Advanced Biomarker Discovery in Imaging Mass Spectrometry Through Interpretable Supervised Machine Learning

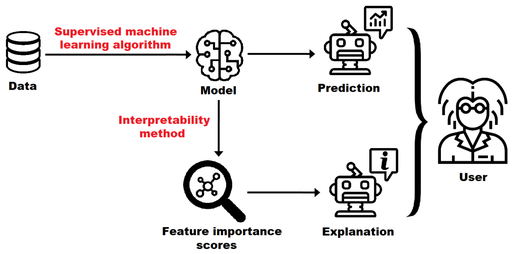
Leonoor E.M. Tideman[1], Lukasz G. Migas[1], Emilio Rivera[2,3], Katerina V. Djambazova[2,4], Elizabeth Neumann[2,3], N. Heath Patterson[2,3], Richard M. Caprioli[2,3,4,5,6], Jeffrey M. Spraggins[2,3,4], Raf Van de Plas[1,2,3]

[1] Delft Center for Systems and Control, Delft University of Technology, Delft, Netherlands
[2] Mass Spectrometry Research Center, Vanderbilt University, Nashville, TN, USA
[3] Department of Biochemistry, Vanderbilt University, Nashville, TN, USA
[4] Department of Chemistry, Vanderbilt University, Nashville, TN, USA
[5] Department of Pharmacology, Vanderbilt University, Nashville, TN, USA
[6] Department of Medicine, Vanderbilt University, Nashville, TN, USA

Introduction

Challenge:
Biomarker discovery in imaging mass spectrometry (IMS) data entails finding, among hundreds of molecular species, a few indicators of a specific biological state (e.g. patho-physiological condition, anatomical structure). Imaging mass spectrometry (IMS) can concurrently map the spatial distribution of thousands of distinct molecular species across the surface of a sample. While this makes IMS particularly well suited for biomarker discovery, it is impractical to manually examine large, high-dimensional IMS datasets in search for highly discriminative molecular species.

Solution:
We propose a novel machine learning (ML) workflow that can automatically narrow down massive lists of potential molecular species to a shortlist of most promising candidate biomarkers. ML interpretability methods computationally estimate the predictive importance of each molecular species with regards to a specific classification task and obtain a ranking of the features in descending order of predictive importance. Ranking the features facilitates the identification of a panel of molecular species that are strongly indicative of disease, and thus have a high likelihood of being good biomarkers.



Icons from https://www.flaticon.com/ and https://icons8.com/

Machine learning model interpretability
Model interpretability is the ability to explain the decision-making process of a supervised ML model by reporting the relative predictive importance of its features [1]. The importance of a feature is the degree to which it influences the model's prediction, considering both its direct effect (i.e. correlation with the prediction) and its indirect effect (i.e. correlation between features).

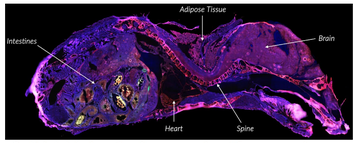
Global interpretability methods investigate the relationship between the features that the model uses as inputs and the model's output, whereas local interpretability methods focus on explaining specific decisions made by the supervised ML model [2]. We use global interpretability methods to discover explanatory principles for how the spatial distribution and relative concentration of certain molecular features relate to the classification of certain regions of a tissue sample. Since biomarker discovery is ultimately about understanding which molecular features drive the biochemical process being modeled, we believe that it is necessary to go beyond the scope of predictive modelling and provide the user with the tools to understand why a supervised ML model makes a certain prediction.

Interpretability methods can be categorized as model-specific or model-agnostic [2] : model-specific interpretability methods derive explanations by examining the model's structure and parameters, whereas model-agnostic interpretability methods treat the model as a black-box and derive post-hoc explanations for its predictions. We focus on model-agnostic interpretability methods that are applicable to any type of supervised ML model.

[1] F. Doshi-Velez and B. Kim. "Towards A Rigorous Science of Interpretable Machine Learning." *arXiv:1702.08608 [cs. stat]*, Feb. 2017.
[2] C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Online book: https://christophm.github.io/interpretable-ml-book/. 2020.
[3] L. Breiman. "Random Forests." *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.
[4] S. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions." *arXiv:1705.07874 [cs. stat]*, May 2017.
[5] S.M. Lundberg et al. "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence*, vol. 2, pp. 56–67, Jan. 2020

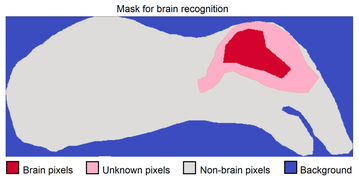
Case study: Classification

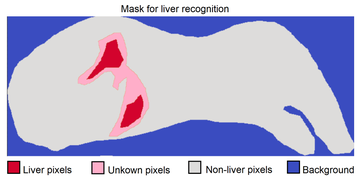
Imaging mass spectrometry data:
Our dataset was acquired by MALDI-TOF MS from a sagittal whole-body section of a mouse-pup using the prototype Bruker timsToF Pro in positive ion mode. A mean mass spectrum of the dataset was generated and peak-picked to produce a feature list of 879 ion species. The m/z acquisition range is 300-1200 and the pixel size is 50×50 µm. The dataset consists of a total of 164,808 pixels. Classification is therefore performed on a matrix of 164,808 instances and 879 features. Our machine learning workflow is implemented in Python 3.7.



Sagittal whole-body section of a mouse-pup by autofluorescence microscopy

Classification task:
Our aim is to recognize the mouse-pup's brain and liver using a one-versus-all classification approach and automatically find biomarkers corresponding to these different organs. We choose to use ensembles of decision trees for the classification of the IMS data because such models are non-linear, computationally efficient, robust to overfitting and scale-invariant. A random forest is used to recognize the mouse-pup's brain and a gradient boosting machine is used to recognize the mouse-pup's liver. The random forest was implemented using the scikit-learn library, whereas the gradient boosting machine was implemented using the xgboost library. Regarding the masks used for training these ML models (see Figure below), the red pixels belong to the brain/liver and are therefore labeled positive; the grey pixels do not belong to the brain/liver and are therefore labeled negative; and the pink pixels cannot be reliably annotated and are therefore excluded from the training and testing sets.





Case study: Interpretability

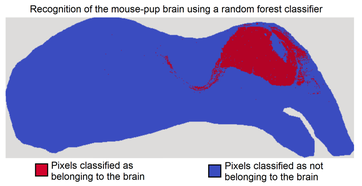
Two model-agnostic interpretability methods, namely permutation importance (PI) and Shapley additive explanations (SHAP), were used to computationally estimate the predictive importance of each molecular species with regards to the recognition of the mouse-pup's brain by a random forest model and the mouse-pup's liver by an XGBoost model, respectively.

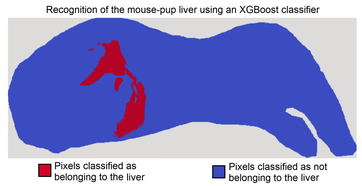
Permutation importance (PI):
PI is a global interpretability method developed by Breiman [3]. The PI of a feature is the average decrease in model accuracy due to randomly permuting its values across all observations. PI measures the degree to which a supervised ML model relies on a specific feature. PI was implemented using the eli5 library.

Shapley additive explanations (SHAP):
SHAP is a local interpretability method developed by Lundberg [4] based on Shapley values from cooperative game theory. SHAP regards the features as players that form coalitions (i.e. ordered subsets) to achieve the supervised ML model's prediction (i.e. payout). The Shapley value of a feature, or local SHAP score, is its contribution to the model's prediction of a specific observation, averaged over all possible feature orderings. SHAP can be used as a global interpretability method for the purpose of biomarker discovery: a global measure of feature importance is obtained by averaging each feature's local SHAP score across all pixels. We use a fast implementation of SHAP for decision-tree based models called TreeExplainer [5], from the shap library.

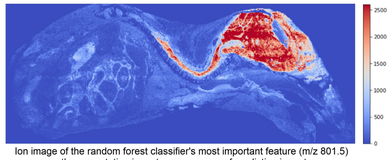
Case study: Results

The classification results are presented hereunder: on the left, a random forest model has recognized the brain (and spinal cord) and, on the right, an XGBoost model has recognized the liver. The two models achieve high predictive performance with 99% accuracy, 99% precision and 98% recall. Note that these measures may slightly overestimate generalization performance because of how the training and testing sets were defined: the red masks do not exhaustively cover all pixels of the target organs. Pixels that are located on the boundaries of the target organs are not included in the training and testing sets because they are difficult to label manually. Yet it is precisely on the boundary pixels that a classifier is most at risk of making erroneous predictions.



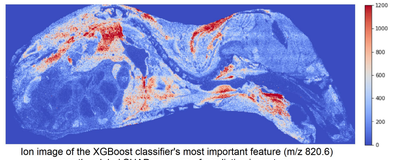


Biomarker discovery by feature ranking:
We translate the problem of biomarker discovery into a feature ranking problem: the molecular features of the random forest and gradient boosting machines are ranked in descending order of relative predictive importance in order to narrow the scope of further clinical investigation from hundreds of candidates down to a short list of highly discriminative features.



Ion image of the random forest classifier's most important feature (m/z 801.5) as per the permutation importance measure of predictive importance

The ion images of two highly discriminative features are presented: the top ion image has a strong influence on the random forest model that recognizes the brain (i.e. maximum PI score), whereas the bottom ion image has a strong influence on the decision-making process of the gradient boosting machine that recognizes the liver (i.e. maximum global SHAP score using TreeExplainer). These ion images are displayed using a pseudo-color scale whose brightness is indicative of the relative molecular concentration measured at a given pixel.



Ion image of the XGBoost classifier's most important feature (m/z 820.6) as per the global SHAP measure of predictive importance

The list of mass-to-charge values hereunder summarizes the results of our ML workflow for biomarker discovery in the mouse-pup dataset. Four top-ranking chemical species are listed by their mass-to-charge ratios in descending order of predictive importance:

- Mass-to-charge ratios of mouse-pup brain biomarkers: m/z 801.5, m/z 740.4, m/z 764.6, m/z 739.4
- Mass-to-charge ratios of mouse-pup liver biomarkers: m/z 820.6, m/z 821.6, m/z 891.6, m/z 892.6

Conclusion
Machine learning interpretability can be an efficient tool enabling us to build a high-performance, user-friendly computational workflow for biomarker discovery in imaging mass spectrometry data. While the case study presented here focuses on the recognition of different anatomical structures, the same approach can be used for disease biomarker discovery. Interpretability methods enable the discovery of explanatory principles for how the spatial distribution and relative concentration of certain molecular features relate to the classification of different regions of the sample.

Global post-hoc interpretability methods, such as permutation importance and Shapley additive explanations, can be used to estimate the predictive importance of each molecular species with regards to a specific classification task and obtain a ranking of the features in descending order of predictive importance. The top-ranking features can then be used as candidate biomarkers worthy of further clinical investigation.

We kindly acknowledge funding from NIH NIGMS 2P41GM103391 and NIH NIAID 1R01AI138581 and NIH NIDDK/OD 1U54DK120058.