Expert Systems With Applications 185 (2021) 115648

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes

Diego Fernández-Edreira [a], Jose Liñares-Blanco [a,b], Carlos Fernandez-Lozano [a,b,*]

[a] Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, Campus Elviña s/n, A Coruña, 15071, Spain

[b] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña, 15071, Spain

ARTICLE INFO

Keywords:
Machine Learning
Diabetes
T1D
Microbiota
Metagenomics
Feature selection
Random forest
Generalized Linear Model

ABSTRACT

Diabetes is a disease that is closely linked to genetics and epigenetics, yet mechanisms for clarifying the onset and/or progression of the disease have sometimes not been fully managed. In recent years and due to the large number of recent studies, it is known that changes in the balance of the microbiota can cause a high battery of diseases, including diabetes. Machine Learning (ML) techniques are able to identify complex, non-linear patterns of expression and relationships within the data set to extract intrinsic knowledge without any biological assumptions about the data. At the same time, mass sequencing techniques allow us to obtain the metagenomic profile of an individual, whether it is a body part, organ or tissue, and thus identify the composition of a given microbe. The great increase in the development of both technologies in their respective fields of study leads to the logical union of both to try to identify the bases of a complex disease such as diabetes. To this end, a Random Forest model has been developed at different taxonomic levels, obtaining results above 0.80 in AUC for families and above 0.98 at species level, following a strict experimental design to ensure that results are compared under equal conditions. It is identified how, in infants, the species Bacteroides uniformis, Bacteroides dorei and Bacteroides thetaiotaomicron are reduced in the microbiota of those with T1D, while, the populations of Prevotella copri increase slightly and that of Bacteroides vulgatus is much higher. Finally, thanks to the more specific metagenomic signature at species level, a model has been generated to predict those seroconverted patients not previously diagnosed with diabetes but who have expressed at least two of the autoantibodies analysed.

1. Introduction

The microbiota is a complex ecosystem of microorganisms composed of bacteria, viruses, protozoa and fungi, which live in different locations in the human body, such as the gastrointestinal tract, skin, mouth, respiratory system and vagina. In the last decade, numerous works have reported certain metabolic activities and host/host interactions that influence the normal physiology of the human being (Belkaid & Hand, 2014; Heijtz et al., 2011; Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012). Over 70% of the microbiota live in the gastrointestinal tract, in symbiosis with human eukaryotic cells. This community consists of about 100 trillion commensal microorganisms, about 10 times the total number of human cells. The difference is much greater if we look at the genome that makes up the set of all these microorganisms, known as the metagenome. Knowing the role played by this entire genomic network is crucial to understanding both their functions and the relationships between the host organism and these microorganisms. This is why the impact of the microbiota on human health is currently one of the greatest challenges in clinical care, drug discovery and biomedicine.

Therefore, disturbances in the composition and/or proportion of the microbiota can result in the development of a significant physiological change or even a pathology. This type of change is known as dysbiosis and can be due to multiple factors, the most common of which are: the individual's own genetics, the diet carried out throughout his or her life, personal hygiene, different infections, uncontrolled intake of drugs and antibiotics or certain medical interventions (Petersen & Round, 2014). These processes of dysbiosis that affect the balance in the microbiota play an important role in complex diseases such as asthma (Petersen & Round, 2014), neurodevelopmental disorders (Hsiao et al., 2013),

* Corresponding author at: Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, Campus Elviña s/n, A Coruña, 15071, Spain.
E-mail addresses: diego.fedreira@udc.es (D. Fernández-Edreira), j.linares@udc.es (J. Liñares-Blanco), carlos.fernandez@udc.es (C. Fernandez-Lozano).
URL: https://cafernandezlo.github.io/es_github_cafernandezlo (C. Fernandez-Lozano).

https://doi.org/10.1016/j.eswa.2021.115648
Received 20 November 2020; Received in revised form 21 June 2021; Accepted 20 July 2021
Available online 29 July 2021

cancer (Garrett, 2015) and diabetes (Karlsson et al., 2013; Petersen & Round, 2014) among others.

Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action or both biological processes. Several pathogenic processes are involved in the development of diabetes. These range from autoimmune destruction of pancreatic cells, with consequent insulin deficiency, to abnormal diseases causing resistance to insulin action. Deficient insulin action is the result of inadequate insulin creation and/or decreased tissue responses to insulin at one or more points along complex hormonal pathways. Impaired insulin secretion and defects in insulin action usually coexist in the same patient and it is often not clear which abnormality, if any, is the main cause of hyperglycaemia (Mellitus, 2005).

Specifically, Type I diabetes (T1D) (Mellitus, 2005) is an organ-specific autoimmune disease due to the infiltration and attack of T-lymphocytes and other immune cells on pancreatic cells β, resulting in the destruction of β cells and progression to insulin deficiency (Boldison & Wong, 2016). Most T1D is of the autoimmune type (TIA), and a smaller proportion is of the non-autoimmune type, also known as idiopathic type I diabetes (TIB). Genetic factors play an important role in the origin of T1D. To date, many loci have been associated with the disease, although there is a lack of knowledge about the environmental factors involved. This is evident from the fact that less than ten percent of individuals who are genetically predisposed end up developing it and there is an increasing frequency of lower-risk genotypes in patients diagnosed with TIB (Zhou et al., 2020).

The environmental factors that the scientific community has reported to be most related to the development of T1D are, for example, the diet of the individual (Norris et al., 2003; Wahlberg, Vaarala, Ludvigsson, Group, et al., 2006; Ziegler, Schmid, Huber, Hummel, & Bonifacio, 2003), different viral infections (Kaufman et al., 1992; Stene & Rewers, 2012) as well as the gut microbiota itself as it is colonized by hundreds of microbial species selected for a mutually beneficial relationship with its host. During infancy and childhood, the intestinal microbiota undergoes constant remodelling and instructs the development of the immune system. Therefore, if this microbiota is affected, it can generate complications in the future, including triggering the onset of T1D. For example, the intestinal microbiota of a newborn will be different depending on whether it is born vaginally or by caesarean section, resulting in a significant difference in the populations of the species that inhabit it. Numerous studies have also reported that there is a difference between the gastrointestinal microbiota of breastfed infants and those not fed by this route, with significant differences being in bifidobacteria, which almost always dominate the gastrointestinal microbiota of infants breastfed at several weeks of age (Favier, Vaughan, De Vos, & Akkermans, 2002; Penders et al., 2006; Stark & Lee, 1982). There are also studies that maintain that the microbiota of infants who have been hospitalized, suffered from fever or had to take antibiotics during the first month, show significant differences from those infants who have been completely healthy during this time (Penders et al., 2006; Schwiertz et al., 2003). All these factors, which modulate the composition of gastrointestinal microbiota, are also risk factors in the development of T1D, so it is reasonable to consider microbiota as a link between these factors and the promotion of the disease.

Currently, the diagnostic form of T1D is mainly concerned with the analysis of antibodies, such as auto-insulin antibodies (AIA), glutamic acid decarboxylase (GADA) antibodies, antibodies to ICA512 or IA-2, a transmembrane protein of the tyrosine phosphatase family and the zinc transporter 8 (ZnT8A) (Boldison & Wong, 2016; Chen et al., 2013; Yi, Huang, & Zhou, 2015). Normally, because the main effect of the disease is changes in glucose levels, other diagnostic tests are available, such as fasting plasma glucose (FPG), 2-h plasma glucose (2-h PG) during an oral glucose tolerance test of 75 g (OGTT) or by the A1C test.

With the gradual incorporation of mass sequencing platforms into the daily routine of a laboratory and the reduction of costs, more and more data is produced every day to study a specific problem. Today, most bacterial identification is already carried out by genome sequencing methods. In addition, major initiatives are making these data available to the scientific community for analysis and different scientific groups are offering new possibilities for analysis and exploitation of the data. The problem has shifted towards managing the analysis of huge amounts of data with their relationships and balances, generating problems that are difficult to address with conventional approaches. At the same time, the rise of Machine Learning (ML) for the computational analysis of complex and high capacity biomedical data makes use of models (Harrington, 2012) with great predictive capacity developed specifically for problems with large amounts of data and with noise. Once trained, the ML algorithm is intended to have the highest possible generalization capacity so that the model works not only with the data it has learned, but also with the data we will obtain in the future (Alpaydin, 2020; Marsland, 2015; Mohri, Rostamizadeh, & Talwalkar, 2018).

The good results obtained in the application of these algorithms in various fields has been the reason why they are also being applied in the field of biomedicine. These algorithms have been used for cancer diagnosis (Huang et al., 2018; Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015; Zhao et al., 2019), to predict the best synergies between drug pairs and biomarkers (Menden et al., 2019), prediction of high anti-angiogenic activity peptides (Liñares-Blanco, Porto-Pazos, Munteanu, Pazos, & Fernandez-Lozano, 2018), neurological diseases (Liu et al., 2014; Ludwig et al., 2019), heart diseases (Guvenir, Acar, Demiroz, & Cekin, 1997; Özçift, 2011) and metabolic diseases (Dugan, Mukhopadhyay, Carroll, & Downs, 2015; LaFreniere, Zulkernine, Barber, & Martin, 2016) among others such as groundwater modelling (Choubin & Rahmati, 2021; Mosavi et al., 2020, 2021). Regarding the analysis of sequencing and ML data, based on clustering techniques in OTUs (Operational Taxonomic Unit) according to their abundance, there are several works (Karlsson et al., 2013) to generate a RF model to predict patients affected by T2D. The prediction power of metagenomic data with different disease states on six available disease-associated datasets using ML and feature selection with AUC scores ranging from 0.65 for obesity to 0.94 for cirrhosis and 0.74 for T2D (Pasolli, Truong, Malik, Waldron, & Segata, 2016). Based on the arrhythmic microbiota throughout the day, attempts have been made to classify individuals according to their risk of suffering T2D (Reitmeier et al., 2020). Supervised ML analysis of relative abundance in T1D an onset showed with sensitivity and specificity of 0.54, 0.62 differences with respect to controls (Biassoni et al., 2020). Also, to look for biomarkers of the association between the microbiome and the development of colorectal cancer (Thomas et al., 2019), irritable bowel syndrome (Fukui et al., 2020) or inflammatory bowel disease (Wingfield, Coleman, McGinnity, & Bjourson, 2018). Furthermore, following this methodology, an attempt has been made to obtain a universal metagenomic signature based on microbiota to predict cirrhosis (Oh et al., 2020). For more information about comparative study of machine learning classifiers for human microbiome data we refer to Wang and Liu (2020).

Due to the above, the aim of this study arises from the potential use of ML models for the diagnosis of T1D from data from the intestinal microbiota of infants to search for species that influence the development of T1D. Furthermore, the analysis of the models will offer the possibility of extracting new knowledge about the possible role played by the microbiota in the development of the disease and at what taxonomic level the greatest amount of information is found, allowing the focus of future research to be on modifying the existing balance to alter a study condition.

The main contributions of this paper can be listed as: first, we have proven that with metagenomic data (sparse matrices) we can build a Machine Learning model capable of obtaining great results; second, we propose a new metagenomic signature highly correlated with T1D diagnosis to be studied by clinicians and as a new reference for future

Table 1
Summary of DIABIMMUNE project samples.

	#Samples	Gender (M/F)	Average age	Std. age
Control	53	25/28	1.70	0.74
Cases	75	30/45	1.56	0.69

work; and third, we present a robust Machine Learning methodology for further research of the diagnosis of T1D in infants.

The current paper is organized as follows: the Materials and Methods section describes the dataset, the technical aspects of the methodology and the machine learning models; the Results section includes a comparison with baseline models, feature selection, best model determination and how we defined the metagenomic signature; the Discussion and the Conclusions sections are presented before a final section with the data and code used in the paper.

2. Materials and methods

2.1. Dataset

The data used in this work was downloaded from the DIABIMMUNE (Kostic et al., 2015) project. This project arises with the objective of testing the hypothesis of hygiene and its role in the development of T1D. For this study, the T1D cohort was used, which aims to compare the microbiome of infants who have developed T1D with healthy controls from the same geographical area. Fecal samples were extracted from each individual and ribosomal 16S RNA sequencing was performed to characterize the metagenomic profile. For this study, data on the relative abundance of each operative taxonomic unit (OTU) of the different infants that make up the cohort were downloaded. The samples have been labelled according to patients and T1D controls. In total, 124 samples have been included for analysis, from a total of 33 infants.

The relative abundance matrix of OTUs presents abundance at the following 6 different taxonomic levels: phylum, order, class, family, genus and species. The general data of the samples analysed are shown in Table 1.

2.2. Machine learning

For the experiments the following ML algorithms have been used in R (R Core Team, 2020): Random Forest (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes & Vapnik, 1995) and Generalized Linear Model (glmnet) (Friedman, Hastie, & Tibshirani, 2010).

2.2.1. Random forest

The RF algorithm consists of a set of independent decision trees based on the random resampling of the variables for the construction of each tree. A search was made for the appropriate values for the hyperparameters *mtry* (number of variables randomly sampled in each data division), *nodesize* (minimum size of the terminal nodes) and number of trees. The range for the number of variables was set between 1 and, as an upper limit, the square root of the number of variables with the largest data set. The minimum size of the terminal nodes was set between 1 and 3. Low values of this parameter provide high growth and depth of each tree, which improves the accuracy of the predictions. In addition, the number of trees was 1000. A large number of trees ensures that each observation is predicted at least several times.

2.2.2. Support vector machines

The SVM algorithm in binary classification tasks tries to find the best hyper plane that separates the two target classes, minimizing the error. Since most of the real problems do not have a linear relation, the SVM algorithm offers the possibility of calculating a kernel function to map the data in a greater number of dimensions, which allows to separate the data linearly. For this study we used the kernel function RBF (radial Gaussian base) and we made a search of the appropriate values for C (penalty for misclassified observations) and *sigma* (standard deviation of Gaussian distribution). The values of the hyperparameters C and *sigma* were investigated, both with a range between 2^{-12} and 2^{12}.

2.2.3. Generalized linear model

The glmnet algorithm is a rapid regularization algorithm that fits a generalized linear model with elastic network penalties. The network penalty depends on two terms: the *ridge* penalty, which aims to reduce the coefficients of the predictors correlated with each other, and the *lasso* penalty, which tends to choose one of them and discard the others. A search was carried out for the appropriate values for *alpha* (controls the penalty of the elastic network) and *lambda* (controls the total strength of the penalty). The values of *alpha* ranged from 0.0001, 0.001, 0.01, 0.1 and 1, while those of *lambda* were 0, 0.15, 0.25, 0.35, 0.5, 0.65, 0.75, 0.85, 1.

2.3. Feature selection

The vast majority of genetic data generated by biomedical research consists of a large number of variables compared to available observations and/or samples, mainly due to their cost. A large number of these variables may not be useful, as they do not provide information to the model, or they cause noise and generate a possible over-fitting. There are different approaches in this field designed to avoid this problem, as it is the case of the selection of characteristics (FS). The aim of these methods is to find a subset of variables that contain the most information from the original data and that by generating a model are able to maintain a similar or higher performance, maintaining the fidelity of the original models, reducing the search times and generating simpler and faster models. All this is achieved by eliminating variables that have a strong correlation and do not propose new information to the model, redundant variables and variables that only increase noise when generating a model (Guyon & Elisseeff, 2003; Liu & Motoda, 2012). In particular the *Filter* methods measure the relevance of variables in relation to the class of output by looking only at the intrinsic properties of the data without taking into account any assumptions about the classification algorithms that will be used later for the analysis. Therefore, this approach is computationally simple and fast and because it is independent of the classification algorithm (Chandrashekar & Sahin, 2014; Saeys, Inza, & Larrañaga, 2007).

2.4. Best model determination

One of the key points in the use of ML techniques is the use of a robust and fair experimental design that allows the identification of the models and configurations that can obtain better generalization results on future unknown data (Fernandez-Lozano, Gestal, Munteanu, Dorado, & Pazos, 2016). That is why a two-level cross validation (CV) was performed for the training of the algorithms. This type of validation consists of two CV processes, an independent internal level (2/3 holdout for training and 1/3 for validation) for the selection of the best hyperparameters of each algorithm and an independent external level (in this case, 5 repetitions of a 10-fold CV) to evaluate the model's capacity for generalization and ensure that there are no biases in the data.

2.5. Performance

For the evaluation of the models, the following three measures have been calculated: accuracy (ACC), area under the ROC curve (AUC) and the mean of the error in classification (MMCE). The ACC is the fraction of predictions that the model has made correctly, in binary classification, we usually talk about the fraction of true positives and negatives. The AUC measure is the area under the ROC curve. A ROC curve represents the rate of true positives (TPR) versus the rate of false positives (FPR) at different classification thresholds, so that, if the classification threshold is reduced, more elements will be classified as positive. This measure refers to the ability of the models to distinguish between classes, the value is between 0 and 1 and the higher it is, in our case, the better the model will be to differentiate between controls and cases. There are also more specific measures for imbalanced data such as concordant partial AUC (Carrington et al., 2020) that could be of interest in the medical domain. The MMCE is the average of the predicted values that are different from the true values.

3. Results

From the original data set new datasets were generated based on relative abundance according to taxonomic level (phylum, class, order, family, genus and species). A pre-processing was then carried out for each dataset in which samples with unavailable values and variables with variance of zero or close to zero were checked and eliminated. The latter was carried out thanks to the function *NearZeroVariance* within the R (Kuhn, 2008) Caret package. In addition, the data were normalized to have mean 0 and standard deviation 1 and those variables that had a value of 0 in at least 95 percent of their observations were eliminated. For each of the six data sets, a univariate FS method (Wilcoxon test) was applied to rank the variables according to statistical significance with the dependent variable. Version R/4.0.2 was used and the training of the models was carried out in a high performance computing (HPC) environment.

3.1. Baseline experiments

Initially three ML models (SVM, RF and glmnet) were trained with data from each of the six taxonomic levels: phylum, class, order, family, genus and species.

In Fig. 1 it is observed how RF presents the best results in AUC for each one of the taxonomic levels in comparison to the other two algorithms. RF also achieves the best result at the species level. It is interesting to stop at the taxonomic level in which the three algorithms start to present good yields. In the case of RF, it is from the family level, with 0.82 in AUC, to 0.987 in AUC at the species level. As for the other two algorithms, glmnet and SVM, both present very similar results although lower in performance. In this case, the two algorithms begin to present significant results at the species level, not being able to model the data at the family or genus level. Furthermore, their best results at species level are comparable to the results obtained by RF at family level.

Furthermore, in view of what these results suggest, we consider it necessary to carry out a more exhaustive analysis of the taxonomic species level, due to its good performance with the three algorithms used. For this purpose, a search and selection of the best characteristics within this abundance matrix will be carried out by means of FS techniques.

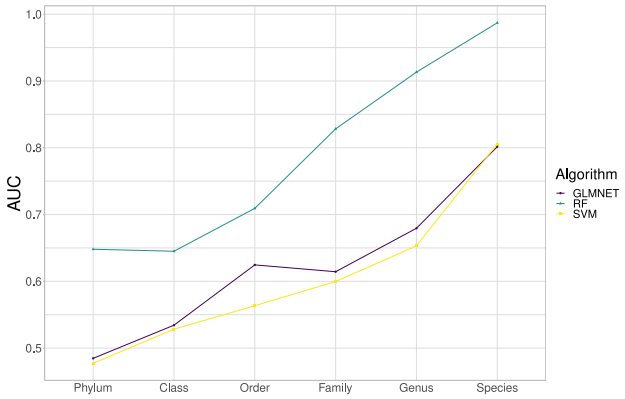

Fig. 1. AUC results according to the algorithm used and taxonomic level.

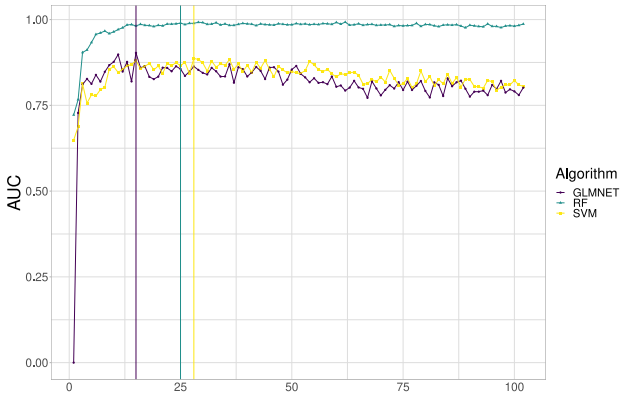

Fig. 2. AUC results at species taxonomic level according to the number of features and algorithm.

3.2. Feature selection

The results reported in the previous section motivated the search for a simpler model capable of obtaining a similar or better performance than the previous one. For this purpose, based on the relative abundance of 102 species, a search was made for characteristics to include in the final model. Each variable was subjected to the Wilcoxon statistical test, in order to order the variables according to their relative importance in the problem. Fig. 2 shows the results according to the AUC of the three algorithms. The results for the other taxonomic levels can also be accessed in Figures S1–S5 (Supplementary Materials).

The results observed in Fig. 2 coincide with those of the previous section, showing the RF algorithm as the one with the best performance. In addition, a stability of the three models is observed as soon as they exceed a certain number of characteristics. This fact is important when selecting the best model, since it tells us which are the characteristics that really present the information. The lines drawn in Fig. 2 show the final model that has been chosen for each algorithm. For RF it was decided to keep 25 characteristics, for the SVM model with 28 characteristics, while for glmnet it was 15 characteristics.

A selection of the best model will then be made based on statistical criteria in order to continue the analysis and deepen the model.

3.3. Best model determination

Table 2 shows the performance of each algorithm according to the three measures reported, as well as the number of characteristics used for their construction.

It is interesting to know if the RF model, besides being better in the average of the performance, is significantly better than the other

Table 2
Results of the best models of each algorithm.

	ACC	AUC	MMCE	Nº features
glmnet	0.834	0.903	0.166	18
RF	0.947	0.99	0.053	25
SVM	0.783	0.887	0.217	28

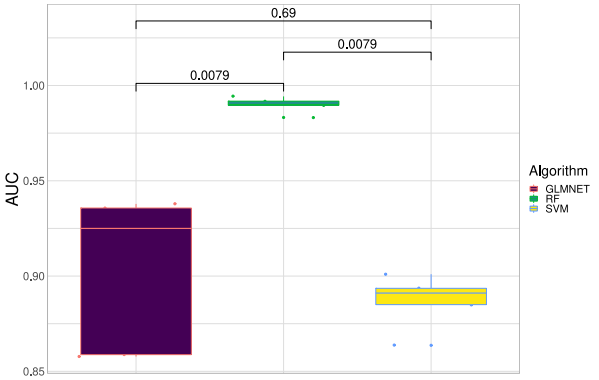

Fig. 3. Comparison of the best models. The p-values were obtained using the Wilcoxon test.

two models as in other domains (Choubin et al., 2019; Rahmati et al., 2019). For this purpose, the box diagram of the performance measures of the three models is shown in the Fig. 3. Because 5 repetitions of a 10 fold CV were performed for the validations of each model, a total of 50 measures of each model were obtained. It can be seen in the figure, how the RF algorithm is the most stable among the three, obtaining all the measures very close to one. In addition, a test of multiple comparisons was made, as it is the Wilcoxon test, being the null hypothesis that both populations have the same probability of presenting greater observations to the other population. A significance value of 0.0079 is observed between the distribution of RF yields and the other two, while there are no significant differences between glmnet and SVM as they have a *p*-value of 0.69. In addition, we can observe how the RF AUC values have very little dispersion since they go from 0.97 to 0.99, so, it remains very stable in the predictions and always maintains precision. On the other hand, glmnet, being the second algorithm with better results, its AUC has a much higher dispersion. Therefore, the RF algorithm with 25 characteristics was selected for a more in-depth study.

3.4. Definition of the metagenomic signature

In accordance with the previous section, a model with a metagenomic signature of 25 characteristics was retained, since, although the stability of the model can be seen from the 15 characteristics, the objective was to analyse a wider and more heterogeneous signature, with a view to the biological discussion of the results and the search for new species that could be related to the disease. In order to observe which characteristics were given more importance by the model, an analysis of their importance was carried out. In Fig. 4 the importance of each species for the winning model is represented. In this case, the importance is based on the Gini impurity index, which is the total of the impurities of the nodes by the division in the variable, averaged over all the trees. The figure shows a species that predominates over the rest, *Bacteroides uniformis*. There are another four species that are very important in comparison with the rest, *Bacteroides dorei, Prevotella copri, Bacteroides vulgatus* and *Bacteroides thetaiotaomicron*.

Fig. 5 shows the abundances of the species belonging to the best model according to whether they are controls or have T1D, in addition, a Wilcoxon test was carried out to compare the populations of both and to check if they had significant differences. By comparing the results

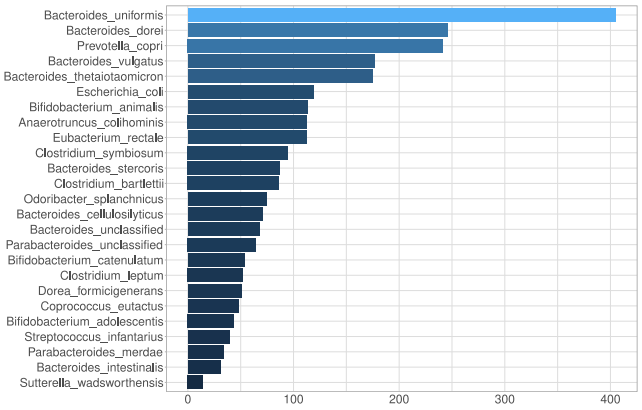

Fig. 4. Importance of the features of the selected model.

shown in this figure with those shown in Fig. 4, it can be seen that the species that present more importance in the model of ML, like *Bacteroides uniformis, Bacteroides dorei* and *Bacteroides thetaiotaomicron* are reduced in the microbiota of the infants with T1D, whereas, the populations of *Prevotella copri* increases slightly and *Bacteroides vulgatus* is much greater. We can also observe in some species that the great majority of the samples present values equal to or very close to zero, but that the samples that present very significant differences to the average belong to a specific class. This fact, which can be observed in species such as *Odoribacter splanchnicus*, *Eubacterium rectale* or *Bifidobacterium adolescentis*, indicates the capacity of the model to find variables that, although not significant with respect to the dependent variable, may be important in the development of the disease.

3.5. Identification of seroconverted patients by metagenomic signature

There are patients who have not been clinically diagnosed with T1D but have expressed at least two of the autoantibodies analysed. These patients, called seroconverts, have a predisposition to present T1D at some point but belong to an intermediate class between healthy and sick. It is interesting to know if our metagenomic signature is able to identify this subgroup of patients, which will have different treatment routes after their diagnosis. For this purpose, the patients were re-labelled in three classes (controls, seroconverts and patients). The selected metagenomic signature was then defined and 90% of the samples were used to train a new RF algorithm to predict which of the three classes each sample belongs to. In the results of the training, following the same methodology as in the previous section, a fairly high and stable performance of the model in terms of ACC is observed, with constant values of 0.94. The remaining 10% of the samples were used as a test to evaluate the model's capacity of generalization. The result of the predictions is shown in Table 3. The model, in this case, was able to classify all the samples correctly, concluding that the model is capable of distinguishing between the three classes, a fact which, at the same time, confirms that the metagenomic signature obtained has a strong relationship with the disease to be treated and has a high specificity when stratifying patients with T1D.

4. Discussion

Previous studies have indicated that gastrointestinal microbiota can play a modulating role in the susceptibility to T1D. This work refers to the differences in gastrointestinal microbiota between cases and controls, in addition to analysing the phenomena that this may cause in the pathogenesis of the disease (Alkanani et al., 2015; Knip & Siljander, 2016; Vaarala, Atkinson, & Neu, 2008; Wen et al., 2008).

The gastrointestinal microbiota of an adult contains approximately 500 to 1000 different bacterial species and is usually dominated by 4

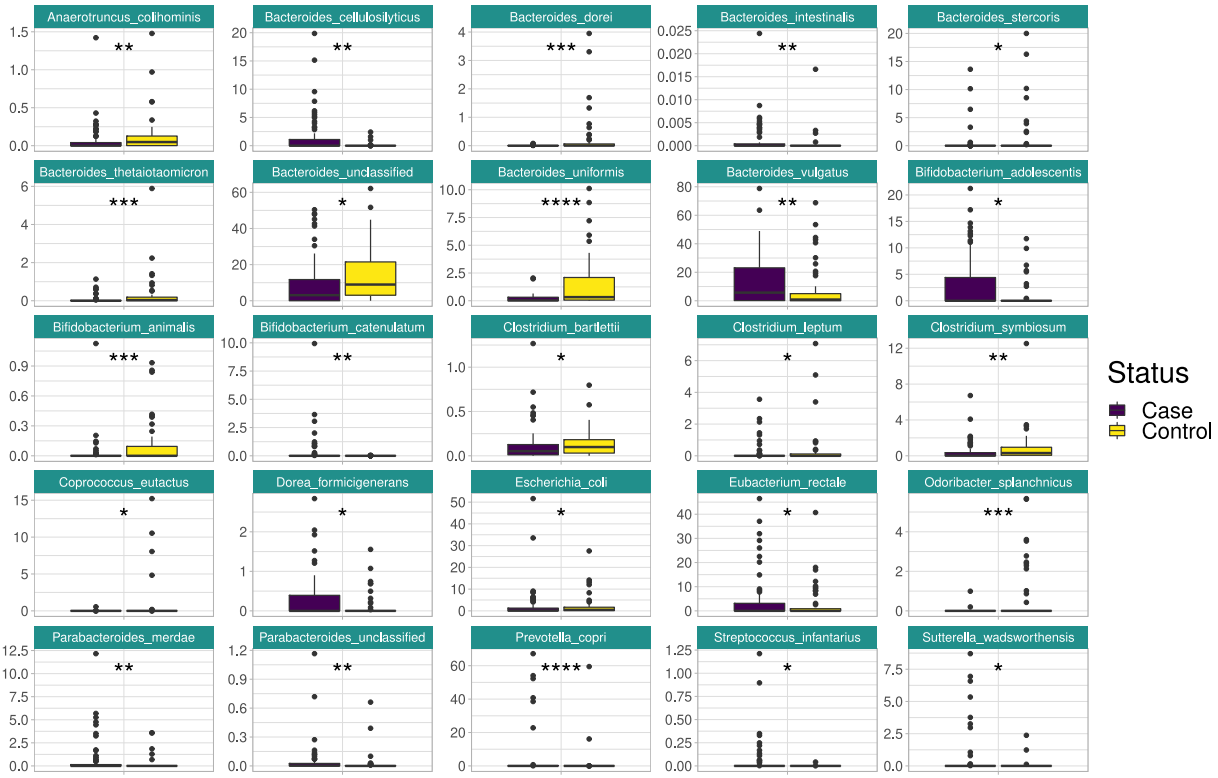

Fig. 5. Comparison of the relative abundances of the species belonging to the metagenomic signature according to the state of the patients. For each box diagram the *p*-value obtained by the Wilcoxon test is shown.

Table 3
Results of the RF model predictions with test data.

Real cases	Prob.Contro	Prob.Serocon	Prob.T1D	Prediction
Control	0.713	0.14	0.147	Control
Serocon	0.196	0.595	0.209	Serocon
Control	0.695	0.148	0.157	Control
Control	0.812	0.074	0.114	Control
T1D	0.273	0.227	0.5	T1D
T1D	0.318	0.21	0.472	T1D
Control	0.644	0.305	0.051	Control
Control	0.536	0.374	0.09	Control
Serocon	0.338	0.562	0.1	Serocon
Serocon	0.392	0.476	0.132	Serocon
Control	0.697	0.17	0.133	Control
Control	0.695	0.128	0.177	Control

phylum, *Firmicutes, Bacteroides, Proteobacteria* and *Actinobacteria* (Bibbò, Dore, Pes, Delitala, & Delitala, 2017; Hooper & Gordon, 2001). In contrast, a child's gastrointestinal microbiota shows greater variability in its composition and remains less stable over time. During the first year, the child's intestinal tract progresses from sterility to extraordinarily dense intestinal colonization, being dominated by two main phyla, *Actinobacteria* and *Proteobacteria*, and finally, by the age of 2 years, the composition, diversity and function of the microbiota is very similar to that of adults, with a predominance of *Firmicutes* and *Bacteroidetes* phyla. (Durazzo, Ferro, & Gruden, 2019; Gülden, Wong, & Wen, 2015; Stark & Lee, 1982).

The baseline results of this work have shown how certain ML models, in particular RF, begin to present significant results from the taxonomic family level. Furthermore, it has been observed that as the metagenomic level is lowered, better model yields are obtained, which indicates the relationship that may exist between the presence or absence of genera and/or species with the development of the pathology. In other words, the development of diabetes, in terms of its relationship with microbiota, is not related to the correct balance between different phyla or orders, but to the presence or absence of certain bacterial species.

These results motivated the search for a metagenomic signature at the species level. The vast majority of studies that have been carried out in relation to the metagenome used data on the relative abundance of the genus, which are not as specific in finding the cause of the disease. For this reason, this work proposes a metagenomic signature based on species with the aim of specifically identifying which are the potential species to play a major role in the development of T1D. The findings are consistent with those of previous studies using RF as classifier with 16S rRNA microbiome data (Corrigan et al., 2018; Roguet, Eren, Newton, & McLellan, 2018; Thompson, Johansen, Dunbar, & Munsky, 2019) or using quantitative metagenomic sequencing (Loomba et al., 2017).

In Table 4, scientific evidence reported in previous works has been gathered that relates the ten most important species to some type of diabetes, immunological pathologies and/or metabolic pathologies. As shown in Table 4 *Bacteroides uniformis* is present in articles related to obesity, which increases a person's risk of suffering from some type of diabetes, and coeliac disease, which together with diabetes itself are both autoimmune diseases, and diabetics are also at risk of coeliac disease. Others such as *Bacteroides dorei, Prevotella copri, Bacteroides vulgatus, Escherichia coli, Bifidobacterium animalis, Eubacterium rectale* and *Clostridium symbiosum* are directly related to diabetes, with populations of these species playing a decisive role in the stratification between controls and cases. There are also some species that are related to metabolic pathways in which glucose is present, such as glucose intolerance or impaired insulin sensitivity, disturbances that can lead to increased blood glucose and result in a diabetic condition.

Fig. 6(a-b) shows the representation of edges and Fig. 6(c-d) shows the genera of the metagenomic signature compared to the total metagenomic population. In terms of the proportion of phyla, the metagenomic signature presented in this work shows that *Verrucomicrobia* is not represented, while there is a large proportional increase in *Bacteroidetes* which becomes the most representative phylum, with a total of twelve

Table 4
Evidence of the first 10 species of the metagenomic signature.

Species	Importance	Type	Gram	Genus	Evidences
Bacteroides uniformis	405.00	*Host-associated*	–	*Bacteroides*	Obesity (Cano, Santacruz, Moya, & Sanz, 2012), Coeliac disease (Sánchez, Donat, Ribes-Koninckx, Calabuig, & Sanz, 2010)
Bacteroides dorei	245.96	*Host-associated*	–	*Bacteroides*	Diabetes (Davis-Richardson et al., 2014; Higuchi et al., 2018; Wu et al., 2019), Coeliac disease (Sánchez et al., 2010)
Prevotella copri	241.60	*Host-associated*	–	*Prevotella*	Diabetes (Higuchi et al., 2018; Kasselman, Vernice, DeLeon, & Reiss, 2018; Leite et al., 2017; Medina-Vera et al., 2019), Insulin sensitivity (Pedersen et al., 2016)
Bacteroides vulgatus	176.44	*Host-associated*	–	*Bacteroides*	Diabetes (Davis-Richardson et al., 2014; Higuchi et al., 2018; Leite et al., 2017), Insulin sensitivity (Pedersen et al., 2016), Coeliac disease (Sánchez et al., 2010)
Bacteroides thetaiotaomicron	174.51	*Host-associated*	–	*Bacteroides*	Coeliac disease (Sánchez et al., 2010)
Escherichia coli	118.80	*Host-associated*	–	*Escherichia*	Diabetes (Qin et al., 2012)
Bifidobacterium animalis	113.29	*Host-associated*	+	*Bifidobacterium*	Diabetes (Amar et al., 2011; Tonucci, Dos Santos, de Oliveira, Ribeiro, & Martino, 2017), Glucose intolerance (Stenman et al., 2014), Coeliac disease (Sánchez et al., 2010)
Anaerotruncus colihominis	112.70	*Host-associated*	+	*Anaerotruncus*	–
Eubacterium rectale	112.48	*Host-associated*	+	*Eubacterium*	Diabetes (Everard & Cani, 2013; Larsen et al., 2010; Murri et al., 2013; Venema, 2010), Chron's disease (Wensinck & Van de Merwe, 1981)
Clostridium symbiosum	94.80	*Host-associated*	+	*Clostridium*	Diabetes (Harsch & Konturek, 2018; Larsen et al., 2010)

species represented, five of which are the most important, indicating the importance of these species in the model (see Fig. 4). On the other hand, the phylum *Actinobacteria* and *Proteinobacteria* maintain a similar proportion in the metagenomic signature. A decrease of *Firmicutes* is detected, being this the second edge with more representation in the signature, but with a minor weight in the model of ML, being its species in low areas in the importance ranking. Therefore, the knowledge embedded in the decision can be extracted from RF by calculating a relative importance value intrinsic to the model, allowing the interpretability and increasing the explainability of the results (Adadi & Berrada, 2018), a key aspect in the medical domain (Holzinger et al., 2019).

With regard to the genera, we first observe that of 42 genera present in the initial dataset, the metagenomic signature remains with only thirteen of them. There is a slight increase in the proportion of all the genera represented in the metagenomic signature in comparison with the original proportions, except for *Streptococcus*, which suffers a decrease in the same, *Eubacterium*, which remains the same, and *Bacteroides*, which suffers a significant increase in comparison with the original. The latter has become the genus with the highest proportion, with eight different species, of which four of the five best species belong to it, and if we look at Fig. 6, we can see that the number of species that have been identified is very high (see Fig. 4). The accumulated importance of these species accounts for most of the weight of the model (as was the case with *Bacteroidetes* in the case of the phylum) and as genera such as *Prevotella, Escherichia, Anaerotruncus* and *Eubacterium*, despite having a very small proportion of species, these are of great importance to the model. The comparison of the proportion for the rest of the metagenomic levels has also been carried out (Supplementary materials Figures S6–S8)

Previous studies have found significant differences in the populations of the genera *Bacteroides* and *Prevotella*, between healthy patients and patients with diabetes (Alkanani et al., 2015; Brown et al., 2011; Dunne et al., 2014; Giongo et al., 2011; Harbison et al., 2019; Mejia-Leon, Petrosino, Ajami, Domínguez-Bello, & Calderon de la Barca, 2014; Murri et al., 2013). Although a direct comparison with our results cannot be made, due to the size and nature of the cohorts (age, culture, country), our results are in line with these statements since the five species where half of the total importance of the model resides, Table 4, are part of these two genera, although in our study, making consensus of all the species that are part of each genus we did not observe a significant difference in the abundance of these genera, but there are individual species that do present a significant difference between the populations of cases and controls, such as *Bacteroides cellulosilyticus, Bacteroides unclassified, Bacteroides uniformis* and *Bacteroides vulgatus*. These observations serve to corroborate that our model has given importance to species that have already been reported with a relationship to T1D.

In Davis-Richardson et al. (2014) a study of the early microbiota of 76 Finnish children at high genetic risk of T1D from stool samples up to 2 years of age was carried out. The results of that study showed that the populations of *Bacteroides dorei* and *Bacteroides vulgatus* were significantly higher in cases than in controls. These two species are found in the metagenomic signature reported in this work. Furthermore, both species are positioned with great importance in the creation of the model, specifically in the second and fourth position, respectively. In Fig. 5 is shown that the abundance of *Bacteroides dorei* is higher in the controls than in the cases, whereas *Bacteroides vulgatus* is significantly higher in the cases. Even so, the positions they occupy after analysing the importance of the model indicate the possible role they may play in the development of the disease.

In Amar et al. (2011) and Tonucci et al. (2017) authors proposed the use of species belonging to the genus *bifidobacterium* to act as a probiotic against diabetes, by controlling bacterial translocation and blood glucose levels. Among the species used, *Bifidobacterium animalis*, being evidently more abundant in specimens whose glycaemia values were more beneficial, in our study the populations of this species are higher in individuals belonging to controls, so they can play an important role against the triggering of the disease.

Reviewing the proposed metagenomic signature, a relationship was observed in terms of the function of *Bacteroides uniformis, Bacteroides dorei, Prevotella copri, Bacteroides vulgatus* and *Bacteroides thetaiotaomicro* with metabolic activities that can be related to T1D. The databases of the Kyoto Encyclopedia at Genes and Genomes (KEGG) and the National Center for Biotechnology Information (NCBI) were consulted. Host and host shared metabolic pathways were detected as the insulin resistance metabolic pathway, the glucagon signalling pathway, glycolysis, gluconeogenesis, the insulin signalling pathway and even in the Type I and Type II diabetes pathways themselves. All these metabolic pathways are involved in the regulation of blood glucose, both its introduction into cells and its metabolism, degradation and storage. These observations show how certain species of gastrointestinal microbiota can play a crucial role in the development of diabetes and be an important source of knowledge for the diagnosis of patients.

Therefore, a stable model has been obtained that achieves a very precise performance. Furthermore, the reported metagenomic signature is capable of differentiating intermediate stages such as seroconversion. The results shown in this work show a starting point in the use of ML models for the diagnosis and prediction of microbiota-related diseases, such as T1D.

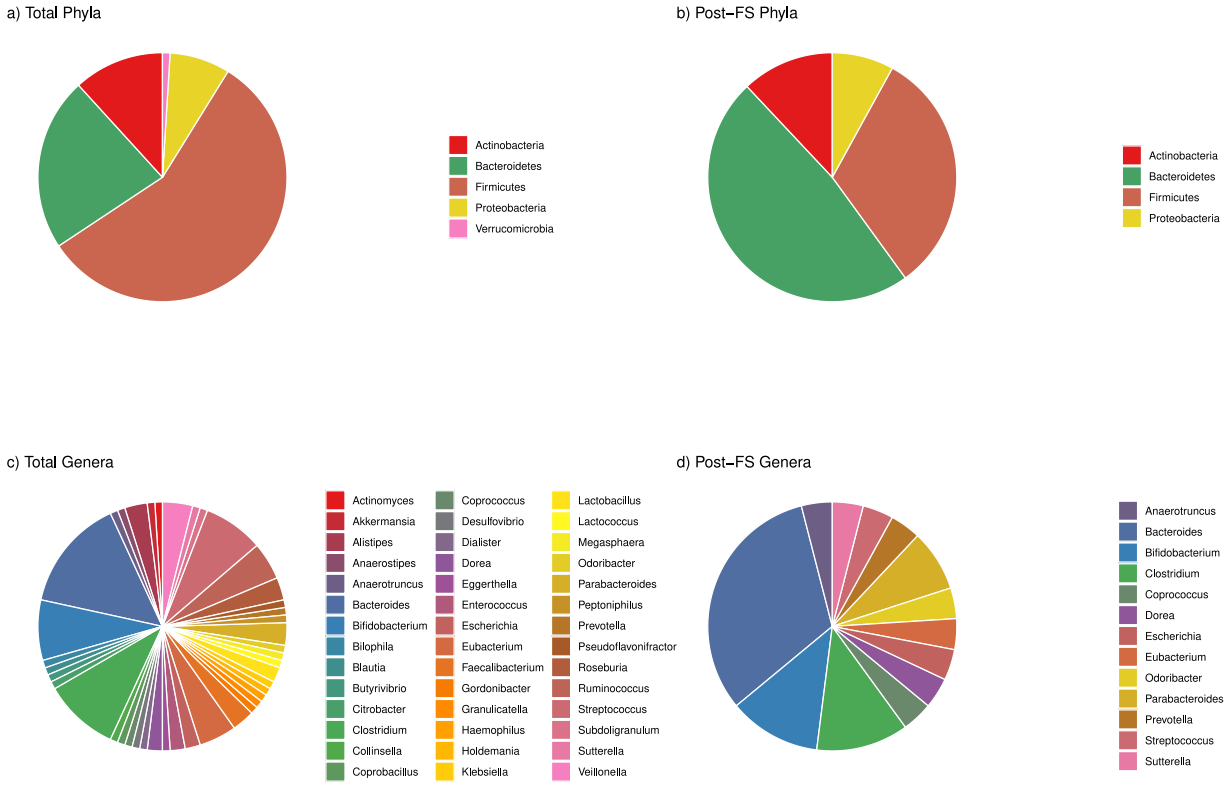

Fig. 6. Comparison of the abundance of phyla and genera in the defined metagenomic signature.

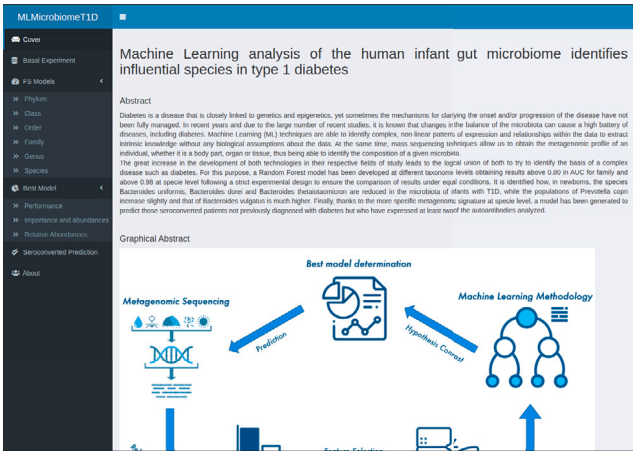

Fig. 7. Screenshot of the developed Shiny application. The application offers the possibility of deepening and observing the results in an interactive way.

Finally, all the results of the tests carried out are available in an interactive Shiny (Chang, Cheng, Allaire, Xie, & McPherson, 2020) app, as shown in Fig. 7 for in-depth analysis and external validation. Analysis code is also available.

5. Conclusions

A metagenomic signature has been obtained that is representative of the influence at the species level on the microbiome of infants with T1D, and after reviewing the literature we have found scientific evidence that confirms that most species have a biological relationship with T1D. In addition, some have been reported to be unrelated, so further study of them should be carried out for the extraction of new knowledge about the disease.

Due to the diversity of other external cohorts and the heterogeneity of their samples, validation of the model has not been possible. Therefore, this study has limitations in terms of its generalization, so it is specific to this cohort with such specific characteristics. Therefore, given the promising results in this pilot study, this methodology should be applied to much larger cohorts for possible transfer to actual clinical practice. It is clear that the methodology that has been carried out works for this type of data and this type of problem. Furthermore, the code of the analyses, the data and the results are available and a Shiny has been published which allows interaction with the experiments carried out and the results obtained.

CRediT authorship contribution statement

Diego Fernández-Edreira: Data curation, Writing – original draft, Software. **Jose Liñares-Blanco:** Conceptualization, Data curation, Methodology, Writing – original draft. **Carlos Fernandez-Lozano:** Conceptualization, Methodology, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability and reproducibility

The source code of the analysis and the Shiny server is available on GitHub: https://github.com/cafernandezlo/MLMicrobiomeT1D. The data used in this study can be downloaded from the DIABIMMUNE project https://diabimmune.broadinstitute.org/diabimmune/

t1d-cohort/resources/16s-sequence-data. The Docker image can be obtained from Docker Hub: https://hub.docker.com/r/cafernandezlo/mlmicrobiomet1d. Therefore, MLMicrobiomeT1D is available as Docker (Merkel, 2014) image which can be downloaded from Docker Hub to interactively explore the results of the analysis locally.

Acknowledgements

This work was supported by the "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe". and the General Directorate of Culture, Education and University Management of Xunta de Galicia, Spain (Ref. ED431D 2017/16), the "Galician Network for Colorectal Cancer Research, Spain" (Ref. ED431D 2017/23) and Competitive Reference Groups, Spain (Ref. ED431C 2018/49). The funding body did not have a role in the experimental design; data collection, analysis and interpretation; and writing of this manuscript. CITIC, as Research Center accredited by Galician University System, is funded by "Consellería de Cultura, Educación e Universidades from Xunta de Galicia, Spain", supported in an 80% through ERDF Funds, Spain, ERDF Operational Programme Galicia 2014–2020, and the remaining 20% by "Secretaría Xeral de Universidades, Spain" (Grant ED431G 2019/01). The funding body did not have a role in the experimental design; data collection, analysis and interpretation; and writing of this manuscript. The calculations were performed on resources provided by the Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER). Funding for open access charge: Universidade da Coruña/CISUG.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2021.115648.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alkanani, A. K., Hara, N., Gottlieb, P. A., Ir, D., Robertson, C. E., Wagner, B. D., et al. (2015). Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. *Diabetes*, *64*(10), 3510–3520.

Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT press.

Amar, J., Chabo, C., Waget, A., Klopp, P., Vachoux, C., Bermúdez-Humarán, L. G., et al. (2011). Intestinal mucosal adherence and translocation of commensal bacteria at the early onset of type 2 diabetes: molecular mechanisms and probiotic treatment. *EMBO Molecular Medicine*, *3*(9), 559–572.

Belkaid, Y., & Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, *157*(1), 121–141.

Biassoni, R., Di Marco, E., Squillario, M., Barla, A., Piccolo, G., Ugolotti, E., et al. (2020). Gut microbiota in T1DM-onset pediatric patients: Machine-learning algorithms to classify microorganisms as disease linked. *The Journal of Clinical Endocrinology & Metabolism*, *105*(9).

Bibbò, S., Dore, M. P., Pes, G. M., Delitala, G., & Delitala, A. P. (2017). Is there a role for gut microbiota in type 1 diabetes pathogenesis? *Annals of Medicine*, *49*(1), 11–22.

Boldison, J., & Wong, F. S. (2016). Immune and pancreatic β cell interactions in type 1 diabetes. *Trends in Endocrinology & Metabolism*, *27*(12), 856–867.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, C. T., Davis-Richardson, A. G., Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., et al. (2011). Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS One*, *6*(10), Article e25792.

Cano, P. G., Santacruz, A., Moya, A., & Sanz, Y. (2012). Bacteroides uniformis CECT 7771 ameliorates metabolic and immunological dysfunction in mice with high-fat-diet induced obesity. *PLoS One*, *7*(7), Article e41079.

Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., et al. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, *20*(1), 4.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). Shiny: Web application framework for R. R package version 1.5.0.

Chen, C., Huang, G., Xia, L., Lin, Y., Jian, L., Ping, J., et al. (2013). Change of glutamic acid decarboxylase antibody and protein tyrosine phosphatase antibody in chinese patients with acute-onset type 1 diabetes mellitus. *Chinese Medical Journal*, *126*(21), 4006–4012.

Choubin, B., Mosavi, A., Alamdarloo, E. H., Hosseini, F. S., Shamshirband, S., Dashtekian, K., et al. (2019). Earth fissure hazard prediction using machine learning models. *Environmental Research*, *179*, Article 108770.

Choubin, B., & Rahmati, O. (2021). 20 - groundwater potential mapping using hybridization of simulated annealing and random forest. In P. Samui, H. Bonakdari, & R. Deo (Eds.), *Water Engineering Modeling and Mathematic Tools* (pp. 391–403). Elsevier.

Corrigan, A., Russell, N., Welge, M., Auvil, L., Bushell, C., White, B. A., et al. (2018). The use of random forests modelling to detect yeast-mannan sensitive bacterial changes in the broiler cecum. *Scientific Reports*, *8*(1), 13270.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., et al. (2014). Bacteroides dorei dominates gut microbiome prior to autoimmunity in finnish children at high risk for type 1 diabetes. *Frontiers in Microbiology*, *5*, 678.

Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine learning techniques for prediction of early childhood obesity. *Applied Clinical Informatics*, *6*(3), 506.

Dunne, J., Triplett, E., Gevers, D., Xavier, R., Insel, R., Danska, J., et al. (2014). The intestinal microbiome in type 1 diabetes. *Clinical & Experimental Immunology*, *177*(1), 30–37.

Durazzo, M., Ferro, A., & Gruden, G. (2019). Gastrointestinal microbiota and type 1 diabetes mellitus: The state of art. *Journal of Clinical Medicine*, *8*(11), 1843.

Everard, A., & Cani, P. D. (2013). Diabetes, obesity and gut microbiota. *Best Practice & Research Clinical Gastroenterology*, *27*(1), 73–83.

Favier, C. F., Vaughan, E. E., De Vos, W. M., & Akkermans, A. D. (2002). Molecular monitoring of succession of bacterial communities in human neonates. *Applied and Environmental Microbiology*, *68*(1), 219–226.

Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J., & Pazos, A. (2016). A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ*, *4*, Article e2721.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., et al. (2020). Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. *Journal of Clinical Medicine*, *9*(8), 2403.

Garrett, W. S. (2015). Cancer and the microbiota. *Science*, *348*(6230), 80–86.

Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., et al. (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *The ISME Journal*, *5*(1), 82–91.

Gülden, E., Wong, F. S., & Wen, L. (2015). The gut microbiota and type 1 diabetes. *Clinical Immunology*, *159*(2), 143–153.

Guvenir, H. A., Acar, B., Demiroz, G., & Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997* (pp. 433–436). IEEE.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Harbison, J. E., Roth-Schulze, A. J., Giles, L. C., Tran, C. D., Ngui, K. M., Penno, M. A., et al. (2019). Gut microbiome dysbiosis and increased intestinal permeability in children with islet autoimmunity and type 1 diabetes: A prospective cohort study. *Pediatric Diabetes*, *20*(5), 574–583.

Harrington, P. (2012). *Machine Learning in Action*. Manning Publications Co..

Harsch, I. A., & Konturek, P. C. (2018). The role of gut microbiota in obesity and type 2 and type 1 diabetes mellitus: new insights into "old" diseases. *Medical Sciences*, *6*(2), 32.

Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., et al. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences*, *108*(7), 3047–3052.

Higuchi, B. S., Rodrigues, N., Gonzaga, M. I., Paiolo, J. a. C. C., Stefanutto, N., Omori, W. P., et al. (2018). Intestinal dysbiosis in autoimmune diabetes is correlated with poor glycemic control and increased interleukin-6: a pilot study. *Frontiers in Immunology*, *9*, 1689.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., et al. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *49*(7), 2401–2414.

Hooper, L. V., & Gordon, J. I. (2001). Commensal host-bacterial relationships in the gut. *Science*, *292*(5519), 1115–1118.

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., et al. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, *155*(7), 1451–1463.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics*, *15*(1), 41–51.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, *498*(7452), 99–103.

Kasselman, L. J., Vernice, N. A., DeLeon, J., & Reiss, A. B. (2018). The gut microbiome and elevated cardiovascular risk in obesity and autoimmunity. *Atherosclerosis*, *271*, 203–213.

Kaufman, D., Erlander, M., Clare-Salzler, M., Atkinson, M., Maclaren, N., Tobin, A., et al. (1992). Autoimmunity to two forms of glutamate decarboxylase in insulin-dependent diabetes mellitus.. *The Journal of Clinical Investigation*, *89*(1), 283–292.

Knip, M., & Siljander, H. (2016). The role of the intestinal microbiota in type 1 diabetes mellitus. *Nature Reviews Endocrinology*, *12*(3), 154.

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe*, *17*(2), 260–273.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26.

LaFreniere, D., Zulkernine, F., Barber, D., & Martin, K. (2016). Using machine learning to predict hypertension from a clinical dataset. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–7). IEEE.

Larsen, N., Vogensen, F. K., Van Den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., et al. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One*, *5*(2), Article e9085.

Leite, A. Z., Rodrigues, N. d. C., Gonzaga, M. I., Paiolo, J. a. C. C., de Souza, C. A., Stefanutto, N. A. V., et al. (2017). Detection of increased plasma interleukin-6 levels and prevalence of prevotella copri and bacteroides vulgatus in the feces of type 2 diabetes patients. *Frontiers in Immunology*, *8*, 1107.

Liñares-Blanco, J., Porto-Pazos, A. B., Munteanu, C., Pazos, A., & Fernandez-Lozano, C. (2018). Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. *Scientific Reports*, *8*(15688), 1–11.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early diagnosis of alzheimer's disease with deep learning. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (pp. 1015–1018). IEEE.

Liu, H., & Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining, Vol. 454*. Springer Science & Business Media.

Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., et al. (2017). Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metabolism*, *25*(5), 1054–1062.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, *489*(7415), 220–230.

Ludwig, N., Fehlmann, T., Kern, F., Gogol, M., Maetzler, W., Deutscher, S., et al. (2019). Machine learning to detect alzheimer's disease from circulating non-coding RNAs. *Genomics, Proteomics & Bioinformatics*, *17*(4), 430–440.

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective*. CRC press.

Medina-Vera, I., Sanchez-Tapia, M., Noriega-Lopez, L., Granados-Portillo, O., Guevara-Cruz, M., Flores-López, A., et al. (2019). A dietary intervention with functional foods reduces metabolic endotoxaemia and attenuates biochemical abnormalities by modifying faecal microbiota in people with type 2 diabetes. *Diabetes & Metabolism*, *45*(2), 122–131.

Mejia-Leon, M., Petrosino, J., Ajami, N., Domínguez-Bello, M., & Calderon de la Barca, A. (2014). HLA DQ/DR prevalence and microbiota disturbance in northwestern mexican children with type 1 diabetes (1118.3). *FASEB Journal*, *28*, 1118.3.

Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, *28*(S37), S5–S10.

Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*, *10*(1), 1–17.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT press.

Mosavi, A., Hosseini, F. S., Choubin, B., Abdolshahnejad, M., Gharechaee, H., Lahijanzadeh, A., et al. (2020). Susceptibility prediction of groundwater hardness using ensemble machine learning models. *Water*, [ISSN: 2073-4441] *12*(10), http://dx.doi.org/10.3390/w12102770.

Mosavi, A., Hosseini, F. S., Choubin, B., Goodarzi, M., Dineva, A. A., & Sardooi, E. R. (2021). Ensemble boosting and bagging based machine learning models for groundwater potential prediction. *Water Resources Management: An International Journal, Published for the European Water Resources Association (EWRA)*, *35*(1), 23–37.

Murri, M., Leiva, I., Gomez-Zumaquero, J. M., Tinahones, F. J., Cardona, F., Soriguer, F., et al. (2013). Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Medicine*, *11*(1), 46.

Norris, J. M., Barriga, K., Klingensmith, G., Hoffman, M., Eisenbarth, G. S., Erlich, H. A., et al. (2003). Timing of initial cereal exposure in infancy and risk of islet autoimmunity. *Jama*, *290*(13), 1713–1720.

Oh, T. G., Kim, S. M., Caussy, C., Fu, T., Guo, J., Bassirian, S., et al. (2020). A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metabolism*.

Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine*, *41*(5), 265–271.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Computational Biology*, *12*(7), Article e1004977.

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, *535*(7612), 376–381.

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., et al. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics*, *118*(2), 511–521.

Petersen, C., & Round, J. L. (2014). Defining dysbiosis and its influence on host immunity and disease. *Cellular Microbiology*, *16*(7), 1024–1033.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55–60.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., et al. (2019). Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. *Science of the Total Environment*, *688*, 855–866.

Reitmeier, S., Kiessling, S., Clavel, T., List, M., Almeida, E. L., Ghosh, T. S., et al. (2020). Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. *Cell Host & Microbe*, *28*(2), 258–272.

Roguet, A., Eren, A. M., Newton, R. J., & McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome*, *6*(1), 185.

Saeys, Y., Inza, I. n., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Sánchez, E., Donat, E., Ribes-Koninckx, C., Calabuig, M., & Sanz, Y. (2010). Intestinal bacteroides species associated with coeliac disease. *Journal of Clinical Pathology*, *63*(12), 1105–1111.

Schwiertz, A., Gruhl, B., Löbnitz, M., Michel, P., Radke, M., & Blaut, M. (2003). Development of the intestinal bacterial composition in hospitalized preterm infants in comparison with breast-fed, full-term infants. *Pediatric Research*, *54*(3), 393–399.

Stark, P. L., & Lee, A. (1982). The microbial ecology of the large bowel of breastfed and formula-fed infants during the first year of life. *Journal of Medical Microbiology*, *15*(2), 189–203.

Stene, L., & Rewers, M. (2012). Immunology in the clinic review series; focus on type 1 diabetes and viruses: the enterovirus link to type 1 diabetes: critical review of human studies. *Clinical & Experimental Immunology*, *168*(1), 12–23.

Stenman, L., Waget, A., Garret, C., Klopp, P., Burcelin, R., & Lahtinen, S. (2014). Potential probiotic bifidobacterium animalis ssp. lactis 420 prevents weight gain and glucose intolerance in diet-induced obese mice. *Beneficial Microbes*, *5*(4), 437–445.

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, *25*(4), 667–678.

Thompson, J., Johansen, R., Dunbar, J., & Munsky, B. (2019). Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS One*, *14*(7), Article e0215502.

Tonucci, L. B., Dos Santos, K. M. O., de Oliveira, L. L., Ribeiro, S. M. R., & Martino, H. S. D. (2017). Clinical application of probiotics in type 2 diabetes mellitus: A randomized, double-blind, placebo-controlled study. *Clinical Nutrition*, *36*(1), 85–92.

Vaarala, O., Atkinson, M. A., & Neu, J. (2008). The "perfect storm" for type 1 diabetes: the complex interplay between intestinal microbiota, gut permeability, and mucosal immunity. *Diabetes*, *57*(10), 2555–2562.

Venema, K. (2010). Role of gut microbiota in the control of energy and carbohydrate metabolism. *Current Opinion in Clinical Nutrition & Metabolic Care*, *13*(4), 432–438.

Wahlberg, J., Vaarala, O., Ludvigsson, J., Group, A.-S., et al. (2006). Dietary risk factors for the emergence of type 1 diabetes-related autoantibodies in 21/2-year-old Swedish children. *British Journal of Nutrition*, *95*(3), 603–608.

Wang, X.-W., & Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. *Medicine in Microecology*, *4*, Article 100013.

Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature*, *455*(7216), 1109–1113.

Wensinck, F., & Van de Merwe, J. (1981). Serum agglutinins to eubacterium and peptostreptococcus species in crohn's and other diseases. *Epidemiology & Infection*, *87*(1), 13–24.

Wingfield, B., Coleman, S., McGinnity, T. M., & Bjourson, A. (2018). Robust microbial markers for non-invasive inflammatory bowel disease identification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(6), 2078–2088.

Wu, Y., Bible, P. W., Long, S., Ming, W.-k., Ding, W., Long, Y., et al. (2019). Metagenomic analysis reveals gestational diabetes mellitus-related microbial regulators of glucose tolerance. *Acta Diabetologica*, 1–13.

Yi, B., Huang, G., & Zhou, Z.-G. (2015). Current and future clinical applications of zinc transporter-8 in type 1 diabetes mellitus. *Chinese Medical Journal*, *128*(17), 2387.

Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Medical & Biological Engineering & Computing*, *57*(4), 901–912.

Zhou, H., Sun, L., Zhang, S., Zhao, X., Gang, X., & Wang, G. (2020). Evaluating the causal role of gut microbiota in type 1 diabetes and its possible pathogenic mechanisms. *Frontiers in Endocrinology*, *11*, 125.

Ziegler, A.-G., Schmid, S., Huber, D., Hummel, M., & Bonifacio, E. (2003). Early infant feeding and risk of developing type 1 diabetes–associated autoantibodies. *Jama*, *290*(13), 1721–1728.

Diego Fernández-Edreira is a M.Sc. bioinformatician at University of A Coruña. Interested in computer science and multi-omics data analysis using Machine Learning.

Jose Liñares-Blanco is a predoc student undertaking a PhD at the University of A Coruña and affiliated researcher of the Centre of Information and Communications Technology Research (CITIC) studying computer science, genetics and machine learning with a particular focus on Cancer.

Carlos Fernandez-Lozano is Assistant Professor in Bioinformatics and Intelligent Systems at the University of A Coruña and affiliated researcher of the Institute of Biomedical Research of A Coruña (INIBIC) and of the Centre for Information and Communications Technology Research (CITIC). His primary research interest is the understanding of the behaviour and biological mechanisms of complex system dynamics using machine learning and the molecular pathways and connections underpinning Cancer Disease.