AI-POWERED DIAGNOSTICS

Factors Influencing Accuracy, Interpretability, And Reproducibility In The Use Of Machine Learning In Biology

Kaitlyn Martinez Los Alamos National Laboratory **Kristen Wilding** Los Alamos National Laboratory **Daniel Jacobsen** Los Alamos National Laboratory **Apoorv Shanker** Los Alamos National Laboratory **Makaela Montoya** Los Alamos National Laboratory **Trent Lllewellyn** UC Santa Barbara **Jessica Kubicek-Sutherland** Los Alamos National Laboratory **Carrie Manore** Los Alamos National Laboratory **Harshini Mukundan** Lawrence Berkley National Laboratory

The innate immune system is capable of differentiating pathogen types and tailoring the immune response accordingly. Therefore, a promising avenue to enable pathogen-agnostic diagnostics of infection lies in identifying signatures of infection types in innate immune signals. The complexity and variability of the relevant signals motivates the use of machine learning methods, which have increasingly been applied to biological data to understand processes and predict outcomes. Simultaneously, the complexity and variability of such data complicates reliable, reproducible, interpretable, and responsible use of such methods, resulting in questionable relevance of the derived outcomes. It is imperative that we develop machine learning model frameworks that are able to reliably parse signals of biological processes for decision support. Using in vitro innate immune data generated by our team, which characterize innate immune responses to pathogen-associated molecular patterns (PAMPs) associated with various infection types, we developed machine learning models to classify PAMP exposure statuses based on cytokine and chemokine data. We evaluated five classical machine learning classifiers (Random Forest, General Linear Models, Neural Networks, Support Vector Machines, and Naive Bayes) and investigated variable importance methods, cross-validation and hyperparameter tuning, and feature selection. Comparing these classifier frameworks, we systematically explored challenges associated with applying machine learning to predict and understand biological processes, such as the type of biochemical signature (transcripts vs. proteins), data curation methods (pre- and post-processing), choice of machine learning classifier, and class of pathogen. Our results show that, while most machine learning classifiers tested were able to select biologically relevant signals in the in vitro data and use these signals to predict in vitro PAMP exposure with high accuracy, choices in data preparation and classifier type greatly impact both the prediction accuracy and the cytokine and chemokine signatures identified from each classifier. These results suggest that ML classifiers can have a valuable role in identifying signatures of infection, but also motivate the use of standardization of data preparation procedures, validation data sets, and the use of multiple classifier structures to improve the reliability and relevance of these classifiers in informing infection response and agnostic disease diagnosis.