Comparison of machine learning algorithms to identify metabolomics features for predictive modeling of beef color

Ranjith Ramanathan[1], Sathyanarayanan Aakur[2], Anjana Suresh[3], Anupam Abraham[1], Frank Kiyimba[1], Gretchen Mafi[1]

[1] Department of Animal and Food Sciences, Oklahoma State University, Stillwater, OK 74078, USA
[2] Department of Computer Science, Oklahoma State University, Stillwater, OK 74078, USA
[3] Data Analyst, Stillwater, OK 74078, USA

Introduction: Studies have shown that both glycolytic and tricarboxylic acid metabolites influence meat color. The application of untargeted metabolomics helps to understand global changes in metabolites. Bioinformatics tools such as principal component analysis, volcano plot, and other analyses help to characterize the overall changes and statistical significance of individual molecules. However, the traditional statistical analysis makes it difficult to interpret feature (metabolites/proteins/genes) contribution to a quality trait. In recent years, the application of machine learning algorithms allows the identification of important features that contribute to dependent variables such as meat quality. Limited research has utilized machine learning algorithms to know how metabolites contribute to beef color changes. Therefore, the objective was to compare machine learning algorithms to identify metabolomics features for predictive modeling of beef discoloration.

Materials and Methods: Untargeted metabolomics approach was used to identify metabolites profile differences in beef psoas major muscles (n = 20 replications) during three days of storage. Steaks were packaged in aerobic polyvinyl packaging, and the surface color was measured using a HunterLab MiniScan spectrophotometer. The metabolites were separated using gas chromatography-based mass spectrometry. The mass spectra were deconvoluted using AMDIS software, and the metabolites were identified using NIST 2017 library. The metabolomics data were analyzed in R using various packages. Day 0 metabolite data were modeled to predict day 3 redness (a* values). A total of four machine learning algorithms such as random forest, Xgboost, KNearestNeighbor, and support vector machine (SVM) and RBF Kernel were compared to find the best performing model. All data reduction, machine learning, and evaluation of predictive models were performed using Python. The performance of the model is improved using feature engineering and hyperparameter tuning. Different error metrics such as mean squared error (MSE), mean absolute error (MAE), and root mean square error (RMSE) was computed to check the model performance.

Results: Psoas major is a color labile muscle, and the a* values changed from 29.5 to 15.2 during three days of storage. Untargeted metabolomics identified 92 metabolites during three days of storage. Analysis of metabolites using machine learning algorithms resulted in MAE, MSE, and RMSE for random forest (1.62, 3.84, 1.96), Xgboost (1.90, 6.11, 2.47), KnearestNeighbor (2.54, 8.27, 2.87), SVM regressor and RBF kernel (3.09, 1.35, 1.75), respectively. When comparing four models, the random forest algorithm resulted in the lowest error.

Conclusions: Metabolomics can generate a large dataset. The application of machine learning helps to utilize metabolomics data to know features that contribute to discoloration. The current research suggests that among various machine learning algorithms, the random forest provides the best fit to predict color. Incorporating a larger data set can increase the accuracy of prediction by lowering the error.