Microbe2Pixel: Taxonomy informed deep-learning models and explanations

Bas Voermans*,1,2, Marcus de Goffau2,3, Max Nieuwdorp1,4, Evgeni Levin2

1Department of Experimental Vascular Medicine, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, Amsterdam, the Netherlands
2HORAIZON Technology BV, Marshallaan 2, Delft, the Netherlands
3Tytgat Institute for Liver and Intestinal Research, Amsterdam University Medical Centers, Meibergdreef 69-71, Amsterdam, the Netherlands
4Department of Vascular Medicine, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, Amsterdam, the Netherlands

Abstract

In recent years, machine learning, especially deep learning, has garnered substantial attention in the biomedical field. For instance, deep learning has become a preferred method for medical image analysis tasks. However, in other areas like fecal metagenomics analysis, the application of deep learning remains underdeveloped. This can be attributed to the tabular nature of metagenomics data, feature sparsity, and the complexity of deep learning techniques, which often lead to perceived inexplicability.

In this paper, we introduce Microbe2Pixel, an innovative technique that applies deep neural networks to fecal metagenomics data by transforming tabular data into images. This transformation is achieved by inferring location from the taxonomic information inherently present in the data. A significant advantage of our method is the use of transfer learning, which reduces the number of samples required for training compared to traditional deep learning. Our method aims to develop a local model-agnostic feature importance algorithm that provides interpretable explanations. We evaluate these explanations against other local image explainer methods using quantitative (statistical performance) and qualitative (biological relevance) assessments.

Microbe2Pixel outperforms all other tested methods from both perspectives. The feature importance values align better with current microbiology knowledge and are more robust concerning the number of samples used to train the model. This is particularly significant for the application of deep learning in smaller interventional clinical trials (e.g., fecal microbial transplant studies), where large sample sizes are unattainable and model interpretability is crucial.

Keywords: metagenomics, interpretable deep learning, local explanations.

1. Introduction

1.1. Background

Deep neural networks (DNNs) have gained immense popularity in recent years and have been successfully applied to various areas of the medical field, such as image classification, object detection, and image segmentation [9]. However, in the field of fecal metagenomics research, which investigates the relationship between numerous aspects of human life (including disease) and the microbial composition of the human gut [11,15], DNNs have yet...
to prove their superiority over traditional machine learning (ML) methods for classification and regression tasks (e.g., gradient-boosting, randomized forests). Several reasons contribute to the unsuitability of DNNs for metagenomics data. Firstly, the data typically consists of tens to hundreds of instances (samples), but can have several thousand features (microbial groups). This is disadvantageous as DNNs perform best with a high instance-to-feature ratio. For example, it is not uncommon for interventional cohorts to have approximately 20 vs. 20 samples and ±3000 features [24, 8]. Increasing sample sizes for these types of trials is challenging due to factors such as funding, participant availability, and ethical approval. Thus, microbiome research often requires extensive feature selection for optimal performance. However, omitting features also risks removing relevant information from the data. Secondly, microbiome data is sparse, as microbiota species are commonly present in only a few instances but may still have a significant mechanistic effect. Thirdly, the model is trained on tabular data without context, even though data on the ecology of microbes is available in metagenomics analysis (i.e., the taxonomic tree, phylogenetic distance, etc.). Especially in smaller datasets, this often results in overfitting and/or the inability to extract meaningful information by focusing on very sparse features. Lastly, DNNs are often perceived as inexplicable black boxes. This perception makes it harder to gain the level of trust needed for adoption in the medical domain. Arguably, in many applications in the medical domain, the primary purpose of ML is not to achieve outright model performance. Instead, ML is primarily used to guide research and treatments of disease. For this purpose, explainability is an essential condition and could be perceived as more important than outright model performance.

1.2. Intuition

Our method aims to enhance the relevance of predictions in smaller datasets by utilizing the ecology of microbiota.

The concept of our method is inspired by the success of transfer learning for smaller image datasets. We hypothesize that transfer-learned DNNs could infer the similarity between microbes if provided, and subsequently make better-informed predictions. Additionally, the explanations generated with such a model would be more relevant concerning the ecology of the human gut microbiome.
1.3. Microbe2Pixel

We propose Microbe2Pixel
\(^1\), a method that transforms tabular fecal metagenomics data into images by leveraging the taxonomy of microbes. Microbe2Pixel consists of three different parts:

- Conversion of metagenomics data from tabular to image data by utilizing the taxonomy of the microbial groups.
- Two steps of transfer-learning a DNN from weights trained on the ImageNet dataset. Initially, we retrain only the top layer, followed by a second round of training to fine-tune the layers in the last convolutional block of the model. We reload the model weights of the epoch that maximized validation performance between both steps.
- Explanation of the DNN using multiple model-agnostic methods for feature importance on the test set.

Figure [1] provides a schematic abstract of all elements of Microbe2Pixel. Note: The taxonomic information used to obtain the coordinates is derived from the sequencing method, as assembling the taxonomic tree (TT) is part of deriving the abundance table process. This makes Microbe2Pixel applicable to both 16S and shotgun sequencing techniques. In principle, any other measure for similarity between microbial species could be used. Here we choose TT as it is easiest to obtain. Another significant advantage of our method is that all features are used to train the model (i.e., no abundance thresholds or feature selection), eliminating the risk of omitting relevant information.

In principle, any other medical data type with an inherent structure can be used to create images. Examples of this include metabolomics with metabolic pathways, transcriptomics/epigenetics with gene networks and proteomics data with protein-protein interaction networks. As such, with minor changes, our method could be applied to other data types with low difficulty, making our results applicable to many field of machine-learning in healthcare.

2. Related Work

To the best of our knowledge, we are the first to combine complex neural networks (>106 parameters) and transfer learning for fecal metagenomics data with a focus on explainable machine learning. In contrast, existing methods typically focus on classification performance and typically use 5-10 layers for the model, while our final model has 813 layers. Moreover, in Microbe2Pixel, feature importances are also informed by taxonomic information, enhancing the ability to extract relevant features for small datasets. The visual nature of Microbe2Pixel also makes it easier for clinicians and other healthcare professionals to understand model explanations. In this section, we provide a brief overview of related approaches and highlight the differences between Microbe2Pixel and existing techniques.

2.0.1. iMic [13]

In this work, the authors investigate the influence of taxonomy in small convolutional neural networks (CNNs) in metagenomics data. There are several key differences between our approach and Imic. For instance, Microbe2Pixel implements transfer learning to reduce the number of training samples required for accurate predictions. Additionally, Microbe2Pixel uses a distance matrix ordinated in 2-dimensions, bringing more fidelity.

\(^1\)We make an implementation of Microbe2Pixel publicly available at: (https://github.com/Bas-Voermans/Microbe2Pixel)
into the embedding of distances between microbes than the discrete 1-dimensional approach of \textit{iMic}. Furthermore, our algorithm only uses species-level relative abundances and adapted \textit{LIME} \cite{lime2016} to assess feature importances. We also evaluate the quality of feature importances both qualitatively and quantitatively, while \textit{Imic} is only compared to the literature and is not rigorously evaluated.

2.0.2. TaxoNN \cite{taxonn2020}

This method uses an ensemble of CNNs stratified per phylum. The convolutional layers are only applied within each phylum, after which the feature maps are concatenated. \textit{TaxoNN} uses small CNNs and is not transfer learned. The method primarily focuses on performance improvement rather than interpretable ML.

2.0.3. PopPhy-CNN \cite{popphy2020}

Similar to \textit{iMic}, the TT is embedded into a 2-dimensional array. Although there are differences in the algorithm of the embedding method, the CNNs of both methods are very similar. Both consist of 2 convolutional layers. A custom method is used to extract feature importances from the model. The relevancy of features is compared to literature and not assessed quantitatively.

2.0.4. Ph-CNN \cite{phcnn2020}

\textit{Ph-CNN} uses the phylogenetic tree to calculate a distance matrix between all microbes. The distance matrix is transformed into Euclidean coordinates using multi-dimensional scaling. \textit{Ph-CNN} proposes a so-called phylo-conv layer, a convolution performed over the K-nearest neighbors of each microbe. This is performed twice, after which a dense layer followed by the output layer is used. \textit{Ph-CNN} has been used in combination with transfer learning. However, this was only performed using synthetic data to pre-train the model weights. Again, the authors focus on model performance rather than interpretation.

3. Methods

3.0.1. Notation

We denote a matrix, a vector and a scalar with capital bold, bold and regular text respectively i.e., $\mathbf{X}$, $\mathbf{x}$, $x$. We denote functions as $f(x)$.

3.1. Image embedding

The taxonomy table of microbes is used to convert the metagenomics data from tabular to image data. Let $\mathbf{T}$ be the taxonomy table matrix in which $\mathbf{T}_{jl} \in \mathbb{R}^{M \times L}$ is the name of species $j$ at the taxonomic level $l$. We encoded 6 taxonomic levels in the dataset (kingdom, phylum, class, order, family, genus and species) as integers in a reverse range (i.e., $t = \{6, 5, \ldots , 0\}$). If present, we omitted the taxon level from the TT.

From $\mathbf{T}$, we calculate the distance matrix of microbes $\mathbf{D} \in \mathbb{R}^{M \times M}$. The linear distance $D_{jk}^{\text{lin}}$ between microbe $j$ and $k$ is defined as the lowest taxonomic level shared between the two species:

$$D_{jk}^{\text{lin}} = \min(t_l : \mathbf{T}_{jl} = \mathbf{T}_{kl}), t = \{6, 5, \ldots , 0\}$$

(1)

For instance, if two species are in the same genus, the linear distance is 1. If they are in the same family, the linear distance is 2, and so on. Note: if $j = k$, $D_{jk}^{\text{lin}} = 0$. 14
Instead of using linear distance, an exponential distance $D^{\text{exp}}$ is used (equation 2). Exponential distances ensure that the change in distance between the microbes diminishes the further up the tree they share a taxonomic level.

$$D^{\text{exp}}_{jk} = \begin{cases} 0, & \text{if: } j = k \\ \log_2(D^{\text{lin}}_{jk}) + 1, & \text{otherwise} \end{cases}$$ (2)

$D^{\text{exp}}$ is converted into two-dimensional coordinates using metric multi-dimensional scaling (MDS). MDS is an ordination method which obtains a low-dimensional representation of the data whilst optimizing for the distances in the original high-dimensional space.

After ordination, we uniformly re-scale image dimensions, ensuring all features have unique integer coordinates. This ensures each microbe falls exactly on a single pixel. Let $x$ and $y$ be the coordinate vectors in a two-dimensional space i.e., $(x, y) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\} \in \mathbb{R}^M$ where $x_j$ and $y_j$ are the integer coordinates of microbe $j$. Note, derivation of the coordinates is not dependent on abundances or samples in the dataset. As a result, we could derive exactly the same coordinates from any one instance in the data. This ensures that there is no data leakage in the form of information from test instances being used to obtain coordinates.

We embed relative abundances in the images as pixel intensities at the coordinates of each microbe. Let $A \in \mathbb{R}^{N\times M}$ be the log-scaled abundance table i.e., $A = \{a_1, a_2, \ldots, a_N\}$ where $a_i$ a vector of relative abundances i.e., $a_i = \{a_{i1}, a_{i2}, \ldots, a_{iM}\}$, with $a_{ij}$ the relative abundance of species $j$ in instance $i$. We use log-scaling because an 8-Bit gray-scale image format is required as input for the pre-trained model and metagenomics data typically has extreme outlier values. Thus, in each image, an intensity of 255 is equal to the maximum relative abundance in the corresponding sample int$\text{st}_{\text{max}}$.

In the end, we obtain a new dataset consisting of an image array $X \in \mathbb{R}^{N\times\max(x)\times\max(y)}$ and a class label vector $c \in \mathbb{R}^N$. $X$ is equal to zero everywhere, except for the locations in $x$ and $y$. For feature $j$ in instance $i$, the pixel value in $X$ is defined as:

$$X_{i, x_j, y_j} = \text{round}\left(\frac{A_{ij}}{\max(a_i)} \cdot \text{int}_{\text{st}_{\text{max}}}\right)$$ (3)

After the image embedding steps, the tabular dataset has been transformed into an image dataset. Each image corresponds to an instance from the tabular data. In the images, coordinates are informed by taxonomy and pixel intensities correspond to the relative abundances from the tabular data which have been transformed to a logarithmic scale. Figure 2 shows one of the images generated with Microbe2Pixel. As can be seen, microbes belonging to the different phyla cluster together, showing that the locations are informed by taxonomy. The locations of the microbes are conserved throughout the dataset (i.e. features have the same locations in all instances). A full summary of the embedding algorithm is given in algorithm 1.

### 3.2. Transfer Learning

As mentioned earlier, we used transfer learning of a complex DNN for the deep-learning method. For the architecture, we used EfficientNetB7 pre-trained on the ImageNet dataset [21]. This particular architecture was chosen as it yielded the highest performance of all architectures directly available in Keras (version 2.4.3.) [4]. In principle, any DNN architecture suitable for transfer learning can be used. The effects of choosing different
DNN architectures and hyperparameter tuning on the performance of Microbe2Pixel is not specifically investigated here. Rather, this research focuses on the concept of using (deep) transfer learning for microbiome data and its effects on model explanations.

The model was trained in two steps: training of the top layer, and fine-tuning of the last convolutional block. Table 1 provides a summary of the (hyper)parameter settings used to train the DNN for each training step.

3.3. Explanations

Due to our choice for complex DNNs (instead of simpler models such as XGBoost and Random Forests), we decrease the direct interpretability of the ML method. At the same time, this also enables the use of explanation models that are typical for medical image analysis. These explanation models are essentially interpretable approximations of complex models. As such, we develop a feature importance algorithm based on Local Interpretable Model-Agnostic Explanations LIME for our specific data type [17]. The main changes in comparison to stock LIME are in the derivation of the superpixels, which is made directly dependent on the locations and abundances of the microbial clusters.

The feature importance algorithm works by generating a weighted linear model around the DNNs prediction on each sample. Fundamentally, we create $P$ perturbed versions of each individual sample in which a set of microbial clusters is removed (set to an abundance of zero). Let $X_{ip}^*$ be the $p$th perturbation of original sample $i$. To generate $X^*$ each image is segmented into $K$ microbial groups. The microbial clusters are obtained using $K$-Means clustering applied to microbes by their (unscaled) relative abundance in each sample and location in the images. In this manner, we derive a unique clustering of the microbes for each individual sample. Next, a $P \times K$ perturbation matrix $P$ is generated by sampling integers from a binomial distribution. From $P$, we define a vector of sample weights $w$ as the inverse of pairwise distance between each perturbation and the original image. Now, let $f(X_{ip}^*)$ be the prediction vector made on perturbed set of sample $i$. Finally, we estimate importances on a superpixel level by fitting a weighted least squares regression model to model predictions based on the perturbation vector:

$$\arg\min_{\beta} w^T (f(X_{ip}^*) - P\beta)^2$$

The model coefficients $\beta$ equal the local feature importance of each superpixel in sample $X_i$. Local importances are stored on a feature level and aggregated to obtain global feature importances across all samples.
Input: taxonomy table $T \in \mathbb{R}^M \times L$, log-scaled relative abundance table $A \in \mathbb{R}^N \times M \in \mathbb{R}$

Output: image dataset $X \in \mathbb{R}^N \times \max(x) \times \max(y)$

$D^{\exp} \gets$ a zero matrix $\in \mathbb{R}^M \times M$

$count \gets 0$

for $j \in \text{range}(M)$ do
  for $k \in \text{range}(count)$ do
    for $level \in \text{reversed}(\text{range}(L - 1))$ do
      if $T_{j, level} == T_{k, level}$ then
        $D^{\exp}_{jk} \gets \log_2(\text{level}) + 1$
        continue
      end
    end
  count $\gets$ count + 1
end

$D^{\exp} \gets D^{\exp} + D^{\exp'}$

$\{x, y\} \gets \text{MDS}(D^{\exp})$

$\text{factor} = 1$, $\text{step} = 0.1$

while $\text{unique}(\{x, y\}) \neq \{x, y\}$ do
  $\text{factor} \gets \text{factor} + \text{step}$
  $\{x, y\} \gets \{x, y\} \cdot \text{factor}$
end

$\{x, y\} \gets \{x - \min(x), y - \min(y)\}$

$X \gets$ a zero matrix $\in \mathbb{R}^N \times \max(x) \times \max(y)$

for $i \in \text{range}(N)$ do
  for $j \in \text{range}(M)$ do
    $X_{i, x_j, y_j} = \text{round}(\frac{A_{ij}}{\max(a_j)} \times 255)$

$255 = \text{max. pixel intensity}.$
end
end

return $X$

Algorithm 1: Microbe2Pixel image embedding algorithm

4. Experiments

4.1. Datasets

We evaluate and validate Microbe2Pixel on two different previously published datasets: a publicly available cirrhosis dataset, and our own bariatric surgery dataset. A summary of the datasets is given in table 2. For an in-depth analysis of the dataset characteristics, the reader is referred to the original papers of the datasets.

4.1.1. Cirrhosis

This is a binary classification fecal metagenomics dataset consisting of 130 Han Chinese subjects. The set holds 68 cirrhosis patients and 62 healthy control subjects. In the subjects, 1342 different species of microbiota are measured. The data is publicly available, and can be obtained from the Knights Lab ML repository [23].

4.1.2. Bariatric surgery

The bariatric surgery dataset is part of the BARIA cohort. It consists of 430 fecal samples. This is another binary classification dataset consisting of of 430 instances. 310 of which are taken from patients pre-bariatric surgery, while 120 samples were taken 1 year post-bariatric surgery. The dataset consists of 3444 features (microbiota species). This dataset
Table 1: Summary of (hyper)parameter settings used to train and fine-tune the DNN.

<table>
<thead>
<tr>
<th>Name</th>
<th>Case</th>
<th>Control</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cirrhosis [15]</td>
<td>68</td>
<td>62</td>
<td>1342</td>
</tr>
<tr>
<td>Bariatric surgery  [12]</td>
<td>310</td>
<td>120</td>
<td>3444</td>
</tr>
</tbody>
</table>

Table 2: Summary of the datasets

is primarily used to obtain an ecological evaluation of feature importances. However, due to privacy restrictions, this dataset will not be made publicly available.

4.2. Experimental setup

We assess the effect of the use of taxonomic information on predictions and feature importances with several experiments performed over 25 stability runs in which we split the data into 80%/20% stratified train/test sets. On the train set, a second 80%/20% stratified training/validation split is performed to monitor overfitting and early stopping criteria. Apart from the weights pre-trained on the ImageNet dataset, no model weights are carried over. We compare the classification performance of Microbe2Pixel to the performance of the popular tree-based Extra-trees (XTR) and XGBoost (XGB) algorithms (with grid-search parameter optimization; see appendix C) and the same method without taxonomic information (i.e., random microbe locations) [3, 14]. These images have the same size as those derived from our Microbe2Pixel method. Thus, we evaluate the effect of taxonomy on performance, as well as comparing our model to the state-of-the-art for machine-learning in our field. Difference in performance is assessed with a two-sample t-test compared to the performance of Microbe2Pixel on the full set of training samples.

4.2.1. Quantitative assessment of feature importance

The feature importances are quantitatively assessed using a method inspired by Covert, Lundberg, and Lee (2020) [5]. Essentially, we compare feature importances acquired from all training samples to those from a training set that is sub-sampled by 50

We compared Microbe2Pixel to several established model-agnostic methods for calculating feature importance. Namely SHAP [10], Grad-CAM [19], and LIME [17]. These methods were chosen as they are popular local model-agnostic methods for feature importances. We consider the use of taxonomy a part of Microbe2Pixel, and therefore calculated the feature importances for the other methods from a DNN transfer-learned in the same way as Microbe2Pixel, but trained on randomly ordinated images. The feature importance from XTR and XGB were obtained from the default feature importance algorithms of the Python packages.

4.2.2. Qualitative assessment of feature importance

Measuring robustness alone is not sufficient to compare methods, since feature importances can be robustly wrong. Therefore, we asked an expert panel to evaluate the quality of
Figure 3: An example of the local explanations obtained from each image explainer method. The explanations are from the same sample. Positive values imply the feature values lead to prediction of class 1 and negative values imply prediction of class 0.

feature importances of Microbe2Pixel. The panel commented on the validity of feature importances compared to the current understanding of the relationship between the gut microbiome and both cirrhosis and bariatric surgery. For the cirrhosis dataset, the feature importances of LIME, SHAP, and Grad-CAM are assessed as well.

4.3. Results

4.3.1. Classification Performance

A fundamental condition for interpreting feature importances is good model performance. After all, if a model cannot predict accurately, its explanations are not representative of the data. Microbe2Pixel achieved a mean AUC equal to 0.85 ± 0.07 and is significantly better than the performance of the same method on images generated without taxonomy (AUC = 0.79 ± 0.09, \( p = 0.011 \)). Thus, incorporating taxonomic distance in the ordination of microbes into images improved classification performance. Moreover, a mean AUC of 0.85 is assessed to be sufficient such that feature importances may be interpreted. However, our Microbe2Pixel does not achieve state-of-the-art performance compared to XTR and XGB, which respectively achieve 0.90 ± 0.06 \( p = 0.0093 \) and 0.92 ± 0.04 \( p = 0.0002 \). Thus, Microbe2Pixel is not preferred if outright performance is the main goal.

4.3.2. Example explanations

An example of an explanation we acquired from each local method is shown in figure 3. This illustrates the results of each technique on the same sample. It is thought that the explanations of Microbe2Pixel require less expertise to interpret than conventional methods (bar graphs, scatter plots, spider plots, etc.). The microbial clusters are easier to identify in the explanation of Microbe2Pixel, and all importances are observable at a glance. This is of high importance, as the results often need to be interpreted by medical professionals who are not well-versed in machine learning.

4.3.3. Quantitative results

In our quantitative experiment, Microbe2Pixel outperforms the other popular model-agnostic feature importance methods we tested. The results of the quantitative assessment of global feature importances are summarized in table 3. \( p \)-values are calculated according to the Two-sample \( t \)-test with respect to Microbe2Pixel. As can be seen in table 3, our method ranks best among all other tested methods. For the local importances, our method observes the same result (table 4). These results illustrate that the incorporation of taxonomy in the derivation of feature importance significantly increases robustness on a
global and local level. Interestingly although our model achieved lower performance than the state-of-the-art tree-based models, we find much higher performance in feature importance robustness. Thus, Microbe2Pixel is the best method when bio-marker discovery is the main goal of the analysis rather than classification performance.

Table 3: SSE (lower is better) of global feature importance values. $p$-values calculated using Two-Sample $t$-Test w.r.t. Microbe2Pixel.

<table>
<thead>
<tr>
<th>Imp. Method</th>
<th>SSE</th>
<th>$p$-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microbe2Pixel</td>
<td>576 ± 314</td>
<td>N/A</td>
</tr>
<tr>
<td>LIME</td>
<td>938 ± 460</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>SHAP</td>
<td>1550 ± 717</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>Grad-CAM</td>
<td>1338 ± 1498</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>Extra-Trees</td>
<td>988 ± 435</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>XGBoost</td>
<td>1056 ± 425</td>
<td>&lt;1E-4</td>
</tr>
</tbody>
</table>

In clinical practice, it is unfeasible inspect the importances of all features in the dataset. Typically, a top ranking of a set number of features is generated as potential relevant biomarkers. Therefore, the robustness of the methods across the highest ranked features is crucial. To assess this robustness, calculating the SSE with respect to the top 10-200 features for both local and global feature importances is a good approach. This range is often chosen for interpretation of fecal metagenomics data by expert panels. The graphs for local and global feature importances (figures 4 and 5) show that Microbe2Pixel consistently has a lower SSE than other methods, indicating its robustness.

Table 4: SSE of local feature importance values. $p$-values calculated using Two-Sample $t$-Test w.r.t. Microbe2Pixel.

<table>
<thead>
<tr>
<th>Imp. Method</th>
<th>SSE</th>
<th>$p$-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microbe2Pixel</td>
<td>685 ± 142</td>
<td>N/A</td>
</tr>
<tr>
<td>LIME</td>
<td>1107 ± 85.5</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>SHAP</td>
<td>863 ± 85.7</td>
<td>&lt;1E-4</td>
</tr>
<tr>
<td>Grad-CAM</td>
<td>782 ± 104</td>
<td>&lt;1E-4</td>
</tr>
</tbody>
</table>

4.3.4. Qualitative assessment

The cirrhosis dataset was chosen because the difference in the microbial profile between cirrhosis patients and healthy controls is very distinct. In cirrhosis patients, a decrease in various commensals is observed. These decreases are large in the absolute sense (% of reads), yet minor per species in a relative sense ($±1.5$-fold decrease). In contrast, the relative increase of specific pathogenic groups that are normally low in healthy controls increase dramatically. The main biological pattern of relevance can be described as a tenfold increase of orally derived pathobionts and epsilon-toxin producing Clostridia [15]. As most *Prevotella* species (in a Western setting) are orally derived, a similar increase of
many members from this genus is observed as well. Abundances of the few typical native gut *Prevotella*, like *P. copri* and *P. stercorea*, remain relatively unchanged in cirrhosis patients.

Figure 6: Top 25 relative feature importances for each local model-agnostic method on the cirrhosis dataset.

The most important features identified by Microbe2Pixel (figure 6) align well with this same biologically relevant pattern. The top features on a genus level, which correspond accurately with the top features on the species level, are *Prevotella*, *Streptococcus*, *Lactobacillus*, *Enterococcus*, *Leuconostoc*, *Porphyromonas*, *Capnocytophaga*, *Treponema* and *Actinomyces* (see 8). All are well-known oral inhabitants. The paper from which this dataset is derived similarly recognizes this oral signature being associated with cirrhosis [15]. This link between the oral microbial signature and cirrhosis is cemented by the association, that has been known for decades, between periodontal disease and cirrhosis [1]. The direction of this relationship remains to be further elucidated and similar oral signatures have also been associated with Colorectal cancer [7]. The top of the feature list furthermore includes *Clostridium*, which encompasses the Epsilon-toxin producing strains like *C. perfringens*, *C. neonatale* and *C. butyricum*. There are numerous reports linking these potentially life-threatening organisms and/or their (endo)toxins with cirrhosis [2]. It additionally lists various viruses associated with *Enterobacteriaceae*, a sign of dysbiosis in cirrhosis patients and the most abundant (declining) commensal group, *Bacteroides* [22].

The lists of features generated by the other three approaches do not convey the main biological patterns of relevance as clearly. Specifically, the main mistakes are characterized by the inclusion of biologically irrelevant features such as *Paenibacillus* and *Vibrio* in Grad-CAM (extremely low abundance and absent in most samples) and a greater focus on general dysbiosis, e.g. *Enterobacteriaceae* as well as other species associated with the *Bacteroides* enterotype in SHAP and LIME.

In the BARIA cohort, which examines the gut microbiome before and after bariatric surgery, the differences in the gut microbiome are massive and all-encompassing, as surgery completely changes the intestinal ecological situation. As a result, the features obtained from Microbe2Pixel focus on microbial groups that become dominant after surgery and, to a degree, on those that disappear. Specifically, species compatible/associated with the *Bacteroides*1 and *Firmicutes*-enriched enterotypes (Western microbial composition types not associated with poor health) decline and are replaced by species that are either associated with the *Prevotella copri* cluster or the *Bacteroides*2 enterotype. Various Bacteroidales, importantly including *Prevotella*, *Alistipes Parabacteroides* and *Bacteroides* increase in abundance as various *Lachnospiraceae* show a large decrease [9]. In short, the
fermentative capability and complexity seem to diminish after surgery as the gastrointestinal tract is shortened by reducing stomach size and bypassing the duodenum (for most patients). In a sense, the large intestine becomes more like the small intestine from a microbial compositional point of view, as fermentable sources have not been depleted to the same degree as would normally occur in a full-length intestinal tract. These are novel insights that are particularly useful for better understanding of metabolic disorders as well as for clinical and nutritional care provided to the patients.

5. Discussion and Conclusion

The use of Deep Neural Networks (DNNs) in microbiome research has not been as prolific as in other areas of the medical domain due to the modality of fecal metagenomics data. We propose a novel method, Microbe2Pixel, to increase the effectiveness of DNNs with the use of the taxonomy of the microbiota. This method also derives feature importances using a local-model agnostic algorithm that is informed by taxonomy. From the results presented in this work, we can conclude that the incorporation of taxonomy has led to an increase in classification performance compared to no taxonomy, robustness/relevance of feature importances, and interpretability of the method. Despite these advantages, our current method does not beat the state-of-the-art of machine learning in fecal metagenomics research in terms of classification performance. However, when inspecting the robustness of feature importances, Microbe2Pixel performs significantly better. This makes Microbe2Pixel preferential over other methods when machine learning is used to guide research and the development of treatments for disease.

The robustness of Microbe2Pixel with smaller sample sizes makes it particularly attractive for clinical research, as (interventional) trials in microbiome research are complicated to scale up. Moreover, the intuitive nature of the explanations makes Microbe2Pixel easier to understand for medical professionals who are not necessarily skilled in machine learning techniques. This is an imperative condition in the application of machine learning in clinical research.

A number of potential improvements can be made to Microbe2Pixel method. For now,
the taxonomic tree was used to calculate the distance between microbes. However, other methods of calculating distance/similarity between microbial species may be applied as well (i.e., phylogenetic distance, trophic networks, functional pathways, etc.). Besides this, Microbe2Pixel is versatile in regards to the specific DNN. All architectures and model types for images which can be transfer learned are possible to use. As a result, it is likely that the performance of Microbe2Pixel can be improved considerably through the investigation of other methods and hyperparameter tuning.

As this work has demonstrated, informing the similarity of features in the embedding of tabular data in images is a viable way of applying transfer learned DNNs to fecal metagenomics data. Microbe2Pixel significantly improves the validity of model-explanations while still yielding good classification performance. This illustrates the potential of our approach in microbiome clinical studies and similar biomedical research areas with high-dimensional and low-sample-size constraints.

6. Declarations and Data Availability

- **Ethics approval and consent to participate**
  This work does not make use any novel data. Two separate published datasets were used, for which data regarding ethics approval can be found in their respective original papers [15][12].

- **Availability of data and materials**
  The raw Illumina read data for all samples of the Cirrhosis trial have been deposited in the European Bioinformatics Institute European Nucleotide Archive under accession number ERP005860 and are publicly available for download.
  BARIA Fecal metagenomic shotgun data have been deposited in the European Nucleotide Archive (ENA: PRJEB47902). Data from this particular clinical trial data can be requested by any qualified researchers who engage in rigorous, independent scientific research and will be provided following review and approval of a research proposal and statistical analysis plan and execution of a data sharing agreement by the corresponding author.

- **Competing interests**
  MN is in the Scientific Advisory Board of Caelus Pharmaceuticals, the Netherlands. Which is deemed not directly relevant to the current paper. There are no patents, products in development or marketed products to declare. The other authors declare no conflicts of interests.

- **Authors’ contributions**
  BV: conception, design, analysis and creation of new software used in this work, as well as Drafting of the manuscript. MG: analysis and interpretation of the data and drafting of the manuscript. MN: acquisition and interpretation of the data. EL: conception and design of the work, as well as drafting of the manuscript.

7. References


of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, 2016.


[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?" explaining the


8. Appendix A: Top 50 Genus importances cirrhosis dataset

The top 50 feature importances on a genus level for each of the methods are plotted in figure 8. These were used by our expert panel to determine the validity of the highest rated features of each method.

9. Appendix B: Top 50 Genus importances BARI dataset

The genera with the highest ranked feature importances in the bariatrix surgery dataset are shown in figure 9.

10. Appendix C: Parameter grids for Extra-Trees and XGBoost

10.1. Extra-Trees

n_estimators : [100, 300, 800, 1000]
max_depth : [2, 5, 7, None]
min_samples_leaf : [1, 5, 10]
bootstrap : [False, True]

10.2. XGBoost

max_depth : [2, 5, 7]
learning_rate : [0.01, 0.1]
Figure 8: Top 50 relative feature importances obtained from the cirrhosis dataset for each local model-agnostic method on a genus level.

n_estimators: [100, 300, 800, 1000, ]
min_child_weight: [1, 5]
gamma: [0.5, 2]
subsample: [0.5, 0.6, 0.8]
colsample_bytree: [0.6, 0.8, 1.0]
Figure 9: Top 50 relative feature importances for Microbe2Pixel obtained from the BARIA cohort on a genus level.