

RESEARCH

Open Access



Multi-class boosting for the analysis of multiple incomplete views on microbiome data

Andrea Simeon^{1*†}, Miloš Radovanović², Tatjana Lončar-Turukalo³, Michelangelo Ceci^{4,5,6}, Sanja Brdar¹ and Gianvito Pio^{4,5*†}

[†]Andrea Simeon and Gianvito Pio have contributed equally to this work.

*Correspondence:
andrea.simeon@biosense.rs;
gianvito.pio@uniba.it

¹ BioSense Institute, University of Novi Sad, dr Zorana Djindjića 1, Novi Sad 21000, Serbia

² Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 3, Novi Sad 21000, Serbia

³ Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad 21000, Serbia

⁴ Department of Computer Science, University Bari Aldo Moro, Via Orabona 4, 70125 Bari, Italy

⁵ Big Data Laboratory, National Interuniversity Consortium for Informatics (CINI), Via Ariosto 25, 00185 Rome, Italy

⁶ Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Abstract

Background: Microbiome dysbiosis has recently been associated with different diseases and disorders. In this context, machine learning (ML) approaches can be useful either to identify new patterns or learn predictive models. However, data to be fed to ML methods can be subject to different sampling, sequencing and preprocessing techniques. Each different choice in the pipeline can lead to a different view (i.e., feature set) of the same individuals, that classical (single-view) ML approaches may fail to simultaneously consider. Moreover, some views may be incomplete, i.e., some individuals may be missing in some views, possibly due to the absence of some measurements or to the fact that some features are not available/applicable for all the individuals. Multi-view learning methods can represent a possible solution to consider multiple feature sets for the same individuals, but most existing multi-view learning methods are limited to binary classification tasks or cannot work with incomplete views.

Results: We propose irBoost.SH, an extension of the multi-view boosting algorithm rBoost.SH, based on multi-armed bandits. irBoost.SH solves multi-class classification tasks and can analyze incomplete views. At each iteration, it identifies one winning view using adversarial multi-armed bandits and uses its predictions to update a shared instance weight distribution in a learning process based on boosting. In our experiments, performed on 5 multi-view microbiome datasets, the model learned by irBoost.SH always outperforms the best model learned from a single view, its closest competitor rBoost.SH, and the model learned by a multi-view approach based on feature concatenation, reaching an improvement of 11.8% of the F1-score in the prediction of the Autism Spectrum disorder and of 114% in the prediction of the Colorectal Cancer disease.

Conclusions: The proposed method irBoost.SH exhibited outstanding performances in our experiments, also compared to competitor approaches. The obtained results confirm that irBoost.SH can fruitfully be adopted for the analysis of microbiome data, due to its capability to simultaneously exploit multiple feature sets obtained through different sequencing and preprocessing pipelines.

Keywords: Multi-view learning, Incomplete views, Multi-armed bandits, Boosting, Microbiome analysis



Introduction

Over the last few years, microbiome dysbiosis has been associated with many diseases, since it tends to occur as an accompanying symptom with a higher prevalence than usual in disease conditions. Microbiome alterations can be related to some types of cancer [1–3] or, through the so-called *microbiota-gut-brain* axis, to neurodevelopmental conditions such as the Autism Spectrum Disorder (ASD) [4, 5].

Sequencing of microbiome samples is considered convenient, especially when other tests are impossible, hard or invasive to perform. However, despite the reduction of sequencing costs, they can still be problematic, specifically for large microbiome studies or whole genome studies. In this context, the Next Generation Sequencing (NGS) technology has revolutionized microbiome sequence analysis. Indeed, while initially the research targeted only hypervariable regions of conserved genes such as 16S ribosomal RNA (rRNA) gene, more recently, it has been expanded on longer reads [6, 7] and on the whole genome. In 16S (*amplicon*) metagenomic sequencing, short chains (< 500 bp) of nucleotides from 16S rRNA gene are amplified by PCR and then read out. This is further used to identify and differentiate multiple microbial species from multiple samples at once. In spite of simplification, this approach suffers from several drawbacks: poor resolution (only up to the level of species); poor diversity detection; sample source dependency, specifically on sampling, sequencing and analysis protocols, such DNA extraction kits and alignment algorithms [6–8]; poor reproducibility [6].

The sequencing of the whole genome solved some of the above-mentioned limitations, among which the most important is identification accuracy. This sequencing technique, called *shotgun*, involves randomly breaking up the genome into small overlapping DNA fragments, that are sequenced individually and virtually reassembled. Even if more precise, it also exhibits some drawbacks: the issue about the sample source dependency is not solved, and it is generally more expensive.

16S amplicon sequencing was initially very popular, and the amount of available data collected with this technique is currently much larger than that of the shotgun counterpart. However, several studies now aim to adopt both approaches [4], even if their simultaneous exploitation in a pipeline of analysis is challenging.

In the literature, we can find several attempts involving the adoption of machine learning (ML) techniques for the analysis of microbiome data [9–11]. However, to the best of the authors' knowledge, there is no work aiming to consider simultaneously 16S and shotgun data in the construction of ML models. A simple solution would consist in the concatenation of the features representing the same individual, obtained through 16S and shotgun techniques. However, this approach can exacerbate the curse of dimensionality [12], which is already present in microbiome data due to the high unbalancing between the (usually low) number of individuals and the (usually high) number of features. In this context, *multi-view learning* approaches [12–14] represent a possible solution, since they mainly follow a *decision level* fusion, based on average/voting approaches, similarly to ensemble techniques. In these methods, different feature sets of the same instances (e.g., pictures of the same objects taken from different perspectives, features representing conceptually different aspects, or features obtained after the application of different preprocessing techniques) are called *views*, and a ML model is learned

on each view separately. Subsequently, their output is combined and iteratively exploited to boost the performance of the learned models.

Among the pioneering techniques, we can mention the co-training framework [15], that works on two views in the semi-supervised setting. In this framework, unlabeled examples are labeled by each of the two classifiers, and the most reliable predictions are fed as new training examples to the classifier working on the other view, for the subsequent iteration(s). This approach was originally limited to working on exactly two views. Different variants have been proposed in the literature, such as the co-regularization technique [16, 17], that was mainly adopted for reducing overfitting issues. In [16], the authors extended it to work with more than two views, while in the [18], the authors proposed a multi-view learning approach able to work in the positive-unlabeled learning setting for gene network reconstruction.

In the work [19], the authors extended the well-known boosting algorithm Adaboost [20] to the multi-view learning setting. In [21], the authors proposed an efficient version using adversarial multi-armed bandits for selecting one of the views based on its usefulness, instead of repeatedly considering all views. Other approaches for multi-view learning exploit kernel-based algorithms. Relevant examples include the system proposed in [22], which uses within- and between-view regularization, and the work [23], where the authors formulate the learning problem on each view as a Multiple Kernel Learning (MKL) problem. Both approaches were compared with four different methods proposed in [12], along with the original Adaboost on concatenated views, and the system Mumbo [24], a multi-view algorithm for multi-class classification that exploits a separate instance weight distribution for each view. Among the considered competitors and the variants proposed in [12], rBoost.SH, which is based on partial information games, obtained the best results.

One of the strongest limitations of the aforementioned approaches is that they can only work when samples are fully observed on all the considered views. However, in real-world situations, some of the available observations may not be represented by all the views (i.e., they are partially observed). A relevant example in the context of microbiome data can be represented by individuals for whom we have a 16S sample but not a shotgun sample, or vice versa. In this scenario, the above-mentioned approaches cannot be adopted without a preprocessing step aimed to fill missing values, possibly introducing excessive approximations in the data. In [25], the authors proposed a multi-view model to complete missing values using a SOR-like optimization algorithm. The issue about incomplete views in multi-view learning has also been tackled in two recent papers: [14] considered data incompleteness for the multi-view clustering task, while [26] adopted a graph-based learning approach for classification with incomplete views.

Another limitation of most of multi-view methods comes from their inability to solve multi-class classification tasks. While some approaches like Mumbo [24] tackled this issue, to the best of the authors' knowledge, none of them can simultaneously solve multi-class classification tasks and work on incomplete views.

In this paper, we aim to fill this gap by proposing a novel multi-view approach that is able to work on incomplete views (see an example in Fig. 1) to solve multi-class classification problems. We adopt the proposed method in the context of microbiome studies for two different scenarios: (i) to classify individuals according to 16S and/or shotgun

COMPLETE VIEWS				INCOMPLETE VIEWS			
	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$		$f_{2,1}$	$f_{2,2}$	$f_{2,3}$
s_1	5	10	2	s_1	10	7	2
s_2	1	1	2	s_2	1	2	2
s_3	7	4	1	s_3	3	4	7
s_4	5	4	3	s_4	5	4	6

Fig. 1 An example showing the difference between complete (on the left) and incomplete (on the right) views: rows correspond to the instances (individuals, in the case of microbiome studies); each view represents a different viewpoint of the instances (e.g., 16S and shotgun data, or features obtained through different preprocessing pipelines). In the case of incomplete views, we can observe that not all instances are represented according to both views (e.g., s_3 is not represented in the first view, while s_2 and s_4 are not represented in the second view)

microbiome data, which represent two, possibly incomplete, different views; (ii) to classify individuals considering multiple preprocessing pipelines to generate different views from the same 16S data. Through the former scenario we demonstrate the ability of our approach to fully exploit both views of the microbiome data, while in the latter we show the applicability of our approach to solve one of the critical issues of the analysis of microbiome data, namely, the selection of the best preprocessing pipeline(s), which in our case is fully automated.

The remainder of the paper is structured as follows: in ‘[Methods](#)’ section we describe the proposed approach in detail, starting from the background concepts it is based on; ‘[Experimental setting](#)’ section presents the considered tasks, together with the adopted datasets and the experimental setting considered; in ‘[Results and discussion](#)’ section we show and discuss the obtained results. Finally, in ‘[Conclusion](#)’ section we conclude the paper and present some avenues for future work.

Methods

The approach we propose in this paper is based on the method [rBoost.SH](#) [12]. Specifically, we significantly extend it to make it able to solve multi-class classification tasks and to work with incomplete views. We call our approach [irBoost.SH](#). In ‘[Boost.SH and rBoost.SH: binary classifier via multi-view boosting from complete views](#)’ section, we first briefly describe the methods [Boost.SH](#) and [rBoost.SH](#), together with the main concepts they rely on. Then, in ‘[The proposed method irBoost.SH: multi-class classifier via multi-view boosting from incomplete views](#)’ section we provide the details about our method [irBoost.SH](#).

Boost.SH and rBoost.SH: binary classifier via multi-view boosting from complete views

In this subsection, we provide a quick overview of *Boost.SH*, which is the simplest version among the multi-view algorithms proposed in [12, 27], and then we describe its extension *rBoost.SH* which is based on adversarial multi-armed bandits. Both *Boost.SH* and *rBoost.SH* adopt a single instance weight distribution that is shared among the

views, unlike the multi-armed version of AdaBoost [21], which allows each view to adopt its own weight distribution. We remind that the approach we propose in this paper is based on *rBoost.SH*, which can originally solve only binary classification tasks from complete views.

Boost.SH starts from training data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i = \{x_i^1, x_i^2, \dots, x_i^K\}$ is a training instance, $x_i^j \in \mathbb{R}^{q_j}$ is the instance x_i represented according to the j -th view in its q_j -dimensional feature space, and $y_i \in \mathcal{Y} = \{-1, +1\}$ is the label of the instance x_i . Instances (\mathbf{x}_i, y_i) are assumed to be independently and identically distributed according to a probability distribution D over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^q$ and $q = \sum_{j=1}^K q_j$. At the beginning, the shared instance weights are uniformly distributed.

For each iteration of the algorithm and for each view, a *weak classifier* is trained and, on the basis of the predictions on training data, an *edge* is computed, that somehow represents the current predictive accuracy of such a view. Afterwards, the algorithm selects the view with the largest edge as the winning view. The edge of the winning view and the corresponding classifier are used to update the shared instance weights for the subsequent iterations. Finally, after T iterations, *Boost.SH* builds a combined classifier as a weighted sum of the winning classifiers over all the T iterations.

Despite optimizing the *consistency* through the shared weight distribution, *Boost.SH* does not promote *diversity* [12]. In general, the algorithm is limited to learn only from the views with the maximal edge. The extreme situation happens when a view wins for all the iterations, that would degenerate to learning from one single view, discarding the possible contribution provided by the other(s).

This limitation has been overcome by its randomized version, called **rBoost.SH**, that is formalized in Algorithm 1. It incorporates *adversarial multi-armed bandits* (AMB) into a boosting process by computing probabilities $\mathbf{p}_{1 \times K}$ for choosing the actions to take (i.e., the views to select as winner).

Auer et al. [28, 29] proposed different versions of AMB. The one of interest here is a partial information game algorithm, called EXP3.P, in which a *player* (an algorithm) and an *adversary* (views) compete, and the reward is the only information propagated to the player. Unlike other problems based on bandits, in this case, no statistical assumptions are made about the process followed to generate the rewards (stochastic bandits, for instance, use the assumption that rewards are generated from a given, pre-determined, distribution). In **rBoost.SH**, the AMB problem is addressed using the exponentially weighted average (EWA) forecaster algorithm [12, 30]. The forecaster algorithm updates the probability distribution over actions/views, such that the probability of choosing an action exponentially depends on the average rewards associated with it [12]. The update of this probability distribution is controlled by two parameters, namely σ and γ .

Contrary to *Boost.SH*, *rBoost.SH* computes the predictions only along the selected view j (see line 5 in Algorithm 1). The reward for such a view is divided by the probability of selecting it (see line 9 in Algorithm 1), which encourages the views with low probability to be selected, promoting diversity.

Algorithm 1 The method rBoost.SH

Input: $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ represented through K views; number of iterations T ; $\sigma > 0$; $\gamma \in (0, 1]$.

- 1: Initialize $w_i^1 = \frac{1}{n}$ for $i = 1, \dots, n$; $p_j^1 = e^{\frac{\sigma\gamma}{3}} \sqrt{\frac{T}{K}}$ for $j = 1, \dots, K$
- 2: **For each** $t = 1, 2, \dots, T$ **do**
- 3: $q_j^t = (1 - \gamma) \frac{p_j^t}{\sum_{k=1}^K p_k^t} + \frac{\gamma}{K}$, $j = 1, \dots, K$
- 4: Let j^* be the view chosen according to \mathbf{q}_t
- 5: Compute the base classifier predictions $h_t^{j^*}(\mathbf{x}^{j^*})$ with \mathbf{w}^t
- 6: Compute the edge $\theta_t^{j^*} = \sum_{i=1}^n w_i^t y_i h_t^{j^*}(x_i^{j^*})$
- 7: Compute the reward $r_t^{j^*} = 1 - \sqrt{1 - (\theta_t^{j^*})^2}$
- 8: **for** $k = 1, \dots, K$ **do**
- 9: $\hat{r}_t^k = \begin{cases} r_t^{j^*}/q_k^t & \text{if } k = j^* \\ 0 & \text{otherwise} \end{cases}$
- 10: $p_{t+1}^k = p_t^k \cdot e^{\left(\frac{\gamma}{3K} \left(\hat{r}_t^k + \frac{\sigma}{q_k^t \sqrt{KT}}\right)\right)}$
- 11: **end for**
- 12: $\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \theta_t^{j^*}}{1 - \theta_t^{j^*}} \right)$
- 13: $w_i^{(t+1)} = \frac{w_i^t}{Z_t^{j^*}} \cdot e^{-\alpha_t y_i h_t^{j^*}(x_i^{j^*})}$
 ($Z_t^{j^*}$ is chosen such that $\sum_{i=1}^n w_i^{(t+1)} = 1$)
- 14: **end for**

Output: $H(\hat{\mathbf{x}}) = \text{sign}(\sum_{t=1}^T \alpha_t h_t^{j^*}(\hat{\mathbf{x}}^{j^*}))$

The proposed method irBoost.SH: multi-class classifier via multi-view boosting from incomplete views

In this subsection, we present our method irBoost.SH, which overcomes the main limitations of rBoost.SH, namely, it can solve multi-class classification tasks and can also work with incomplete views, i.e., when some instances are not represented according to all the views.

The proposed method irBoost.SH is formalized in Algorithm 2. In this case, the set S consists of views of possibly different shapes: each instance can be present in one or more views, but not necessarily in all the views, and the number of features in each view can be different.

For each view j , with $j = 1, 2, \dots, K$, we denote with \mathcal{N}_j the set of instances that are present in view j , and with $\mathcal{N} = \bigcup_{j=1}^K \mathcal{N}_j$ the set of all the instances. Accordingly, $|\mathcal{N}|$ denotes the total number of instances. Every instance is associated with a class label $y_i \in \mathcal{Y} \subset \mathbb{N}^+$, which corresponds to a multi-class classification task, with each different integer representing a different class.

The algorithm starts by initializing the weights w_i for each instance $i \in \mathcal{N}$, and probabilities p_j for each view $j = 1, 2, \dots, K$ (Algorithm 2, line 1). Lines 3 and 4 of Algorithm 2 are the same as lines 3 and 4 of Algorithm 1 (rBoost.SH). Besides computing the predicted label, the weak classifier in irBoost.SH outputs a *class probability vector* \mathbf{v}_i for each instance (Algorithm 2, line 5), that is a confidence vector with components in $[0, 1]$ for each class $l \in \mathcal{Y}$. The predicted label is used in the computation of the edge (Algorithm 2, line 7), thus indirectly also for the computation of the reward (Algorithm 2, line 8) and for the update

of the shared weights (Algorithm 2, line 14). Unlike rBoost.SH, irBoost.SH uses class probabilities in the output prediction.

Because of the possibility to have incomplete views, we define the modified predicted label $\hat{h}^j(x_i^j)$ and the modified class probability vector $\hat{\mathbf{v}}^j(x_i^j)$ for $i \in \mathcal{N}$, as follows:

$$\hat{h}^j(x_i^j) = \begin{cases} h^j(x_i^j) & \text{if } i \in \mathcal{N}_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\hat{\mathbf{v}}^j(x_i^j) = \begin{cases} \mathbf{v}^j(x_i^j) & \text{if } i \in \mathcal{N}_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Note that 0 in Eq. (1) does not correspond to a class, but it is an artificial label (note that $\mathcal{Y} \subset \mathbb{N}^+$) for unknown class label prediction for instances not contained in that view j . In that case, the artificial prediction, and hence unknown instance representation, will not contribute to the computation of the edge nor to the update of the weights. Accordingly, in Eq. (2), 0 represents the unknown probability vector, which is, however, used in the prediction phase.

We also update the computation of the edge towards the multi-class classification task. In the original rBoost.SH, the multiplication of real and predicted labels $y_i \cdot h_t^j$ leads to 1 when they agree, i.e., $1 \cdot 1$ or $(-1) \cdot (-1)$, and to -1 when they disagree, i.e., $(-1) \cdot 1$ or $1 \cdot (-1)$. Analogously, we define the function \mathbb{I} , that generalizes such agreement, as follows:

$$\mathbb{I}[a = b] = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Accordingly, we also modify the sum of multiplications of the original rBoost.SH ($\sum_{i \in \mathcal{N}} y_i \cdot h(x_i)$, where $y_i, h(x_i) \in \{-1, +1\}$) as $2 \cdot \sum_{i \in \mathcal{N}} (\mathbb{I}[y_i = h(x_i)] - 0.5)$, where \mathbb{I} is computed through Eq. (3), and $y_i, h(x_i) \in \mathbb{N}$. More formally, we derive the updated sum of multiplications as follows:

$$\begin{aligned} \sum_{i \in \mathcal{N}} y_i h(x_i) &= \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 - \sum_{\substack{y_i \neq h(x_i) \\ i \in \mathcal{N}}} 1 = \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 - \left(n - \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 \right) \\ &= \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 - \left(\sum_{i \in \mathcal{N}} 1 - \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 \right) = 2 \sum_{\substack{y_i = h(x_i) \\ i \in \mathcal{N}}} 1 - \sum_{i \in \mathcal{N}} 1 \\ &= 2 \sum_{i \in \mathcal{N}} \mathbb{I}[y_i = h(x_i)] - \sum_{i \in \mathcal{N}} 1 = \sum_{i \in \mathcal{N}} 2\mathbb{I}[y_i = h(x_i)] - \sum_{i \in \mathcal{N}} 1 \\ &= \sum_{i \in \mathcal{N}} (2\mathbb{I}[y_i = h(x_i)] - 1) = \sum_{i \in \mathcal{N}} 2\left(\mathbb{I}[y_i = h(x_i)] - \frac{1}{2}\right) \\ &= 2 \sum_{i \in \mathcal{N}} \left(\mathbb{I}[y_i = h(x_i)] - \frac{1}{2}\right) \end{aligned} \tag{4}$$

The above term is almost the same used in the computation of the edge (Algorithm 2, line 7), except for the fact that it is weighted. Since \mathbb{I} compares real and predicted values, and outputs values in $\{0, 1\}$, now the class label can be any positive natural number. In the case of incomplete views, the predicted and the real label for an instance that is absent in the chosen view would never be the same, because of the definition given in Eq. (1) and the assumption that class labels are from a subset of positive natural numbers. Accordingly, it will not contribute to the total sum. Formally:

$$\sum_{i \in \mathcal{N}} y_i \hat{h}(x_i) = 2 \sum_{i \in \mathcal{N}} (\mathbb{I}[y_i = \hat{h}(x_i)] - 0.5)$$

As a final remark, we want to point out that the approach followed by irBoost.SH (but also its predecessor rBoost.SH) exhibits an interesting advantage: the number of times each view is selected as the winning view can be seen as an indicator of its relevance. As emphasized in the introduction, this aspect is very useful when views correspond to different preprocessing pipelines in order to identify the most useful ones for the predictive task at hand.

Algorithm 2 The proposed method irBoost.SH

Input: $S = \{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{N}}$ represented through K views, where \mathcal{N}_j is the set of instances in j -th view, and $y_i \in \mathcal{Y} \subset \mathbb{N}$; number of iterations T ; $\sigma > 0$; $\gamma \in (0, 1]$

- 1: Initialize $w_i^1 = \frac{1}{|\mathcal{N}|}$ for $i \in \mathcal{N}$; $p_j^1 = e^{\frac{\sigma\gamma}{3}} \sqrt{\frac{T}{K}}$ for $j = 1, \dots, K$
- 2: **For each** $t = 1, 2, \dots, T$ **do**
- 3: $q_j^t = (1 - \gamma) \frac{p_j^t}{\sum_{k=1}^K p_k^t} + \frac{\gamma}{K}, j = 1, \dots, K$
- 4: Let j^* be the view chosen according to \mathbf{q}_t
- 5: Compute base classifier predictions $h_t^{j^*}(x_i^{j^*})$
 and class probabilities $\mathbf{v}_t^{j^*}(x_i^{j^*})$ with w_i^t , for $i \in \mathcal{N}_{j^*}$
- 6: Set $\hat{h}_t^{j^*}(x_i^{j^*})$ with Equation (1) and $\hat{\mathbf{v}}_t^{j^*}(x_i^{j^*})$ with Equation (2)
- 7: Compute edge $\theta_t^{j^*} = 2 \sum_{i \in \mathcal{N}} w_i^t (\mathbb{I}[y_i = \hat{h}_t^{j^*}(x_i^{j^*})] - 0.5)$
 where $\mathbb{I}[\cdot = \cdot]$ is given in Equation (3)
- 8: Compute reward $r_t^{j^*} = 1 - \sqrt{1 - (\theta_t^{j^*})^2}$
- 9: **for** $k = 1, \dots, K$ **do**
- 10: $\hat{r}_t^k = \begin{cases} r_t^{j^*}/q_k^t & \text{if } k = j^* \\ 0 & \text{otherwise} \end{cases}$
- 11: $p_{t+1}^k = p_t^k e^{\left(\frac{\gamma}{3K} \left(\hat{r}_t^k + \frac{\sigma}{q_k^t \sqrt{KT}}\right)\right)}$
- 12: **end for**
- 13: $\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \theta_t^{j^*}}{1 - \theta_t^{j^*}}\right)$
- 14: $w_i^{(t+1)} = \frac{w_i^t}{Z_t^{j^*}} \cdot e^{-2 \alpha_t (\mathbb{I}[y_i = \hat{h}_t^{j^*}(x_i^{j^*})] - 0.5)}$
 ($Z_t^{j^*}$ is chosen such that $\sum_{i \in \mathcal{N}} w_i^{(t+1)} = 1$)
- 15: **end for**

Output: $H(\hat{\mathbf{x}}) = \operatorname{argmax}_{l \in \mathcal{Y}} \frac{\sum_{t=1}^T \alpha_t \hat{\mathbf{v}}_{t,l}^{j^*}(\hat{\mathbf{x}}^{j^*})}{C}$,
 (C is chosen such that $\sum_{l \in \mathcal{Y}} \sum_{t=1}^T \alpha_t \hat{\mathbf{v}}_{t,l}^{j^*}(\hat{\mathbf{x}}^{j^*}) = 1$)

Experimental setting

In this section, we describe the experimental setting of the evaluation we performed to assess the effectiveness of the proposed method irBoost.SH. As introduced at the end of section ‘Introduction’, we considered two different scenarios: *i*) the classification of individuals according to 16S and/or shotgun microbiome data, which represent two, possibly incomplete, different views; *ii*) the classification of individuals considering multiple preprocessing pipelines and hyper-parameter settings to generate multiple views from the same 16S data. For the former, we consider data related to the (binary) classification of individuals in terms of presence or absence of Autism Spectrum Disorders (ASD); for the latter, we consider 16S microbiome data related to ASD and 16S microbiome data for the (multi-class) classification of individuals according to the degree (or absence, in case of healthy individuals) of the Colorectal Cancer (CRC) disease.

In the following subsections, we first describe the contexts and the datasets; then we describe the pipelines followed to define the views; finally, we describe the evaluation setting and the considered competitors.

Autism spectrum disorder (ASD)

The Autism Spectrum Disorder (ASD) is a severe neurodevelopmental disorder that is primarily characterized by abnormal behavioral symptoms: social interaction impairment, stereotyped behavior, and restricted interests. Recent studies have shown a significant association between this disease and gut microbiome through the so-called microbiota-gut-brain axis [4, 31, 32].

The considered dataset is publicly available at the NCBI repository¹ and consists of a cohort of 143 children with clinical diagnosis of ASD, aged 2–13 years old, and age- and sex-matched typically developing (TD) individuals (average age 5.189 ± 0.170 ; 127 males and 16 females), who attended annual physical examinations. 16S rRNA sequencing of feces samples was performed for all individuals, while shotgun metagenomic sequencing was performed only for 30 ASD and 30 TD individuals. We considered two multi-view versions of the dataset:

- **MV-ASD**, that contains pre-computed 16S rRNA and shotgun Operational Taxonomic Unit (OTU) tables as available at the Kaggle repository,² where 16S data represent the first view and shotgun data represent the second view (see quantitative information in Table 1³);
- **ASD-16S**, that we constructed by downloading the raw 16S rRNA sequences from NCBI and by preprocessing them using the pipelines described in ‘Data preprocessing pipelines’ section, obtaining 40 different views for all children in the study (143 ASD + 143 TD).

¹ <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA453621>

² <https://www.kaggle.com/datasets/antaresnyc/human-gut-microbiome-with-asd>

³ Note that the number of available TD individuals in the dataset available at the Kaggle repository is slightly less than the original dataset available at NCBI (28 missing TD individuals). This is possibly due to some preprocessing steps applied by the authors of the dataset available at the Kaggle repository.

Table 1 Quantitative details of the MV-ASD dataset

	# Individuals			# Features
	ASD	TD	Total	
16S	143	111	254	1322
Shotgun	30	30	60	5619
16S Only	113	85	198	1322
Shotgun Only	0	4	4	5619
16S + Shotgun	30	26	56	6941

Table 2 Experimental settings for the MV-ASD dataset

Setting	# Individuals	Features	Classifier
MV-ASD-1 (incomplete)	254	16S	RF (SV)
	254	16S+shotgun	Concat-RF (MV)
	254	16S+shotgun	rBoost.SH (MV)
	254	16S+shotgun	irBoost.SH (MV)
MV-ASD-2 (incomplete)	60	shotgun	RF (SV)
	60	shotgun+16S	Concat-RF (MV)
	60	shotgun+16S	rBoost.SH (MV)
	60	shotgun+16S	irBoost.SH (MV)
MV-ASD-3 (complete)	56	16S+shotgun	Concat-RF (MV)
	56	16S+shotgun	rBoost.SH (MV)
	56	16S+shotgun	irBoost.SH (MV)

It is noteworthy that MV-ASD is inherently incomplete: primarily, incompleteness comes from a much lower number of individuals represented through the shotgun view with respect to those represented through the 16S view; moreover, some 16S samples were discarded during the preprocessing of the dataset (see footnote 3). Therefore, in the dataset we have individuals that are represented with (i) only 16S features, (ii) only shotgun features, or (iii) both 16S and shotgun features. Accordingly, we defined three different experimental settings, that are summarized in Table 2. The first setting (**MV-ASD-1**) considers only individuals that are represented according to (at least) the 16S view: the representation according to the shotgun view is considered only for those individuals that also have a representation in the 16S view. The second setting (**MV-ASD-2**) considers only individuals that are represented according to (at least) the shotgun view: the representation according to the 16S view is considered only for those individuals that also have a representation in the shotgun view. Finally, the third setting (**MV-ASD-3**) only focuses on those individuals who are represented according to both 16S and shotgun views. Therefore, in this setting, views are complete.

Colorectal cancer (CRC)

Colorectal cancer (CRC) is the second leading cause of death among cancers in the USA [1]. Several studies suggest that gut microbiota may represent a reservoir of biomarkers that would complement existing non-invasive methods, such as the fecal immunochemical test (FIT) [1–3].

The considered dataset, hereafter called **CRC**, consists of OTU tables computed from 16S rRNA sequences for the prediction of the CRC condition between *CRC* (191 samples), *Adenoma*⁴ (241 samples) or *Control* (277 healthy samples). Sequences were downloaded and merged from three repositories^{5,6,7} selected according to the guidelines⁸ produced within ML4Microbiome COST Action CA18131,⁹ and preprocessed following the pipelines described in ‘Data preprocessing pipelines’ section, leading to a total of 40 different views.

Data preprocessing pipelines

Amplicon sequence data preprocessing was performed using QIIME 2 [33], version 2021.2. Forward sequences were denoised with DADA2 [34], truncated at 240 bp and closed-reference clustered against SILVA database by varying the similarity threshold in {90%, 94%, 97%, 99%}. Each of the produced OTU tables after clustering was then subjected to feature filtering, that has been performed according to the following criteria:

- by *frequency*, i.e., on the minimum number of individuals for whom a given feature must be present, in {2, 5, 10, 50};
- by *samples*, i.e., on the minimum percentage of the individuals for whom a given feature must be present, in {2%, 5%, 10%, 20%, 50%}.

Moreover, we also considered unfiltered OTU tables. In total, we obtained 40 different views, considering 4 different similarity thresholds in the clustering phase, multiplied by 10 different filtering configurations, i.e., 4 (filtering by frequency) + 5 (filtering by samples) + 1 (unfiltered).

For the ASD-16S dataset (‘Autism spectrum disorder (ASD)’ section), average reads quality scores were very good (higher than 30), except for one point in the region between 240 and 260 bp. This was the reason for truncating at 240bp in DADA2 denoising. On average, 88% of input ASD-16S sequences passed denoising and, in total 2517 of features were identified in Amplicon Sequence Variant (ASV) table, subsequently used for the clustering phase. The final number of features ranged from 219 to 1737, depending on the pipeline.

For the CRC dataset (‘Colorectal cancer (CRC)’ section), the quality of the paired-end sequences (both forward and reverse) was not particularly high (especially after 240bp). On average, around 70% of input sequences passed DADA2 denoising (trimmed at 240bp) and merging for paired-end reads. After denoising, 14055 ASV features were found, but this number was reduced by the clustering into OTUs and by the filtering step, leading to a final number of features ranging from 69 to 4402.

⁴ A tumor that is not cancer. It starts in gland-like cells of the epithelial tissue (thin layer of tissue that covers organs, glands, and other structures within the body).

⁵ <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA290926>.

⁶ <https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6070>.

⁷ <http://mothur.org/MicrobiomeBiomarkerCRC/>.

⁸ <https://doi.org/10.5281/zenodo.7382814>.

⁹ <https://www.cost.eu/actions/CA18131/>.

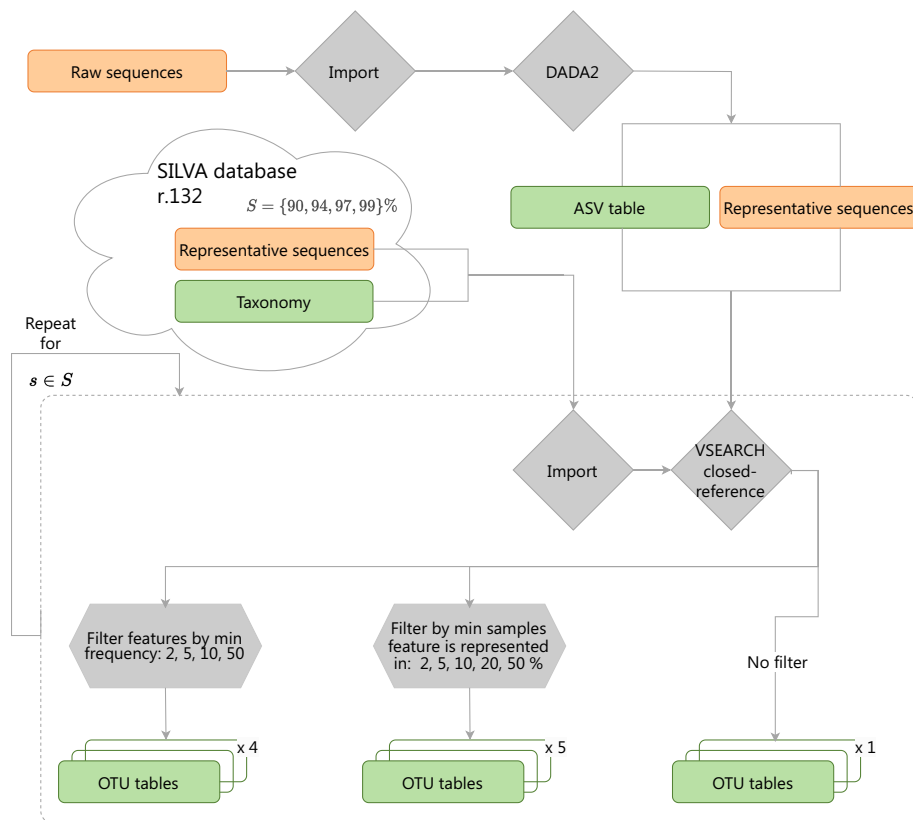


Fig. 2 QIIME 2 data preprocessing pipeline

A complete overview of the workflow followed for sequence preprocessing is provided in the Fig. 2.

Model evaluation and comparison

For irBoost.SH, we set the parameters σ and γ as suggested in [12] for rBoost.SH, namely, $\sigma = 0.15$ and $\gamma = 0.3$. As base classifiers we considered single decision trees (DT) and random forests (RF) with 50 trees. Since after 50 iterations the results appeared to be comparable, in the following, we only show those obtained through RF.

In order to actually assess the contribution provided by irBoost.SH in capturing the information conveyed by multiple views, we also performed the experiments with some competitor approaches. Specifically, we adopted the standard (single-view) version of RF, hereafter denoted with **RF (SV)**, for MV-ASD-1 and MV-ASD-2, as well as for all the 40 views of ASD-16S and CRC, that are considered separately. This comparison allows us to understand if the multi-view model learned by irBoost.SH outperforms all the models (2 in the case of MV-ASD-1 and MV-ASD-2, 40 in the case of ASD-16S and CRC) learned from a single view.

We also ran the experiments with the original version of rBoost.SH (hereafter denoted with **rBoost.SH (MV)**), on all the considered datasets except for the CRC dataset, on which it cannot be run due to the presence of multiple classes. Moreover, to overcome the incompleteness of the views for MV-ASD-1 and MV-ASD-2, we adopted a standard

Table 3 Results obtained by all the methods on all the considered datasets, in terms of Precision, Recall, F1-score, Accuracy and AUC

	Method	Prec. (%)	Rec. (%)	F1 (%)	Acc. (%)	AUC (%)
MV-ASD-1	RF (SV)	91.31	90.35	90.70	90.94	95.67
	Concat-RF (MV)	92.78	92.05	92.34	92.52	96.27
	rBoost.SH (MV)	91.07	87.59	88.41	88.98	–
	irBoost.SH (MV)	98.31	97.75	97.99	98.03	99.89
MV-ASD-2	RF (SV)	76.79	76.67	76.64	76.67	88.89
	Concat-RF (MV)	87.33	86.67	86.61	86.67	92.78
	rBoost.SH (MV)	98.39	98.33	98.33	98.33	–
	irBoost.SH (MV)	100.00	100.00	100.00	100.00	100.00
MV-ASD-3	Concat-RF (MV)	89.74	89.74	89.29	89.29	94.94
	rBoost.SH (MV)	98.39	98.08	98.20	98.21	–
	irBoost.SH (MV)	98.15	98.33	98.21	98.21	100.00
ASD-16S	RF (SV) Worst	88.84	88.81	88.81	88.81	95.48
	RF (SV) Average	91.49	91.45	91.45	91.45	96.80
	RF (SV) Best	93.74	93.71	93.71	93.71	97.78
	Concat-RF (MV)	90.98	90.91	90.91	90.91	96.52
	rBoost.SH (MV)	98.97	98.95	98.95	98.95	–
	irBoost.SH (MV)	99.31	99.30	99.30	99.30	99.99
CRC	RF (SV) Worst	38.16	35.97	35.54	37.71	56.17
	RF (SV) Average	46.59	42.53	42.98	43.77	61.56
	RF (SV) Best	51.63	46.52	47.31	47.46	64.16
	Concat-RF (MV)	47.51	44.11	44.73	45.06	62.17
	irBoost.SH (MV)	96.15	95.83	95.97	95.90	99.69

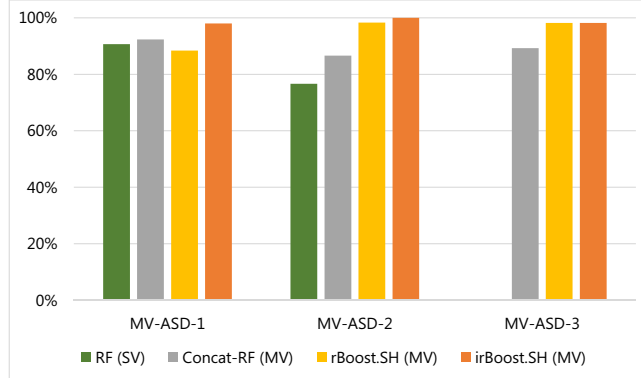
The best result for each measure and dataset is highlighted in bold

approach based on filling missing values with the mean value observed for each feature. Finally, we ran a baseline multi-view version of RF, hereafter denoted with **Concat-RF (MV)**, where the features coming from all the views are concatenated; also in this case, to manage the possible incompleteness of the views, we replaced missing values with the mean value observed for each feature.

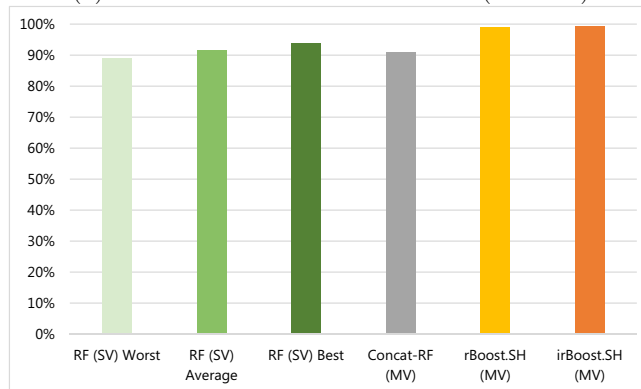
We want to remark that, although other base learners could have been considered in our experiments (either plugged into our framework or as standalone single-view systems), they were not included since the primary focus was on the assessment of the contribution provided by the multiple available (possibly incomplete) views, rather than on the comparison of the performance achieved through different base learners.

As the evaluation strategy we adopted 10-fold cross-validation. In the case of incomplete views, the partitioning was made such that each view is split into 10 equally-sized folds. The adopted performance measures are precision, recall, accuracy, F1-score and area under the ROC curve (AUC). Note that it was not possible to collect the AUC for rBoost.SH (MV) since it does neither return a probability distribution for labels nor a score/confidence, but directly the predicted label.

(a) F1-scores on the MV-ASD datasets (16S and shotgun views)



(b) F1-scores on the ASD-16S dataset (40 views)



(c) F1-scores on the CRC dataset (40 views)

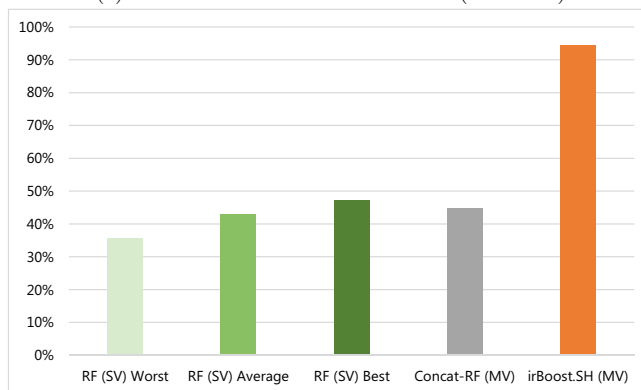


Fig. 3 F1-scores obtained on all the analyzed datasets by irBoost.SH, and by competitor single-view (SV) and multi-view (MV) approaches

Results and discussion

In Table 3, we show the results of the considered approaches on all the datasets in terms of all the evaluation measures, while in Fig. 3 we graphically summarize the results in terms of F1-scores. In Table 3 and Fig. 3, *RF (SV) Worst*, *RF (SV) Average* and *RF (SV) Best* denote, respectively, the worst, the average and the best result achieved among all the constructed views, by the single-view Random Forest model. The whole set of results obtained by each single view is reported in Additional file 1, while some confusion matrices are shown in Additional file 2. In the following, we mainly discuss the results in

terms of F1-scores, but similar conclusions can be drawn considering the other evaluation measures.

Focusing on the MV-ASD datasets, we can observe that simultaneously considering both 16S and shotgun views is generally beneficial, even if they are exploited through a simple concatenation strategy, as done by Concat-RF (MV). However, in this case, the improvement obtained by adding shotgun data about 60 individuals over 254 total individuals represented through 16S data (MV-ASD-1) led to a less sensible improvement with respect to adding 16S data about 56 individuals over 60 total individuals represented through shotgun data (MV-ASD-2). The results obtained by rBoost.SH exhibit this phenomenon in a much more evident manner: on MV-ASD-1, rBoost.SH obtains worse results than RF (SV), while on MV-ASD-2 it exhibits interesting improvements. This phenomenon is possibly due to the fact that the number of missing values (replaced by the mean value of each feature) is, in the case of MV-ASD-1, much higher than in MV-ASD-2. This is supported by the fact that on MV-ASD-3, where views are complete, rBoost.SH obtains results that are comparable with those achieved by our method. This observation confirms the effectiveness of irBoost.SH in handling the incompleteness of the views.

In summary, we achieved an improvement of 8% over RF (SV), 6.1% over Concat-RF (MV), and 10.8% over rBoost.SH on MV-ASD-1; an improvement of 30.5% over RF (SV), 15.5% over Concat-RF (MV), and 1.7% over rBoost.SH on MV-ASD-2; an improvement of 10% over Concat-RF (MV) and a tie with rBoost.SH on MV-ASD-3.

The exploitation of the complementarity of the information conveyed by 16S and shotgun views, performed by irBoost.SH, is confirmed by the fact that the 16S view was selected as the winner in 51% of the iterations, while the shotgun view was selected as the winner in 49% of the iterations, emphasizing an almost equal importance and contribution of the views.

Focusing on the results on the other datasets, we can observe that the selection of the preprocessing pipeline in a single-view setting influences the results. Indeed, for the ASD-16S dataset, the F1-score ranges from about 89% (in the worst case) to about 94% (in the best case), while for the CRC dataset, it ranges from about 35% (in the worst case) to 47% (in the best case). The multi-view approach based on the concatenation, i.e., Concat-RF (MV), always performed better than the worst single-view pipeline, but was not able to reach the best single-view configuration. On the contrary, irBoost.SH provided a significant boost, with irBoost.SH reaching a F1-score of more than 99% on ASD-16S and more than 95% on CRC, achieving an improvement over Concat-RF (MV) of 9% and 114%, respectively. rBoost.SH was able to obtain good results on ASD-16S, with a disadvantage of about 0.3% over our method irBoost.SH. Like MV-ASD-3, this dataset was naturally complete, therefore, in these scenarios, the only difference with respect to irBoost.SH is the different way they combine the output. On the other hand, as already stated in 'Model evaluation and comparison' section, rBoost.SH cannot be run at all on the CRC dataset since it is multi-class.

In order to further confirm the significance of the obtained results, we computed the following three Wilcoxon signed-rank tests:

Table 4 P values obtained by the Wilcoxon signed-rank tests, after the application of the False Discovery Rate (FDR) correction [35]

	RF (SV) Best	Concat-RF (MV)	rBoost.SH (MV)
irBoost.SH versus	<0.0001	<0.0001	0.00164

- **irBoost.SH** versus **RF (SV) Best**, on all the datasets except for MV-ASD-3, on which RF (SV) Best is not applicable;
- **irBoost.SH** versus **Concat-RF (MV)**, on all the datasets;
- **irBoost.SH** versus **rBoost.SH (MV)**, on all the datasets except for CRC, on which rBoost.SH (MV) cannot be run since it is multi-class.

Considering that we ran multiple tests, we corrected the obtained p values with the False Discovery Rate (FDR) correction proposed by Benjamini and Hochberg [35], obtaining the results reported in Table 4. As it can be observed from the table, irBoost.SH outperforms all the considered competitors at a significance level $\alpha = 0.01$.

For the ASD-16S dataset, on average, each view built through a frequency-based filtering, sample-based filtering, and no filtering was selected as the winner in 2.8%, 2.3%, and 2.1% of the iterations, respectively. For the CRC dataset, on average, each view built through a frequency-based filtering, sample-based filtering, and no filtering was selected as the winner in 2.5%, 2.75%, and 2.44% of the iterations, respectively. An overview of the percentage of iterations for which each view was selected is provided in Additional file 3.

The huge improvement in terms of all the evaluation measures provided by irBoost.SH, together with the (almost) equal distribution of the times each type of views was selected, confirms that the features constructed by multiple pipelines describe the phenomenon from different, complimentary, viewpoints, that are fruitfully captured by irBoost.SH.

Conclusion

In this paper, we proposed a novel method, called irBoost.SH, to solve multi-class classification tasks from multiple, possibly incomplete, views. Instead of discarding the information conveyed by incomplete views or filling in missing values, irBoost.SH exploits the available information of all the views, without introducing excessive approximation in the data, through a boosting process based on multi-armed bandits.

The proposed algorithm is motivated by the challenges that arise in the analysis of microbiome data, where the presence of multiple, possibly incomplete, views is very common. The obtained experimental results emphasized that the models learned by irBoost.SH were able to outperform all the models learned from each single view independently, as well as a baseline multi-view approach based on the concatenation of the features of all the views. The superiority of irBoost.SH was observed in two different tasks, i.e., in the prediction of the presence of Autism Spectrum Disorders (ASD) and in the prediction of the presence of the Colorectal Cancer (CRC) disease, where views were constructed considering 16S and shotgun data (2 views) or by applying several

preprocessing pipelines (40 views). The obtained results confirmed that irBoost.SH can fruitfully be adopted for the analysis of microbiome data, to simultaneously exploit 16S and shotgun data, and to solve the issues coming from the identification of the most proper preprocessing pipeline, that in this case is fully automated.

As future work, we will adapt irBoost.SH to solve other tasks, such as regression and multi-target classification/regression. Moreover, we will investigate the possibility to make it able to work in the semi-supervised learning setting, where the class label is not available for all the training instances. Finally, we will evaluate the effectiveness of irBoost.SH in capturing the information conveyed by multiple, possibly incomplete, views also in other application domains.

Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-024-05767-w>.

Additional file 1. Full set of results. Description: Complete results in terms of precision, recall, F1-score, accuracy and AUROC for all the considered datasets and classifiers

Additional file 2. Confusion matrices. Description: Confusion matrices for all the considered datasets and classifiers

Additional file 3. Selected views. Description: Overview of the percentage of chosen views over iterations in irBoost.SH for all the considered datasets

Acknowledgements

The research on the topic started during Short-term Scientific Mission (STSM) within ML4Microbiome COST Action. The first author wishes to thank Gianvito Pio, PhD for his warm patience and immense help during her visit to Bari, Italy, and for the support during the whole process concluding with this paper. This endeavor would not be possible without Tatjana Lončar-Turukalo, PhD and Michelangelo Ceci, PhD. Warmest thanks to Sanja Brdar, PhD and Miloš Radovanović, PhD for continuous support throughout all these years.

Author Contributions

AS and GP conceived the task and designed the solution from a methodological point of view. AS, SB and TLT developed the workflow for microbiome data preprocessing. AS implemented the model, ran the experiments and collected the results. AS, GP, MC and MR interpreted and discussed the results. GP, MC and MR supervised the whole research. All the authors contributed to the manuscript drafting and approved the final version of the manuscript.

Funding

This publication is based upon work from COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies", supported by COST (European Cooperation in Science and Technology). This work was also partially supported by the following projects under the NRRP MUR program funded by the NextGenerationEU: Future AI Research (PE00000013), Spoke 6 - Symbiotic AI; PRIN 2022 "BA-PHERD: Big Data Analytics Pipeline for the Identification of Heterogeneous Extracellular non-coding RNAs as Disease Biomarkers", grant n. 2022XABBMA, CUP: H53D23003690006; CN3 RNA - "National Center for Gene Therapy and Drugs based on RNA Technology", CUP: H93C22000430007.

Availability of data and materials

The datasets considered in this study for the prediction of the Autism Spectrum Disorder are available at the following links: <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA453621> and <https://www.kaggle.com/datasets/antaresnyc/human-gut-microbiome-with-asd>. The datasets considered in this study for the prediction of the Colorectal Cancer Disease are available at the following links: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA290926>, <https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6070>, and <http://mothur.org/MicrobiomeBiomarkerCRC>. All the datasets in the pre-processed form are publicly available at: https://figshare.com/collections/Processed_16S_rRNA_sequencing_multi-view_datasets/7066355. The source code of the proposed method irBoost.SH is publicly available at: <https://github.com/AndreaMSBios/Multi-class-boosting-for-the-analysis-of-multiple-incomplete-views-on-microbiome-data>.

Declarations

Ethics approval and consent to participate

The data used in this study was publicly available for research purposes and collected during previous studies. Prior consent to the use of such data for research purposes was already collected in the original studies.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Received: 16 August 2023 Accepted: 4 April 2024

Published online: 14 May 2024

References

- Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 2016;8(1):37.
- ...Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Nhieu JTV, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, Doeberitz MK, Sobhani I, Bork P. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol.* 2014;10:766.
- Zackular JP, Rogers MA, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res.* 2014;7:1112–21.
- Dan Z, Mao X, Liu Q, Guo M, Zhuang Y, Liu Z, Chen K, Chen J, Xu R, Tang J, Qin L, Gu B, Liu K, Su C, Zhang F, Xia Y, Hu Z, Liu X. Altered gut microbial profile is associated with abnormal metabolism activity of autism spectrum disorder. *Gut Microbes.* 2020;11:1246–67.
- Rosenfeld CS. Microbiome disturbances and autism spectrum disorders. *Drug Metab Dispos.* 2015;43:1557–71.
- Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochem Biophys Res Commun.* 2016;469:967–77.
- Mas-Lloret J, Obón-Santacana M, Ibáñez-Sanz G, Guinó E, Pato ML, Rodríguez-Moranta F, Mata A, García-Rodríguez A, Moreno V, Pimenoff VN. Gut microbiome diversity detected by high-coverage 16s and shotgun sequencing of paired stool and colon sample. *Sci Data.* 2020;7(1):92.
- Shaffer JP, Carpenter CS, Martino C, Salido RA, Minich JJ, Bryant M, Sanders K, Schwartz T, Humphrey G, Swafford AD, Knight R. A comparison of six DNA extraction protocols for 16s, ITS and shotgun metagenomic sequencing of microbial communities. *BioTechniques.* 2022;73:34–46.
- ...Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G, Adilovic M, Aydemir O, Bakir-Gungor B, Santa Pau ECD, D'Elia D, Desai MS, Falquet L, Gundogdu A, Hron K, Klammsteiner T, Lopes MB, Marcos-Zambrano LJ, Marques C, Mason M, May P, Pašić L, Pio G, Pongor S, Promponas VJ, Przymus P, Saez-Rodríguez J, Sampri A, Shigdel R, Stres B, Suharoschi R, Truu J, Truica CO, Vilne B, Vlachakis D, Yilmaz E, Zeller G, Zomer AL, Gómez-Cabrero D, Claesson MJ. Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front Microbiol.* 2021. <https://doi.org/10.3389/fmicb.2021.635781>.
- ...D'Elia D, Truu J, Lahti L, Berland M, Papoutsoglou G, Ceci M, Zomer A, Lopes MB, Ibrahim E, Gruca A, Nechyporenko A, Frohme M, Klammsteiner T, Pau ECDS, Marcos-Zambrano LJ, Hron K, Pio G, Simeon A, Suharoschi R, Moreno-Indias I, Temko A, Nedyalkova M, Apostol ES, Truica CO, Shigdel R, Telalovic JH, Bongcam-Rudloff E, Przymus P, Jordamovic NB, Falquet L, Tarazona S, Sampri A, Isola G, Pérez-Serrano D, Trajkovic V, Klucar L, Loncar-Turukalo T, Havulinna AS, Jansen C, Bertelsen RJ, Claesson MJ. Advancing microbiome research with machine learning: key findings from the ml4microbiome cost action. *Front Microbiol.* 2023. <https://doi.org/10.3389/fmicb.2023.1257002>.
- ...Marcos-Zambrano LJ, López-Molina VM, Bakir-Gungor B, Frohme M, Karadzovic-Hadziabdic K, Klammsteiner T, Ibrahim E, Lahti L, Loncar-Turukalo T, Dharmo X, Simeon A, Nechyporenko A, Pio G, Przymus P, Sampri A, Trajkovic V, Lacruz-Pleguezuelos B, Aasmets O, Araujo R, Anagnostopoulos I, Aydemir O, Berland M, Calle ML, Ceci M, Duman H, Gundogdu A, Havulinna AS, Kaka Bra KHN, Kalluci E, Karav S, Lode D, Lopes MB, May P, Nap B, Nedyalkova M, Paciência I, Pasic L, Pujolassos M, Shigdel R, Susin A, Thiele I, Truica CO, Wilmes P, Yilmaz E, Yousef M, Claesson MJ, Truu J, de Santa Pau EC. A toolbox of machine learning software to support microbiome analysis. *Front Microbiol.* 2023. <https://doi.org/10.3389/fmicb.2023.1250806>.
- Peng J, Aved AJ, Seetharaman G, Palaniappan K. Multiview boosting with information propagation for classification. *IEEE Trans Neural Netw Learn Syst.* 2018;29:657–69.
- Nguyen ND, Wang D. Multiview learning for understanding functional multiomics. *PLOS Comput Biol.* 2020;16:e1007677.
- Yang M, Li Y, Hu P, Bai J, Lv J, Peng X. Robust multi-view clustering with incomplete information. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(1):1055–69.
- Blum A, Mitchell T, in *Proc. of the 11th annual conference on Computational learning theory - COLT'98* (ACM Press, 1998), pp. 92–100
- Sindhwani V, Niyogi P, Belkin M. In: *Proceedings of the workshop on learning with multiple views, 22nd (ICML, Bonn, 2005)*
- Wang J, Luo S, Li Y. *Advances in Neural Networks - ISNN 2010*. Berlin: Springer; 2010. p. 444–9.
- Ceci M, Pio G, Kuzmanovski V, Džeroski S. Semi-supervised multi-view learning for gene network reconstruction. *PLoS ONE.* 2015;10(12):1–27.
- Xu Z, Sun S. *Neural information processing. Theory and algorithms*. Berlin: Springer; 2010. p. 355–62.
- Freund Y, Schapire RE. *Lecture notes in computer science*. Berlin: Springer; 1995. p. 23–37.
- Busa-Fekete R, Kegli B. In: *Proceedings of the 27th international conference on machine learning, Haifa, Israel. 2010*, pp 143–150
- Hà Quang M, Bazzani L, Murino V. In: *Proceedings of the 30th international conference on machine learning. 2013*; vol 28, pp 100–108
- Xu X, Li W, Xu D, Tsang IW. Co-labeling for multi-view weakly labeled learning. *IEEE Trans Pattern Anal Mach Intell.* 2016;38:1113–25.
- Koco S, Capponi C. *Machine learning and knowledge discovery in databases*. Berlin: Springer; 2011. p. 209–28.
- Xu C, Tao D, Xu C. Multi-view learning with incomplete views. *IEEE Trans Image Process.* 2015;24:5812–25.
- Zhu P, Yao X, Wang Y, Cao M, Hui B, Zhao S, Hu Q. Latent heterogeneous graph network for incomplete multi-view learning. *IEEE Trans Multimed.* 2022.

27. Peng J, Barbu C, Seetharaman G, Fan W, Wu X, Palaniappan K. Machine learning and knowledge discovery in databases. Berlin: Springer; 2011. p. 597–612.
28. Peter Auer YF, Cesa-Bianchi Nicolo, Schapire RE, Gambling in a rigged casino: the adversarial multi-armed bandit problem. Tech. rep., University of Technology Graz/Universita di Milano/AT & T Labs, Florham Park, NJ, 1998
29. Auer P, Cesa-Bianchi N, Freund Y, Schapire RE. The nonstochastic multiarmed bandit problem. *SIAM J Comput.* 2002;32:48–77.
30. Cesa-Bianchi N. Prediction, learning, and games. London: Cambridge University Press; 2010.
31. Liu Z, Mao X, Dan Z, Pei Y, Xu R, Guo M, Liu K, Zhang F, Chen J, Su C, Zhuang Y, Tang J, Xia Y, Qin L, Hu Z, Liu X. Gene variations in autism spectrum disorder are associated with alternation of gut microbiota, metabolites and cytokines. *Gut Microbes.* 2021;13:1854967.
32. Pulikkan J, Mazumder A, Grace T. Advances in experimental medicine and biology. Berlin: Springer; 2019. p. 253–69.
33. ...Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Silva RD, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciorek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimy AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hoof JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* 2019;37:852–7.
34. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from illumina amplicon data. *Nat Methods.* 2016;13:581–3.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57(1):289–300.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.