Robust Microbial Markers for Non-Invasive Inflammatory Bowel Disease Identification

Benjamin Wingfield, Sonya Coleman, *Member, IEEE,* TM McGinnity, *Senior Member, IEEE,* and AJ Bjourson

Abstract—Inflammatory Bowel Disease (IBD) is an umbrella term for a group of inflammatory diseases of the human gastrointestinal tract, including Crohn's Disease (CD) and ulcerative colitis (UC). Changes to the intestinal microbiome, the community of micro-organisms that resides in the human gut, have been shown to contribute to the pathogenesis of IBD. IBD diagnosis is often delayed due its non-specific symptoms (e.g. abdominal pain) and an invasive colonoscopy is required for confirmation. Delayed diagnosis is linked to poor growth in children and worse treatment outcomes. Microbial communities are extremely complex and feature selection algorithms are often applied to identify key bacterial groups that drive disease. It has been shown that aggregating Ensemble Feature Selection (EFS) can be used to improve the robustness of feature selection algorithms. The robustness of a feature selector is defined as the variation of feature selector output caused by small changes to the dataset. Typical feature selection algorithms can be used to help build simpler, faster, and easier to understand models - but suffer from poor robustness. Having confidence in the output of a feature selector algorithm is key for enabling knowledge discovery from complex biological datasets. In this work we apply a two-step filter and an EFS process to generate robust feature subsets that can non-invasively predict IBD subtypes from high-resolution microbiome data. The predictive power of the robust feature subsets is the highest reported in literature to date. Furthermore, we identify five biologically plausible bacterial species that have not previously been implicated in IBD aetiology.

Index Terms—Feature evaluation and selection, machine learning algorithms, pattern recognition, medicine

1 Introduction

Metagenomics is the study of genetic material sourced from environmental samples, which allows microbial genomes to be detected and analysed [1]. Metagenomics overcomes the flaws that traditional microbiology suffers from when identifying and analysing micro-organisms. Only a small proportion of micro-organisms can be cultured in growth media in standard laboratory conditions [2]. High throughput sequencing has enabled culture free detection of microbial communities, and increased the resolution and power of downstream analysis [3].

Inflammatory Bowel Disease (IBD) is a group of disorders that cause chronic inflammation of the gut. IBD caused 53,000 deaths worldwide in 2013 and its prevalence is increasing, particularly in western countries [4]. The symptoms of IBD are non-specific and diagnosis is typically confirmed via colonoscopy. However, the invasiveness of the colonoscopy procedure can introduce a delay in diagnosis; delayed diagnosis is common [5] and linked to poor treatment outcomes, particularly in children [6].

Metagenomic analysis of the intestinal microbiome, the community of micro-organisms that live in the small intestine and colon, requires DNA to be taken from an environmental sample. For example, to study the human intestine an intestinal tissue sample or faecal sample could be taken. DNA is then isolated from the sample, purified, and sequenced. Large amounts of DNA are present in such environmental samples, and marker gene surveys are a cost effective protocol that sequence small sections of DNA (markers) from such samples. The 16S ribosomal RNA (16S rRNA) gene is widely used as a universal marker gene for bacteria, and can be used to create a bacterial census [7]. The microbial communities present in the human body range from simple (e.g. skin) to complex (e.g. in the gut), and metagenomic data reflects this complexity; metagenomic data sampled from the human gut is highly dimensional [7]. Highly dimensional data often contains redundant data, which can hinder knowledge discovery. It is common for data to be reduced to remove irrelevant features. A popular approach to data reduction is a selection based strategy, known as feature selection.

Feature selection is a common preprocessing step in machine learning applications. Feature selection is useful for optimising the performance of a model while finding the smallest subset of features, which can improve model performance and lower computational complexity [8]. Feature selection also enables knowledge discovery from high dimensional data [9]. Domain experts are often interested in experimentally validating feature subsets, which is an expensive proposition for biological data. Feature selection algorithms can return different feature subsets from the same input data; different feature subsets can be equally optimal, particularly if a high degree of redundancy is present in the dataset [10]. Feature selection algorithms can also return significantly different feature subsets from input data that has been changed slightly (e.g. by removing a

- *B. Wingfield and S. Coleman are with the Intelligent Systems Research Centre, Ulster University, Derry, United Kingdom (e-mail: wingfield-b@ulster.ac.uk).*
- *TM McGinnity is with the School of Science and Technology, Nottingham Trent University, Nottingham, UK and the Intelligent Systems Research Centre, Ulster University, Derry, United Kingdom.*
- *AJ Bjourson is with the Northern Ireland Centre for Stratified Medicine, Ulster University, Derry, United Kingdom.*

Manuscript received XXX-XXX

sample or after adding noise to a feature). Domain experts will have more confidence in feature selection algorithms that generate consistent (robust) feature subsets.

Ensemble Feature Selection (EFS) can generate robust feature subsets [11]. EFS is inspired by ensemble learning, where the output of multiple weaker classifiers can be combined to outperform a single strong model. It has been shown that combining the output of multiple unstable feature selectors can create a robust consensus feature ranking [11]. Typically filter, wrapper, and embedded feature selection methods that do not consider the robustness of output have been previously applied to microbiome data [12]. Random Forests have been widely applied for supervised classification of IBD from microbiome data and the feature rankings have been reported for knowledge discovery purposes [13], [14], [15]; rankings are often combined with a recursive feature elimination procedure to generate a feature subset. Recently an EFS approach was used to generate a feature subset for the non-invasive prediction of advanced fibrosis in non-alcoholic fatty liver disease [16]. However, this approach does not employ an aggregation paradigm and instead uses an ensemble of Random Forests to select the top features from an optimal model. This approach discards the feature ranks from every non-winning model, and does not consider the robustness of the selected feature subset.

In this work we investigate if the application of EFS to high-resolution microbiome count data can improve classification of IBD from stool samples to enable non-invasive prediction of IBD subtype. Non-invasive prediction of IBD subtype has been attempted with some success before [14], [15] and We also aim to investigate if the robust feature subsets generate new insights with regards to the composition of the intestinal microbiome in subjects with IBD. We employ an EFS technique that uses similarity measures to aggregate feature ranks into a final consensus list to improve confidence for future experimental validation. Our rationale for this approach lies in the nature of the high-resolution microbiome count data: we define bacteria as specific 16S sequence variants called Denoised Sequence Variants (DSVs) rather than traditional fuzzy clusters of sequences. Recent work [17] has shown that defining bacteria with the DSV paradigm provides a host of benefits compared with traditional fuzzy clusters, described further in Section 2. The DSV approach has not been applied to date with regards to IBD classification, and thus our approach provides a promising area for knowledge discovery.

The remainder of this paper is organised as follows: Section 2 describes the background of defining robust feature subsets, estimating the robustness of a feature subset, EFS, and the rationale behind using 16S sequence variants rather than traditional fuzzy clusters. An overview of related research in feature selection applied to metagenomics, feature selection applied to biological data with regards to IBD, and the application of classification algorithms to metagenomic data is also presented in Section 2. Section 3 describes the publicly available dataset that was analysed, the experimental procedure, and the aggregating EFS algorithm. The microbial markers generated by EFS and the performance of the models fitted to the feature subsets are reported, discussed, and compared with other reported results in Section 4. Conclusions, limitations, and plans for future work are discussed in Section 5.

2 Background

2.1 Measuring the robustness of feature selectors

The robustness of a feature selector can be defined by the variation of feature subset output caused by small changes to the input [18]. In this work the input data were modified by instance perturbation (removing or adding features) via resampling with replacement (bootstrapping). Modification can also be done at the feature level (e.g. by adding random noise to a feature or group of features) or by a combination of instance and feature level perturbation.

To measure the overall effect of bootstrapping on feature stability, [18] proposed a similarity measure based approach. In this approach the stability was measured by averaging the pairwise similarity comparison of feature subset output for k bootstraps, which was defined as:

$$S_{\text{global}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S\left(f_i, f_j\right)}{k\left(k-1\right)} \tag{1}$$

where f_i is the feature selector output applied to bootstrap i, and $S\left(f_i, f_j\right)$ is a similarity measure between f_i and f_j. In this work we use the Jaccard Index as a similarity measure:

$$S\left(f_i, f_j\right) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} = \frac{\sum_l I(F_i^l = f_j^l = 1)}{\sum_l I(F_i^l + f_j^l > 0)} \tag{2}$$

where the function I returns 1 if its argument is true and 0 if its argument is false.

2.2 Aggregating Ensemble Feature Selection

It has been shown that an aggregating EFS approach can improve the robustness of feature selectors [18]. Ensemble models are capable of outperforming single models because if a group of different but equally good hypotheses exist it is less likely that an ensemble will pick the wrong hypothesis. Furthermore, algorithms can end up in different local optima enabling an ensemble to better approximate a true function. Finally it is known that EFS can achieve greater robustness because it expands the hypotheses space [19]. EFS has two stages: choosing a set of feature selection algorithms, and combining the feature subsets into a final consensus ranked list. In this work we combine the feature subsets via complete linear aggregation:

$$f^l = \sum_{i=1}^{s} f_i^l \tag{3}$$

where an ensemble contains s feature selectors $F_1, \ldots, F_s$. Each feature selector outputs a feature ranking $f_i = f_i^l, \ldots, f_i^N$. Feature selection must always be combined with an evaluation of classification performance: domain experts will not be interested in a stable feature subset that has poor predictive performance. In this work we apply embedded feature selection algorithms, such as Random Forests [20] and linear Support Vector Machines (SVM). Embedded feature selection algorithms provide feature ranking during training which decreases the computational complexity of the EFS process, and embedded feature selection is discussed further in Section 2.4.1. Random Forests are an ensemble of decorrelated decision trees [21]; feature rankings are calculated by randomly permuting a feature in the out-of-bag samples and calculating the mean change in impurity or accuracy compared with the out-of-bag rate with unpermuted features. Linear SVMs can rank features from the absolute value of the weight vector of the hyperplane [22]; a process known as Recursive Feature Elimination (RFE) is used to reduce the size of the feature subsets by iteratively removing the poorest 10% of features until the subset is empty. In order to effectively evaluate which feature selector should be chosen for a particular classification problem it is necessary to use a metric that balances the classification performance of a feature aggregation and the stability of the aggregated features. The robustness-performance trade-off (RPT) [11] is a metric that does this; it is a variant of the widely used F1-score which is the harmonic mean of the precision and recall [23]. The RPT is defined as:

$$\mathrm{RPT}_{\beta} = \frac{(\beta^2 + 1) \cdot S_{\mathrm{global}} \cdot P}{\beta^2 \cdot S_{\mathrm{global}} \cdot P} \tag{4}$$

where P is the prediction accuracy of the classification model trained on the robust feature subset. β is a parameter used to weight the relative importance between robustness and classification performance. In this work we use $\beta = 1$ to give equal importance to classification performance and robustness.

2.3 High resolution microbiome count data

Raw 16S data typically consists of millions of short sequences (typically less than 400 nucleotides long). Conventionally the sequence reads are clustered according to fixed similarity thresholds; typically sequences that are more than 97% similar are binned into an Operational Taxonomic Unit (OTU), which approximates a bacterial species [3], [24]. A clustering strategy is required because during amplification and sequencing significant noise is introduced into the set of sequence reads [17] (e.g. insertion, deletion, or substitution sequencing errors). A range of new methods [25], [26], [27] has been developed that is capable of removing this noise from the set of sequence reads. These methods are capable of resolving DSVs to a single-nucleotide resolution, which removes the need for arbitrary similarity thresholds. These high-resolution methods have better specificity and sensitivity compared with OTU clustering algorithms [27], and are better at identifying patterns of community similarity because DSVs are much less likely to be ecologically mixed units [25]. It is important to note that although the term OTU can apply to DSVs (the definition of OTU is intentionally vague and simply means "the thing(s) being studied" - one proposed term for DSVs is zero-radius OTU [28]) it is useful to consistently use different terminology to avoid confusion as the underlying paradigms are so different (clustering versus denoising). DSVs offer increased taxonomic resolution, are defined independently of any reference database, have consistent labels, and can be re-used across studies [17]. In a conventional clustering approach, units are defined according to a reference database (reuse is possible but uncharacterised organisms will be omitted) or in a *de novo* fashion (*de novo* OTUs can include uncharacterised organisms but lack consistent labels and cannot be re-used across studies [17]). OTUs must be mapped to a taxonomy in order to provide consistent labels, while DSVs are independent of taxonomy and represent true biological variation [17]. Predictive microbial markers of disease are only useful if they can be applied to new data, which is our rationale for combining the DSV paradigm with an aggregating EFS strategy. In this work we use the `dada2` software package [27] to implement the DSV paradigm on a publicly available dataset of treatment-naïve children with IBD.

Microbiome count data consist of an N by M matrix with N bacterial units (e.g. DSVs) and M samples. The count data are often normalised into proportions (creating relative abundance data) or randomly subsampled (rarefaction) to counteract uneven library sizes per sample. The total sum of sequence reads per sample (library size) can differ by orders of magnitude across a sequencing run [7], which distorts measures of bacterial abundance. Both types of normalisation are flawed and may not be appropriate for machine learning applications as they do not resolve the underlying heteroscedasticity present in the data [29] which has been shown to violate the assumptions of models such as regression trees [30]. In this work we apply a variance stabilising transformation [31] to avoid this problem, which is recommended for machine learning applications [29].

2.4 Related work

2.4.1 Feature selection on metagenomic data

Feature selection has been broadly applied to metagenomic data to identify a subset of microbes with predictive power for a particular phenotype, including diabetes [32], obesity [33], liver cirrhosis [34], non-alcoholic liver cirrhosis [16], pregnancy [35], psoriasis [36], and IBD [13]. The implemented algorithms cover all the major feature selection algorithm categories: filter, wrapper, and embedded. The metagenomic datasets described above are labelled (e.g. sample A has IBD, sample B is a control), so we only consider supervised feature selection algorithms, discussed below. Semi-supervised and unsupervised feature selection algorithms are available for data that are fully or partially missing labels or for experiments that aim to investigate the structure of the data [37].

Filter feature selection algorithms select features without building a model, and aim to reduce dimensionality by directly operating on the dataset with criteria such as correlation, redundancy, or information gain [8]. Filter methods are quick and relatively simple to implement at the expense of model performance. The Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm [38], originally developed for gene expression data, has been applied to identify 50 microbial gene markers from the intestinal microbiome that can be used to successfully classify type 2 diabetes (AUROC: 0.81 [32]). The MRMR approach has also been applied to identify 15 microbial gene markers from the gut microbiome, to predict liver cirrhosis with an SVM (AUROC: 0.918 [34]). A multivariate filter algorithm called Generalized Local Learning (GLL) was used as the main feature selection technique for the prediction of psoriaris from skin microbiome samples with a good classification performance (between 0.85 – 0.90 AUROC [36]).

Wrapper feature selection algorithms use a multi-objective optimisation approach to maximise model performance and minimise feature subset size [8]. Wrappers search through the space of possible feature subsets using the constructed model as a performance measure (e.g. classification accuracy). The search method can range from simple (combinatorial) to complex (computational intelligence approaches such as genetic algorithms). Although wrapper methods provide better results than filter methods they have a high computational cost and tend to overfit [8]. Genetic algorithms have been used to identify a subset of microbes present in the vaginal microbiome that can be used to predict bacterial vaginosis with an accuracy of between 97-99% [39]. An exhaustive exploration of bacterial combinations was used with a ternary regression model to identify a set of weights that could be used to predict obesity from the gut microbiome [33]. Although no feature elimination was done unweighted bacteria could be considered to be unselected.

Embedded feature selection algorithms use internal data from the classification model to enable feature selection (e.g. feature rankings of Random Forests). Embedded methods provide a balance between computational complexity and performance [8]. A comprehensive review of multi-class classification and feature selection algorithms found that embedded feature selection algorithms performed best across 8 metagenomic datasets [12]. Feature rankings are sometimes used only for knowledge discovery purposes; a feature elimination step is not always applied [14], [15]. A Random Forest paired with the Boruta feature selection algorithm [40] was used to identify a subset of differentially abundant bacteria present in the vaginal microbiome of pregnant women [35]. When no feature selection was used pregnancy could be successfully predicted from the vaginal microbiome (Scott's pi index up to 0.8). A Nearest Shrunken Centroids model [41] was used to select 30 bacterial genera which could predict IBD with 70% accuracy from the gut microbiome of subjects in remission [13].

An ensemble of Random Forests has been used to identify the top features of the best performing model for the supervised classification of non-alcoholic fatty liver disease [16]. This is a type of non-aggregating EFS, but is more similar to the wrapper approach in that a comprehensive search is undertaken to identify an optimal model. Ensembles are powerful because they fuse decisions [42]. This approach performs no fusion and instead discards feature ranks from all non-winning models. These discarded data could be valuable, and the robustness of the final feature subset is not considered.

2.4.2 *Feature selection on biological data for IBD*

Here we consider feature selection techniques applied to biological data for the purpose of diagnosing IBD. The data described below include gene expression, proteomic, metabolomic, and imaging data. A filter was applied to a proteomics dataset in order to predict IBD at 66% accuracy from immune response to Escherichia coli proteins [43]. The dataset consisted of 4,256 Escherichia coli K12 proteins in a microarray which were tested against blood serum from around 100 subjects. A univariate statistical test called Significance Analysis of Microarrays (SAM) was used to find proteins that exhibited a statistically significant immunogenic response. Significant proteins with a false discovery rate of 0 were used as input for k-nearest neighbours and SVM classification models.

Wrapper approaches have been more commonly applied to biological datasets with lower dimensionality because the computational complexity of wrapper algorithms does not scale well for highly dimensional data [8]. An extensive feature selection pipeline was implemented in order to predict Crohns Disease Endoscopic Index of Severity (CDEIS) on a dataset of 30 patients from Magnetic Resonance Imaging (MRI) images [44]. The implemented approach differs from the other work described in the literature because human experts (radiologists) manually performed feature extraction prior to feature selection. The raw images were first inspected by radiologists to define a set of 17 features which were exhaustively searched in a combinatorial manner and fitted to a linear regression model via a wrapper approach. An optimal feature subset was able to achieve a correlation of up to $r^2 = 0.65$. In a metabolomics dataset a genetic algorithm was used to determine optimal spectra that aimed to identify bowel diseases including IBD from faecal samples with nuclear magnetic resonance spectroscopy [45]

A two-step filter and embedded feature selection process was used in a Genome Wide Association Study (GWAS) to predict IBD [46]. The rationale behind this approach was the extreme size and dimensionality of the dataset, which consisted of nearly 180,000 single nucleotide polymorphisms across 44,000 samples. The filtering stage reduced the number of features from approximately 180,000 to around 10,000. The embedded stage used L1 (lasso) logistic regression. L1 penalised models assume only a small proportion of features will be relevant, and many of the estimated coefficients can be zero [47]. Feature selection can be achieved by ignoring features with a coefficient of zero. The two-stage feature selection process managed to achieve an AUROC of approximately 0.85 for predicting ulcerative colitis and Crohns disease.

3 Methods

3.1 Dataset

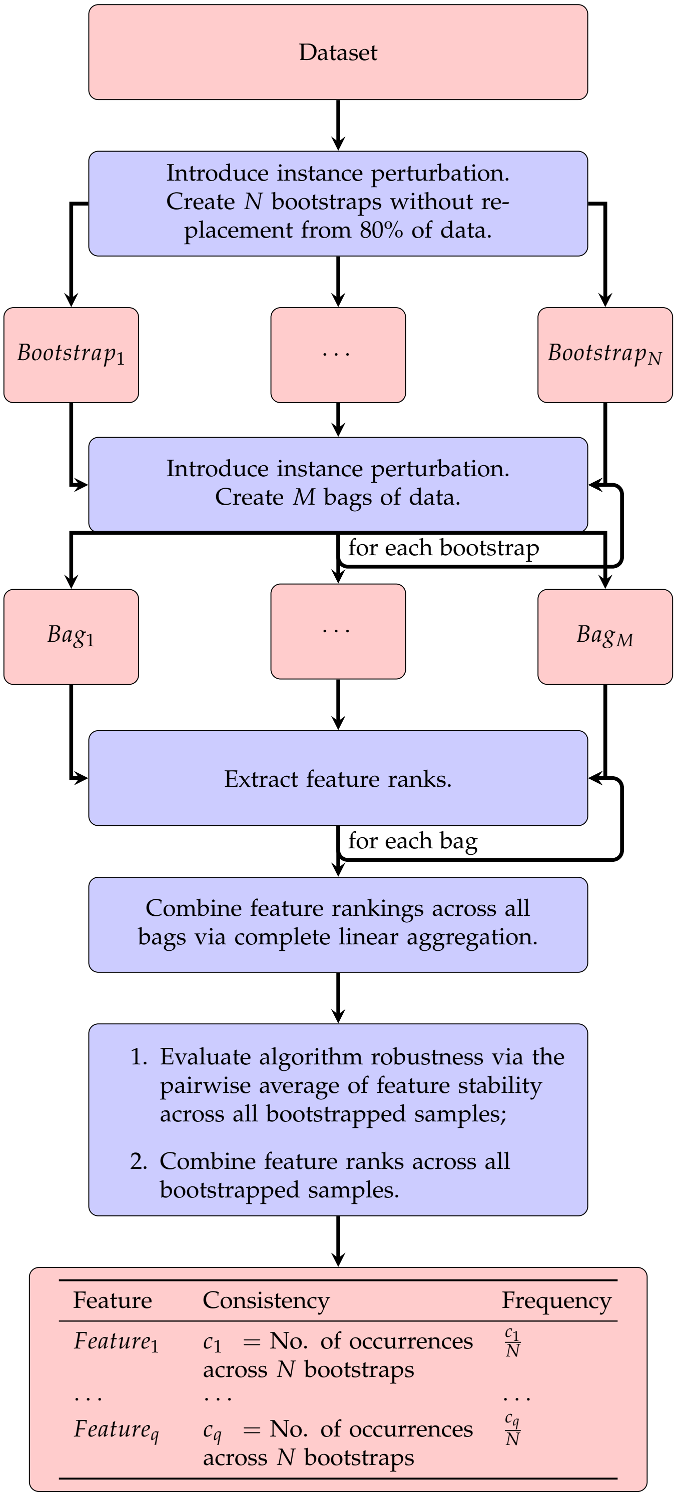
We applied a high-resolution microbiome pipeline, described further in Section 3.2, to a publicly available dataset [15] which consisted of 1643 samples collected from treatment-naïve children and adults diagnosed with IBD and controls. However, only the paediatric data was used in this study as there were not enough adults for analysis; children were defined as being $\leq$ 16 years old (the A1 Montreal classification of IBD [48]). Samples were collected at disease onset at the time of diagnosis, so IBD was in an active state. In this work we focus on stool samples in order to develop a set of robust markers that can be used to non-invasively predict IBD, so all biopsy samples were discarded, leaving 311 stool samples. Classes were defined according to an IBD subtype: control versus ulcerative colitis (UC) or control versus Crohn's Disease (CD). Although they fall under the umbrella term IBD the subtypes have significant biological differences [49], [50], which is our rationale for choosing an IBD subtype to define classes.

3.2 Experiment procedure

A reproducible computational workflow was implemented with `Docker` and `nextflow` [51], which is available at https://github.com/nebfield/crohnsemble. `Docker` is an open source container platform. A container bundles together all of the data, software, and library dependencies necessary to run a piece of software into an image, similar to a very efficient virtual machine. `Docker` helps to improve reproducible research by solving "dependency hell", poor documentation (`docker` images are self-documenting), and code rot [52]. The dataset was downloaded using `esearch` [53], `sra-tools` [54], and `GNU Parallel` [55]. Microbiome count data were generated with `dada2` [27] and processed with `phyloseq` [56] according to a standard operating protocol [57]. A variance stabilising transformation [31] was applied to the microbiome count data to normalise the uneven library sizes and heteroscedasticity in the data, which has been recommended for machine learning applications [29]. We implemented aggregating EFS, described further in Section 3.3, using the `OmicsMarkeR` package [58]. The Synthetic Minority Over-sampling Technique [59] (SMOTE) was used to mitigate the class imbalance present in the dataset. SMOTE is a powerful synthetic sampling technque that has been successfully applied for a variety of applications (including biomedical data) [60]. Imbalanced data can be significantly more difficult to learn, decreasing model performance [60], [61]. The distribution of microbial markers was visualised with `Venny` [62].

3.3 Ensemble feature selection

Prior to applying EFS a simple filter was applied to remove extremely rare DSVs. DSVs present in less than 5% of samples were removed, as this study aims to find microbial markers that are present across a broad population. Prior to EFS 20% of data were retained from the dataset for independent validation of the final model. In the first stage of EFS, a portion of the data (20%) is retained in order to test the performance of the model trained on the remainder of the data (see Figure 1). The training data were repeatedly sampled with replacement (bootstrapped). For each bootstrap bag a SVM and Random Forest were fit, and recursive feature elimination was applied to each bag. Feature ranks were extracted across all of the bags, and merged via complete linear aggregation [11] to form a single feature ranking list. Each ranked list was combined across all of the bootstraps to form a final feature subset, along with frequency and consistency measurements (see supplement). The RPT was calculated for both models from the classification performance of the model on the test data and the global similarity measure across all feature lists. Random Forests were used to validate the generalisation ability of the microbial markers as they had the highest RPT for both CD and UC. All classification results reported are from the Random Forest model.


Dataset
Introduce instance perturbation. Create N bootstraps without replacement from 80% of data.
$Bootstrap_1$
...
$Bootstrap_N$
Introduce instance perturbation. Create M bags of data.
for each bootstrap
Bag_1
...
Bag_M
Extract feature ranks.
for each bag
Combine feature rankings across all bags via complete linear aggregation.
1. Evaluate algorithm robustness via the pairwise average of feature stability across all bootstrapped samples;
2. Combine feature ranks across all bootstrapped samples.

Feature	Consistency	Frequency
$Feature_1$	c_1 = No. of occurrences across N bootstraps	$\frac{c_1}{N}$
...	...	...
$Feature_q$	c_q = No. of occurrences across N bootstraps	$\frac{c_q}{N}$

Fig. 1. Ensemble feature selection workflow.

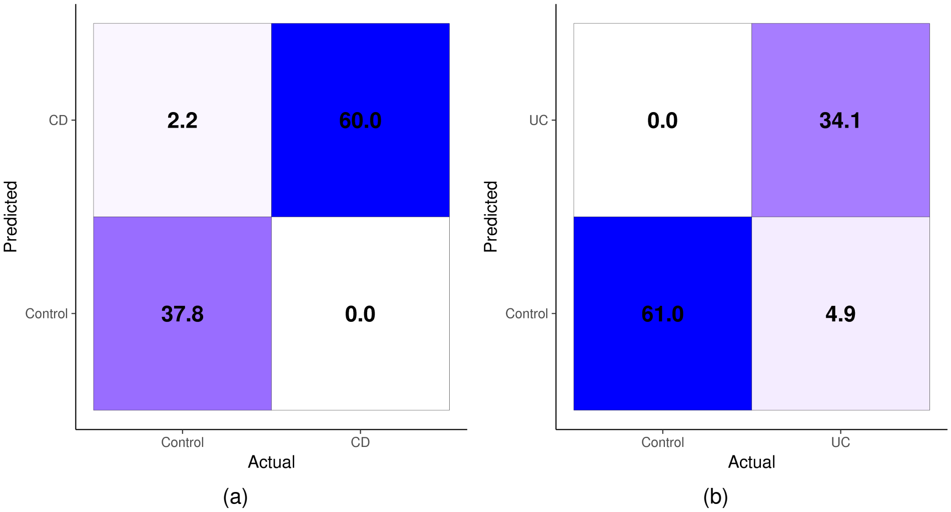
4 Results and Discussion

4.1 Comparison of model performance

We begin by evaluating the performance of the classification models after aggregating EFS is applied to the dataset, as a robust set of microbial markers must have strong predictive power to be of value for knowledge discovery and further investigation by domain experts. The classification and feature selection ability of Random Forests and SVMs were

TABLE 1
Classification performance of feature subset

Classification problem	Data split	Sensitivity	Specificity	PPV	NPV
Crohn's disease	Testing	94.5%	90.9%	87.6%	96.1%
Ulcerative colitis		100%	94.5%	94.5%	100%
Crohn's disease	Validation	100%	94.4%	96.4%	100%
Ulcerative colitis		87.5%	100%	100%	92.6%



(a)

(b)

Fig. 2. Confusion matrices of models fitted to feature subsets on paediatric validation data; Crohn's disease (left) and Ulcerative colitis (right). Each cell contains a percentage of samples assigned to it: light colours represent a small percentage, and darker colours represent a large percentage.

tested. Random Forests were chosen as the final model for both IBD subtypes as they had the highest RPT (a balanced metric of classification performance and aggregated feature robustness, see Table 2). The final models were used to validate the feature subsets against independent validation data. The dataset was split into two cohorts according to IBD subtype; the classification task was to distinguish between control and disease subjects (two class classification). Our rationale for this approach stems from the important biological differences present in the pathophysiology of UC and CD, which represents an interesting area for knowledge discovery to be derived from consensus feature subsets.

TABLE 2
An ensemble of Random Forests were chosen for both classification problems as they had the highest Robustness-Performance Tradeoff (RPT) measure.

Classification task	Model	RPT	No. features retained
Crohn's disease	**Random Forest**	**0.60**	**20**
	SVM	0.58	20
Ulcerative colitis	**Random Forest**	**0.70**	**17**
	SVM	0.48	17

Non-invasive prediction of both IBD diagnosis and IBD subtype from stool samples has been previously attempted, including from this dataset [15]. In [15] IBD was predicted from biopsies of the terminal ileum (mean AUC: 0.85) and rectum (mean AUC: 0.78) with good performance. Prediction from stool samples was less successful (mean AUC: 0.66 with much lower consistency). The models used relative microbial abundance data agglomerated to a genus level. In [14] classification performance was reported at two different thresholds: in the first, a sensitivity of 80.3% and a specificity of 69.7% was reported. The second reported a sensitivity of 45.8% and a specificity of 92.4%. It is important to note that the patient cohort used in [14] had a mean disease duration of 34.8 months, while the publicly available dataset used in this work consists of samples collected at time of diagnosis. Due to this lengthy disease duration many of the patients in the cohort had been treated with anti-inflammatory drugs or other pharmacological interventions which may have impacted the composition of the microbiome — the data used in this work do not suffer from this limitation. Prediction of IBD in an adult cohort from biopsies achieved an accuracy of up to 70% [13] using nearest shrunken centroid classification [63].

We report here the highest performance of non-invasive IBD classification from stool samples described in the literature to date. The classification performance of both feature subsets was excellent. CD was classified with a Positive Predictive Value (PPV) of 87.6% in the testing set and 96.4% in the validation set, and a Negative Predictive Value (NPV) of 97.1% in the testing set and 100% in the validation set. UC was predicted with a PPV of 94.5% in the testing set and 100% in the validation set, and a NPV of 100% in the testing set and 92.6% in the validation set (see Table 1). This is significantly better than performance metrics reported in [13], [14], [15].

4.2 Robust microbial sequence markers of IBD

Approximately 0.5% of DSVs were retained after a two-stage filter and aggregating EFS feature selection strategy (see Table 1). Nearly 4500 DSVs were identified from the stool samples: a simple filter was applied to remove any DSVs that were not present in at least 5% of samples. After this process, aggregating EFS was successfully applied to the remaining features (around 250 prevalent DSVs). The overlap of DSVs across IBD subtypes is low - 12.1% of DSVs were shared across the CD and UC subsets (see Figure 3) - which reflects the distinct biological differences between the two subtypes.

The stability of a selected feature can be measured by its frequency, which is the number of times a feature appears in each bootstrap divided by the total number of bootstraps (see Figure 1). Perfectly robust features have a frequency of 1 while the least robust features will only be present in a single bootstrap; in this work we used 5 bootstraps so features with a frequency of 0.2 are the least stable. In the CD cohort 3 DSVs had a perfect frequency, and in the UC cohort 4 features had a perfect frequency (see Tables 3–4). It is important to note that the DSV paradigm reveals greater differences than would otherwise be reported by a clustering approach. OTUs are generally capable of being matched to taxonomic databases at the level of family or genus [15]; all other IBD classification work agglomerated OTUs into genus-level relative abundance data to represent the microbiome. DSVs are capable of resolving separate bacterial strains (e.g. at a higher than species). However, because DSVs are relatively short fragments of the full 16S rRNA gene, taxonomic assignment is sometimes limited to higher ranks. The agglomeration process will discard bacteria that do not meet a defined phylogenetic or taxonomic threshold. For example, if a genus-level agglomeration is chosen then OTUs or DSVs that only match to the family level or higher will be discarded. In this work we chose not to agglomerate DSVs into specific taxonomic ranks as biological phenomenon (e.g. IBD subtype) may not be accurately modelled according to human-defined taxonomic hierarchies. DSVs have been shown to accurately represent true biological variation independently of any taxonomic reference database [17]. Of the most robust features for CD prediction two could be mapped to genus (*Bacteroides* and *Haemophilus*) and one to family (*Lachnospiraceae*). One of the most robust features for UC prediction could be mapped to species (*Bacteroides vulgatus*), two to genus (*Pediococcus*) and *Ersyipelotrichaceae*), and one to family (*Ruminococcaceae*).

4.3 Knowledge discovery

Every denoised microbial sequence marker we have described is novel, as previous work has relied on analysis of fuzzy clusters (see Tables 3–4). We have reported a set of 16S exact sequence variants (DSVs) that can non-invasively predict IBD with the highest reported accuracy to date, which have innate biological meaning and do not rely on reference databases or taxonomic assignments. The behaviour of DSVs that match the same species can be markedly different [25], which demonstrates the limitations of human-defined taxonomic systems. In order to compare our DSVs to previous work we have mapped the DSVs to the SILVA taxonomic database [69]. Below we describe elements of the robust microbial marker set that have been found previously in the literature, and then move on to bacterial species that have not been previously implicated in IBD pathogenesis. It is important to note fuzzy clusters, under normal circumstances, are limited to resolving bacteria at high taxonomic ranks such as Order, Family, or Genus. All of the identified DSVs have been previously reported in the literature as biomarkers for IBD at high taxonomic ranks which confirms that our aggregating EFS process has selected biologically plausible markers. One of the many advantages of the denoising DSV paradigm is increased taxonomic resolution; as the resolution increases, previously undescribed microbial markers emerge. The previously described markers below are gathered from differential abundance statistical tests and machine learning algorithms (e.g. Random Forest ranks). The reported biomarkers are from samples gathered from the entire gastrointestinal tract, including stool, rectal or ileal biopsies.

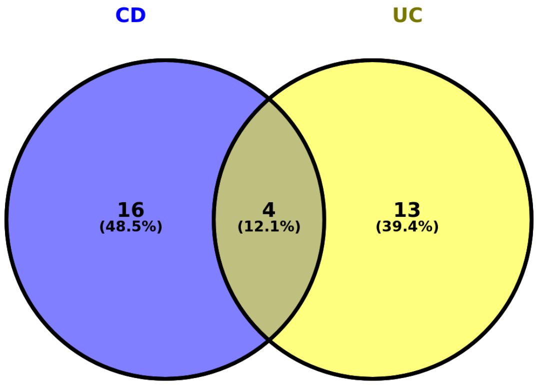


Fig. 3. Venn diagram of DSV microbial marker distribution by cohort (CD: Crohn's disease, UC: ulcerative colitis).

We begin by considering the biomarkers originally reported in [15]: *Blautia*, *Ruminoccous*, Pasteurellaceae, Erysipelotrichales, and Veillonellaceae are repeatedly observed in our set of robust markers [15]. Enterobacteriaceae, Bacteroidales, and Clostridiales have been repeatedly identified across the literature as IBD biomarkers [14], [15], [70], and all are strongly represented in our set of microbial markers. Fusobacterium has been previously reported as a biomarker for a number of conditions including IBD [71] and colorectal cancer [72]; the risk of developing colorectal cancer in IBD patients is significantly increased [73]. Lachnospiraceae, including the Roseburia genus specifically, is differentially abundant in IBD subjects [70]. *Faecalibacterium prausnitzii* is an anti-inflammatory organism and is associated with health [68]. *Parasutterella excrementihominis* has been observed to be unique to a cohort of treatment-naïve children [64]. Bacillales [74] and Bifidobacterium [67] have also been found to be IBD biomarkers.

When the taxonomic resolution is increased, bacterial species previously unassociated with IBD begin to emerge. *Actinomyces graevenitzii* is capable of infecting humans in combination with other bacterial species. Copathogens such

TABLE 3
Taxonomy of Robust Microbial Markers of Crohn's disease.

Frequency	Order	Family	Genus	Species	Previously reported?
1	Bacteroidales	Bacteroidaceae	Bacteroides		✓ [15]
1	Pasteurellales	Pasteurellaceae	Haemophilus		✓ [15]
1	Clostridiales	Lachnospiraceae			✓ [15]
0.8	Actinomycetales	Actinomycetaceae	Actinomyces	graevenitzii	✗
0.8	Clostridiales	Lachnospiraceae	Roseburia		✓ [15]
0.8	Clostridiales	Peptostreptococcaceae	Intestinibacter	bartlettii	✗
0.8	Clostridiales	Ruminococcaceae	Ruminococcaceae UCG-002		✓ [15]
0.6	Erysipelotrichales	Erysipelotrichaceae	Erysipelatoclostridium		✓ [15]
0.4	Clostridiales	Lachnospiraceae	Roseburia	inulinivorans	✗
0.4	Bacteroidales	Bacteroidaceae	Bacteroides	vulgatus	✓ [15]
0.4	Burkholderiales	Alcaligenaceae	Parasutterella	excrementihominis	✓ [64]
0.4	Pasteurellales	Pasteurellaceae	Actinobacillus		✓ [15]
0.2	Selenomonadales	Veillonellaceae	Megamonas	funiformis	✗
0.2	Fusobacteriales	Fusobacteriaceae	Fusobacterium		✓ [15]
0.2	Bacteroidales	Bacteroidaceae	Bacteroides		✓ [65]
0.2	Pasteurellales	Pasteurellaceae	Haemophilus	influenzae or parainfluenzae	✓ [15]
0.2	Clostridiales	Ruminococcaceae	Ruminiclostridium 5		✓ [15]
0.2	Enterobacteriales	Enterobacteriaceae	Escherichia /Shigella		✓ [15]
0.2	Clostridiales	Ruminococcaceae	Ruminococcus 2	bromii	✓ [66]
0.2	Clostridiales	Lachnospiraceae	Blautia		✓ [15]

TABLE 4
Taxonomy of Robust Microbial Markers of ulcerative colitis.

Frequency	Order	Family	Genus	Species	Previously reported?
1	Clostridiales	Ruminococcaceae			✓ [15]
1	Bacteroidales	Bacteroidaceae	Bacteroides	vulgatus	✓ [15]
1	Lactobacillales	Lactobacillaceae	Pediococcus		✓ [67]
1	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae UCG-003		✓ [15]
0.8	Clostridiales	Lachnospiraceae	Anaerostipes	hadrus	✗
0.8	Clostridiales	Peptostreptococcaceae	Intestinibacter	bartlettii	✗
0.8	Lactobacillales	Streptococcaceae	Streptococcus		✓ [15]
0.6	Enterobacteriales	Enterobacteriaceae			✓ [15]
0.6	Clostridiales	Lachnospiraceae			✓ [15]
0.6	Lactobacillales	Streptococcaceae	Lactococcus		✓ [15]
0.6	Lactobacillales	Lactobacillaceae	Lactobacillus		✓ [15]
0.2	Bacillales	Family_XI	Gemella		✓ [15]
0.2	Clostridiales	Lachnospiraceae			✓ [15]
0.2	Bacteroidales	Bacteroidaceae	Bacteroides		✓ [15]
0.2	Clostridiales	Ruminococcaceae	Faecalibacterium	cf./prausnitzii	✓ [68]
0.2	Bacteroidales	Bacteroidaceae	Bacteroides		✓ [15]
0.2	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium		✓ [15]

as *A. graevenitzii* rely on other bacterial species to inhibit the host immune system or to reduce the amount of oxygen in the local environment before infection can occur [75]; *A. graevenitzii* has been implicated in coinfection with tuberculosis [75]. In active IBD localised areas of the gut are hypoxic due to metabolic demand outpacing supply [76]: the IBD gut appears to provide ideal conditions for *A. graevenitzii* to grow. *A. graevenitzii* is a strong biomarker for CD, with a frequency of 0.8 (see Tables 3–4). *Intestinibacter bartlettii* has only been very recently defined, and its role in the human gut and human health is uncertain; recent work shows that *I. bartlettii* is thought to be resistant to oxidative stress and is involved with mucus degradation [77]. Oxidative stress is significantly increased in areas of mucosal inflammation in IBD [76]. *I. bartlettii* is a robust biomarker for both of the CD and UC cohorts, with a frequency of 0.8. Both *Anaerostipes hadrus* and *Roseburia inulinivorans* are lactate utilising butyrate-producing bacteria, which have been proposed as potential probiotics because butyrate promotes gut health [78]. *A. hadrus* is a strong biomarker for UC only with a frequency of 0.8, and *R. inulinivorans* is a moderate marker for CD with a frequency of 0.4. *Megamonas funiformis* is a weak biomarker for CD (with a frequency of 0.2) and was originally isolated from human faeces. Its role in the human gut or health is currently unclear [79]. In summary, we present a group of previously undescribed biologically plausible bacterial species that are robust microbial markers for IBD. The group includes gut health promoting bacteria, bacterial species that thrive in the inflammatory environment of an IBD gut and possibly exacerbate the disease, and other bacterial species with unclear roles in human health.

5 Conclusion

In this paper we reported use of a two-stage feature selection process of prevalence filtering and aggregating EFS to identify a robust set of 16S exact sequence variants (DSVs) that can non-invasively predict IBD. The development of an accurate non-invasive test for IBD could decrease time to diagnosis, improving patient outcome. The DSV paradigm offers a wide variety of benefits over the fuzzy clustering OTU approach, including increased taxonomic resolution [17]; previous work has focused on higher taxonomic ranks (family or genus). In the aggregating EFS paradigm, robustness is defined as the variation of feature selector output caused by small changes to the input. Merging the output of multiple weaker feature selectors has been shown to improve the robustness of feature selectors, in a manner similar to ensemble learning [18]. We test the classification performance of the robust feature subsets and find their predictive power is the highest reported in the literature to date. The generalisation ability of the robust feature subset was also verified against a validation dataset (a 20% hold-out partition). A robust feature subset is also valuable for knowledge discovery which domain experts can use to plan future experiments. The majority of DSVs in the feature subset have been previously implicated in IBD, which validates the biological plausibility of the EFS output. Here we identify five stable bacterial species that have not been previously implicated in the pathogenesis of IBD: *Actinomyces graevenitzii*, *Intestinibacter bartlettii*, *Megamonas funiformis*, and *Anaerostipes hadrus*. For the majority of the novel DSVs it is biologically plausible that they are involved with the pathophysiology of IBD. Evidence in the literature has shown the novel DSVs thrive in conditions low in oxygen, high in oxidative stress, or produce substrates that promote gut health.

Limitations of this work include that the anatomical location of the disease was not taken into account. IBD can be confined to the ileum, rectum, or be present across both. Our rationale for this was that stool samples will act as a proxy for the entire gastrointestinal tract, so disease will naturally be reflected in the composition of the microbiome sampled from stool. Additionally, there was a significant imbalance in the size and structure of the dataset in that relatively few controls were present.

Overcoming these limitations will require expanding the cohort (in particular creating cohorts with disease limited to certain areas of the gut) to further validate the robust microbial markers in future work. Additionally, the microbiome does not exist in isolation. Taking into account the human host (e.g. incorporating data that describe the host genome and genetic or epigenetic predispositions) could enable more holistic modelling of the gastrointestinal microbiome, and applying aggregating EFS to this type of data could generate new insights into the aetiology of IBD. The use of DSVs significantly increases the clinical utility of the identified feature subsets. The specific sequence of the selected microbes associated with disease are known, and in future probes could be designed for *in vitro* or *in vivo* validation of the microbial sequence markers. For example, the relative fluorescence of real-time PCR probes that target the DSV subsets could be used to measure the abundance of the DSVs. From these data, new diagnostic tests could be created and validated.

Acknowledgments

This work was completed under a PhD studentship supported by the Northern Ireland Department for Employment and Learning (DEL).

References

[1] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Comput Biol*, vol. 6, no. 2, p. e1000667, 2010.

[2] M. S. Rappé and S. J. Giovannoni, "The uncultured microbial majority," *Annual Reviews in Microbiology*, vol. 57, no. 1, pp. 369–394, 2003.

[3] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon *et al.*, "QIIME allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[4] N. A. Molodecky, S. Soon, D. M. Rabi, W. A. Ghali, M. Ferris, G. Chernoff, E. I. Benchimol, R. Panaccione, S. Ghosh, H. W. Barkema *et al.*, "Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review," *Gastroenterology*, vol. 142, no. 1, pp. 46–54, 2012.

[5] M. J. Carter, A. J. Lobo, and S. P. Travis, "Guidelines for the management of inflammatory bowel disease in adults," *Gut*, vol. 53, no. suppl 5, pp. v1–v16, 2004.

[6] C. Spray, G. Debelle, and M. Murphy, "Current diagnosis, management and morbidity in paediatric inflammatory bowel disease," *Acta Paediatrica*, vol. 90, no. 4, pp. 400–405, 2001.

[7] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4516–4522, 2011.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[9] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012, vol. 454.

[10] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.

[11] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[12] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. J. Blaser, C. F. Aliferis, and A. V. Alekseyenko, "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, vol. 1, no. 1, p. 11, 2013.

[13] M. Tong, X. Li, L. W. Parfrey, B. Roth, A. Ippoliti, B. Wei, J. Borneman, D. P. McGovern, D. N. Frank, E. Li *et al.*, "A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease," *PloS one*, vol. 8, no. 11, p. e80702, 2013.

[14] E. Papa, M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram *et al.*, "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease," *PloS one*, vol. 7, no. 6, p. e39242, 2012.

[15] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour *et al.*, "The treatment-naive microbiome in new-onset crohns disease," *Cell host & microbe*, vol. 15, no. 3, pp. 382–392, 2014.

[16] R. Loomba, V. Seguritan, W. Li, T. Long, N. Klitgord, A. Bhatt, P. S. Dulai, C. Caussy, R. Bettencourt, S. K. Highlander *et al.*, "Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease," *Cell Metabolism*, vol. 25, no. 5, pp. 1054–1062, 2017.

[17] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *ISME J*, Jul 2017, perspective. [Online]. Available: http://dx.doi.org/10.1038/ismej.2017.119

[18] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Machine learning and knowledge discovery in databases*. Springer, 2008, pp. 313–325.

[19] T. G. Dietterich *et al.*, "Ensemble methods in machine learning," *Multiple classifier systems*, vol. 1857, pp. 1–15, 2000.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning*. Springer, 2012, pp. 307–323.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[23] C. Van Rijsbergen, *Information Retrieval*. Butterworths, 1979. [Online]. Available: https://books.google.co.uk/books?id=t-pTAAAAMAAJ

[24] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[25] A. M. Eren, L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin, "Oligotyping: differentiating between closely related microbial taxa using 16s rrna gene data," *Methods in Ecology and Evolution*, vol. 4, no. 12, pp. 1111–1119, 2013.

[26] A. M. Eren, H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin, "Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences," *The ISME journal*, vol. 9, no. 4, p. 968, 2015.

[27] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: high-resolution sample inference from Illumina amplicon data," *Nature methods*, 2016.

[28] R. C. Edgar, "Unoise2: improved error-correction for illumina 16s and its amplicon sequencing," *bioRxiv*, p. 081257, 2016.

[29] P. J. McMurdie and S. Holmes, "Waste not, want not: why rarefying microbiome data is inadmissible," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003531, 2014.

[30] W. Ruth and T. Loughin, "The effect of heteroscedasticity on regression trees," *arXiv preprint arXiv:1606.05273*, 2016.

[31] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[32] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[33] E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy *et al.*, "Richness of human gut microbiome correlates with metabolic markers," *Nature*, vol. 500, no. 7464, pp. 541–546, 2013.

[34] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu *et al.*, "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, p. 59, 2014.

[35] K. Aagaard, K. Riehle, J. Ma, N. Segata, T.-A. Mistretta, C. Coarfa, S. Raza, S. Rosenbaum, I. Van den Veyver, A. Milosavljevic *et al.*, "A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy," *PloS one*, vol. 7, no. 6, p. e36466, 2012.

[36] A. Statnikov, A. V. Alekseyenko, Z. Li, M. Henaff, G. I. Perez-Perez, M. J. Blaser, and C. F. Aliferis, "Microbiomic signatures of psoriasis: feasibility and methodology comparison," *Scientific reports*, vol. 3, 2013.

[37] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971–989, 2016.

[38] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[39] J. Carter, D. Beck, H. Williams, G. Dozier, and J. A. Foster, "Ga-based selection of vaginal microbiome features associated with bacterial vaginosis," in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. ACM, 2014, pp. 265–268.

[40] M. B. Kursa, W. R. Rudnicki *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[41] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[42] R. Polikar, "Ensemble based systems in decision making," *Circuits and systems magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[43] C.-S. Chen, S. Sullivan, T. Anderson, A. C. Tan, P. J. Alex, S. R. Brant, C. Cuffari, T. M. Bayless, M. V. Talor, C. L. Burek *et al.*, "Identification of novel serological biomarkers for inflammatory bowel disease using escherichia coli proteome chip," *Molecular & Cellular Proteomics*, vol. 8, no. 8, pp. 1765–1776, 2009.

[44] P. J. Schüffler, D. Mahapatra, J. A. Tielbeek, F. M. Vos, J. Makanyanga, D. A. Pendsé, C. Y. Nio, J. Stoker, S. A. Taylor, and J. M. Buhmann, "A model development pipeline for crohns disease severity assessment from magnetic resonance images," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*. Springer, 2013, pp. 1–10.

[45] T. Bezabeh, R. L. Somorjai, and I. C. Smith, "Mr metabolomics of fecal extracts: applications in the study of bowel diseases," *Magnetic Resonance in Chemistry*, vol. 47, no. S1, 2009.

[46] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. M. Visscher *et al.*, "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," *The American Journal of Human Genetics*, vol. 92, no. 6, pp. 1008–1012, 2013.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[48] M. S. Silverberg, J. Satsangi, T. Ahmad, I. D. Arnott, C. N. Bernstein, S. R. Brant, R. Caprilli, J.-F. Colombel, C. Gasche, K. Geboes *et al.*, "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a working party of the 2005 montreal world congress of gastroenterology," *Canadian Journal of Gastroenterology and Hepatology*, vol. 19, no. Suppl A, pp. 5A–36A, 2005.

[49] A. N. Ananthakrishnan, "Epidemiology and risk factors for ibd," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 4, pp. 205–217, 2015.

[50] R. B. Sartor, "Mechanisms of disease: pathogenesis of crohn's disease and ulcerative colitis," *Nature clinical practice Gastroenterology & hepatology*, vol. 3, no. 7, pp. 390–407, 2006.

[51] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.

[52] C. Boettiger, "An introduction to docker for reproducible research," *ACM SIGOPS Operating Systems Review*, vol. 49, no. 1, pp. 71–79, 2015.

[53] J. Kans, "Entrez direct: E-utilities on the UNIX command line," 2013, Available from: https://www.ncbi.nlm.nih.gov/books/NBK179288/, accessed: 2017-06-29.

[54] R. Leinonen, H. Sugawara, M. Shumway, and I. N. S. D. Collaboration, "The sequence read archive," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D19–D21, 2010.

[55] O. Tange, "Gnu parallel - the command-line power tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: http://www.gnu.org/s/parallel

[56] P. J. McMurdie and S. Holmes, "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data," *PLoS ONE*, vol. 8, no. 4, p. e61217, 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0061217

[57] B. J. Callahan, K. Sankaran, J. A. Fukuyama, P. J. McMurdie, and S. P. Holmes, "Bioconductor workflow for microbiome data analysis: from raw reads to community analyses," *F1000Research*, vol. 5, 2016.

[58] C. E. Determan Jr, "Optimal algorithm for metabolomics classification and feature selection varies by dataset," *International Journal of Biology*, vol. 7, no. 1, p. 100, 2015.

[59] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[60] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[61] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[62] J. Oliveros, "Venny. an interactive tool for comparing lists with venn diagrams. 2007," 2015.

[63] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids, with applications to dna microarrays," *Statistical Science*, pp. 104–117, 2003.

[64] P. Ricanek, S. M. Lothe, S. A. Frye, A. Rydning, M. H. Vatn, and T. Tønjum, "Gut bacterial profile in patients newly diagnosed with treatment-naive crohns disease," *Clinical and experimental gastroenterology*, vol. 5, p. 173, 2012.

[65] L. Chen, W. Wang, R. Zhou, S. C. Ng, J. Li, M. Huang, F. Zhou, X. Wang, B. Shen, M. A. Kamm *et al.*, "Characteristics of fecal and mucosa-associated microbiota in chinese patients with inflammatory bowel disease," *Medicine*, vol. 93, no. 8, 2014.

[66] A. Swidsinski, J. Weber, V. Loening-Baucke, L. P. Hale, and H. Lochs, "Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease," *Journal of clinical microbiology*, vol. 43, no. 7, pp. 3380–3389, 2005.

[67] W. Wang, L. Chen, R. Zhou, X. Wang, L. Song, S. Huang, G. Wang, and B. Xia, "Increased proportions of bifidobacterium and the lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease," *Journal of clinical microbiology*, vol. 52, no. 2, pp. 398–406, 2014.

[68] H. Sokol, B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermúdez-Humarán, J.-J. Gratadoux, S. Blugeon, C. Bridonneau, J.-P. Furet, G. Corthier *et al.*, "Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients," *Proceedings of the National Academy of Sciences*, vol. 105, no. 43, pp. 16 731–16 736, 2008.

[69] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[70] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper *et al.*, "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome biology*, vol. 13, no. 9, p. R79, 2012.

[71] J. Strauss, G. G. Kaplan, P. L. Beck, K. Rioux, R. Panaccione, R. DeVinney, T. Lynch, and E. Allen-Vercoe, "Invasive potential of gut mucosa-derived fusobacterium nucleatum positively correlates with ibd status of the host," *Inflammatory bowel diseases*, vol. 17, no. 9, pp. 1971–1978, 2011.

[72] A. D. Kostic, D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero *et al.*, "Genomic analysis identifies association of fusobacterium with colorectal carcinoma," *Genome research*, vol. 22, no. 2, pp. 292–298, 2012.

[73] J. K. Triantafillidis, G. Nasioulas, and P. A. Kosmidis, "Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies," *Anticancer research*, vol. 29, no. 7, pp. 2727–2737, 2009.

[74] S. Hourigan, L. Chen, Z. Grigoryan, G. Laroche, M. Weidner, C. Sears, and M. Oliva-Hemker, "Microbiome changes associated with sustained eradication of clostridium difficile after single faecal microbiota transplantation in children with and without inflammatory bowel disease," *Alimentary pharmacology & therapeutics*, vol. 42, no. 6, pp. 741–752, 2015.

[75] A. Tietz, K. E. Aldridge, and J. E. Figueroa, "Disseminated coinfection with actinomyces graevenitzii and mycobacterium tuberculosis: case report and review of the literature," *Journal of clinical microbiology*, vol. 43, no. 6, pp. 3017–3022, 2005.

[76] S. P. Colgan, V. F. Curtis, and E. L. Campbell, "The inflammatory tissue microenvironment in ibd," *Inflammatory bowel diseases*, vol. 19, no. 10, p. 2238, 2013.

[77] K. Forslund, F. Hildebrand, T. Nielsen, G. Falony, E. Le Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H. K. Pedersen *et al.*, "Disentangling the effects of type 2 diabetes and metformin on the human gut microbiota," *Nature*, vol. 528, no. 7581, p. 262, 2015.

[78] S. H. Duncan and H. J. Flint, "Probiotics and prebiotics and health in ageing populations," *Maturitas*, vol. 75, no. 1, pp. 44–50, 2013.

[79] H. Sakon, F. Nagai, M. Morotomi, and R. Tanaka, "Sutterella parvirubra sp. nov. and megamonas funiformis sp. nov., isolated from human faeces," *International journal of systematic and evolutionary microbiology*, vol. 58, no. 4, pp. 970–975, 2008.

Benjamin Wingfield Benjamin Wingfield has a BSc (Hons.) degree in Molecular and Cellular Biology and an MSc degree in Synthetic Biology from Newcastle University, UK. He is a cross faculty student working toward a PhD degree in the School of Computing and Engineering and the Northern Ireland Centre for Stratified Medicine at Ulster University. His research interests include personalised medicine, human microbiomes, and computational intelligence.

Sonya Coleman Professor Sonya Coleman is a Professor in the ISRC, the Cognitive Robotics team leader. She has a first class honours degree in Mathematics, Statistics and Computing, and a doctorate from the University of Ulster. She has 150+ publications in image processing, pattern recognition, computational intelligence and robotics. She has substantial experience of managing research grants (with respect to technical aspects and personnel) both as a principal and co-investigator on research grants funded by EPSRC, The Leverhulme Trust and the Nuffield Foundation. Additionally, she was co-investigator on the EU FP7 funded project RUBICON, the FP7 project VISUALISE and is currently co-investigator in the FP7 SLANDIAL project. She is also secretary of the Irish Pattern Recognition and Classification Society.

TM McGinnity T. Martin McGinnity (SMIEEE, FIET) received a First Class (Hons.) degree in Physics in 1975, and a Ph.D degree from the University of Durham, UK in 1979. He is currently Pro Vice Chancellor for Student Affairs and Head of the College of Science and Technology at Nottingham Trent University, UK. Formerly he was Professor of Intelligent Systems Engineering and Director of the Intelligent Systems Research Centre in the Faculty of Computing and Engineering, University of Ulster. He is the author or coauthor of over 300 research papers and has attracted over £25 million in research funding. His research interests are focused on computational intelligence, computational neuroscience, modelling of biological information processing and cognitive robotics.

AJ Bjourson Professor Tony Bjourson is Director of the Northern Ireland Centre for Stratified Medicine which he established at Altnagelvin (C-TRIC) in 2013 with an £11.5M investment. He is also PI on a new grant aimed at establishing an 8.6M Centre for Personalised Medicine Patient Safety and Clinical Decision Making due to commence in mid 2017. He obtained his MSc in Biological Sciences from the University of Ulster and his PhD from Queens University Belfast. He has over 30 years of research experience. Prior to joining Ulster in 2001, he established and managed genomic programmes for the DARDNI and Queens University Belfast and participated in the first international eukaryotic genome project (EU Yeast Genome sequencing program 1994-1996). After joining Ulster he led the Pharmaceutical Biotechnology Research Group and subsequently established and led the Biomedical Genomics Research Group. He has extensive experience of managing large research projects. He was founder and serves as a Director on the board of the Clinical Translation Research & Innovation Centre (C-TRIC) based in L/Derry aimed at translating biomedical research outputs from laboratory bench to patient bedside. He also served on the Board of Directors of the Ulster venture company Innovation Ulster Ltd. He is a Steering Committee member of the Northern Ireland Biobank, a Council Member of the Irish Society for Clinical genetics. He has personally secured in excess of £28M in research grants and has supervised >22 PhD students to successful completion. His own current research is focused on stratified and personalised medicine in the area of immune/autoimmune disease. He is highly committed to economic regeneration in Northern Ireland and the North West region in particular; and is currently focused on attracting major pharma and diagnostic industry location to that region.