Article

BiGAMi: Bi-Objective Genetic Algorithm Fitness Function for Feature Selection on Microbiome Datasets

Mike Lenske [1], Francesca Bottacini [2], Haithem Afli [1,*] and Bruno G. N. Andrade [1,*]

[1] Department of Computer Sciences, Munster Technological University, MTU/ADAPT, Cork, Ireland
[2] Department of Biological Sciences, Munster Technological University, MTU, Cork, Ireland
* Correspondences: Haithem.Afli@mtu.ie, Bruno.Andrade@mtu.ie

Abstract: The relationship between the host and the microbiome, or the assemblage of microorganisms (including bacteria, archaea, fungi, and viruses), has been proven crucial for its health and disease development. The high dimensionality of microbiome datasets has often been addressed as a major difficulty for data analysis, such as the use of Machine Learning (ML) and Deep Learning (DL) models. Here we present BiGAMi, a bi-objective genetic algorithm fitness function for feature selection in microbial datasets to train high-performing phenotype classifiers. The proposed fitness function allowed us to build classifiers that outperformed the baseline performance estimated by the original studies by using as few as 0.04% to 2.32% features of the original dataset. In 19 out of 21 classification exercises, BiGAMi achieved its results by selecting 6-68% fewer features than the highest performance of a Sequential Forward Feature Selection algorithm. This study showed that the application of a bi-objective GA fitness function against microbiome datasets succeeded in selecting small subsets of bacteria whose contribution to understood diseases and the host state was already experimentally proven. Applying this feature selection approach to novel diseases is expected to quickly reveal the microbes most relevant to a specific condition.

Keywords: microbiome; genetic algorithm; feature selection; human health; machine learning

1. Introduction

The past 10 years have shown a gradual introduction of classical and advanced Machine Learning (ML) techniques to bioinformatics, which enabled some of the first successes in diagnosing host conditions or phenotypes from the host's microbiota [1-4]. In general, conditions of interest to be predicted by ML models include, but are not limited to: age; health state and diseases; dietary preferences; quality of meat; and methane emissions. A general property of microbiome datasets is their tendency to include a fairly small sample size (often a few dozens to a few hundred), due to sampling and sequencing costs, while being highly dimensional with hundreds to thousands of features to be analysed. Features represent clustered genetic markers into operational taxonomic units (OTUs) and their mapping to microbial taxonomy classification for metabarcoding studies. Alternatively, features can also represent single genes, genes functions, or functional pathways investigated in metagenomic studies. The high dimensionality of microbiome datasets has often been addressed as a major difficulty in the application of ML algorithms. Feature Selection is a commonly used methodology to improve ML algorithm performance in classification and regression. However, algorithms like Forward Selection or Backward Elimination [5] quickly result in an unmanageable computational complexity due to a large number of microbial features. Other Dimensionality Reduction algorithms, such as Principal Component Analysis (PCA) transform the data in a way that makes the prediction results harder to be explained.

To address the high dimensionality of 'omics datasets few research groups have implemented a GA to search for subsets of microbiome features leading to high predictive performance. Unlike classical feature selection methods, such as Forward Selection and

Backward Elimination [5], which focus on sequential addition or removal of features, a GA-based approach has the potential to evaluate more complex feature interactions. Carter et al. [6] applied a Genetic and Evolutionary Feature Selection search on a vaginal microbiome dataset to detect Bacterial Vaginosis. Zhang et al. [7] applied a Genetic Algorithm after a PCA-based dimensionality reduction with a fixed number of principal components. Chiesa et al. [8] demonstrated the applicability of using a GA to select a fixed number of highly predictive features from small, medium, and large-sized 'omics datasets.

Herein we addressed the dimensionality problem of microbiome datasets through the application of a Bi-Objective Genetic Algorithm to select subsets of microbiome features for classification models. We tested our method using 4 publicly available datasets, with 3 different dataset representations, raw counts, relative abundance and Centered Log-Ratio (CLR) transformed compositions, and compared our results with the baseline scores provided by Vangay et al. [9] as well as the performance of a classical Sequential Forward Feature Selection.

Our method, called "BiGAMi - Bi-Objective Genetic Algorithm Fitness Function for Feature Selection on Microbiome Datasets" is implemented in Python and released under the open-source MIT license on GitHub (https://github.com/mikeleske/BiGAMi).

2. Materials and Methods

2.1. Data retrieval and pre-processing

The microbiome datasets were retrieved from the Microbiome Learning [9], and included microbiome matrices containing Operational Taxonomic Units (OTU) abundance counts and sample metadata for each dataset. The datasets were generated by studies of microbiome and health, such as the investigation of the relationship between microbiome, colorectal cancer, and cirrhosis [10, 11], and to investigate differences in the vaginal microbiota of human populations [12]. The OTU matrices were transformed from 16s rRNA counts to ratios, such as relative abundance and Centered-Log Ratio (CLR), through the use of custom Python scripts and the Scikit-bio package (https://github.com/biocore/scikit-bio). Next, the datasets were scaled using the Scikit-learn [13] MinMax scaling operation, followed by a selection of the 128 most important features using the Scikit-learn SelectKBest operation based on Chi-squared statistics. Lastly, each dataset was split into 6 folds, where folds 1 to 5 belong to the training set, and fold 6 was treated as a hold-out test set. Figure 1 shows the data preprocessing workflow. For each selected classification task and its associated OTU mapped abundance counts (Greengenes 97 [14] and NCBI RefSeq [15]) our GA-based feature selection method was executed with transformed datasets based on the original (raw abundance counts), as well as relative abundances and CLR-transformed counts (Table 1). Next, the performance of the same set of experiments was executed using a Sequential Forward Feature Selection algorithm provided by the *MLxtend* Python library [16]. Vangay et al. [9] also applied classical ML techniques to each classification task, which will be used as performance baselines using the Area-under-the-Curve (AUC) metric.

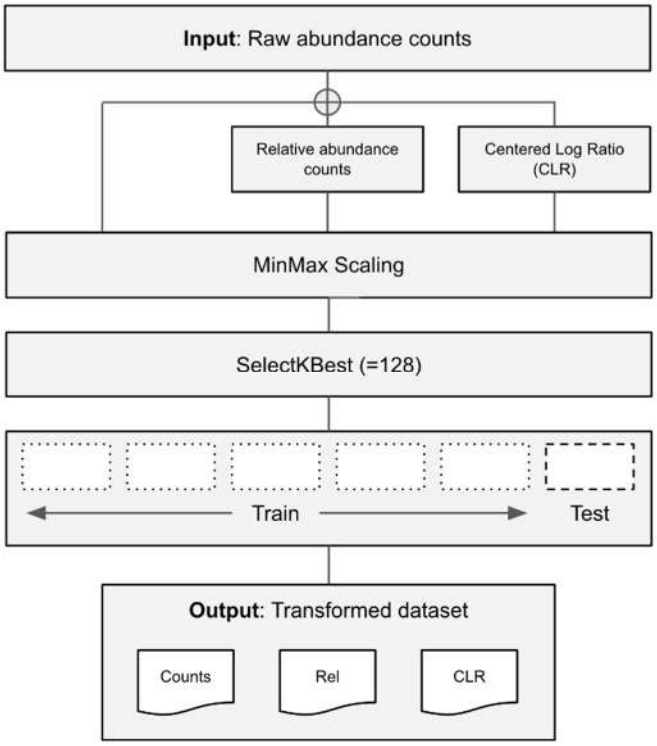

Figure 1. Data preprocessing flow from raw abundance counts input data to the transformed datasets used in classification tasks.

Table 1. Data pre-processing.

Input data	Pre-processing
Counts	Raw / Absolute abundance counts
Rel	Relative abundance counts
CLR	Centered-Log Ratio (CLR)

2.2. Bi-Objective Genetic Algorithm Fitness Function and Implementation (BiGAMi)

The Python package *DEAP* [17] was used as the core framework for the Genetic Algorithm and evolutionary search process. The central goal of GA-based feature selection is to identify the smallest feature subset resulting in the highest fitness metric. Performing such a feature selection relies on encoding the GA individuals' chromosomes as a binary string. A gene is set to the value 1, if the associated feature is included in the fitness evaluation process, and set to 0 otherwise. We propose a sparse chromosome initialization strategy for the initial GA generation that activates only a small fraction of the dataset features per chromosome. A random initialization of the chromosome results in unnecessarily large feature subsets and the activation of mostly irrelevant features. Activating 10 features per chromosome has shown successful results during this study. For each chromosome, a separate Scikit-learn SGD Classifier (Stochastic Gradient Descent Classifier) was trained using k-fold cross-validation. The algorithm's feature coefficients table was used to reset chromosome genes to 0 whose associated features have shown no significant relevance on the classification task. The crossover operation used the one-point mating option to minimize the risk of producing exact copies of the mating individuals due to their chromosome sparsity. A custom mutation operation was implemented to provide equal chances to either switch off an active gene or switch on an inactive gene. Lastly, a bi-objective fitness function was used to evaluate a chromosome's fitness which not only considered the classification metric but penalized the usage of larger feature subsets. The bi-objective fitness function was implemented using the following formulae:

fitness score = x * metric + y * (selected features/total features) (1)
metric = avg(cv_score) + min(cv_score) (2)

Where metric represents the sum of the average and minimum k-fold cross-validation scores, AUC, the quotient *selected features* / *total features* stands for the ratio of the dataset features (Counts, Relative Abundance and CLR abundances) a GA chromosome uses, x=1 and y=-1. For each input data, 25 GA runs with a population size of 300 were executed for 10 generations. For each separate GA run the 6-fold data split remained consistent across its generations. In each generation, the individuals' performances were evaluated by a 5-fold CV using an SGDClassifier on the training data (folds 1 to 5) and the sum of the 5-fold average and minimum metric was reported as this individual's score, which provides a larger optimization (maximization) space compared to relying on the k-fold average alone. Each SGDClassifier used the "log" loss function, the "l1" penalty, and was restricted to 500 iterations. After each generation, the best-performing individual (according to 5-fold CV using the bi-objective fitness function) was evaluated against the hold-out test set (6th fold), logged with its test set performance and selected feature subset, and copied unaltered into the next generation (elitism concept). After each GA search the individual with the highest bi-objective evaluation metric on the hold-out test set was subsequently selected, resulting in 25 high-performance individuals per input data.

Individual selection for crossover was based on Tournament Selection with size 3. The crossover operation was set to one-point to reduce the risk of selecting only patches of 0-genes due to chromosome sparsity. The probability of 2 selected individuals producing offsprings was set to 0.8. Likewise, the probability of offsprings undergoing a mutation was set to 0.8. A custom mutation operation was implemented to give equal chances to either activate an inactive gene or to deactivate an active gene. Due to the sparsely activated genes within the chromosomes default mutations operations would have had a bias towards activating inactive genes.

2.3. Sequential Forward Feature Selection

To compare the efficiency of our BiGAMi algorithm a classical Sequential Forward Feature Selection (SFFS) method using 5-fold cross-validation (based on ⅝ of the dataset size) was applied using the *MLxtend* Python library for the same input data. 25 SFFS runs were executed per input data to identify stepwise the 1 to 32 most important features using the training dataset (fold 1 to 5) and subsequently evaluated on the hold-out test set (6th fold). For each SFFS run, the best-performing feature subset was reported including the test set performance and the selected features.

3. Results

The use of the BiGAMi method to reanalyze consolidated data allowed us to select highly informative microbiome OTUs (features) which, when used for classification tasks, achieved at least the performance of the original studies level or greatly improved it. The results listed here represent the average performance and OTU subset sizes of the 25 best-performing solutions found across the 25 BiGAMi runs. For the Kostic Colorectal Cancer dataset (task I), we reduced the average number of OTUs to 16-18/3228 (GG97) and to 8-17/908 (RefSeq) while increasing the AUC score from 0.74 to 0.93-0.95 and from 0.69 to 0.85-0.86. For the Vaginal Nugent Category dataset (II), the baseline AUC score was already 0.99, however, we achieved similar scores of 0.99 using 8-9/1083 (GG97) and 0.98 using 7-8/586 (RefSeq) OTUs on average used in the original study. The same behavior was observed when using the same dataset to classify the host's ethnicity in black or white groups (III), an increase of 0.64 to 0.73-0.77 (GG97) and from 0.70 to 0.73-0.78 (RefSeq) while using 11-16/1083 and 10-14/586 OTUs on average and when classifying healthy vs patients with cirrhosis (IV), increasing the AUC score from 0.92 to 0.93-0.94, while using 4-12/8483 OTUs on average (Table 2).

Table 2. Performance results per classification task. Per input data, the average AUC score and the average number of OTUs of the 25 best-performing individuals are provided. (I) Kostic Colorectal Cancer Healthy/Tumor GG97, (II) Ravel Vaginal Nugent Category, (III) Ravel Vaginal Black/White, (IV) Qin Cirrhosis RefSeq. Bold entries were selected for further analysis due to combined metric and OTU selection preference.

Task	Database	Total OTUs	Baseline AUC	Input Data	BiGAMi (ours) AUC	BiGAMi (ours) OTUs
(I)	GG97	3228	0.74	Counts	0.94	15.9 (0.49%)
				Rel	0.95	18.1 (0.56%)
				CLR	**0.94**	**12.3 (0.38%)**
	RefSeq	908	0.69	Counts	0.85	14.9 (1.64%)
				Rel	0.86	16.8 (1.85%)
				CLR	0.84	8.2 (0.90%)
(II)	GG97	1093	0.99	Counts	0.99	9.0 (0.82%)
				Rel	0.98	9.2 (0.84%)
				CLR	0.99	8.6 (0.79%)
	RefSeq	586	0.99	Counts	0.98	8.0 (1.37%)
				Rel	0.98	6.6 (1.13%)
				CLR	**0.98**	**6.5 (1.11%)**
(III)	GG97	1093	0.64	Counts	0.76	13.6 (1.24%)
				Rel	0.77	16.2 (1.48%)
				CLR	0.73	11.1 (1.02%)
	RefSeq	586	0.70	**Counts**	**0.78**	**12.8 (2.18%)**
				Rel	0.75	13.6 (2.32%)
				CLR	0.73	10.2 (1.74%)
(IV)	RefSeq	8483	0.92	Count	0.94	12.2 (0.14%)
				Rel	0.93	11.6 (0.14%)
				CLR	**0.93**	**4.3 (0.05%)**

While the Kostic Colorectal Cancer GG97 dataset led to better predictive performances than the RefSeq-mapped dataset for this classification task, the performances for both GG97 and RefSeq-mapped datasets for both Vaginal classification tasks were on par despite the lower number of overall OTUs within the RefSeq datasets. The Cirrhosis dataset was only provided with a RefSeq mapping. Differences in the input data (Counts, REL or CLR) showed a minor impact on the BiGAMi performance. In almost all classification tasks, the highest performance scores were displayed by the use of the input data Counts or Rel.

Using CLR-transformed data input mostly led to comparable average predictive performances of the SGD classifier, but it allowed the GA search to achieve this by selecting consistently smaller OTU subsets across all classification tasks (Table 2 and Figure 2). Especially for classification task (IV) Qin Cirrhosis RefSeq the CLR-transformed dataset resulted in a 2.6-2.8x lower number of selected OTUs.

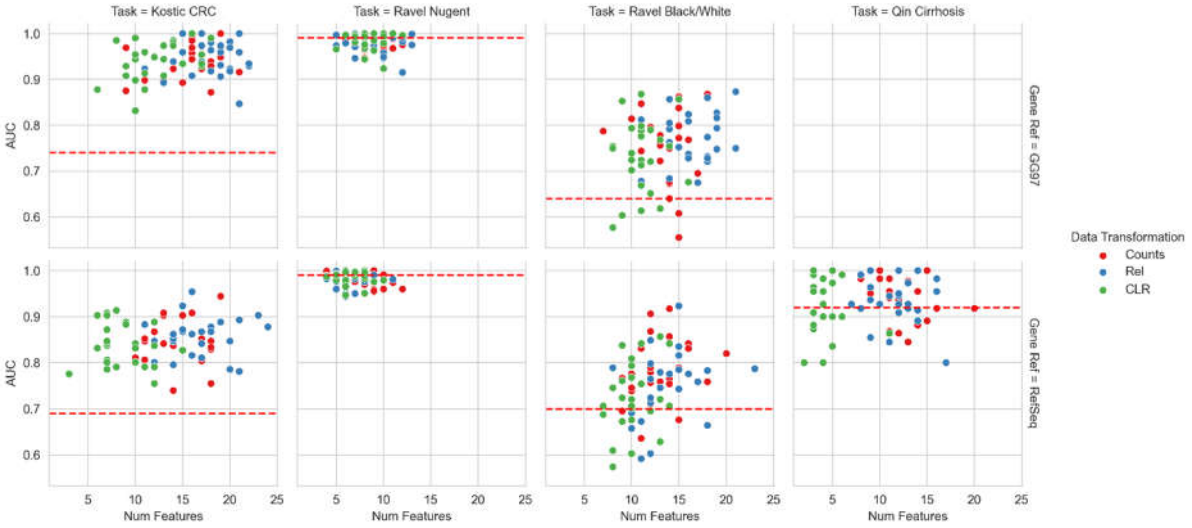

Figure 2. Metric plots to evaluate the GA performance against the four classification tasks. Subplots display the achieved metric vs. number of OTUs used by the 25 best performing individuals per GA run per input data after test set evaluation.

The taxonomic annotation up to the species level was further analyzed to identify and explore the microorganisms selected as important for the classification of their host's state for tasks I-IV (Figure 3). Species that appeared in less than 3 of a task's best-performing GA individuals have been excluded. For task II, *Gardnerella vaginalis* and *Lactobacillus vaginalis* account for 67% of the selected species, accompanied by *Gemella asaccharolytica* (13%), *Prevotella timonensis* (13%), and *Prevotella amnii* (7%). For task III, a set of 6 geni (*Anaerococcus*, *Corynebacterium*, *Streptococcus*, *Atopobium*, *Lactobacillus*, and *Blautia*) account for 74% of the selected species with special emphasis on *Anaerococcus hydrogenalis*, *Atopobium vaginae*, and *Blautia luti*. Species selected for task IV were dominated by *Megasphaera micronuciformis* (48%), accompanied by *Oribacterium sinus* (20%), *Lactobacillus salivarius* (13%), *Anaeroglobus geminatus* (11%), and *Fusobacterium periodonticum* (8%).

For task 1, where the classification performance based on the Greengenes97 dataset outperformed the RefSeq dataset, the translation of OTU IDs to taxonomic annotation often resulted in missing genus and species information. On a family level, *Lachnospiraceae* (including *Blautia* and *Coprococcus*), *Ruminococcaceae* (including *Oscillospira*), and *Veillonellaceae* (including *Veillonella dispar*) account for 63% of the selected OTUs, accompanied by *Fusobacteriaceae* (13%), *Rikenellaceae* (9%), *Bacteroidaceae* (9%), *Enterobacteriaceae* (6%) and *Methylobacteriaceae* (4%).

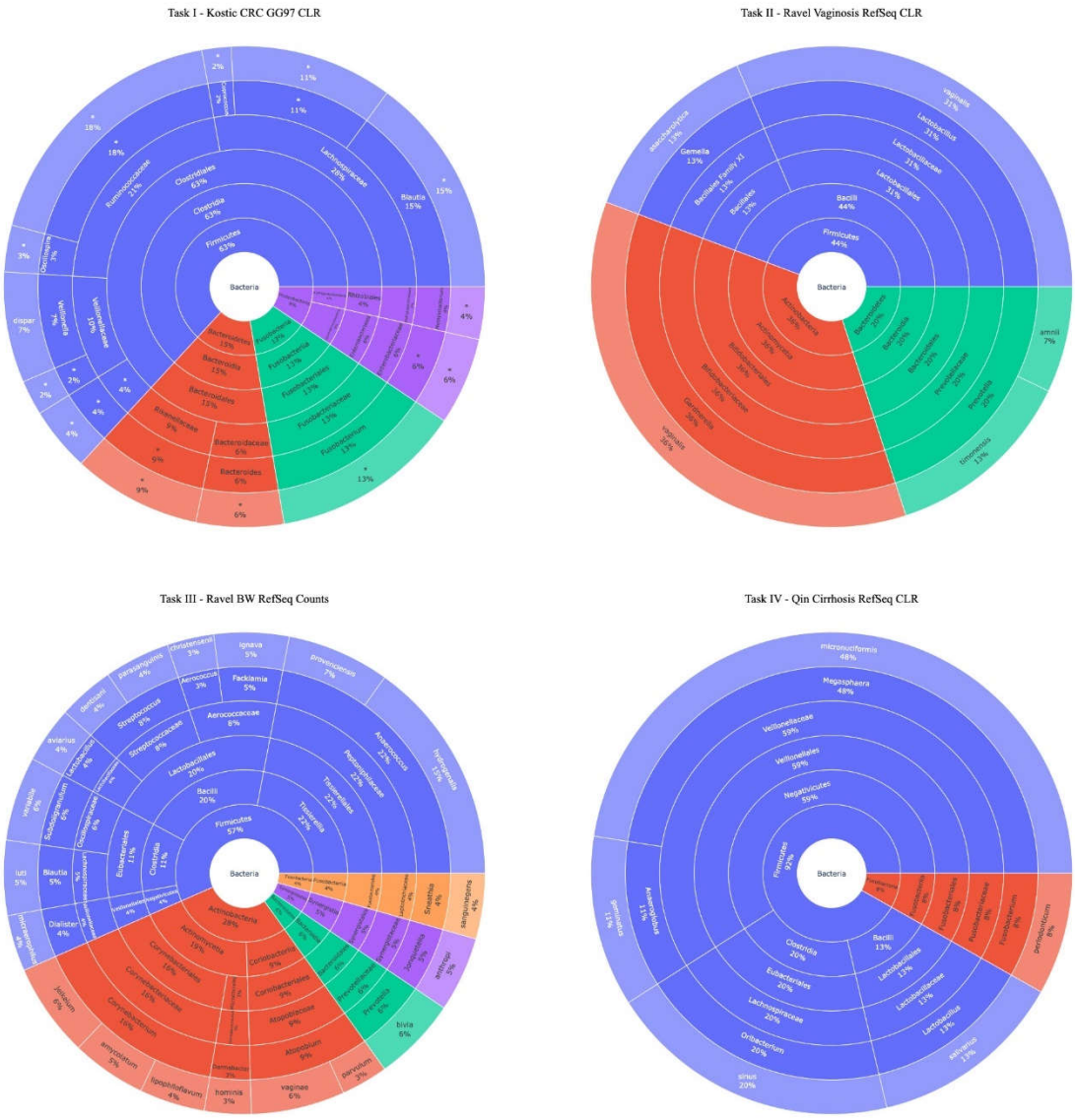

Figure 3. Overview of the bacteria selected by the 25 best-performing GA individuals for all classification tasks. (I) Kostic Colorectal Cancer Healthy/Tumor GG97 Counts, (II) Ravel Vaginal Nugent Category RefSeq CLR, (III) Ravel Vaginal Black/White RefSeq Counts, (IV) Qin Cirrhosis RefSeq CLR.

4. Discussion

Microbiome datasets are sparse, with a high number of zero counts since most microorganisms are not present in all samples, and highly dimensional, being the number of OTUs normally exceeding the number of samples by an order of magnitude, which are known challenges for Machine Learning applications. Herein we present BiGAMi, a new OTU selection method for Microbiome data using Genetic Algorithms to tackle the dimensionality burden by reducing the number of OTUs used in ML classification tasks, while retaining a high classification score. We also compare its results with the Sequential Forward Feature Selection (SFFS) method provided by the *MLxtend* library and baseline results for each study. Out of 21 distinct input data experiments across the 4 different classification tasks, BiGAMi achieved on 18 a performance score better than or equal to the SFFS method (Table 3). Only for 3 input data, the SFFS performance was marginally higher than BiGAMi's performance metric at the cost of a larger OTU subset. For 19 input data experiments, BiGAMi achieved its classification performance with up to 68% fewer

OTUs than the SFFS method. Performance and OTU selection results for classification tasks like (II) Ravel Vaginal Nugent Category where even the baseline result achieved a metric of 0.99 only differed marginally between SFFS and BiGAMi. Both algorithms were able to identify the very limited number of bacteria needed to reliably classify samples into the correct categories. While both SFFS and BiGAMi achieved the average classification performance per data input with comparable 99% confidence intervals (Figure 4), BiGAMi identified its 25 best-performing OTU subsets more consistently around the average number of OTUs selected per data input (Figure 5 and Figure 6). Using an SGDClassifier without any form of OTU selection expectedly resulted in less performant classification results than those achieved with BiGAMi or SFFS.

Table 3. Performance results per classification task. Per input data, the average AUC score and the average number of OTUs for SFFS and BiGAMi solutions are provided. (I) Kostic Colorectal Cancer Healthy/Tumor GG97, (II) Ravel Vaginal Nugent Category, (III) Ravel Vaginal Black/White, (IV) Qin Cirrhosis RefSeq.

Task	Database	Total OTUs	Baseline AUC	Input Data	SGD AUC	SFFS AUC / OTUs	BiGAMi AUC / OTUs (ours)	Gain BiGAMi AUC / OTUs (ours)
(I)	GG97	3228	0.74	Counts	0.86	0.96 / 22.4	0.94 / 15.9	-2.08% / 29.02%
				Rel	0.85	0.94 / 21.6	0.95 / 18.1	1.06% / 16.20%
				CLR	0.90	0.94 / 19.0	0.94 / 12.3	0.00% / 35.26%
	RefSeq	908	0.69	Counts	0.83	0.84 / 23.6	0.85 / 14.9	1.19% / 36.86%
				Rel	0.80	0.87 / 22.4	0.86 / 16.8	-1.03% / 25.00%
				CLR	0.85	0.85 / 10.0	0.84 / 8.2	-1.18% / 18.00%
(II)	GG97	1093	0.99	Counts	0.95	0.98 / 9.6	0.99 / 9.0	1.02% / 6.25%
				Rel	0.96	0.98 / 9.8	0.98 / 9.2	0.00% / 6.12%
				CLR	0.97	0.99 / 6.6	0.99 / 8.6	0.00% / -30.30%
	RefSeq	586	0.99	Counts	0.95	0.98 / 8.7	0.98 / 8.0	0.00% / 8.05%
				Rel	0.96	0.98 / 7.4	0.98 / 6.6	0.00% / 10.81%
				CLR	0.96	0.98 / 5.5	0.98 / 6.5	0.00% / -18.18%
(III)	GG97	1093	0.64	Counts	0.64	0.73 / 22.1	0.76 / 13.6	4.11 % / 38.46%
				Rel	0.65	0.71 / 19.4	0.77 / 16.2	8.45% / 16.49%
				CLR	0.63	0.73 / 16.2	0.73 / 11.1	0.00% / 31.48%
	RefSeq	586	0.70	Counts	0.60	0.72 / 17.6	0.78 / 12.8	5.56% / 27.27%
				Rel	0.60	0.75 / 19.4	0.75 / 13.6	0.00% / 29.90%
				CLR	0.61	0.73 / 14.4	0.73 / 10.2	0.00% / 29.17%
(IV)	RefSeq	8483	0.92	Count	0.83	0.91 / 14.7	0.94 / 12.2	3.30% / 17.01%
				Rel	0.82	0.92 / 16.4	0.93 / 11.6	1.09% / 28.27%
				CLR	0.83	0.93 / 13.5	0.93 / 4.3	0.00% / 68.15%

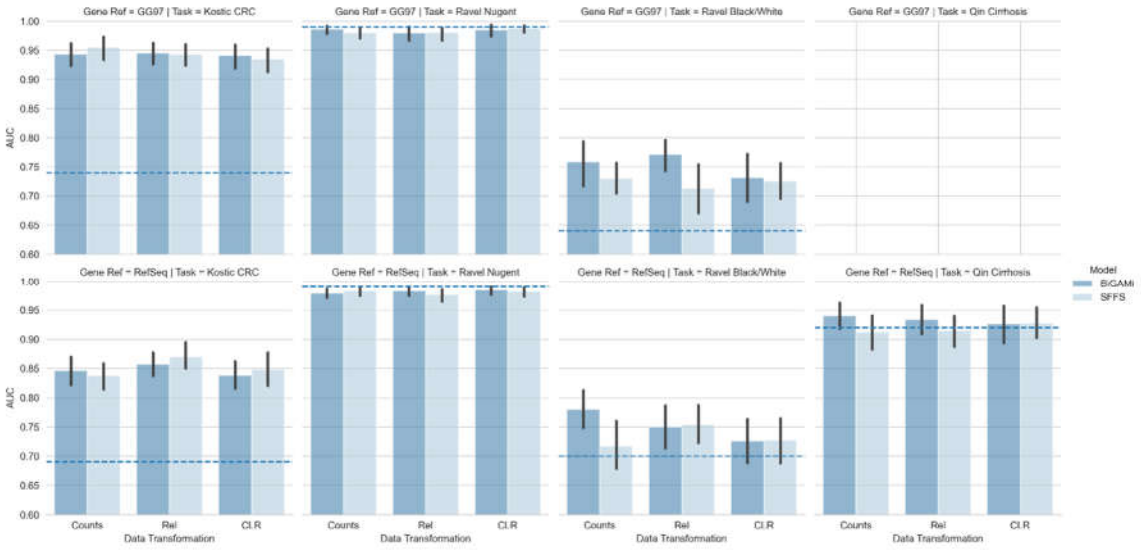

Figure 4. Average classification performance (including 99% confidence interval) for each data input achieved by BiGAMi and SFFS methods. Top: Performance results for GG97-based data input. Bottom: Performance results for RefSeq-based data input.

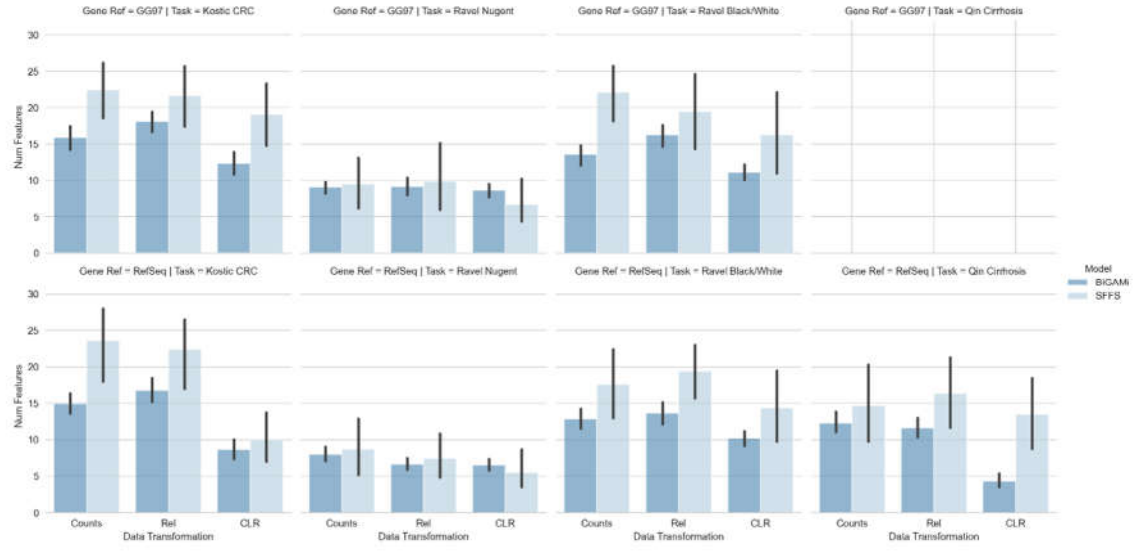

Figure 5. Average number of selected OTUs (including 99% confidence interval) for each data input achieved by BiGAMi and SFFS methods. Top: Selected OTUs for GG97-based data input. Bottom: Selected OTUs for RefSeq-based data input.

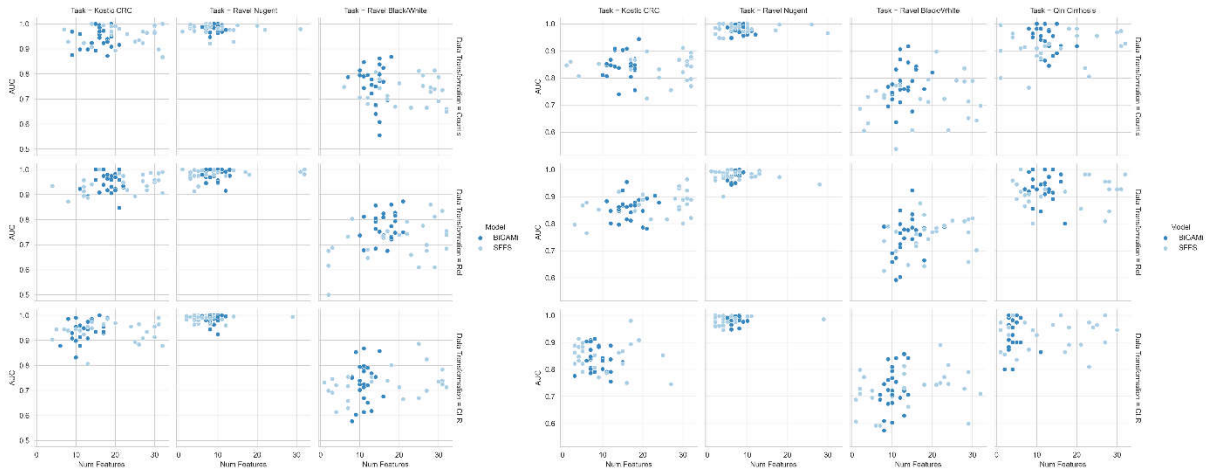

Figure 6. Scatter plot representing classification performance and number of selected OTUs for each data input achieved by BiGAMi and SFFS methods. **Left**: Results for GG97-based data input. **Right**: Results for RefSeq-based data input.

The results have shown that a bi-objective Genetic Algorithm fitness function helps in building and training well-performing host-state classifiers using a minimized subset of OTUs that improve predictive performance over the baseline models by Vangay et al. [9] and also exceed or matches the performance of the SFFS algorithm on almost all data inputs. At the same time, BiGAMi achieves these classification performance results by drastically reducing the number of predictive OTUs compared to SFFS. The use of a fitness function that merges the actual classification performance and the chromosome size of an individual into a single metric is essential for guiding the GA search towards finding high-performance OTU subsets and has proven to work efficiently on microbiome datasets.

Literature review confirmed that the proposed method indeed extracts microbiologically relevant OTUs per classification task for well-known diseases, making this method suitable for reliable identification of relevant microbes for novel diseases or phenotypes in non-disease tasks:

- Task I (Kostic Colorectal Cancer): The several selected microbial families are well known biomarkers for the detection of colorectal cancer [18-22]. Zhong et al. [18] describes the relation of *Collinsella aerofaciens* and *Bacteroides*, among others, to the development of colorectal cancer. Gao et al. [19] discovered that, among others, *Blautia* were significantly reduced in cancer patients, while Bacteroides fragilis and *Fusobacterium nucleatum* were enriched. Flemer et al. [20] document that cancer patients display an increased abundance of *Ruminococcus*.
- Task II (Ravel Vaginal Nugent Category): Multiple recent publications confirm that the selected OTUs, including *Gardnerella vaginalis* and *Lactobacilli*, are related to the development of Bacterial Vaginosis which itself is diagnosed by a high vaginal Nugent score [23-25].
- Task III (Ravel Vaginal Black/White): Fettweis et al. [26] already documented that Caucasian females have a vaginal microbiome dominated by *Lactobacillus crispatus*, among others, whereas women of African heritage show higher abundances of *Anaerococcus* and *Atopobium*. Women of European ancestry, when diagnosed with bacterial vaginosis were more likely to be colonized by *Corynebacterium*.
- Task IV (Qin Cirrhosis): According to Chen et al. [27] at the genus level *Megasphaera*, among others, shows higher abundance counts in cirrhosis duodenum, while *Lachnospiraceae* show decreased abundances in the salivary microbiome of cirrhotic patients. Yang et al. [28] document the protective effect of *Lactobacillus salivarius* on liver injuries. Jensen et al. [29] discovered that *Fusobacterium periodonticum* is enriched in cirrhosis patients.

This study only evaluated the effectiveness of the GA-based OTU selection on classification problems. It is expected that additional regularization operations are required to trade off the regression metric and the number of selected OTUs. General GA search parameters like the number of generations and population size have not been deeply assessed but selected in a way that limited compute capacity led to good results. It remains for future research to define parameter guidelines that produce equally good results with reduced computational cost.

5. Conclusions

This study demonstrated the successful application of a Genetic Algorithm with a bi-objective fitness function to select the most predictive combination of OTUs from microbiome datasets to classify host phenotypes. It has been shown that such a GA evolutionary search for the most predictive feature (OTU) subset improves classification performance for all classification problems. Where classifiers without a feature selection already achieved almost perfect results, our proposed BiGAMi method performed 'on par'. Furthermore, BiGAMi achieved its results by selecting up to 68% fewer OTUs than a Sequential Forward Feature Selection method we compared our results to. The proposed method to select the most predictive features per classification task resulted in the selection of <2.5% of the original number of OTUs on average across the different classification tasks and data transformations. 10 out of 21 experiments even resulted in selecting <1% of the original number of OTUs on average reducing the feature space by 2 orders of magnitude. Compared to methods relying on the adoption of Deep Learning and Variational Autoencoders, this feature space reduction helps simpler classifiers to find patterns in the data more easily and improve the interpretability of the classification results. Especially for machine learning models used for medical diagnoses, this is an important capability.

Author Contributions: ML was responsible for data curation, formal analysis, investigation, methodology and writing. FB was responsible for conceptualization and writing. HA and BGNA were responsible for supervision, conceptualization, funding acquisition, project administration and writing.

Funding: This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106. This research has partially received funding by the Horizon 2020 projects STOP Obesity Platform under Grant Agreement No. 823978.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used for this study are accessible via the following links:
https://github.com/knights-lab/MLRepo/tree/master/datasets/kostic
https://github.com/knights-lab/MLRepo/tree/master/datasets/ravel
https://github.com/knights-lab/MLRepo/tree/master/datasets/qin2014

The code generated for the BiGAMi framework is available on GitHub:
https://github.com/mikeleske/BiGAMi

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Knights, D.; Costello, E.K.; Knight, R. Supervised Classification of Human Microbiota. *FEMS Microbiol. Rev.* **2011**, *35*, 343–359.
2. Statnikov, A.; Henaff, M.; Narendra, V.; Konganti, K.; Li, Z.; Yang, L.; Pei, Z.; Blaser, M.J.; Aliferis, C.F.; Alekseyenko, A.V. A Comprehensive Evaluation of Multicategory Classification Methods for Microbiomic Data. *Microbiome* **2013**, *1*, 11.
3. Min, S.; Lee, B.; Yoon, S. Deep Learning in Bioinformatics. *Brief. Bioinform.* **2016**, bbw068.

4. Zhou, Y.-H.; Gallins, P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Front. Genet.* **2019**, *10*, 579.
5. Bang, S.; Yoo, D.; Kim, S.-J.; Jhang, S.; Cho, S.; Kim, H. Establishment and Evaluation of Prediction Model for Multiple Disease Classification Based on Gut Microbial Data. *Sci. Rep.* **2019**, *9*, 10189.
6. Carter, J.; Beck, D.; Williams, H.; Foster, J.; Dozier, G. GA-Based Selection of Vaginal Microbiome Features Associated with Bacterial Vaginosis. *Genet. Evol. Comput. Conf.* **2014**, *2014*, 265–268.
7. Zhang, P.; West, N.P.; Chen, P.-Y.; Thang, M.W.C.; Price, G.; Cripps, A.W.; Cox, A.J. Selection of Microbial Biomarkers with Genetic Algorithm and Principal Component Analysis. *BMC Bioinformatics* **2019**, *20*, 413.
8. Chiesa, M.; Maioli, G.; Colombo, G.I.; Piacentini, L. GARS: Genetic Algorithm for the Identification of a Robust Subset of Features in High-Dimensional Datasets. *BMC Bioinformatics* **2020**, *21*, 54.
9. Vangay, P.; Hillmann, B.M.; Knights, D. Microbiome Learning Repo (ML Repo): A Public Repository of Microbiome Regression and Classification Tasks. *Gigascience* **2019**, *8*.
10. Kostic, A.D.; Gevers, D.; Pedamallu, C.S.; Michaud, M.; Duke, F.; Earl, A.M.; Ojesina, A.I.; Jung, J.; Bass, A.J.; Tabernero, J.; et al. Genomic Analysis Identifies Association of Fusobacterium with Colorectal Carcinoma. *Genome Res.* **2012**, *22*, 292–298.
11. Qin, N.; Yang, F.; Li, A.; Prifti, E.; Chen, Y.; Shao, L.; Guo, J.; Le Chatelier, E.; Yao, J.; Wu, L.; et al. Alterations of the Human Gut Microbiome in Liver Cirrhosis. *Nature* **2014**, *513*, 59–64.
12. Ravel, J.; Gajer, P.; Abdo, Z.; Schneider, G.M.; Koenig, S.S.K.; McCulle, S.L.; Karlebach, S.; Gorle, R.; Russell, J.; Tacket, C.O.; et al. Vaginal Microbiome of Reproductive-Age Women. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108 Suppl 1*, 4680–4687.
13. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. **2012**.
14. McDonald, D.; Price, M.N.; Goodrich, J.; Nawrocki, E.P.; DeSantis, T.Z.; Probst, A.; Andersen, G.L.; Knight, R.; Hugenholtz, P. An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J.* **2012**, *6*, 610–618.
15. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44*, D733-45.
16. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* **2018**, *3*, 638.
17. De Rainville, F.-M.; Fortin, F.-A.; Gardner, M.-A.; Parizeau, M.; Gagné, C. DEAP: A python framework for evolutionary algorithms. In Proceedings of the Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference companion - GECCO Companion '12; ACM Press: New York, New York, USA, **2012**.
18. Zhong, M.; Xiong, Y.; Ye, Z.; Zhao, J.; Zhong, L.; Liu, Y.; Zhu, Y.; Tian, L.; Qiu, X.; Hong, X. Microbial Community Profiling Distinguishes Left-Sided and Right-Sided Colon Cancer. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 498502.
19. Gao, R.; Gao, Z.; Huang, L.; Qin, H. Gut Microbiota and Colorectal Cancer. *Eur. J. Clin. Microbiol. Infect. Dis.* **2017**, *36*, 757–769.
20. Yang, J.; McDowell, A.; Kim, E.K.; Seo, H.; Lee, W.H.; Moon, C.-M.; Kym, S.-M.; Lee, D.H.; Park, Y.S.; Jee, Y.-K.; et al. Development of a Colorectal Cancer Diagnostic Model and Dietary Risk Assessment through Gut Microbiome Analysis. *Exp. Mol. Med.* **2019**, *51*, 1–15.
21. Flemer, B.; Lynch, D.B.; Brown, J.M.R.; Jeffery, I.B.; Ryan, F.J.; Claesson, M.J.; O'Riordain, M.; Shanahan, F.; O'Toole, P.W. Tumour-Associated and Non-Tumour-Associated Microbiota in Colorectal Cancer. *Gut* **2017**, *66*, 633–643.
22. Xu, K.; Jiang, B. Analysis of Mucosa-Associated Microbiota in Colorectal Cancer. *Med. Sci. Monit.* **2017**, *23*, 4422–4430.
23. Chee, W.J.Y.; Chew, S.Y.; Than, L.T.L. Vaginal Microbiota and the Potential of Lactobacillus Derivatives in Maintaining Vaginal Health. *Microb. Cell Fact.* **2020**, *19*, 203.
24. Morrill, S.; Gilbert, N.M.; Lewis, A.L. Gardnerella vaginalis as a Cause of Bacterial Vaginosis: Appraisal of the Evidence From in vivo Models. *Front. Cell. Infect. Microbiol.* **2020**, *10*.
25. Diop, K.; Dufour, J.-C.; Levasseur, A.; Fenollar, F. Exhaustive Repertoire of Human Vaginal Microbiota. *Hum. microbiome j.* **2019**, *11*, 100051.
26. Fettweis, J.M.; Brooks, J.P.; Serrano, M.G.; Sheth, N.U.; Girerd, P.H.; Edwards, D.J.; Strauss, J.F.; The Vaginal Microbiome Consortium; Jefferson, K.K.; Buck, G.A. Differences in Vaginal Microbiome in African American Women versus Women of European Ancestry. *Microbiology* **2014**, *160*, 2272–2282.
27. Chen, Y.; Ji, F.; Guo, J.; Shi, D.; Fang, D.; Li, L. Dysbiosis of Small Intestinal Microbiota in Liver Cirrhosis and Its Association with Etiology. *Sci. Rep.* **2016**, *6*, 34055.
28. Yang, L.; Bian, X.; Wu, W.; Lv, L.; Li, Y.; Ye, J.; Jiang, X.; Wang, Q.; Shi, D.; Fang, D.; et al. Protective Effect of Lactobacillus Salivarius Li01 on Thioacetamide-Induced Acute Liver Injury and Hyperammonaemia. *Microb. Biotechnol.* **2020**, *13*, 1860–1876.
29. Jensen, A.; Ladegaard Grønkjær, L.; Holmstrup, P.; Vilstrup, H.; Kilian, M. Unique Subgingival Microbiota Associated with Periodontitis in Cirrhosis Patients. *Sci. Rep.* **2018**, *8*.