regression, in conjunction with ReliefF, to diagnose BV. Aeroccocus, Atopobium, Dialister, Eggerthella, and Gardnerella were identified as the most important bacteria associated with BV in their findings. Pérez-Gómez et al (2020) used a decision tree and the ReliefF algorithm as feature selectors, as well as the support vector machine and the logistic regression algorithm as classifiers to identify bacteria associated with BV. Loquet et al. (2021) designed classification and regression trees for BV to diagnosis in pregnant women. These works fall into the category of discovering BV-related feature bacteria (or OTU, Operational taxonomic unit).

Existing research indicates that BV is a systemic abnormality caused by multiple bacteria and that interactions between bacteria also play a role in the onset of BV (Srinivasan et al. 2010; White et al. 2011; Ravel et al., 2011, 2013; Gajer et al. 2012; Romero et al. 2014; Doyle et al. 2018). As a result, studying bacterial interactions is required to gain insight into BV signaling pathways. The bacterium can be defined as a network node, and interactions between bacteria can be defined as network edges. The challenge now is to identify network edges (interactions between two bacteria) that can characterize the state of the vaginal microbiome. Efforts to find reliable feature edges rely on information about bacterial interactions, so temporal sample datasets are required. The dataset reported by Ravel et al. (2013) provides ideal material to investigate this topic. In this paper, we model each temporal dataset of the vaginal community from Ravel et al. to a network, and then we build 25 networks. We apply supervised machine learning methods to 25 networks to find feature edges that are related to BV. We hope that these feature edges will aid in the diagnosis of BV and promote research into the pathogenesis of BV.

Materials & Methods

Vaginal Microbiome Dataset

The dataset was originally reported by Ravel et al. (2013). Ravel et al. (2013) sequenced vaginal communities collected daily for ten weeks from 25 women diagnosed with symptomatic BV (SBV: n = 15 women), asymptomatic BV (ABV: n = 6), or healthy (HEA: n = 4). In total, Ravel et al. (2013) sequenced 1,657 samples (median = 67 per woman) and obtained 420 8,757,681 high-quality sequenced reads of the V1–V3 hypervariable region of 16S-rRNA genes, with a median of 5,093 reads per sample. The dataset is freely accessible to the public (Ravel et al., 2013).

Feature Selection Algorithms

Feature selection aims to find the optimal subset of features. Feature selection can be used to eliminate irrelevant or redundant features, reduce the number of features, filter out features related to class information, and improve model accuracy. The general process of feature selection:

1. Generate subsets: search for feature subsets and provide feature subsets for the evaluation function;
2. Evaluation function: evaluate the quality of the feature subset;
3. Stopping criteria: related to the evaluation function, generally a threshold, the search can be stopped after the evaluation function reaches a certain standard;
4. Verification process: verify the validity of the selected feature subset on the verification data set.

Decision tree (Bramer 2007) and ReliefF (Robnik-Šikonja et al., 2003) are used in this work, they are belonging to the surprised feature selection method. These methods are implemented function by function in the Python modules skfeature (Li et al, 2018) and sklearn.

Classification Algorithms

A classification algorithm has two phases: learning and classification. The classification model is trained on the given dataset and its label information during the learning phase; during the classification phase, the classification model assigns the label to the new dataset. The classification model in this paper uses logistic regression (Han et al., 2011) and support vector machine (SVM, Wang et al., 2018), both of which are classic binary classification models that are widely used in a variety of fields.

Given a training dataset of feature space $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where $x_i \in R^n$, $y_i \in \{+1, -1\}$, $i = 1, 2, ..N$, x_i is i th feature vector, y_i is the class label. For a given dataset T and hyperplane $w \cdot x + b = 0$, the distance between the sample point and hyperplane can be defined as

$\gamma_i = y_i\left(\frac{w}{\|w\|}\cdot x_i + \frac{b}{\|w\|}\right)$. The minimum value of the geometric interval of the hyperplane with respect to all sample points is $\gamma = \min_i \gamma_i$. According to the above definition, SVM can be represented as

$$\max_{w,b} \gamma \quad s.t. y_i\left(\frac{w}{\|w\|}\cdot x_i + \frac{b}{\|w\|}\right) \geq \gamma . i = 1,2,...,N .$$

In logistic regression, for a given dataset T, the aim of the algorithm is still to find the decision boundary $w\cdot x + b = 0$ (named hyperplane in SVM). Based on the likelihood theory in statistics, the optimization model of the model is

$L(w) = \prod[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}$, where $p(x) = \frac{1}{1+e^{-(w^T x+b)}}$.

Leave-One-Out Validation

The dataset will divide into the training set and validation set. The training set is used to train the model, while the validation set is used to assess the model's generalizability. If the size of the dataset D is N, then use $N-1$ pieces of data for training, and use the remaining piece of data as validation. A total of times N are calculated for each group taken from D as the verification set until all samples have been verified as the set, and finally the verification error is averaged. This method is called leave-one-out cross-validation (Torgo 2010).

Performance Measures

Assume there are only two categories (positive and negative, usually the class of interest is the positive class, and the other classes are the negative class). The confusion matrix is as follows (table 1):

Accuracy: the ratio of correctly classified samples to total samples, is calculated as $\frac{TP+TN}{P+N}$.

Precision: the ratio of the number of true positive cases to the number of positive cases judged as positive, is calculated as $\frac{TP}{TP+FP}$.

Recall: the ratio of the number of positive cases correctly determined to the total number of positive cases, is calculated as $\frac{TP}{P}$.

Experimental Studies

We construct the networks and perform machine learning algorithms to find BV related feature edges.

1.The OTU in each table is taken as the network node for the OTU table of 25 vaginal microbiomes; the Spearman's rank correlation coefficients are calculated as the weight of edge between the OTUs, then 25 networks $A_i, i = 1, 2, \mathrm{L}\ 25$ are obtained. Each network assigned labels SBV, ABV, and H according to the diagnosis of the corresponding women.

2. Divided the 25 networks into four groups according to the research intention (BV = ABV+SBV vs. H; SBV vs. H; ABV vs. H; SBV vs. ABV)

3. To find feature edges in each group, use a feature selection and classification algorithm. The specific procedure is as follows: the significance of each edge is scored using a feature selection algorithm under leave-one-out cross-validation, and the scores are recorded in each run. After leave-one-out cross-validation, the importance scores of each edge are averaged. Edges are sorted in descending order by mean importance score. Again, using cross-validation, according to the mean value of the edge's importance score, a select subset of edges as feature edges to train the classification model on the training set and classify on the prediction set. The indicators (accuracy, precision, and recall) are used to evaluate classification performance after cross validation. The process is depicted in the diagram below.

Results

We performed the results of the four groupings, as shown in Table 2.

From the calculation results, we get the following conclusions.

1. Machine learning can distinguish different vaginal microbiome states (BV, ABV, SBV, H) based on bacterial interaction. It captures the difference between BV, SBV, ABV, and H better than that between SBV and ABV is weak.

2. Selecting the top 5 feature edges of importance can achieve the best accuracy for the feature selection and classification model. In some cases, the increase of the number of feature edges will reduce the performance of the classification algorithm.

3. The feature edges selected by decision tree outperform those selected by ReliefF in terms of classification algorithm logistic regression and SVM performance; however, there is almost no difference between classification algorithm logistic regression and SVM on the same feature edges.

4. The two feature selection algorithms have great differences in the importance of ranking of edges. Using the top 5 edge set as an example, the feature edges chosen by the two algorithms have almost no intersection.

Discussion

The feature edges that we discovered can distinguish the state of the vaginal microbiome (BV vs. H; SBV vs. H; ABV vs. H); however, the ability to distinguish between SBV and ABV is limited. In conclusion, our results show that there are differences in the expression of feature edges (interaction between the bacteria) under different vaginal environmental conditions. As a result, these feature edges may be useful in the diagnosis of BV. The feature edges chosen by different feature selection algorithms are inconsistent, a problem that has also been observed in previous studies (Baker et al., 2014; Beck and Foster, 2015). This adds to the complexity of the interpretability of feature edges. Similarly, Ma et al., (2021a) found 15 different types of network markers (motif, interactions among three species) that were present only in the BV microbiome and absent in the healthy microbiome, and which were validated on other BV datasets. We take the result of the decision tree algorithm to compare with the result of Ma et al. (2021a). We found that there was no overlap between them. It implies that the identification of BV associate feature edges may not be unique and that finding universal feature edges is difficult and complex, necessitating the mining of more sample data.

Further insights can be shed on the ecological mechanisms of BV by distinguishing key bacteria, or by the identifications of the critical pathway of interactions. However, BV is still poorly understood. However, the BV “single causative agent” theory is no longer widely accepted. Alternatively, BV is thought to be polymicrobial in nature. There are evidences that interspecies interactions characterize the vaginal microbiota with BV. Gardnerella spp. may provide a favorable environment for the growth of other BV-associated bacteria during the onset of BV, according to Pybus and Onderdonk (1997). Srinivasan and Fredricks (2008) proposed that BV occurs when BV-associated bacteria enter the vagina and displace lactobacilli. Furthermore, BV-associated bacteria (Bacteroides spp., Enterococcus faecalis, Vaginal G., Mobiluncus spp., and Peptoclococcus spp.) can inhibit Lactobacillus growth. And in a healthy vaginal environment, lactobacillus species produce hydrogen peroxide (H_2O_2) to inhibit the overgrowth of anaerobic bacteria. The reduction of Lactobacillus spp. was therefore considered to indicate vaginal dysbiosis. Those arguments imply, logically, that interactions between certain bacteria are related to BV. Feature edges (interaction between bacteria) have the potential to reveal the dysbiosis pathway and signaling associated with BV. However, several risk factors have been identified in the pathogenesis of BV, such as age, socio-economic status, antibiotic usage, sexual behavior, and ethnicity (Brumley, 2012; Singh et al., 2015; Ranjit et al., 2018). As a result, while the road to discovering the full face of BV remains long, our research provides important candidate materials (feature edges) and tools to further our understanding of BV risk and etiology.

Conclusion

The feature edges discovered by the machine learning algorithm can accurately distinguish BV and the health status of the vaginal microbiome. These features can also help reveal the pathogenesis of BV. However, different machine learning algorithms find different feature

edges, which increases the complexity of feature interpretation. Furthermore, the data set used in the study is insufficient, and the sample size is unbalanced. Because only the Spearman correlation coefficient is used when building the sample network, more work is required. In the future, we will also try to use different correlation measures to build a sample network, collect more data, and consider sample balance for research, in the hopes of obtaining more reliable results and promoting BV diagnosis and pathogenesis.

Acknowledgements

The authors thank Ravel et al. for their data support to this study.

References

Koumans, EH, Kendrick, JS. 2001. Preventing adverse sequelae of bacterial vaginosis: a public health program and research agenda. Sexually Transmitted Diseases, 28(5):292–297.

Hay, P. E., Lamont, R. F., Taylor-Robinson, D., Morgan, D. J., Ison, C., and Pearson, J. 1994. Abnormal bacterial colonisation of the genital tract and subsequent preterm delivery and late miscarriage. BMJ 308, 295–298.

Ness, R. B., Kip, K. E., Hillier, S. L., Soper, D. E., Stamm, C. A., Sweet, R. L., et al. 2005. A cluster analysis of bacterial vaginosis-associated microflora and pelvic inflammatory disease. Am. J. Epidemiol. 162, 585–590.

Ma, B, LJ Forney, J. Ravel. 2012. The Vaginal Microbiome: Rethinking Health and Disease. Annual Review of Microbiology. 66:371-389.

Sha, B. E., Zariffard, M. R., Wang, Q. J., Chen, H. Y., Bremer, J., Cohen, M. H., et al. 2005a. Female genital-tract HIV load correlates inversely with Lactobacillus species but positively with bacterial vaginosis and Mycoplasma hominis. J. Infect. Dis. 191, 25–32.

Atashili, J., Poole, C., Ndumbe, P. M., Adimora, A. A., and Smith, J. S. 2008. Bacterial vaginosis and HIV acquisition: a meta-analysis of published studies. AIDS 22, 1493–1501.

van de Wijgert, J. H. H. M., Morrison, C. S., Cornelisse, P. G., Munjoma, M., Moncada, J., Awio, P., et al. 2008. Bacterial vaginosis and vaginal yeast, but not vaginal cleansing, increase HIV-1 acquisition in African women. J. Acquir. Immune Defic. Syndr. 48, 203–210.

Redelinghuys MJ, Geldenhuys J, Jung H, Kock MM. Bacterial Vaginosis: Current Diagnostic Avenues and Future Opportunities. Front Cell Infect Microbiol. 2020 Aug 11; 10:354.

Nugent, R. P., Krohn, M. A., and Hillier, S. L. 1991. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. J. Clin. Microbiol. 29, 297–301.

Amsel, R., Totten, P. A., Spiegel, C. A., Chen, K. C. S., Eschenbach, D., and Holmes, K. K. 1983. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. Am. J. Med. 74, 14–22.

Klebanoff, M. A., Schwebke, J. R., Zhang, J., Nansel, T. R., Yu, K. F., and Andrews, W. W. 2004. Vulvovaginal symptoms in women with bacterial vaginosis. Obstet. Gynecol. 104, 267–272.

Modak, T., Arora, P., Agnes, C., Ray, R., Goswami, S., Ghosh, P., et al. 2011. Diagnosis of bacterial vaginosis in cases of abnormal vaginal discharge: comparison of clinical and microbiological criteria. J. Infection Dev. Countries 5, 353–360.

Adzitey, F., Huda, N., and Ali, G. R. R. 2013. Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. Biotechnology 3, 97–107.

Srinivasan S, Liu C, Mitchell CM, Fiedler TL, Thomas KK, Agnew KJ, et al. 2010. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. 2010, PloS One 2010, 5:e10197.
Ravel J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., et al. 2011. Vaginal microbiome of reproductive-age women. Proceedings of the US National Academy of Sciences, 108 Suppl. 1, 4680–4687.
Ravel, J. Jacques Ravel, Rebecca M Brotman, Pawel Gajer, Bing Ma, Melissa Nandy, Douglas W Fadrosh, Joyce Sakamoto, Sara SK Koenig, Li Fu, Xia Zhou, Roxana J Hickey, Jane R Schwebke, Larry J Forney. 2013. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome 2013, 1:29.
White, BA, D. J. Creedon, K. E. Nelson, B. A. Wilson. 2011. The vaginal microbiome in health and disease. Trends in Endocrinology and Metabolism, 2011, 22(1): 389-393.
Gajer, P., R. M. Brotman, G. Bail, J. Sakamoto, U. M. E. Schütte, X. Zhong, S. S. Koeng, L. Fu, Z. Ma, X. Zhou, Z. Abdo, L. J. Forney and J. Ravel. 2012. Temporal Dynamics of the Human Vaginal Microbiota. Science Translational Medicine, 4(132):132ra52.
Hickey RJ, Zhou X, Pierson JD, Ravel J, Forney LJ. 2012. Understanding vaginal microbiome complexity from an ecological perspective. Transl. Res, 2012, 160:267-282.
Romero, R, SS Hassan, P. Gajer, AL Tarca, DW Fadrosh, L. Nikita, M. Galuppi, RF Lamont, P. Chaemsaithong, J. Miranda, T. Chaiworapongsa and J. Ravel. 2014. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome, 2014, 2:4.
Doyle R, Gondwe A, Fan Y, et al. 2018. A Lactobacillus-Deficient Vaginal Microbiota Dominates Postpartum Women in Rural Malawi. Applied and Environmental Microbiology, 84(6): e02150-17.
Baker, Y.S.; Beck, D.; Agrawal, R.; Dozier, G.; Foster, J.A. Detecting Bacterial Vaginosis using machine learning. In Proceedings of the 2014 ACM Southeast Regional Conference, Kennesaw, GA, USA, 28–29 March 2014.
Beck, D.; Foster, J.A. Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData Min. 2015, 8, 23.
Pérez-Gómez, J.F., Canul-Reich, J., Hernández-Torruco, J., & Hernández-Ocaña, B. 2020. Predictor Selection for Bacterial Vaginosis Diagnosis Using Decision Tree and Relief Algorithms. Applied Sciences.
Loquet A, Le Guern R, Grandjean T, Duployez C, Bauduin M, Kipnis E, Brabant G, Subtil D, Dessein R. Classification and Regression Trees for Bacterial Vaginosis Diagnosis in Pregnant Women Based on High-Throughput Quantitative PCR. J Mol Diagn. 2021 Feb;23(2):234-241.
Bramer, M. Principles of Data Mining; Springer: London, UK, 2007; Volume 180.
Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 2003, 53, 23–69.
Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (November 2018), 45 pages.
Han, J.; Pei, J.; Kamber, M. Data Mining: Concepts and Techniques, 3rd. ed.; University of Illinois at Urbana-Champaign: Champaign, IL, USA; Simon Fraser University: Burnaby, BC, Canada; Elsevier: Amsterdam, The Netherlands, 2011; ISBN 978-0-12-381479-1.
Wang, H.; Zheng, B.; Yoon, S.W.; Ko, H.S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. EJOR 2018, 267, 687–699.

298 Torgo, L. Data Mining with R: Learning with Case Studies; eBook; Chapman and Hall/CRC: New York,
299 NY, USA, 2010; ISBN 9780429292859.
300 Ma, Z. S., and Ellison, A. M. (2021a). In silico trio-biomarkers for bacterial vaginosis revealed by species
301 dominance network analysis. Comput. Struct. Biotechnol. J. 19, 2979–2989.
302 Brumley, J. (2012). Testing a model of bacterial vaginosis among black women. (PhD Dissertation).
303 University of South Florida. Available online at: https:// scholarcommons.usf.edu/etd/3995 (accessed 16
304 December 2019).
305 Singh, H. O., Singh, A., Dhole, T. N., and Sumitra, N. (2015). Factor associated to bacterial vaginosis in
306 non-pregnant women of North Indian population. J. Biotechnol. Biomater. 05:195.
307 Ranjit, E., Raghubanshi, B. R., Maskey, S., and Parajuli, P. (2018). Prevalence of bacterial vaginosis and
308 its association with risk factors among nonpregnant women: a hospital-based study. Int. J. Microbiol.
309 2018:8349601.

Table 1(on next page)

The indicators to evaluate the classification performance.

* The AB mode in the table: the first indicates whether the predicted result was correct or incorrect, and the second indicates the predicted category. For example, TP means True Positive, that is, the correct prediction is a positive class; FN means, False Negative, that is, the wrong prediction is a negative class.

1

Actual category	Prediction category		
	Positive	Negative	Summarize
Positive	TP	FN	P (Actually positive)
Negative	FP	TN	N (Actually negative)

2

Table 2(on next page)

Tables corresponding to calculation results.

* The results of feature selection only list the top 5 feature edges of importance.

1

Groups	Feature selection	Performance of classifiers
BV vs. H	Table 3	Table 4
SBV vs. H	Table 5	Table 6
ABV vs. H	Table 7	Table 8
SBV vs. ABV	Table 9	Table 10

2

Table 3(on next page)

Top 5 feature edges of importance for BV vs. H group selection by Decision tree and ReliefF feature selection algorithms.

1

Decision tree		ReliefF	
Feature edges	Import value	Feature edges	Import value
Streptococcus anginosus Veillonellaceae	0.38	Lactobacillus iners Lactobacillus crispatus	1429.8
Actinomycetales Prevotella buccalis	0.228	Atopobium vaginae Megasp3aera	1340.52
Peptonip3ilus Stap3ylococcus	0.04	Atopobium vaginae Stap3ylococcus aureus	1311.08
Streptococcus anginosus Prevotella buccalis	0.04	Lactobacillus iners Atopobium vaginae	1273.84
Atopobium vaginae Parvimonas micra	0.03	Clostridiales Prevotella	1251.24

2

Table 4(on next page)

The performance of two classification algorithms on different quantities of features for BV vs. H.

1

Feature Selection	Decision tree/ReliefF					
Classifiers	Logistic regression			SVM		
Feature number	Acc	Pre	Recall	Acc	Pre	Recall
5	0.92/0.84	0.95/0.84	0.95/1	0.92/0.84	0.95/0.84	0.95/1
10	1.00/0.84	1.00/0.84	1.00/1	1.00/0.84	1.00/0.84	1.00/1
15	1.00/0.84	1.00/0.84	1.00/1	1.00/0.84	1.00/0.84	1.00/1
20	1.00/0.84	1.00/0.84	1.00/1	1.00/0.84	1.00/0.84	1.00/1
25	0.96/0.84	0.95/0.84	1.00/1	0.96/0.84	0.95/0.84	1.00/1
30	0.96/0.84	0.95/0.84	1.00/1	0.96/0.84	0.95/0.84	1.00/1
50	0.96/0.84	0.95/0.84	1.00/1	0.96/0.84	0.95/0.84	1.00/1
80	0.84/0.84	0.84/0.84	1.00/1	0.84/0.84	0.84/0.84	1.00/1
100	0.84/0.84	0.84/0.84	1.00/1	0.84/0.84	0.84/0.84	1.00/1
200	0.84/0.84	0.84/0.84	1.00/1	0.84/0.84	0.84/0.84	1.00/1
Mean	0.93/0.84	0.93/0.84	1.00/1	0.93/0.84	0.93/0.84	1.00/1

2

Table 5(on next page)

Top 5 feature edges of importance for SBV vs. H group selection by Decision tree / ReliefF feature selection algorithms.

1

Decision tree		ReliefF	
Feature edges	**Import value**	**Feature edges**	**Import value**
Lactobacillus iners Bifidobacteriaceae	0.21	Lactobacillus jensenii Streptococcus salivarius	1347.42
Actinomycetales Prevotella buccalis	0.17	Stap3ylococcus Eggert3ella	1315.68
Lactobacillus iners Eggert3ella	0.13	Lactobacillus iners Atopobium vaginae	1314.79
Streptococcus anginosus Veillonellaceae	0.09	Lactobacillus iners BVAB2	1314.26
Megasp3aera sp. type 2 Streptococcus anginosus	0.05	Lactobacillus iners Lactobacillus jensenii	1299.16

2

Table 6(on next page)

Top 5 feature edges of importance for ABV vs. H group selection by Decision tree/ReliefF feature selection algorithms.

1

Feature selection	**Decision tree/ReliefF**					
Classifiers	**Logistic regression**			**SVM**		
Feature number	Acc	Pre	Recall	Acc	Pre	Recall
5	0.95/0.95	0.94/0.94	1/1	0.95/0.95	0.94/0.94	1/1
10	0.95/0.89	0.94/0.88	1/1	0.95/0.89	0.94/0.88	1/1
15	0.95/0.89	0.94/0.88	1/1	0.95/0.89	0.94/0.88	1/1
20	0.95/0.89	0.94/0.88	1/1	0.95/0.89	0.94/0.88	1/1
25	0.95/0.89	0.94/0.88	1/1	0.95/0.89	0.94/0.88	1/1
30	0.95/0.84	0.94/0.83	1/1	0.95/0.84	0.94/0.83	1/1
50	0.95/0.79	0.94/0.79	1/1	0.95/0.79	0.94/0.79	1/1
80	0.79/0.79	0.79/0.79	1/1	0.79/0.79	0.79/0.79	1/1
100	0.79/0.79	0.79/0.79	1/1	0.79/0.79	0.79/0.79	1/1
200	0.79/0.79	0.79/0.79	1/1	0.79/0.79	0.79/0.79	1/1
Mean	0.90/0.85	0.90/0.85	1/1	0.90/0.85	0.90/0.85	1/1

2

Table 7 (on next page)

Top 5 feature edges of importance for ABV vs. H group selection by Decision tree / ReliefF feature selection algorithms.

1

Decision tree		ReliefF	
Feature edges	**Import value**	**Feature edges**	**Import value**
Sneat3ia sanguinegens Megasp3aera	0.2	Megasp3aera sp. type 1 Megasp3aera	1431.7
Lactobacillus iners Peptostreptococcus	0.1	BVAB2 Peptonip3ilus asacc3arolyticus	1384
Bacteria Lactobacillus vaginalis	0.1	Lactobacillus crispatus Gammaproteobacteria	1377.4
Dialister sp. type 2 Sneat3ia sanguinegens	0.1	Megasp3aera sp. type 2 Streptococcus	1374.3
BVAB2 Clostridiales	0.1	Atopobium vaginae Prevotella bivia	1352.7

2

Table 8(on next page)

The performance of two classification algorithms on different quantities of features for ABV vs. H.

1

Feature Selection	Decision tree/ReliefF					
Classifiers	Logistic regression			SVM		
Feaure number	Acc	Pre	Recall	Acc	Pre	Recall
5	1/0.3	1/0.43	1/0.5	1/0.3	1/0.43	1/0.5
10	1/0.3	1/0.43	1/0.5	1/0.3	1/0.43	1/0.5
15	1/0.5	1/0.56	1/0.83	1/0.5	1/0.56	1/0.83
20	1/0.5	1/0.56	1/0.83	1/0.5	1/0.56	1/0.83
25	1/0.5	1/0.56	1/0.83	1/0.5	1/0.56	1/0.83
30	1/0.6	1/0.6	1/1	1/0.6	1/0.6	1/1
50	1/0.6	1/0.6	1/1	1/0.6	1/0.6	1/1
80	1/0.6	1/0.6	1/1	1/0.6	1/0.6	1/1
100	0.7/0.6	0.67/0.6	1/1	0.7/0.6	0.67/0.6	1/1
200	0.6/0.6	0.6/0.6	1/1	0.6/0.6	0.6/0.6	1/1
Mean	0.93/0.51	0.93/0.55	1/0.85	0.93/0.51	0.93/0.55	1/0.85

2

Table 9(on next page)

Top 5 feature edges of importance for SBV vs. ABV group selection by Decision tree/ReliefF feature selection algorithms.

1

Decision tree		ReliefF	
Features	**Import value**	**Features**	**Import value**
Megasp3aera sp. type 2 Stap3ylococcus	0.29	BVAB1 Firmicutes	1246.48
Megasp3aera sp. type 2 Enterococcus faecalis	0.22	Prevotella genogroup 1 Prevotella buccalis	1240.57
Megasp3aera sp. type 2 Actinomycetales	0.19	Gemella Sneat3ia sanguinegens	1240.10
Stap3ylococcus aureus Megasp3aera	0.05	Actinobacteria .class. Clostridiales Family XI. Incertae Sedis	1224.48
Prevotella buccalis Bifidobacterium	0.05	Eggert3ella Prevotella genogroup 3	1219.38

2

Table 10(on next page)

The performance of two classification algorithms on different quantities of features for SBV vs. ABV

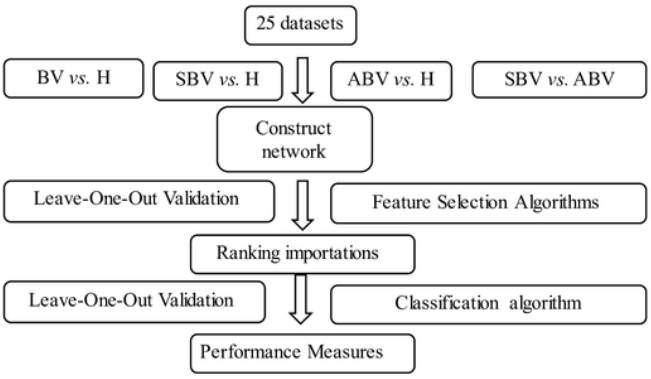
1

Feature Selection	**Decision tree/ReliefF**					
Classifiers	**Logistic regression**			**SVM**		
Feature number	Acc	Pre	Recall	Acc	Pre	Recall
5	0.86/0.67	0.93/0.70	0.87/0.93	0.86/0.67	0.93/0.70	0.87/0.93
10	0.67/0.67	0.7/0.70	0.93/0.93	0.67/0.67	0.7/0.70	0.93/0.93
15	0.62/0.71	0.68/0.71	0.87/1	0.62/0.71	0.68/0.71	0.87/1
20	0.57/0.71	0.67/0.71	0.8/1	0.57/0.71	0.67/0.71	0.8/1
25	0.57/0.71	0.67/0.71	0.8/1	0.57/0.71	0.67/0.71	0.8/1
30	0.62/0.71	0.68/0.71	0.87/1	0.62/0.71	0.68/0.71	0.87/1
50	0.67/0.71	0.7/0.71	0.93/1	0.67/0.71	0.7/0.71	0.93/1
80	0.67/0.71	0.7/0.71	0.93/1	0.67/0.71	0.7/0.71	0.93/1
100	0.71/0.71	0.71/0.71	1/1	0.71/0.71	0.71/0.71	1/1
200	0.71/0.71	0.71/0.71	1/1	0.71/0.71	0.71/0.71	1/1
Mean	0.67/0.70	0.72/0.71	0.9//0.99	0.67/0.70	0.72/0.71	0.9//0.99

2

Figure 1

Conceptual diagram of experimental process


25 datasets
BV vs. H
SBV vs. H
ABV vs. H
SBV vs. ABV
Construct network
Leave-One-Out Validation
Feature Selection Algorithms
Ranking importations
Leave-One-Out Validation
Classification algorithm
Performance Measures