Feature Impact Assessment: A New Score to Identify Relevant Metabolomics Features in Artificial Neural Networks

Danhui Wang[1,2], Peyton Greenwood[1], Matthias S. Klein[1,*]

[1] Department of Food Science and Technology, The Ohio State University, Columbus, OH 43210, USA

[2] Department of Food Science, University of Massachusetts, Amherst, MA 01003, USA

* Correspondence: klein.663@osu.edu

Supplemental Materials

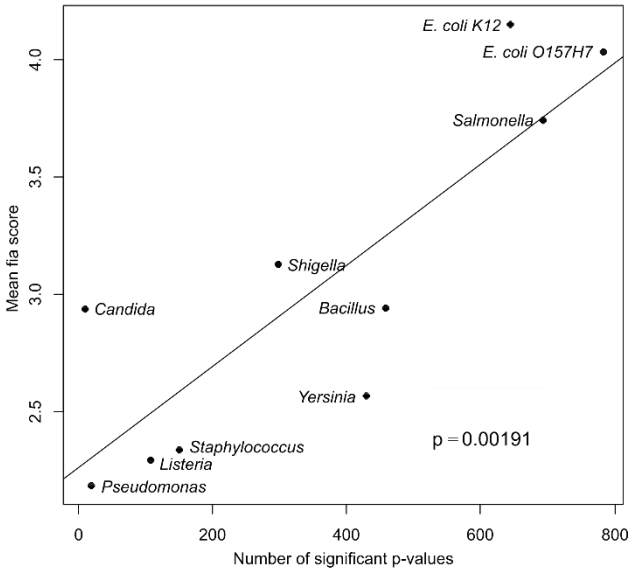

Fig. S1 Relationship of mean FIA score to the number of significant *p*-values when using the top 10 fia scores. A slightly lower *p*-value is observed as compared to using the top 100 shown in Fig. 2.

Interpretation of FIA Scores

Table S1 shows a summary of how to interpret fia scores, based on our experiences.

Table S1 Summary of various FIA scores and the suggested interpretation.

FIA score	Interpretation
1.0	Maximum impact, changing it can change the prediction outcome in all samples in the dataset
1.01–1.99	Very strong impact, but only in part of the samples
2.0–3.99	Strong impact, multiple features need to be changed simultaneously in order to change the prediction outcome
≥4	Medium to low impact, these results may be hard to interpret in metabolomics datasets; however, if the lowest observed FIA score is ≥4, this indicates that the predictive model is very stable and many features need to be changed to disrupt the prediction outcome. Still, features with the lowest FIA values will have highest impact for these models, even if the FIA score is >4.

Algorithm Description

Checking prediction outcomes by varying all possible combinations of features would be a time-consuming effort. Instead, we suggest a faster approach to screen the search space for potential hits and then narrowing down the results. Key elements are shown in Fig. S2.

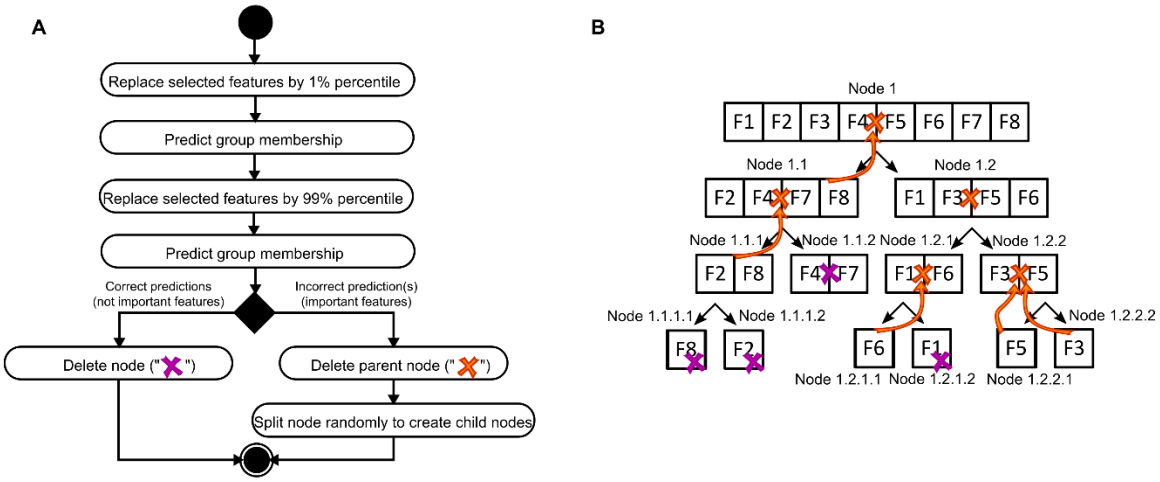

Fig. S2 Workflow of a fast approach to find combinations of features that are able to change the prediction results. (**A**) Workflow diagram of the code applied to each "node". (**B**) Example of a simple analysis using only 8 features (F1-F8).

Initially, samples that are correctly predicted are identified. For the first sample, a single node containing all features is created. Fig. S2A shows the workflow of the code applied to each node. A new data set is generated by switching all features in the node to the 99 percentile, while keeping all other features (if any) at their original levels. This process is repeated using the 1 percentile. If any of these two data sets have an incorrect prediction outcome, child nodes are generated by randomly dividing the node's features into two new nodes. If the prediction is correct in both cases, the features seem unimportant and the node is deleted. This procedure is repeated until all nodes have been covered. Nodes that have not been deleted in the end are assumed to contain important features.

Fig. S2B shows a simple example. Initially, only node 1 is present, containing all eight features (F1 to F8). After applying the code (A) on node 1, nodes 1.1 and 1.2 are created. After applying (A) on each node until no additional nodes are left, only nodes 1.1.1, 1.2.1.1, 1.2.2.1, and 1.2.2.2 are left. This result means features F3, F5, and F6 have a raw FIA score 1, and F2 and F8 have a raw FIA score 2 in this example.

To ensure that the random splitting of nodes does not affect the results, the final nodes are analyzed again using different seeds for the random number generator multiple times (the so-called *inner loop*). The features contained in the final nodes are each assigned their respective node length as their raw FIA score.

This whole workflow is repeated for every sample in the data set. It is then repeated using different seeds for the random number generator to avoid missing combinations of features.

In the end, a list of features is generated for each group. These features are assigned their lowest FIA score found in the results, plus the percentage of samples in which this raw FIA value was not observed for the respective feature.

R Code Example

The FIA algorithm was implemented in the function FIA in the R package mrbin, version 1.6.2 or higher, available at https://CRAN.R-project.org/package=mrbin. Sample R code to perform FIA analysis on an ANN using keras is shown below. Please be advised that keras and tensorflow require additional software to be installed to run properly, please see the package manuals for advice.

```
library(keras)
library(tensorflow)
library(mrbin)
#load data
dataSet<-as.matrix(read.csv("binDataMicrobes.csv", row.names=1,
check.names=FALSE))
#load factors with group memberships
factors0<-read.csv("factorsMicrobes.csv",as.is=TRUE)
factors<-factor(factors0[,2])
names(factors)<-factors0[,1]
#load pre-trained ANN model
model<-load_model_tf(filepath="model",custom_objects=NULL,compile=TRUE)
#calculate fia scores
fiaResults<-fia(model=model,dataSet=dataSet, factors=factors,nSeed=6)
#display some fia scores
fiaResults$scores[[1]][1:100]
fiaResults$scores[[4]][1:100]
```

Validation in a Separate Model and Dataset

The FIA method was tested on an additional published metabolomics dataset, namely a study of cardiovascular disease risk (Shearer *et al.*, 2021). The data set consisted of serum metabolite data as measured by 1D ^{1}H NMR. Details on the population can be found in the original publication. For testing FIA score analysis, we selected only maternal serum samples and stratified them into a “Lean” group and an “Overweight” group (BMI>25). For this analysis, no noise removal was performed, leaving 1462 features and 37 samples in the dataset.

Analyses were performed in R (3.5.1) using packages keras (2.6.0) and tensorflow (2.6.0). ANN models using a single hidden layer of 800 neurons, ReLU activation, and the Adam optimizer were trained on the dataset to predict group membership. After training, all samples in the dataset were correctly predicted by the model. FIA scores were calculated using the *fia* function in mrbin (1.6.3). Initially, only very few FIA scores less than 100 were observed, therefore, the parameter *innerLoop* was increased from 100 to 300. This parameter controls how often remaining nodes are randomly split into child nodes to narrow down the list of features of impact. Two-tailed, unpaired *t*-tests assuming unequal variances were used as an alternate way of scoring group differences.

Fig. S3 shows a comparison of *p*-values for the FIA top 10 features versus the rest of the features. It is obvious that the features with best FIA scores were significantly associated with lower *p*-values (p = 0.039).

It should be noted that ANN models are not optimized for finding *all* features of importance. Instead, depending on the outcome of the training run, the ANN might use only a subset of important features that deliver high accuracy results. In this sense, it is to be expected that low p-values will be observed even among features that are of no importance in the ANN model and thus do not receive top FIA scores. This is clearly visible by the whiskers of the right boxplot in Fig. S3 reaching down close to the value of 0. This expected behavior is anticipated to negatively impact the significance of the comparison of top FIA scores with other features.

Still, the results of the FIA analysis of this dataset show that FIA scoring was able to select features that are highly differentiating between the observed outcomes.

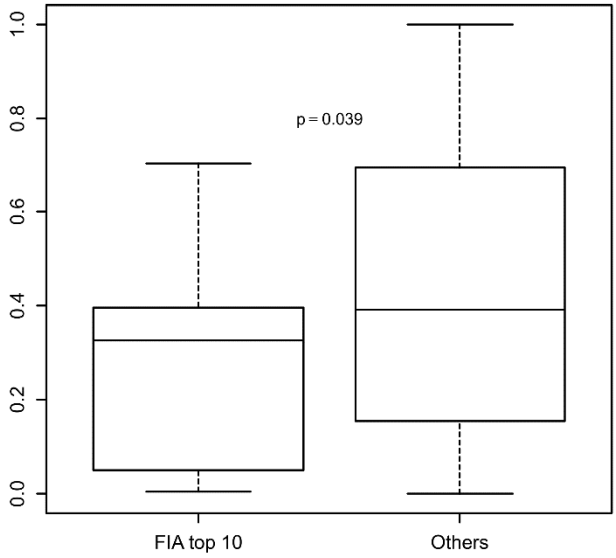

Fig. S3 Boxplots of *p*-values of the top 10 FIA features compared to all other features.

Fig. S4 shows a heatmap of the signal intensities of the top 10 FIA features. Further analyzing these features revealed that several of the top 10 FIA features were also identified as features of interest in the original publication, including β-hydroxybutyrate (BHB), acetoacetate, and pyruvate. In addition, other top FIA features are visibly strongly correlated to the phenotype but were not detected at the significance level in the original publication. Some signals in the heatmap apparently exhibit strong correlations within subgroups of the dataset. In this sense, FIA analysis could help identify additional features of high interest that might be missed by other data analysis approaches.

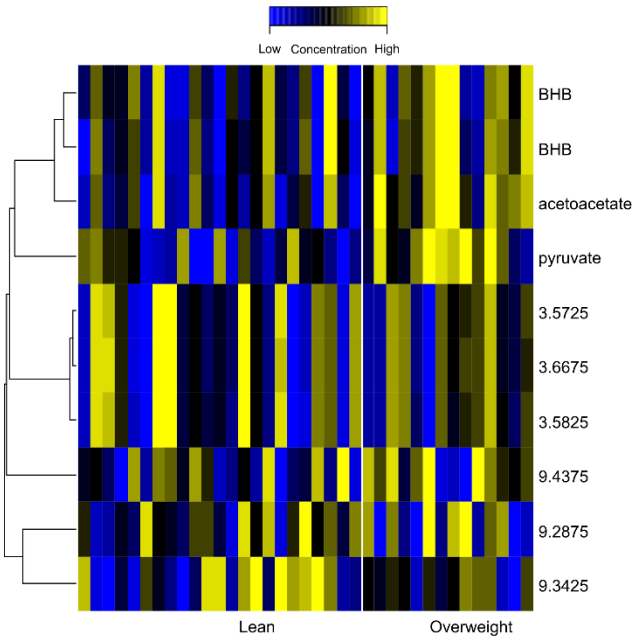

Fig. S4 Heatmap of the features with top 10 FIA scores. BHB: β-hydroxybutyrate.

References for Supplemental Materials

Shearer, J., Klein, M.S., Vogel, H.J., Mohammad, S., Bainbridge, S. and Adamo, K.B. (2021) Maternal and Cord Blood Metabolite Associations with Gestational Weight Gain and Pregnancy Health Outcomes. *Journal of Proteome Research* **20,** 1630-1638.