**The topic is important, but neither reviewer felt that the paper was in the top 10%. There is much jargon that needs to be explained, and I also feel that the paper would benefit from a parallel analysis of another data set--which could be one that is publicly available. I also wonder whether with all that is needed to be done, the paper can fit inside the size guidelines of mBio.**

We appreciate the comments from the Editor and two reviewers. All line numbers, figures, and tables below refer the marked-up version of the manuscript.

We appreciate the hesitation to accept a manuscript that is not perceived to be in the top 10%. However, it is well established that we are generally unable to predict the long term importance of a manuscript [see Casadevall & Fang, <https://doi.org/10.1128/mBio.01593-15>]. For what it’s worth, the preprint version of this manuscript (<https://doi.org/10.1101/816090>) has been accessed 1,856 times since it was initially submitted in October and it has an Altmetric score of 37, placing it in the top 5% of all research outputs scored by Altmetric. It has also already been cited in another preprint (<https://doi.org/10.1101/2019.12.31.891978>). We would suggest that the initial enthusiasm for a preprint version of this manuscript suggests that it will have both impact and importance to the microbiome research community.

One recommendation from the Editor that is referenced by the first reviewer is to include an additional dataset. As the Editor points out, it is unlikely that we can add an additional dataset within the space constraints of *mBio*. With this in mind, we have decided not to add another dataset. Furthermore, this somewhat misses the point of the manuscript which is to show a framework for using machine learning techniques to analyze microbiome data. The dataset that’s used or the results of analyzing that dataset really aren’t important. Rather, the discussion of using multiple modeling techniques and interpretability are the important points in this manuscript. Throughout the manuscript, including the title, we have attempted to highlight that the point of the paper is the framework rather than the biological findings.

**Reviewer #1**

**This is an interesting albeit not exactly novel or general in terms of the number of datasets analyzed benchmark of ML methods for classification with microbiome data. Please see the criticism below. I hope that it will help in improving and tightening the authors' message.**

**Effective application of machine learning to microbiome-based classification problems**

**Running title: Machine learning methods applied to microbiome studies**

**The title is misleading. It reads as if the article provides evidence based approach to machine learning generalized on many data; however, the analysis involves a single dataset. I strongly suggest renaming the article to highlight that its focus is primarily an analysis of a single dataset.**

We appreciate the reviewer’s concern that readers might make the wrong impression from the title. We have revised the title to highlight that the manuscript describes a framework for applying machine learning practices to microbiome data. Repeating the analysis for other datasets will not change the framework, except to suggest that different modeling approaches are preferred over others depending on the datasets; however, this is already one of the suggestions we make in the manuscript and why we suggest using multiple methods and comparing the results based on performance and interpretability. We rephrased L75-76 to clarify that we are using the CRC dataset as a case study to establish reproducible ML methods.

**The authors mention some previous work that yielded essentially similar results in terms of the performance of machine learning methods. For example, Statnikov A et al. 2013 is based on a benchmark of several datasets and show that RF and regularized regression show best performance in ML without preceding feature selection. That study also demonstrates that feature selection generally helps improve model predictivity regardless of classification technique used. This undermines the value of the present manuscript as a novel benchmark. However, there is still scientific value in the analyzed data.**

Statnikov et al 2013 provided a comparative analysis of machine learning algorithms used with microbiome data. However, the comparison was performed to find the ML model with highest accuracy for each of the 8 different classification problems. That study also does not discuss problems of interpretability. Our study, builds upon this earlier work and aims to showcase a standardized ML pipeline and a set of guidelines that works for all the classification algorithms. We utilize a single case study to provide a commentary on the interpretability, speed, and predictive performance of these algorithms in comparison to one another. Finally, we also provide an open-source pipeline that can be used with any microbiome dataset for a binary classification problem while providing food for thought on model selection. We added sections in Introduction (L68-69, 83-86) to emphasize the goal of the study.

**The Importance statement mentions health outcomes however the dataset analyzed is observational and does not include an outcome, just an incident diagnosis.**

We have changed L21 and 27 to indicate that we are predicting a diagnosis.

**It has been shown that rarefaction (i.e. "subsampled to the size of the smallest sample") is an inadmissible statistical practice for microbiome data analysis (**[**https://doi.org/10.1371/journal.pcbi.1003531**](https://doi.org/10.1371/journal.pcbi.1003531)**). The authors use this approach in feature preparation. Further this approach may affect generalizability of the models to new data that was no co-processed with this training-testing dataset. How would a new sample be normalized if the feature preparation included a step that "normalized across samples"?**

As discussed in the Discussion (L286-295) there are still a number of issues that need to be resolved before an ML model can be deployed in the clinic or applied to other datasets. We would foresee subsampling the new data to the same number of reads found in the training dataset. The concern of the McMurdie paper the reviewer references is that by subsampling we remove information about the variation in the data that is gained by having the non subsampled data. We are unaware of any ML approach that takes into account variation in the number of observations per sample. Furthermore, the approach that was advocated for in that paper assumes that all OTUs are present in all communities and ignores the ecological reality that microbial communities are patchy. It should also be mentioned that not subsampling data to a common number of reads leads to a risk that an ML model would classify samples based on the abundance of rare taxa found in well-sequenced samples. Far from being “inadmissible”, this is a fairly controversial opinion in the microbial ecology literature.

**The authors state that they "used L2-regularization to keep all potentially important features"; however L2 penalty does not shrink the coefficients to exactly zero, therefore it cannot be used for feature selection. It is not clear what the authors had in mind.**

We developed linear ML models with two different regularization methods (i.e. L1-regularized SVM with linear kernel and L2-regularized logistic regression) to compare the effects of L1 and L2 regularization in training the models. The feature coefficients (regardless of the regularization method) can be evaluated to interpret the model. The features with the largest coefficients (and in the case of L1-regularization, the coefficients that are not 0) are chosen as the “most important” for both models. To avoid confusion, we removed the L96-100 where different regularization methods were described.

**line 128. "though not significantly (p=0.5)" state the statistical test used.**

More description is added to the text (L131 and 364-366) and the caption of FigureS4.

**The methodology describes taxonomy assignment procedures used for the OTUs, however, the main text of the manuscript does not mention a single important taxon. More insight into what these are will make the manuscript more interesting. This is especially important for the author's claim that interpretable models are more preferred.**

It is important to note that this manuscript is not intended to be an exposition of the taxa that are important to diagnosing people with screen relevant neoplasias in their colons. It is meant to be a description and application of a framework for evaluating ML models. The discussion of such features has already been described in previous studies (e.g. Baxter, ref 5). We added L265-266 and included a new citation to emphasize that a model that can provide accurate prediction of an outcome does not necessarily mean that the predictors used by the model cause the outcome. The features defined as important by the proposed interpretation methods in our study must be further analyzed in the physical world with controlled experiments.

To give more flavor for the types of features the models select, we have added Figure S8 to show the abundances of the most important 20 OTUs/features with taxonomy information. There were minimal differences in relative abundances between the healthy and SRN groups across OTUs. This supports our claim that it is not possible to differentiate disease vs healthy states by focusing on individual taxa. For example, this figure shows *Peptostreptococcus* (OTU00367), which was identified as important across models, has a broad distribution of relative abundances in the healthy and disease states. We interpret this to mean that the importance of these taxa is context dependent.

**Reviewer #2 (Comments for the Author):**

**Summary:**

**The authors sought to demonstrate the importance of justification and documentation in choosing appropriate machine learning models for microbiome analyses. Here they addressed the temptation among microbiome researchers to adopt increasingly complex machine learning methods, by performing side-by-side AUROC and feature selection comparisons of both linear and non-linear models in binary classification. Importantly, differences in AUROC between the first (random forest - non-linear) and second (L2-regularized logistic regression - linear) highest-scoring models were not statistically significant (p=0.5). The authors emphasized that complexity does not necessarily yield more accurate results, and that great care should be taken when choosing models and hyperparameters for different datasets. In addition, immediate interpretability and time-consumption should also be considered.**

**In total, seven machine learning models (three linear, and four non-linear) were trained using existing fecal 16S rRNA sequence data to predict the presence of colonic screen relevant neoplasias (SRNs). Of the 490 samples used, SRNs were present in 229 cases; cross-validated and test subsets were stratified to retain this ratio. While this paper used only 16S data, the authors stressed that best practices and major takeaways may be applied to all microbiome studies, both environmental and clinical.**

**Opinion:**

**In general, this paper addresses an often-overlooked and poorly documented variable in microbiome research - parameters and interpretation in data processing - and provides steps toward standardization. It introduces a healthy amount of skepticism when interpreting classification data and feature selection, and represents a starting point for further discussion. The paper is well written, documentation and methods are clear, and it does excel at providing fodder for individual thought. However, it would benefit from a more clear call to action. For readers not familiar with machine learning techniques, clearly showing the abundances of the OTUs/features displayed in Figs 3 and 4 would help uncover the black box of the ML outputs. Furthermore, examples of when these ML models have been successful in the applications proposed here (e.g. biomarker identification) and information about the data required for success would be very interesting to include in the intro or discussion.**

**Furthermore, examples of when these ML models have been successful in the applications proposed here (e.g. biomarker identification) and information about the data required for success would be very interesting to include in the intro or discussion.**

We appreciate the reviewer’s enthusiasm for this manuscript and our approach to making the material and methods transparent. Their concern that the manuscript would benefit from a more clear call to action is valued. We have added a figure (Figure S1) that shows an NMDS ordination of Bray-Curtis distances between the 490 samples. It should be clear from this figure that inter-personal variation is large in human microbiome samples and that there is no clear “healthy” or “SRN” microbiome. We have also added strip charts showing the relative abundances of the most important OTUs for the best models as Figures S8-A and S-8B. Those stripcharts make it clear that there are no clear differences between diagnosis groups when OTUs are treated separately. Rather, each OTU is important in a specific context. We have added text at L38-40 to strengthen the rationale for why ML approaches are needed. To strengthen the “call to action” more, we also added text to the Introduction (L68-69, 83-86) and Discussion (Lxxx) to urge the audience to use reproducible methods, justify their model selection, be diligent with their study design, model development, evaluation and interpretation.

**Major comments:**

**1. For microbiologists familiar with multivariate stats or beta diversity analyses often used in microbiome studies, it might help to justify initial sample size and address inherent variation and familiarize the readers with the overall structure of the data using a familiar statistical approach like an NMDS ordination with Bray-Curtis distances and a PERMANOVA test. How much variance is associated with the dataset/batch, the individual the sample came from, and the health vs disease state? This would also help put the features chosen by the ML approaches in perspective.**

**2. Line 41: how does ML overcome the interpersonal variation in microbiome composition?**

Please see the discussion in the previous comments to Reviewer 2.

**3. Showing the absolute abundance of the OTUs that were determined to be important would help make the feature weights (e.g. Figs 3 and 4) less of a black box and more tangible. Were they high or low abundance? Were increases or decreases associated with health vs disease, across models? For example, was the Peptostreptococcus (OTU00367) that was found to be important across models shown in Fig 4 abundant or present across many of the healthy and diseased samples?**

We addedFigure S8 to clearly show the abundances of the most important 20 OTUs/features in L2-regularized logistic regression and random forest models. We added L209-213 to explain the relative abundance differences for these OTUs. There were minimal relative abundance differences between the 2 diagnoses for the 20 OTUs. This supports our claim that it is not possible to differentiate disease vs healthy states by focusing on individual taxa by traditional statistical analyses. This figure also shows Peptostreptococcus (OTU00367)which was identified as important across models and how the relative abundances cannot be differentiated between healthy and disease states.

**4. Would using 90% OTUs (or another method for grouping similar taxa together) change which features are selected? (As discussed, taxa that are highly correlated may have similar functions, and it may make sense to group them together.)**

It is likely that model performance will vary with the taxonomic level being used to assign sequences to features. It could be that a model trained at the family level might outperform one trained at OTUs or that using amplicon sequence variants (ASVs) might perform better. As the reviewer points out higher taxonomic levels could aggregate microbial guilds but lower taxonomic levels could provide more resolution by splitting apart taxa. It might be useful o train models that include all levels of taxa. Unfortunately, this question is beyond the scope of the current manuscript. It is an area of active research for our group.

**5. Permutation importance and perfect correlation were difficult to interpret for someone unfamiliar with the techniques. Lines 176-178 were particularly confusing to read. Two separate explanations or figures for permutation importance and perfect correlation grouping could help (Figure 1 was incredibly easy to read).**

We edited L176-178 to improve their clarity. We also generated a Figure S6 as a diagram to explain how permutation importance is performed for individual features and a group of perfectly correlated features.

**6. How many features actually had perfect correlation, and why was perfect correlation chosen over something more common, like 0.9?**

We added L250-253 to explain our reasoning and urge the reader to consider correlation structures of their own datasets. We found 432 of the 6,920 OTUs were perfectly correlated with each other.

**7. Tree models have a built-in feature selection (e.g. gini drop). Why was permutation importance chosen over these built-in methods? Advantages of permutation importance would be worth highlighting to an audience using RF packages in R.**

Though built-in feature selection methods are used to grow the tree they do not necessarily have the same meaning nor are they guaranteed to return the same results as permutation importance. Permutation importance is a model-agnostic tool that can be used to compared feature importance across different models. We added L253-260 to discourage the audience from using package built-in interpretation methods. We added a new citation (ref. 38) to a previous study which point out the biases in random forest variable importance measures such as gini drop.

**8. Weight feature selection and permutation importance were compared for linear models (lines 196-198), but linear and non-linear permutation importance were not. Is there a significant amount of overlap between the two? Should variation in features selected by linear and non-linear models be considered, and does accuracy (AUROC) hold any weight in which models' features should be chosen?**

We added L209-213 to include a comparison between L2-logistic and tree-based models when permutation importance is used as the interpretation method. There were a few OTUs that overlapped when a single interpretation method was used.

We added L265-266 and included a new citation to emphasize that a model that can provide accurate prediction of an outcome does not mean that the predictors used by the model are causes of the outcome. The features defined as important in the prediction must be further analyzed in the physical world with controlled experiments.

**9. One huge advantage in interpretability for linear models is the inherent negative or positive sign associated with each feature weight (lines 170-173). Regardless of arguments for or against the accuracy of weight feature selection, this inherent "risk" value should be emphasized in the discussion as an immediately interpretable characteristic, unique to simple linear models. This would support the argument for using weights over permutation importance or other methods of feature selection.**

We rephrased L173-176 to underscore this argument.

**10. In general, these models did not seem to vary much in their predictive power. Would the authors choose one method over another for any given purpose? For instance, if they were "researchers trying to identify the microbiota associated with a disease" (line 64), would they choose the L2-regularized regression's weight feature selection, its permutation importance, or the random forest's permutation importance? If they chose L2-regularized regression for its speed and interpretability, would they go on to perform permutation importance with groups made from Spearman's rank-order correlations, and would the benefits of this convenience be lost in doing so?**

As the reviewer pointed out, except decision tree and L1-regularized SVM with linear kernel, the models tested in this study do not vary much in predictive performance. This is exactly why we underscored the importance of choice of modeling approach based on the goal of the study (L65-69).

If the researchers are trying to identify the microbiota associated with disease using linear models, we recommend researchers to use both of the methods described in the study. We have added L253-257 and added new citations (ref. 37, 38) to emphasize this point. More importantly, we also want to discourage the audience from using package built-in interpretation methods, as they are often not reproducible and the documentation is opaque.

**11. In general, this paper does an excellent job of providing reasons to delve deeper into results to avoid misinterpretation, but by the end, this big picture principle is largely lost among vague statements that might have been more useful in the introduction (lines 275-282). In the discussion, a final interpretation of the data would help generate a call to action (e.g. variance in feature selection for high AUROC models demonstrates the importance of considering correlation structures; variation in accuracy across all models during hyperparameter selection highlights a need to document all steps in data processing; etc.).**

We added sections in Introduction (L68-69, 83-86) and Discussion (Lxxx) to urge the audience to use reproducible methods, justify their model selection, be diligent with their study design, model selection, development, evaluation and interpretation.

**Minor Comments:**

**1. Lines 52-54: Difficult to understand. "ways to better use ML tools" a little vague.**

Edited for clarification.

**2. Lines 100-102: Awkward to read.**

To avoid confusion, we removed the lines that describe L1 and L2 regularization methods.

**3. Lines 143-144: Are there generally accepted thresholds for detection of overfitting?**

In theory, if the validation data are representative of the test data then there should be no statistically significant difference in performance between the cross-validation and test performance. A statistically significant difference could indicate a flaw in the cross-validation setup that has led to overfitting to the validation set. Though there is no accepted ‘threshold’ this phenomenon can be identified by observing the training curves (e.g., number of iterations versus training/validation loss) which we have shown by plotting all iteration AUROCs for cross-validation and testing in Figure 2. Alternatively, one might identify overfitting by examining the curves resulting from hyperparameter selection (e.g., sweeping the hyperparameter C) which we have performed as shown in Figures S2 and S3.

**4. Line 222: Why is a random forest not interpretable?**

We were unable to find a statement in the manuscript where we indicate that random forest is not interpretable. In the section of text the reviewer cites, we point out that linear methods are *easier* to interpret, but not that random forest is *not* interpretable.

**5. Line 501: "ensemble"**

We fixed this typo.

**6. Figure 4: Colors too similar. Blue missing from legend. Green inaccurate in legend (listed as XGBoost and decision tree, but actually XGBoost, decision tree, and RBF SVM).**

We changed the colors so that they are easier to differentiate. We also made sure the legend has all the groupings for overlapping OTUs.

**7. Line 542: "reflective of the population to which the model will be applied". Inconsistent use of "e.g." and "e.g.,".**

We fixed both.

**8. Figure S3: Check typos and variation in use of "data-splits" both in figure and in caption. Line 557: "were" -> "is" (referring to "percentage", and inconsistent tense). Line 558: "models". In general, this figure is difficult to interpret. Consider adding, very explicitly, that the x-axis is AUROC of random forest - AUROC of L2-regularized log regression, and that each "data-split" is one of your 100 rounds. Caption is confusing - much easier to say something like, "In 75% of data-splits, the AUROC of random forest was greater than that of L2-regularized log regression," in place of lines 555-557. Line 555: what is the number of data-splits in each bin? Mention that this represents how often random forest performed more accurately than did L2-regularized log regression. The red line may not be necessary. Also, for those who are not statistically savvy, what is a double tail event?**

We rewrote Figure S3 caption for clarification. We removed the red line and corrected typos in the figure.