Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation

Analytica Chimica Acta 1183 (2021) 339001

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation

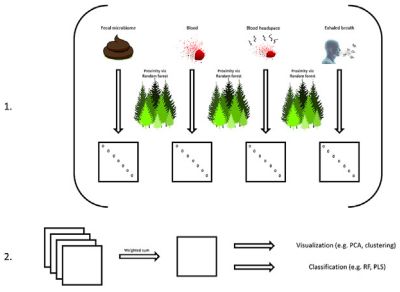
Georgios Stavropoulos [a], Robert van Vorstenbosch [a], Daisy M.A.E. Jonkers [b], John Penders [c], Jane E. Hill [d], Frederik-Jan van Schooten [a], Agnieszka Smolinska [a, *]

[a] Department of Pharmacology and Toxicology, NUTRIM School of Nutrition and Translational Research, Maastricht University, Maastricht, the Netherlands
[b] Division of Gastroenterology and Hepatology, NUTRIM School of Nutrition and Translational Research, Maastricht University, Maastricht, the Netherlands
[c] Department of Medical Microbiology, NUTRIM School of Nutrition and Translational Research, Maastricht University, Maastricht, the Netherlands
[d] Department of Chemical and Biological Engineering, School of Biomedical Engineering, The University of British Columbia, Vancouver, Canada

HIGHLIGHTS

- Random forest proximities of various data platforms can be fused via a weighted sum to increase prediction accuracy in complex biological data.
- The problem of variable interpretation and examination when working with proximities or kernels is tackled by implementing the pseudo-sample principle.
- Random forest proximities fusion can outperform the traditional ways of fusion as well as demonstrate the contribution of every platform in the outcome.
- The pseudo-sample principle allows for identification of relations among variables from different data platforms.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 5 November 2020
Received in revised form
24 August 2021
Accepted 25 August 2021
Available online 28 August 2021

Keywords:
Data fusion
Proximities
Stacking

ABSTRACT

Data fusion has gained much attention in the field of life sciences, and this is because analysis of biological samples may require the use of data coming from multiple complementary sources to express the samples fully. Data fusion lies in the idea that different data platforms detect different biological entities. Therefore, if these different biological compounds are then combined, they can provide comprehensive profiling and understanding of the research question in hand. Data fusion can be performed in three different traditional ways: low-level, mid-level, and high-level data fusion. However, the increasing complexity and amount of generated data require the development of more sophisticated fusion approaches. In that regard, the current study presents an advanced data fusion approach (i.e. proximities stacking) based on random forest proximities coupled with the pseudo-sample principle. Four different data platforms of 130 samples each (faecal microbiome, blood, blood headspace, and exhaled breath

Abbreviations: RF, random forest: CD; Crohn's disease: OTUs, operation taxonomic units: VOCs; volatile organic compounds: PCA, principal component analysis: OOB; out of bag: ROC, receiver operating characteristic.

* Corresponding author.
E-mail address: a.smolinska@maastrichtuniversity.nl (A. Smolinska).

https://doi.org/10.1016/j.aca.2021.339001
0003-2670/

Variable behaviour
Crohn's disease
Classification

samples of patients who have Crohn's disease) were used to demonstrate the classification performance of this new approach. More specifically, 104 samples were used to train and validate the models, whereas the remaining 26 samples were used to validate the models externally. Mid-level, high-level, as well as individual platform classification predictions, were made and compared against the proximities stacking approach. The performance of each approach was assessed by calculating the sensitivity and specificity of each model for the external test set, and visualized by performing principal component analysis on the proximity matrices of the training samples to then, subsequently, project the test samples onto that space. The implementation of pseudo-samples allowed for the identification of the most important variables per platform, finding relations among variables of the different data platforms, and the examination of how variables behave in the samples. The proximities stacking approach outperforms both mid-level and high-level fusion approaches, as well as all individual platform predictions. Concurrently, it tackles significant bottlenecks of the traditional ways of fusion and of another advanced fusion way discussed in the paper, and finally, it contradicts the general belief that the more data, the merrier the result, and therefore, considerations have to be taken into account before any data fusion analysis is conducted.

1. Introduction

Data fusion has gained much attention in the field of, among others, life sciences [1–10], and this is because analysis of biological samples may require the use of data coming from multiple complementary sources to express the samples fully. The principle behind data fusion lies in the idea that different data platforms, such as gas chromatography-mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR) detect different biological entities. Therefore, if these different biological compounds are then combined, they can provide comprehensive profiling and understanding of the research question in hand [2]. Theoretically, one would imagine that the more data generated per biological sample, the merrier since different data platforms demonstrate different strengths. Practically, this is not always the case; considerations have to be made regarding the research question, and the nature of the samples before any data fusion analysis is conducted. Data fusion can be performed in three different ways: low-level, mid-level, and high-level data fusion [5]. At the low-level, the various data platforms are fused at a data level, whereas in the mid-level, the platforms are fused at a data level of selected variables or features of the original data. At the high-level, the platforms are fused at a prediction level, meaning that each platform gives predictions individually and then, these individual predictions are combined to get the final prediction.

Recently, a more sophisticated way of data fusion was introduced that can also be seen as a modified version of mid-level fusion [1]. Smolinska et al. introduced the fusion of kernels of the individual platforms rather than the important variables, features or latent variables of the platforms. More specifically, they mapped each platform to a higher-dimensional feature space with the use of a kernel function, and they then fused all the individual kernels by using a weighted sum. Kernel functions transform the data in such a way that they result in non-negative square matrices, and these matrices can be seen as measures of similarity/dissimilarity of samples; therefore, when one works with kernels, they work with samples rather than variables. This approach holds great potential when it comes to unravelling trends in data or getting predictions of data since it considers both linear and nonlinear relations amongst data, and most of the biological systems reveal nonlinear characteristics [1]. Another advantage of working with kernels, and therefore samples, rather than variables/features is that scaling issues are overcome. For example, in a mid-level fusion approach, scaling of the original variables is required before any data from different sources are concatenated since the magnitude of the data coming from different sources is most likely different. To find the optimal scaling parameter that would suit all the data might be not an easy task to perform, and on top of that, if the data being concatenated are of different type (i.e. quantitative or discrete), then this issue gets even more challenging. The major disadvantage of working with kernels is that information about the importance/contribution of variables of the dataset in the model performance is lost due to the transformation of variables to distance or similarity measures among samples, and it can be challenging to trace back these variables. Nonlinear bi-plots introduced by Gower at al. have been further modified and developed the idea of pseudo-samples by Krooshof et al. [11,12] and Smolinska et al. [13], to overcome this bottleneck. The pseudo-sample principle uses the transformed data (i.e. the square matrices) to illustrate not only the importance of the original variables but also the original variable trajectory (i.e. how the variables behave amount-wise) in the samples of interest, which are both essential assets when it comes to drawing safe conclusions on the study results.

Proximity matrices are actual measures of similarity/dissimilarity of samples, and they are non-negative square matrices [14]. Originally, the term proximity means “closeness” or “nearness” between pairs, and it is calculated by using traditional distance measures such as Euclidean distance or Gaussian distance. The closer to zero the proximity of two samples is, the more similar these two samples are; this is why the diagonal of a proximity matrix always consists of zeros. The square matrix has a size of $n \times n$ (where n is the number of samples in the original dataset) since proximities imply similarities amongst samples. Moreover, proximities do not consist of transformed data, which is the case with kernels (e.g. the original dataset is transformed using the radial basis function), but instead of newly generated data (i.e. distances in space among samples). Random forest (RF) also returns a proximity matrix of the data that it is run on; although, the proximity matrix here is calculated differently [15]. The RF proximity matrix is indicative of the number of times that samples ended up in the same terminal node rather than a demonstration of the actual distance in the space of samples. More details on how the proximity via RF is calculated are shown in the materials and methods section. Recently, Blanchet et al. [16] published a tutorial where they illustrate the successful implementation of the RF proximities along with the pseudo-sample principle to visualize variable importance. However, to the best of the authors' knowledge, proximity matrices, and mainly RF proximity matrices, have not been examined before in terms of data fusion to check their performance on predicting and investigating complex biological samples.

In this research, Crohn's disease (CD) serves as a case study to demonstrate the utility of data fusion using proximities. CD is a complex biological metabolic disorder. CD is a chronic inflammatory process with no known cause (idiopathic) that can affect any part of the gastrointestinal tract, from the mouth to the anus [17]. More specifically, CD causes muscle hypertrophy, it changes the colon to a cobblestone appearance, it creates fissures in the colon, and it also covers the colon with fat. Colonoscopy has been the gold standard to diagnose and monitor the disease activity; therefore, alternative ways (e.g. biological biomarkers) to diagnose and monitor the disease activity are needed since colonoscopy is a considerably invasive and costly technique. Previous research focused on identifying CD biomarkers in either human blood (i.e. metabolites) or faeces (i.e. bacterial species) [18–21] to diagnose and monitor the disease activity. All studies demonstrated promising results as far as prediction accuracy is concerned; although, each of these studies examined one data platform only to draw their conclusions. Consequently, the aim of the present study is to propose a new, advanced fusion approach based on RF proximities, as well as to see whether prediction accuracy of CD can be increased if more data are concatenated along with potential biomarker behaviour examination using pseudo-sample principle. To illustrate that this new fusion approach performs well, it is advantageous over the currently existing fusion approaches, and that it can be implemented in biomedical data, it is compared against the current ways of data fusion and the individual platforms used in the present study.

2. Materials and methods

2.1. Data used and data preprocessing

Four different data platforms were used: faecal microbiome, blood, blood headspace, and exhaled breath samples from patients suffering from CD. The CD patients were categorized into two classes based on the disease activity: remission and active cases of CD. The criteria used to classify the patients as being in either remission or active stage can be found elsewhere [19]. In the present study, 130 CD patients were sampled, of which 66 were patients in the remission stage of the disease, and the remaining 64 were patients in the active stage of the disease. Initially, all the raw data were preprocessed before the actual analysis took place. Data preprocessing diminishes the effect of possible instrumental artefacts that can occur during the analysis. Each data platform followed a different preprocessing strategy.

The faecal microbiome samples were treated and sampled as described elsewhere [18], and they were analysed by employing 16S ribosomal RNA pyrosequencing. The faecal microbiome was analysed in terms of operational taxonomic units (OTUs). The raw microbiome pyrosequencing reads were, first, preprocessed by means of quality filters to reduce the error rate, and de-multiplexed and clustered into OTUs based on a 97% similarity—the entire prepossessing procedure that was followed is described elsewhere [18].Then, they were transformed into continuous data. This is because preprocessing of the pyrosequencing reads results in data counts (i.e. OTUs per sample) which cannot be used for multivariate analysis purposes; the transformation was done by employing the inverse hyperbolic sine [22]. Next, the exclusion of zeros followed. The majority of bacterial species (OTUs) is not present in all the samples; consequently, only those that are present in a specified per cent of the samples are kept. Here, species that were found in at least 35% [18] of the samples were retained. As a final preprocessing step, microbiome data were logarithmically transformed since the log transformation accounts for high skewness in the data.

The blood was treated and sampled as described elsewhere [23], and the blood sample metabolites were analysed by using NMR Bruker 600 MHz with a cryoprobe. In the blood NMR data, first, the water peak was removed, and then, baseline correction via P-splines [24], misalignment correction via correlation optimized warping [25], and peak picking in the form of binning via adaptive intelligent binning [26] were performed. Moreover, normalization via a reference peak (i.e. trimethylsilyl-propanoic acid—TSP) as well as via probabilistic quotient normalization [27] followed. Normalization via the TSP peak is done to enhance the signal comparison among the samples, whereas probabilistic quotient normalization accounts for dilution effects, effect size, among the samples. Finally, the blood data were logarithmically transformed.

The blood headspace was treated and sampled as described elsewhere [28]; in short, the blood headspace samples were measured by utilizing gas chromatography/gas chromatography-*time-of-flight*-mass spectrometry (GC × GC-*tof*-MS; Pegasus 4D, LECO Corporation, St Joseph, MI, USA). Blood headspace was analysed in terms of volatile organic compounds (VOCs). The blood headspace data were initially preprocessed as discussed elsewhere [28], and in the end, the exclusion of zeros followed. As with the microbiome data, the majority of VOCs does not occur in all the samples; therefore, only those found present in at least 20% [29] of the samples coming from the same class were kept for further analysis. In the end, a logarithmic transformation was performed.

Finally, the exhaled breath was captured as described elsewhere [19], and the exhaled breath samples were analysed by using GC-*tof*-MS. Breath was analysed in terms of VOCs as well. The exhaled breath data were preprocessed as described elsewhere [19], and as an extra preprocessing step, these data underwent exclusion of zeros (compounds found in at least 20% of each class [29] of the samples were retained) and logarithmic transformation.

2.2. Data fusion approaches

Data platforms can traditionally be fused at three different levels: low-level, mid-level, and high-level data fusion [5]. Low-level data fusion refers to concatenation of the whole data platforms, sample-wise, into a single matrix that consists of as many rows as the number of samples, and as many columns as the total number of variables from all different data platforms. Low-level fusion attempts were not tried here because this would affect the degrees of freedom of the data, and thus, making the concatenated matrix challenging to deal with and the analysis results untrustworthy; readers interested in low-level fusion applications are referred to Ref. [5].

2.2.1. Mid-level fusion

Mid-level data fusion can be divided into two categories: the concatenation of either important/significant variables or features of the different platforms. A variety of ways exists to find important variables or features. For example, variables can be found by using, among others, RF [15], partial least squares based variable selection [30], or even significance multivariate correlation [31], whereas features (or latent variable space) can be found by implementing principal component analysis (PCA) (and use the principal components) [32], recursive feature elimination [33], or partial least squares analysis (and use the latent variables) [34,35]. Then, all these variables or features are concatenated, sample-wise, to create the single fused matrix to be used for further analysis. In the present study, RF was used to find the most important variables per platform. A schematic representation of the mid-level fusion approach is given in Fig. 1.

2.2.2. High-level fusion

High-level data fusion refers to a combination of the outcome of

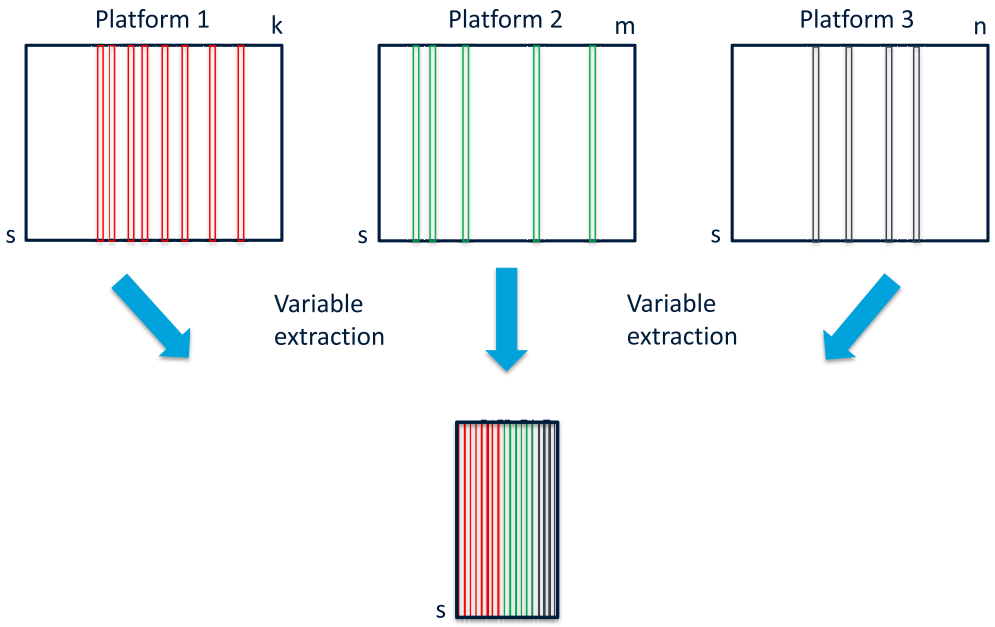


Fig. 1. Schematic representation of the mid-level fusion approach of three datasets. RF is run on each of the datasets to get their most important variables. Then, all the important variables of all three datasets are fused sample-wise to get the final fused matrix.

the individual platforms; this is why it is also called as decision-level fusion. Specifically, a classification or regression model is built for each one of the available data platforms, and the results from each model are combined to obtain the final decision for every sample of interest. The outcome of each model is given as either a class label or a set of probabilities; therefore, one can choose to either use majority voting [36] or adjusted probabilities to get the final decision for the samples of interest. In the current study, adjusted probabilities via the Bayes' theorem [37] were used to get the final decisions, and the optimal decision threshold was found from a loop of 100 cross-validation iterations; in every iteration, the data were randomly split into training and validation sets, and therefore, each iteration used different training and validation samples. Bayes' theorem is also called Bayesian integration because it provides the ability to define probability models for disparate or independent types of data. More specifically, RF was used on every single platform to get the sets of initial likelihood probabilities (i.e. prior probabilities) for every sample of interest, and then, these probabilities were transformed into posterior probabilities. A detailed description of the implementation of the Bayes' theorem in biological data can be found elsewhere [38]. A schematic representation of the high-level fusion approach is shown in Fig. 2.

2.2.3. Proximities stacking

The current study implemented a modified version of mid-level fusion. This approach makes use of proximity matrices ($\boldsymbol{P}_i$) of the original platforms, which are then arranged one on top of each other; consequently, this approach was called as proximities stacking. The proposed approach consists of two steps: the creation of the ($\boldsymbol{P}_i$) matrices of each used data platforms, and the discovery of an optimal set of weights w with which the platforms are combined in a weighted linear parameterized way to create a new single proximity matrix $\boldsymbol{K}$ that is used for further analysis.

$\boldsymbol{P}_i$ matrices are square distance matrices that show how similar or dissimilar the data are. RF returns $\boldsymbol{P}_i$ matrices of the data that the algorithm was run on, and these proximities were used here; however, these $\boldsymbol{P}_i$ matrices are not calculated by using a distance measure [15]. More specifically, for every pair of samples, the proximity indicates the percentage of the times these two samples ended up in the same terminal node. For instance, if the RF consists of 1000 trees and the pair of samples ended up in the same terminal node in 100 of the 1000 trees, then the proximity for this pair of samples is 100/1000 = 0.1. As a result, the higher the proximity, the more similar the objects are. This means that the diagonal of the RF $\boldsymbol{P}_i$ matrix is filled with ones rather than zeros (as an actual proximity matrix); therefore, the RF $\boldsymbol{P}_i$ matrix is subtracted from one to, ultimately, transform it to an actual distance/proximity matrix ($\boldsymbol{P}_{trans,i}$). Equation (1) depicts a toy example of such transformation.

$$\boldsymbol{P}_{trans,i} = 1 - \boldsymbol{P}_i = 1 - \begin{pmatrix} 1 & x_{1,2} & x_{1,3} \\ x_{2,1} & 1 & x_{2,3} \\ x_{3,1} & x_{3,2} & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1-x_{1,2} & 1-x_{1,3} \\ 1-x_{2,1} & 0 & 1-x_{2,3} \\ 1-x_{3,1} & 1-x_{3,2} & 0 \end{pmatrix} \quad (1)$$

Subsequently, the $\boldsymbol{P}_{trans,i}$ matrices of the present study were combined in a weighted linear parameterized combination to create the new single proximity matrix $\boldsymbol{K}$ that was used for further analysis. This linear combination can be expressed as follows:

$$\boldsymbol{K} = \sum_{i=1}^{m} w_i \times \boldsymbol{P}_{trans,i} \quad (2)$$

where m is the total number of $\boldsymbol{P}_{trans,i}$ matrices (here, m equals four), and w_i is the weight or importance of the $\boldsymbol{P}_{trans,i}$ matrix in the new $\boldsymbol{K}$ matrix. The set of weights w can be found by applying regularisation methods such as L_1 or L_2 norm. Regularisation methods are processes that introduce additional information to prevent over-fitting. L_2 norm is applied when the data platforms are complementary to each other because it avoids the possibility of shrinking the importance of any of the platforms; L_2 norm [1] was used here, and it is expressed as follows:

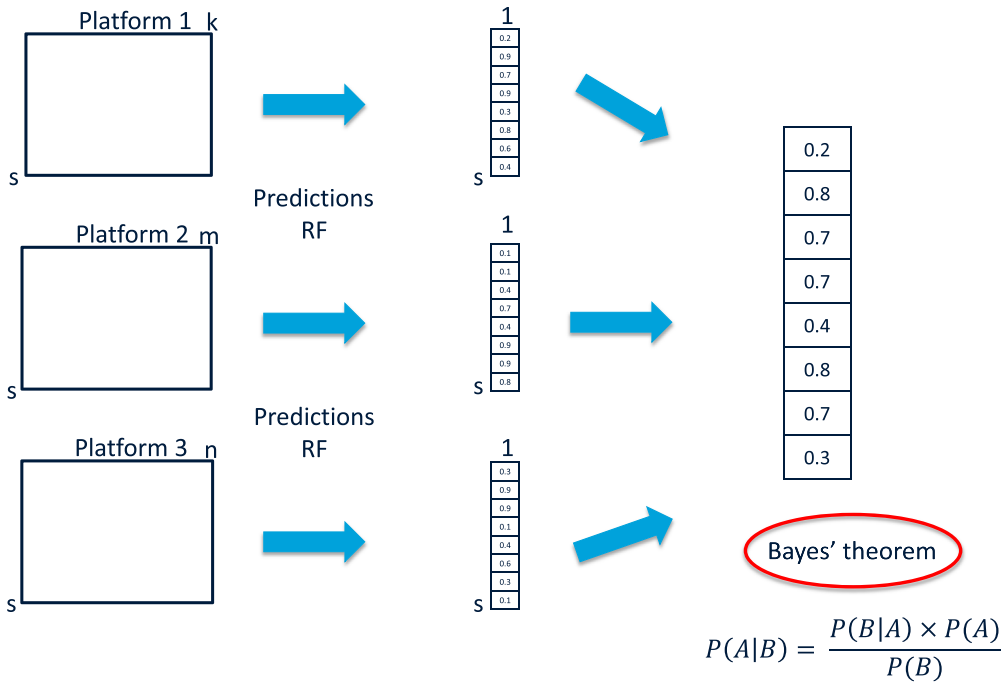


Fig. 2. Schematic representation of the high-level fusion approach of three datasets. RF is run on each of the datasets to get their predictions (i.e. classification probabilities here), which are then adjusted via the Bayes' theorem to get the outcome. The Bayes' theorem formula is depicted at the bottom right corner of the figure, where P(A) and P(B) are the probabilities of observing the events A and B respectively, P(B|A) is the probability of event B occurring given that event A is true, and P(A|B) is the probability of event A occurring given that event B is true.

$$w = \sqrt{\sum_{i=1}^{m} w_i^2} = 1 \tag{3}$$

where m is the total number of $\boldsymbol{P}_{trans,i}$ matrices, and w_i is the weight of the $\boldsymbol{P}_{trans,i}$ matrix.

The optimal set of weights was selected in two steps approach. In the first step, ten sets of numbers that fulfilled the equation (3) were generated via grid search. Then, the weight values of every w_i were shuffled to create a total of 40 different possible combinations of w since there were four data platforms available in this study. The $w_{optimal}$ that maximizes classification accuracy of the model was found from a loop of 100 cross-validation iterations. A schematic representation of the proximities stacking fusion is pictured in Fig. 3.

2.3. Pseudo-sample principle

The pseudo-sample principle was employed to explore the behaviour and importance of the original variables (i.e. bacterial species, metabolites, and VOCs) in the final classification model in the proximities stacking fusion approach [11]. A pseudo-sample is a matrix that has the values of one particular variable from an entire dataset (e.g. $\boldsymbol{A} = (n \times p)$, where n is the number of samples, and p is the number of variables) sorted out in one column, and the rest of its columns are filled in with zeros. For every original variable in the $\boldsymbol{A}$ matrix, a $\boldsymbol{B} = (k \times p)$ pseudo-sample matrix is created, where k is the number of points that one chooses to spread the range of the values of that particular variable on. Based on existing literature [1,11], k usually ranges from 20 to 40, to properly represent the range of the values of each variable—the present study used 40 points. Then, this $\boldsymbol{B}$ matrix is predicted using RF, which results in obtaining its corresponding pseudo-sample proximity matrix. In the end, one gets as many pseudo-sample proximity matrices as the total number of the original variables (i.e. p pseudo-sample proximity matrices to be analysed, in total). A graphical illustration of how a single pseudo-sample proximity matrix is created is shown in Fig. 4. As a final step, principal coordinate analysis (PCoA) is run on the proximity matrix of the original dataset, and subsequently, all the pseudo-sample proximity matrices are projected onto the PCoA space of the proximity of the original dataset since they can be treated as any other subject/patient sample.

2.4. Conceptual flowchart and fusion approach optimization

2.4.1. Variable selection and RF model optimization

Each data platform underwent preprocessing, and then, its samples were divided into the training and validation samples (i.e. 104 samples, of which 57 were remission and 47 were active), and independent internal test set samples (i.e. 26 samples, of which 11 were remission and 15 were active). The division between the training and validation, and the independent internal test samples was achieved by employing the Duplex algorithm [39] since Duplex algorithm aims to maintain a comparable diversity between the sets. The URF/RF model parameters (i.e. number of trees, predictors, and samples per tree terminal leaf per RF model), as well as to the number of variables to be kept per platform were optimized within a 1000-iteration loop—the number of samples per tree terminal leaf accounts for overfitting minimization and model complexity reduction.

For each iteration, the training and validation set samples were randomly split (80% of the 104 samples were used as training samples, and the remaining 20% of the 104 samples were used as validation samples), an RF model was built, and the importance of every variable was found. By default, a variable is considered important if its importance value is positive; however, here, a variable was considered as important if its importance value was equal or higher than 30% of the amount of the highest variable

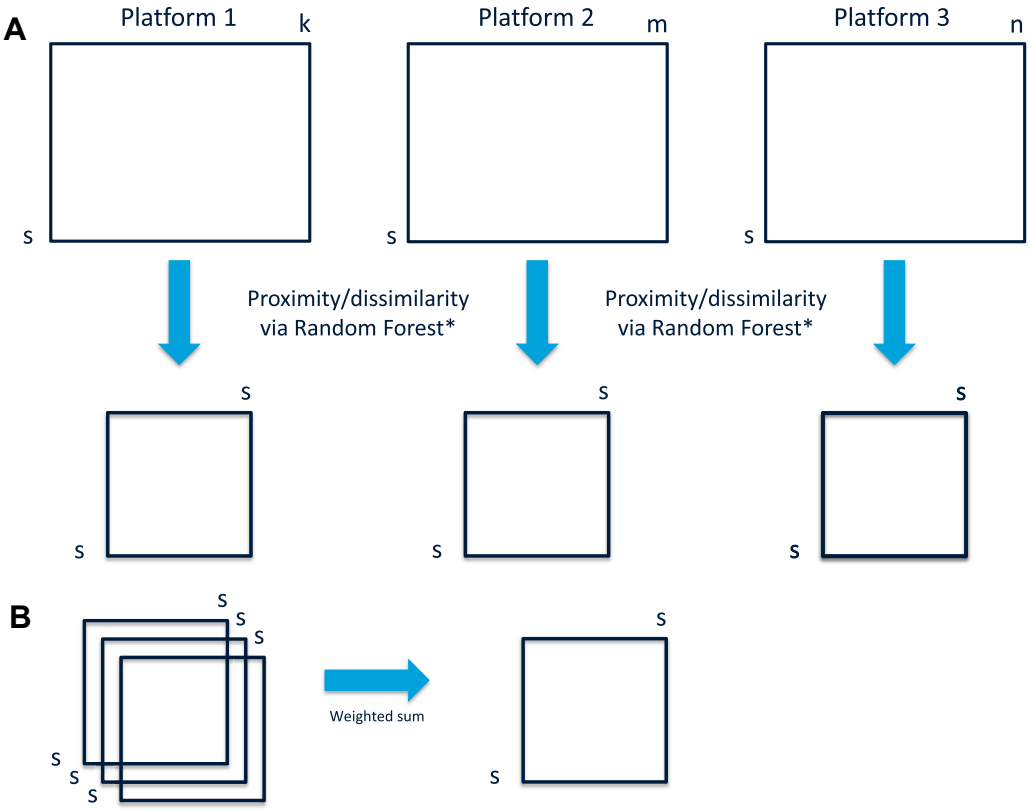


Fig. 3. Schematic representation of the proximities stacking fusion approach of three datasets. RF is run on each of the datasets to get their proximity matrix. Then, all three proximity matrices are stacked one on top of each other, and via a weighted sum, they create the final single proximity matrix K. *The proximity/dissimilarity matrices can also be created via unsupervised random forest, and these proximities were used in the present study. More details on the matter can be found in section 2.4.3.

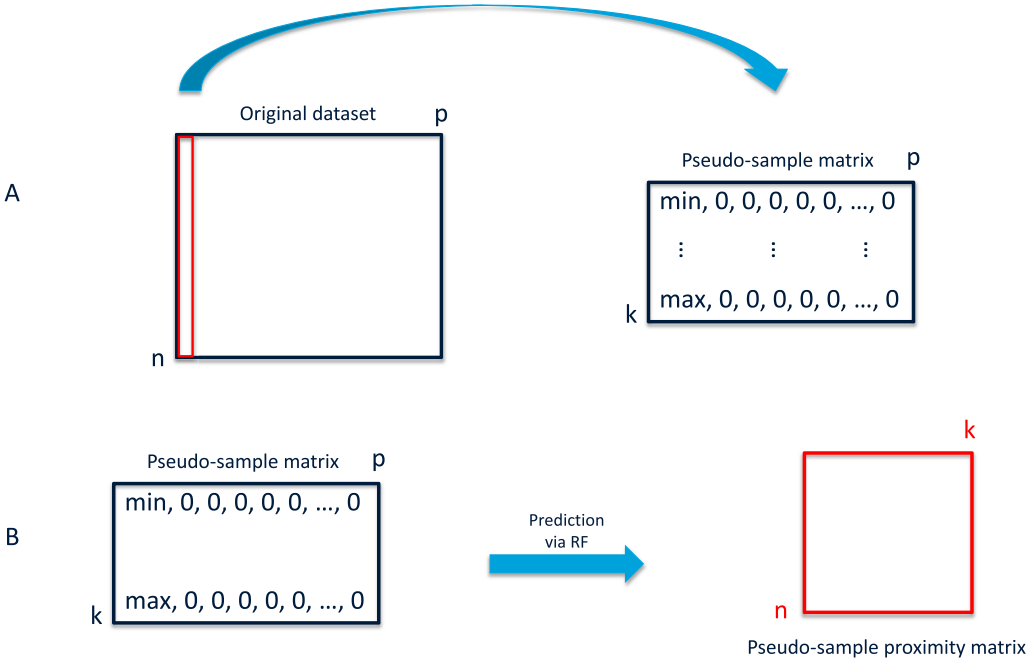


Fig. 4. Graphical representation of how the pseudo-sample proximity matrix of the very first original variable is created. A) First, all the values of variable one are sorted out and placed in column one of the pseudo-sample matrix, whereas the remaining of the columns are filled in with zeros. B) Then, RF is run on the pseudo-sample matrix to obtain the pseudo-sample proximity matrix, which ultimately holds information of the very first variable only. n is the number of samples, p is the number of the original variables, and k is the number of points that the user chooses to spread the range of the values of variable one on.

importance value found in the RF model. Next, the number of times that every single variable had been found as important in all the 1000 RF models was calculated (i.e. counts per variables), and in the end, the variables that had the most counts were kept. The threshold which determined the optimal number of variable counts to be kept for further analysis differed per platform since the data platforms contained different types of data. For each of the 1000 iterations, a one-by-one backwards variable elimination procedure was performed, and every time a variable was eliminated, the root-mean-square-error-prediction (RMSEP) value was calculated. The number of variables that gave the lowest RMSEP value was considered as optimal. Each of the 1000 iterations gave its own optimal number of variables, and by averaging them out, the optimal number of variables per platform (i.e. counts per variable) was found. The RF model parameters were optimized with a similar way too. The RF model optimization, i.e. number of trees and the number of samples to be kept per tree terminal leaf, was done using the out-of-bag error of the model. As far as the number of predictors to be used in the bootstrapping procedure goes, the square root of the total number of predictors present in the data was used. Finally, a 1000-iteration permutation test was run to confirm that the selection of the RF parameters was indeed optimized. In the present study, 4000 trees per model were used, and at the same time, the minimum number of samples per tree leaf for every tree in each model was set to eight. Ultimately, a new optimized RF model was built by using the 104 training and validation samples to predict the independent internal test set samples. Its performance was assessed by calculating the sensitivity and specificity for the independent internal test set.

2.4.2. Mid-level and high-level fusion

In the mid-level fusion case (Fig. 1), the variables with the most counts (found as described in section 2.4.1) from all the platforms were fused, sample-wise, and then, a single optimized RF model was built by using all the 104 samples. Its performance was assessed by calculating the sensitivity and specificity for the independent internal test set and visualized by subsequently performing PCA on the RF proximity matrix of the training samples, where then the independent internal test samples were also projected.

In the high-level fusion case (Fig. 2), optimization of the classification probability threshold within a 100-iteration loop followed the optimization of the variable selection and the RF model parameters (section 2.4.1). For each of the 100 iterations, the 104 samples were randomly split into training and validation sets, and individual platform predictions were made. Then, the individual platform classification probabilities were adjusted via the Bayes' theorem, and the receiver operating characteristic (ROC) curve was plotted to find the classification probability threshold that maximized both sensitivity and specificity of the model. The average of all the optimal thresholds of all the 100 models was calculated, and this threshold was then considered optimal. In the end, all 104 samples were used once again to build the final optimized RF model, whose performance was then assessed by predicting the independent internal test set, in terms of sensitivity and specificity.

2.4.3. Proximities stacking fusion

As mentioned already in the data fusion approaches paragraph (paragraph 2.2), first, ten sets of numbers that fulfilled the equation (3) were found. Then, these sets were shuffled to give 40 different possible combinations of sets. A table with all the sets of weights w can be found in the supplementary materials.

For each of the 100 iterations, the 104 samples were randomly split into training and validation sets, and the proximity matrices of these sets and for all the data platforms were obtained by unsupervised random forest (URF) (i.e. four training and four validation proximity matrices) [40]. URF is the unsupervised version of RF that assumes that if there is any structure hidden in the data, it should be possible to distinguish them from a randomly generated version of themselves. URF was employed to get the proximity matrices instead of RF to limit possible overfitting when a small number of samples is used (Fig. 3). The use of RF proximities may result in possible overfitting even though an optimization of the RF model has been performed due to the supervised nature of RF. The sample classes of the training data are embedded in the RF model by definition, and when the number of samples is small, it can lead to overfitting and to unnecessarily complex or nonflexible models. The use of URF proximities should suffice for improving classification accuracy; however, if the user does not achieve a fair classification accuracy by using the URF proximities, RF proximities may also be used. In each iteration and for every set of weights (i.e. w, found via equation (3)), the training proximities were stacked as well as the validation proximities (Fig. 3). This resulted in 40 training proximities with their corresponding validation proximities. PCA was applied to every training proximity, and its related validation proximity was projected onto its training proximity PCA space, and based on how well the validation samples were projected on the training sample PCA space, the best set of weights w for this particular set of training and validation samples was found. The number of times that every set of weights w_i was found as the optimal one out of all the 100 iterations was calculated. In brief, an AUC was calculated for every set of weights w to find the optimal one per iteration; the PCo1 scores of each iteration validation set were used to calculate these AUCs. The set of weights w that gave the highest AUC was considered as the optimal. The final classification model was assessed by calculating the sensitivity and specificity for the independent internal test set.

All data analyses were performed by using MatLab R2016b version—the Statistics and Machine Learning Toolbox. For the RF models, the TreeBagger function was used, whereas for the URF models, the code was found elsewhere [40].

3. Results

The raw microbiome data consisted of 6629 variables, whereas the raw blood data consisted of 32768 variables. The raw blood headspace data consisted of 2549 variables, while the raw exhaled breath data consisted of 545 variables. After data preprocessing and data reduction steps, microbiome matrix was left with 734 variables, blood matrix with 423, blood headspace matrix with 531, and exhaled breath matrix with 256. The optimal number of variables per platform (found via the platform optimization process described in section 2.5.1) to be used for both individual and fused matrices predictions were 58 for the microbiome (the threshold was 50%, meaning that the variables that found as important in more than 50% of the total number of iterations were kept), 19 for blood (with a threshold of 35%), 14 for blood headspace (with a threshold of 40%), and 16 for exhaled breath (with a threshold of 40%). At the same time, all four data platforms consisted of 130 samples, of which 66 were remission cases, and the remaining 64 were active cases of the disease. Notably, 104 samples were used to build and validate the models, whereas the remaining 26 were used to validate the models independently.

Mid-level, high-level, proximities stacking data fusion, as well as individual platform RF models were built, and their performance was assessed by calculating the sensitivity and specificity for the independent internal test set. Furthermore, for the individual platform cases, the mid-level, and the proximities stacking fusion cases, PCA was performed on the training sample proximity matrices, where the independent internal test samples were projected for visualisation purposes. The mid-level case gave a

Table 1
Sensitivities and specificities of all the fusion and all the individual platform cases for the external test set. The numbers in the parentheses show the actual number of the correctly classified patients; the number of patients in each individual platform differed from either other and from the number of samples present in the external test set in the fused cases. This is because some patients provided all three samples (i.e. faeces, blood, breath), whereas some others only provided one (meaning either only breath, or faeces, or blood) or two samples (meaning either blood and faeces, or faeces and breath, or breath and blood).

	Sensitivity	Specificity
Mid-level fusion	67% (10/15)	91% (10/11)
High-level fusion	27% (4/15)	100% (11/11)
Proximities stacking fusion	93% (14/15)	100% (11/11)
Microbiome	95% (19/20)	94% (15/16)
Blood	21% (3/14)	93% (13/14)
Blood headspace	35% (6/17)	47% (8/17)
Exhaled breath	85% (17/20)	50% (8/16)

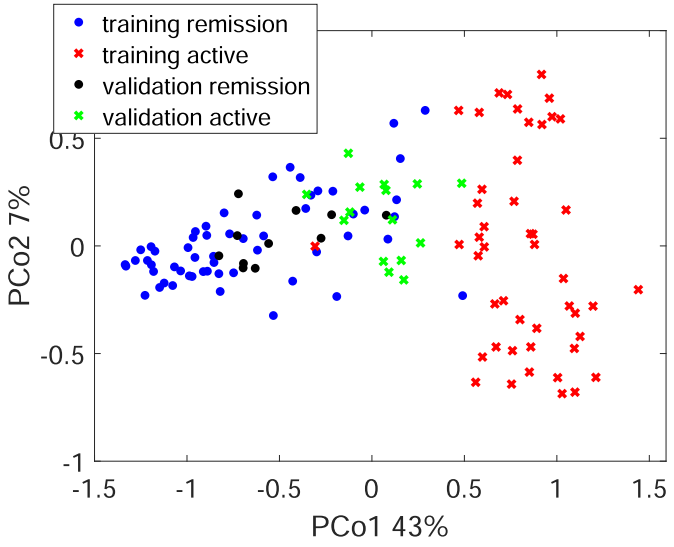


Fig. 5. Score plot of the training (i.e. 104) and validation (i.e. 26) samples of the RF model in the mid-level fusion case. The blue dots represent the remission training samples, whereas the black dots represent the remission validation samples. The red crosses represent the active training samples, while the greens crosses represent the active validation samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

sensitivity of 67% and a specificity of 91% (Table 1) and its corresponding score plot can be seen in Fig. 5, while the high-level case gave a sensitivity of 27% and a specificity of 100% (Table 1). In the proximities stacking attempt, the optimal set of weights w was $[m_1 = 0.900\ m_2 = 0.200\ m_3 = 0.0100\ m_4 = 0.3872]$, which shows the contribution of the microbiome, blood, blood headspace, and exhaled breath in the final RF model, respectively; the final RF model gave a sensitivity of 93% and a specificity of 100% (Table 1). The proximities stacking corresponding score plot is illustrated in Fig. 6. The sensitivities and specificities of the individual platforms are summarised in Table 1, and their corresponding score plots can be found in the supplementary materials. The proximities stacking approach outperformed both the mid-level and high-level fusion approaches, as well as all the individual platform results in terms of sensitivity and specificity except for the microbiome, which performed equally well.

The results of the pseudo-sample principle applied in the proximities stacking case are shown in Fig. 7 and Fig. 8. In particular, Fig. 7 shows the importance of the two most important variables per platform: the first two variables (i.e., variables number 17

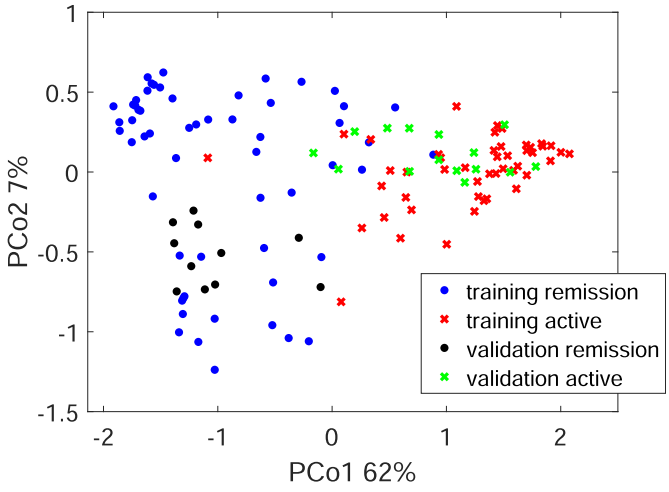


Fig. 6. Score plot of the training (i.e. 104) and validation (i.e. 26) samples of the RF model in the proximities stacking fusion case. The blue dots represent the remission training samples, whereas the black dots represent the remission validation samples. The red crosses represent the active training samples, while the greens crosses represent the active validation samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

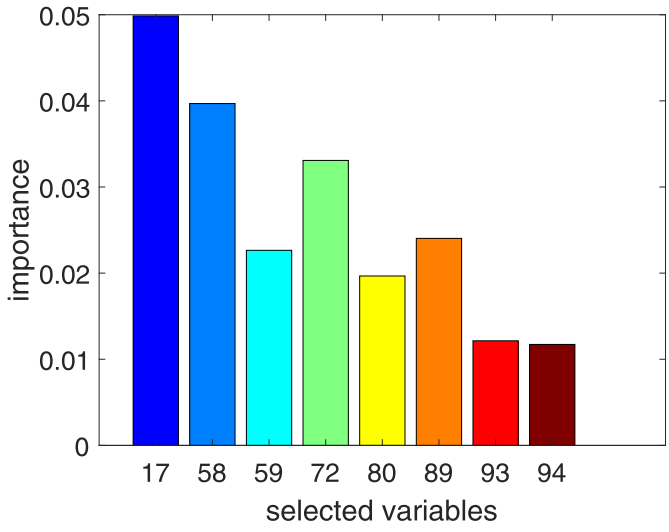


Fig. 7. Bar plot depicting the importance of the two most important variables per platform (in total, there were 107 fused variables). The variables 17 and 58 come from the microbiome, the variables 59 and 72 come from blood, the variables 80 and 89 from blood headspace, and the variables 93 and 94 come from exhaled breath. The variable indices come from the RF model, and they represent the position of each variable in the dataset. The different colours are used for illustrative purposes only. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and 58 out of all the 107 that were fused) come from the microbiome, the next two variables (i.e., variables number 59 and 72) come from blood, the following two (i.e., 80 and 89) variables come from blood headspace, and the last two (i.e., 93 and 94) variables come from exhaled breath. It should be mentioned here that the variable numbers represent the position of each variable in the original variable concatenated dataset, and that the importance of each variable was found via the pseudo-samples projected onto the PCoA space, and it is calculated by using the maximum absolute value of the loadings of the original variables trajectories. Fig. 8 represents the trajectory plot of two selected variables (i.e. those with the highest importance) per data platform. The variables are

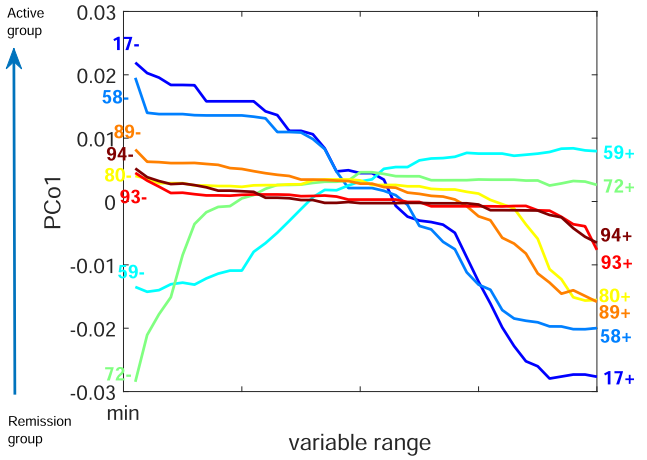


Fig. 8. Trajectory plot of the two most important variables per platform. More specifically, variables 17 and 58 come from microbiome and in active groups, they are present in very low relative abundances, while in remission cases, they show their highest relative abundances. Variables 80 and 89 come from blood headspace, and they show the same trend as the ones coming from microbiome. The same holds for variables 93 and 94 that come from exhaled breath, whereas variables 59 and 72 that come from blood, they are present in very low relative concentrations in remission groups; when these groups become active, these variables show their highest relative abundances. The different colours are used for illustrative purposes only. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

colour-coded with the same colours in both figures to provide better illustrative comparisons. More specifically, Fig. 8 shows the relation between the top two variables per platform and their relative amount change in the active and remission groups. One can see that the relative amounts of variables 59 and 72 exhibit downregulation in the remission group in comparison to the active group. The other way around holds for the other six variables (i.e., 17, 58, 80, 89, 93, and 94) coming from the microbiome, blood headspace, and exhaled breath. These particular variables are present in very low relative abundance amounts in active cases of the disease, but when these cases become remission, these variables show their highest relative abundance amount.

4. Discussion

The current study investigated the potential of fusing RF proximities of various datasets (i.e. proximities stacking) to ultimately increase the prediction accuracy of disease activity in CD cases, and compared its performance against traditional ways of data fusion in terms of sensitivity and specificity of an external test set. Proximities stacking demonstrated an excellent classification of the independent internal test samples (Fig. 6), whereas mid-level fusion (Fig. 5) gave a fair classification accuracy of the independent internal test samples. Proximities stacking significantly outperformed all individual platform results as well except for the microbiome case, which performed equally well (Table 1 and supplementary materials). Concurrently, this study also applied the pseudo-sample principle that helped discover and examine possible biomarker behaviour in CD patients in the proximities stacking fusion case (Figs. 7 and 8).

Data fusion has proved to be a valuable asset not only in computer science domains but also in life science fields (e.g. metabolomics) too [1–10] as a result of the vast amount of data that are generated nowadays. High-level fusion is rightfully considered as, perhaps, the most potent traditional way of data fusion when it comes to high prediction accuracy due to the way it is defined: many models are combined to get the final predictions instead of one model. The various model outcomes can be combined by using either class labels (i.e. majority voting [36]) or adjusted probabilities. The substantial advantage of choosing adjusted probabilities over majority voting is that one can find how sure the individual models are about their decisions on the samples of interest. Another advantage of high-level fusion is that if a new dataset for the problem in hand becomes available, it can be used to improve the versatility of the decision process too. The major disadvantage, however, of high-level fusion is that it does not give any information about variables/compounds that are important in classifying/predicting samples since it only works with outcomes and not variables. However, in the present study, the high-level fusion results (via the Bayes' theorem) did not demonstrate the best performance with a sensitivity and specificity of 27% and 100%, respectively, which may be due to the limited number of platforms and therefore models that were combined to get the fused outcome. Mid-level fusion can, possibly, increase prediction accuracy when compared to individual platform predictions, as well as it gives the ability to biomarker discovery since it works with either variables or features. In life science fields, and the metabolomic world more specifically, an at least fair prediction accuracy along with biomarker identification are sought; this is why mid-level fusion has become the most broadly implemented fusion approach. Here, the mid-level fusion results (Fig. 5) were inferior to the proximities stacking results (Fig. 6), and superior to both the high-level fusion and the individual platform results (Figs. S1–S4) except for the microbiome case, achieving a sensitivity and specificity of 67% and 91%, respectively. Further variable importance in the mid-level fusion results (e.g. compound behaviour in the CD samples) was not conducted. Low-level fusion is the least applied approach in the metabolomic world, and as it was mentioned in the 2.3 section already, the degrees of freedom of the data play a crucial role in this. In low-level fusion approach, the error degrees of freedom is negative since the number of variables is almost always a lot bigger than the number of samples; leading to challenges in proper model optimization and development. Metabolomic data are high-dimensionality data on their own (i.e. the number of variables far exceeds the number of samples); therefore, fusing already high-dimensionality data creates matrices of hundreds or thousands of variables which are challenging to be dealt with. This is why low-level fusion was not applied in the present study. Furthermore, all individual platform results (Figs. S1–S4) were inferior to the proximities stacking fusion results (Fig. 6), except for the microbiome case and superior to the high-level fusion results. The sensitivities and specificities of the individual platforms are summarised in Table 1—the microbiome was the only platform that outperformed the mid-level fusion results.

The fusion of RF/URF proximities by using a weighted sum (i.e. proximities stacking) has not been performed before to the best of the authors' knowledge, and the current study results showed that they could be successfully implemented in complex biological samples, such as CD cases. In particular, proximities stacking demonstrated excellent performance in classifying the external CD cases (Fig. 6). The optimal set of weights w was $[m_1 = 0.900\ m_2 = 0.200\ m_3 = 0.0100\ m_4 = 0.3872]$, which shows the contribution of every platform in the final model. On the one hand, the microbiome contributed the most, and then breath and blood followed. On the other hand, blood headspace contribution was the least. The low contribution of blood headspace contradicts the general belief that the more data, the merrier the result, and as it has been stated already in the introduction, considerations have to be taken before any data fusion analysis is conducted. For example, if the aim of a study is to explore the biology of a system, then the more data gathered would be beneficial; however, if the aim is biomarker

discovery, the more data gathered is not always beneficial. The contribution of each platform provided by the set of weights *w* was to be expected given the individual platform performances. The pseudo-sample principle results illustrated the importance of the original variables in classifying the CD cases (Fig. 7), as well as the original variable behaviour in the samples for two selected variables per platform (Fig. 8). Fig. 7 supports the optimal set of weights *w* since one can see in the figure that the most high-importance variables are the microbiome variables. In Fig. 8, one can see that the blood selected variables are present in very low relative abundances in the remission cases of CD, and they reach their highest relative abundances in the active CD cases—the other way around holds for the microbiome, blood, and exhaled breath selected variables. Most importantly, Fig. 8 helps demonstrate changes that occur in the variable relative abundances. For example, the breath variables (i.e. variable numbers 93 and 94) exhibit an instant increase in their relative abundance when going from active to remission. The same holds for the blood headspace variables (i.e. variable numbers 80 and 89) as well; however, the blood headspace variables exhibit a slower pace increase right before they reach their highest relative abundances. This similar behaviour amongst the blood headspace and exhaled breath variables indicates a connection of these four compounds coming from different sources, and therefore, it can also help dive deeper into the CD pathophysiology.

URF/RF proximities, in terms of fusion, would be of added value in the field of metabolomics and data science, in general. This is because the URF/RF proximities stacking, combined with the pseudo-sample principle approach, has several strengths to show over the other traditional ways of fusion. First of all, it proved that it significantly outperforms the other traditional fusion ways in terms of sample classification, and when compared against the mid-level fusion, it also solves the variable scaling problem since proximities make use of samples rather than variables [5]. Moreover, when compared against the high-level fusion, it solves the variable examination problem that occurs since high-level only uses model outcomes rather than variables [5]. Most importantly, URF/RF proximities stacking, via the weighted sum, also demonstrates the contribution of every platform in the final model, something that no other traditional fusion approach does. The proximities stacking approach also permits the fusion of any type of data (i.e. continuous or discrete), which has proved to be an issue when different data sources are used for a question in hand. It should also be noted here that the URF/RF proximities stacking approach illustrates an essential advantage over the approach reported by Smolinska et al. [1] as well. Smolinska et al. [1] fused kernels instead of proximities. Their approach was successfully applied in metabolomics data, however, finding the optimal kernel for the analysis in hand might be a challenging task to conduct because it requires variable scaling beforehand, and a rather extensive optimization process. In a fused kernel approach, the user has to select and optimize the type of the kernel and the corresponding parameters, such as the polynomial order if the kernel used is the polynomial or the distribution width if the kernel used is the radial basis function. Finally, it has to be mentioned that in the proximities stacking approach, the final fused matrix (i.e. all the individual proximities combined via the weighted sum) can be used for visualisation purposes of the data as well by directly applying PCA, for instance. In the present study, this fused matrix was used for classification purposes of the independent internal test set samples instead (Fig. 6). Linear supervised approaches such as partial-least-squares (PLS) [35] analysis may also be used for either classification or visualisation purposes.

The present study validated its results by using an independent internal test set, thus strengthening its validity even more; nonetheless, the present study also demonstrates some limitations that have to be addressed. The current study did not perform the low-level fusion. Although, it is considered highly unlikely that low-level fusion would have been of any added value to the study since the dimensionally of the data was high, and low-level fusion cannot cope with high dimensionality data. One can also argue that the present analysis lacks variable/compound identification since the pseudo-sample principle permits for compound identification. This was not performed due to the nature of the paper, which is to present the proximities stacking approach rather than identify biomarkers for the disease activity. Lastly, the authors acknowledge the fact that the study results might be seen as accidental since the proposed fusion approach was applied only on one disease data; to prove that the presented approach works on other datasets as well, a simulation analysis was also performed, and it can be found in the supplementary materials. Briefly, four data platforms consisting of 250 samples and 50 variables each were generated. Proximities stacking achieved the best classification results, and the contribution of each simulated platform provided by the set of weights *w* was to be expected given the individual simulated platform performances. Nevertheless, the proposed fusion approach should be tried on other real data fusion occasions as well to further confirm its strength over the currently available fusion ways.

5. Conclusion

In conclusion, URF/RF proximities stacking fusion coupled with the pseudo-sample principle approach proved to outperform the traditional ways of fusion significantly, overcame essential drawbacks of the current fusion methods, and helped examine variable behaviour and relations; therefore, establishing itself as a new, powerful data fusion tool that can be implemented in any scientific domain. Data fusion keeps gaining a lot of attention in various scientific fields since combining different types of data can yield higher model performance. However, this is not always the case, and considerations have to be taken into account before any analysis is conducted based on the type of study and the ultimate analysis aim. For example, the data have to be complementary for data fusion to work successfully, and as the present study demonstrates, the more data used or fused does not necessarily mean the merrier the result. The traditional ways of fusion (i.e., low-level, mid-level, and high-level) have been successfully implemented [1–10] so far, but as complexity and amount of data increase along with the complexity of the question in hand, more advanced and sophisticated fusion ways are needed.

Authors' contribution

G.S. – statistical analysis, results interpretation, and manuscript writing, R.V.V. – MatLab code provision, D.M.A.E.J, J.P., and J.E.H – data collection, F.J.V.S – study management, A.S. – data collection, study management, concept and design, and results interpretation. All the authors have critically read and revised the present manuscript, and approved its final version to be published.

Declaration of competing interest

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

Acknowledgements

The present study was supported by the VENI grant, Netherlands organization for scientific research (NWO) no. 016 VENI 178.064.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.339001.

References

[1] A. Smolinska, et al., Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis, PloS One 7 (6) (2012).
[2] E. Acar, et al., Forecasting chronic diseases using data fusion, J. Proteome Res. 16 (7) (2017) 2435–2444.
[3] L. Blanchet, et al., Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis, BMC Bioinf. 12 (1) (2011) 254.
[4] A.K. Smilde, et al., Fusion of mass spectrometry-based metabolomics data, Anal. Chem. 77 (20) (2005) 6729–6736.
[5] E. Borràs, et al., Data fusion methodologies for food and beverage authentication and quality assessment–A review, Anal. Chim. Acta 891 (2015) 1–14.
[6] M. Silvestri, et al., A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemometr. Intell. Lab. Syst. 137 (2014) 181–189.
[7] L. Vera, et al., Discrimination and sensory description of beers through data fusion, Talanta 87 (2011) 136–142.
[8] W. Sun, et al., Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb, Spectrochim. Acta Mol. Biomol. Spectrosc. 171 (2017) 72–79.
[9] C. Malegori, et al., A modified mid-level data fusion approach on electronic nose and FT-NIR data for evaluating the effect of different storage conditions on rice germ shelf life, Talanta 206 (2020) 120208.
[10] C. Márquez, et al., FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80–86.
[11] P.W. Krooshof, et al., Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification, Anal. Chem. 82 (16) (2010) 7000–7007.
[12] G. Postma, P. Krooshof, L. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Anal. Chim. Acta 705 (1–2) (2011) 123–134.
[13] J. Gower, S. Harding, *Nonlinear biplots.* Biometrika 75 (3) (1988) 445–455.
[14] G. Upton, I. Cook, A Dictionary of Statistics 2 Rev, 2008.
[15] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[16] L. Blanchet, et al., Constructing bi-plots for random forest: tutorial, Anal. Chim. Acta 1131 (2020) 146–155, https://doi.org/10.1016/j.aca.2020.06.043.
[17] D.C. Baumgart, W.J. Sandborn, Crohn's disease, Lancet 380 (9853) (2012) 1590–1605.
[18] D.I. Tedjo, et al., The fecal microbiota as a biomarker for disease activity in Crohn's disease, Sci. Rep. 6 (2016) 35216.
[19] A.G. Bodelier, et al., Volatile organic compounds in exhaled air as novel marker for disease activity in crohn's disease: a metabolomic approach, Inflamm. Bowel Dis. 21 (8) (2015) 1776–1785.
[20] J. Jansson, et al., Metabolomics reveals metabolic biomarkers of Crohn's disease, PloS One 4 (7) (2009).
[21] U. Daniluk, et al., Untargeted metabolomics and inflammatory markers profiling in children with Crohn's disease and ulcerative colitis—a preliminary study, Inflamm. Bowel Dis. 25 (7) (2019) 1120–1128.
[22] J.B. Burbidge, L. Magee, A.L. Robb, Alternative transformations to handle extreme values of the dependent variable, J. Am. Stat. Assoc. 83 (401) (1988) 123–127.
[23] A. Smolinska, et al., Simultaneous analysis of plasma and CSF by NMR and hierarchical models fusion, Anal. Bioanal. Chem. 403 (4) (2012) 947–959.
[24] P.H. Eilers, A perfect smoother, Anal. Chem. 75 (14) (2003) 3631–3636.
[25] G. Tomasi, F. Van Den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemometr.: A Journal of the Chemometrics Society 18 (5) (2004) 231–241.
[26] T. De Meyer, et al., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm, Anal. Chem. 80 (10) (2008) 3783–3790.
[27] F. Dieterle, et al., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics, Anal. Chem. 78 (13) (2006) 4281–4290.
[28] C.A. Rees, A. Smolinska, J.E. Hill, The volatile metabolome of Klebsiella pneumoniae in human blood, J. Breath Res. 10 (2) (2016), 027101.
[29] A. Smolinska, et al., Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis, J. Breath Res. 8 (2) (2014), 027105.
[30] Z.X. Wang, Q.P. He, J. Wang, Comparison of variable selection methods for PLS-based soft sensor modeling, J. Process Contr. 26 (2015) 56–72.
[31] T.N. Tran, et al., Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemometr. Intell. Lab. Syst. 138 (2014) 153–160.
[32] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (9) (2014) 2812–2831.
[33] I. Guyon, et al., Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[34] S. Wold, et al., The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (3) (1984) 735–743.
[35] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometr. 17 (3) (2003) 166–173.
[36] L.S. Penrose, The elementary statistics of majority voting, J. Roy. Stat. Soc. 109 (1) (1946) 53–57.
[37] D.V. Lindley, Fiducial distributions and Bayes' theorem, J. Roy. Stat. Soc. B (1958) 102–107.
[38] B.-J.M. Webb-Robertson, et al., A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections, Biocomputing (2009) 451–463, 2009, World Scientific.
[39] W. Wu, et al., A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks, Water Resour. Res. 49 (11) (2013) 7598–7614.
[40] N.L. Afanador, et al., Unsupervised random forest: a tutorial with case studies, J. Chemometr. 30 (5) (2016) 232–241.