DLSU Research Congress 2024
De La Salle University, Manila, Philippines
June 20 to 22, 2024

Random Forest Machine Learning for the Geographical Identification of Economically Important Fruit Crop

Christian D. Laurio, Drexel H. Camacho*
Chemistry Department, De La Salle University
**Corresponding Author: drexel.camacho@dlsu.edu.ph*

Abstract: Economically important fruit crops help improve the economy. However, problems in food fraud and origin mislabeling affects the market performance. An innovative solution at the postharvest stage is highly desired. We report herein the use of artificial intelligence to classify and authenticate fruit crops using mango as a test case. Mango samples collected from Guimaras and Zambales, which are known sources of high-value mango fruits were analyzed for its metal ion composition using a validated method of Inductively-coupled Plasma Mass Spectrometry (ICP-MS). Seventeen metals were identified in various concentration. The data were subjected to partial least square–discriminant analysis (PLS-DA) which discriminated the samples distinctly. The scores obtained in the variable importance of projection (VIP) and F analysis agreed that the metals contributed significantly to the origin discrimination identifying the following metals Ni, V, Mn, Ca, Ba, Fe, and Cu as chemical markers. The data were used to develop a Random Forest (RF) machine learning classification model. Results showed that the predictive accuracy of RF model reached 87.5% allowing for a fairly reliable authentication of mango origin. This study illustrates the use of machine learning in determining the fruit's geographical traceability that should be useful in the agriculture sector.

Key Words: Mango, geographical origin, ionomics, metallomics, machine learning

1. INTRODUCTION

Artificial intelligence (AI) describes "*the capability of a machine to perform human-like intelligence functions such as learning, adapting, reasoning, and self-correction.*" (Kingston & Kingston, 1994). Machine learning is a subset of artificial intelligence that deals with "*vast data in the most intelligent fashion by developing algorithms to derive actionable insights where the algorithms are expected to learn without being explicitly programmed*" (https://www.analyticsvidhya.com). In the context of food, machine learning is seen as the way forward for easy authentication and assessment of food quality.

Food origin classification and authentication based on chemical composition is an interesting problem that provides useful information for a variety of purposes, including recognition of geographical origin, authenticity, product characteristics, quality control, preservation, and category differentiation (Maione et al., 2016). It also addresses the issues of food fraud, which involves deliberate alteration, misrepresentation, substitution, or mislabeling of a food product (Smith, 2018). Among the economically important fruit crops in the Philippines, mango is often a victim of fraud considering that superior mango varieties abound, and certain geographical locations offer premium quality that are distinct in terms of taste, flavor, texture, and sweetness. Thus, the quality of mango based on its provenance poses a risk among consumers. How can consumers confirm

the origin claims of mango sellers? There is a need to establish a science-based system to authenticate the geographical origins of economically important fruit crops. Moreover, the Philippine mango industry recognizes the dismal export performance in the world market despite claims of best mango varieties and realizes that there is a lack of strategic and long-term R&D efforts towards innovative technology that can enhance competitiveness. This gap can be addressed by incorporating AI in the agriculture sector.

Fruits, in general, differ from one another by four factors: (1) genetic make-up, (2) farming practice, (3) type of soil, and (4) climate. Because of these factors, fruits obtain a unique fingerprint of elemental and chemical composition depending on the conditions of the location where they originated. The inorganic elemental composition in an organismal system is part of its total uniqueness as they form the mineral, nutrient, and trace element composition of the fruit. Analysis of the metal ion contents in a fruit will generate vast data and is ripe for incorporation of a machine leaning model in the hope of improving the food authentication and food quality assessment in the agriculture sector. This AI approach using the elemental composition data of mangoes is seen as potential solution to address the innovation gap in the mango industry and the agriculture sector, in general. The aim of this work is to identify the macro-, micro-minerals, and trace elements in mango samples and to build machine learning binary classification model using random forest to classify mango samples.

2. METHODOLOGY

2.1 Materials, reagents, and sample preparation

Carabao mango samples were collected from different farms located in Zambales and in Guimaras in the months of April - May 2022. Samples from local markets were likewise obtained. Type I Ultrapure water (18.2 MΩ*cm, Merck Millipore Integral 3, USA), concentrated nitric acid (65% HNO_3, Merck SupraPur), and hydrogen peroxide (30% H_2O_2, Merck) for the digestion of mango samples were obtained. Standard calibration solutions were prepared with high purity ICP multi-elements calibration solution traceable to NIST SRM (Ag, Al, As, B, Ba, Be, Ca, Cd, Ce, Co, Cr, Cu, Fe, Hg, K, Li, Mg, Mn, Na, Ni, P, Pb, Se, Sr, Ti, V, Zn).

The mango fruits were washed thoroughly, dried, and sliced using a ceramic knife excluding the kernel portion. The pulp with the skin was cut into cubes and placed in a plastic container, frozen in an ultralow freezer at -80° C overnight, then lyophilized using a CHRIST Gamma 2016 LSC freeze dryer. The freeze-dried mangoes were grounded using a mortar and pestle one distinct sample at a time. Each powdered mango samples were subjected to microwave-assisted closed vessel acid digestion using Milestone Ethos UP equipped with MAXI-44 rotor vessels (Milestone Ethos UP, Italy) and subjected to Inductively-coupled Plasma Mass Spectrometer (Shimadzu ICPMS-2030 LF ICP-MS, Japan) for the simultaneous metal analysis.

2.2 Machine Learning

To perform multivariate analysis on the metal signatures of the samples such as PCA and PLS-DA, MetaboAnalyst version 5.0 was used. To build, train, cross-validate, and test machine learning classification model i.e Random Forest, open-source Google Colaboratory with Python 3.10.11 was employed. In the training of the machine learning classification model, the dataset was split into two: The train set and the Test set. The Test set was first set aside then X% from the Train dataset was randomly chosen to be the actual Train set and the remaining set (100-X) % was the Validation set, where X is a fixed number (i.e., 80%). The model was then cross validated by iterative training and validation on these different sets. The training set was used to generate multiple splits of the Train and Validation sets (Shah, 2017).

Random forest (RF) is a tree-based supervised machine learning algorithm that is based on ensemble of decision trees. It combines decision trees and then trains each on a different sample of observations (David, 2020). To build an RF model, using Google Colab, random samples were selected from the metal data set of Guimaras and Zambales mangoes. From these selected random samples, decision trees were created, and each decision tree gave its prediction. Voting was then performed for every predicted result. The algorithm selected the average of the most voted result as the final prediction. To evaluate the performance of the model, criteria such as accuracy, sensitivity, specificity, and area under curve (AUC) value were evaluated. To achieve more stable predictions and due to the limitations brought by the small size of the dataset, 10-fold cross-validation was performed in the machine learning classification model. To evaluate the classification performance of the trained machine learning models, the Test Set,

which is the other portion of the metal dataset generated from the mango samples purchased from the local market with claims of provenance, was subjected to the trained model. Each prediction was tallied in a Confusion Matrix and the model's accuracy, sensitivity, and specificity were calculated as well as the graphs of the area under curve (AUC) of receiver operating characteristics (ROC) analysis.

3. RESULTS AND DISCUSSION

The mean values of the metals determined from the mango samples using the validated method of ICPMS is shown in Table 1. The reported values were moisture-corrected by 14 – 16% dry weight. The %water was 82-87 % for Guimaras samples and 73 - 83% water for Zambales samples.

Table 1 Average concentration of the metals in mango samples between the two locations.

Ionomes	Guimaras (n=35)	Zambales (n=35)	p-value
Macro-minerals			
K (mg g^{-1})	15.6	13.5	0.174
Mg (mg g^{-1})	1.32	0.989	0.000
P (mg g^{-1})	0.801	0.809	0.399
Ca (mg Kg^{-1})	195	74.1	<0.001
Na (mg Kg^{-1})	36.3	57.0	0.289
Micro-minerals			
Ba (mg Kg^{-1})	1.01	0.222	<0.001
B (mg Kg^{-1})	6.36	6.18	0.135
Cu (mg Kg^{-1})	7.41	5.85	<0.001
Fe (mg Kg^{-1})	6.93	8.18	0.008
Mn (mg Kg^{-1})	19.6	3.63	<0.001
Ni (mg Kg^{-1})	0.136	2.50	<0.001
Sr (mg Kg^{-1})	4.65	2.67	0.018
Zn (mg Kg^{-1})	5.49	4.76	0.005
Trace Elements			
As (µg Kg^{-1})	11.5	8.95	0.001
Cd (µg Kg^{-1})	5.92	7.21	0.004
Pb (µg Kg^{-1})	12.6	29.6	0.002
V (µg Kg^{-1})	55.1	99.3	<0.001

The differences between the metal concentrations in Guimaras and Zambales mango samples could be attributed to their different geographical features. Guimaras is a small, secluded island between the islands of Negros and Panay. Zambales, on the other hand, has a long coastal line facing the West Philippine Sea and has mountainous range and volcanic feature on its eastern to southeastern side. Previous volcanic activities and eruptions of Mt. Pinatubo located in Botolan, Zambales, may have contributed to the elevated concentrations of these trace elements. The coastal nature of both locations may be a contributory factor as the metal cations from the sea may have migrated into the farmlands.

Using MetaboAnalyst 5.0 and the preprocessed metal data from both locations afforded a PLS-DA scores plot with distinct clusters of Guimaras and Zambales groups clearly defined The Zambales group, however, was more spread out than the Guimaras group and percent of variation for PC1 is 50.6% while PC2 is 13.2%.

Figure 1 shows the score values assigned to each metal, which represent their relative importance following the F-scores calculation as described by Maione et al. (2016). In this work, nickel has the highest score among the rest of the metals, which means that this has the largest discriminating power for the mango samples and thus was a very important candidate for the classification model learning process. It was followed by Ca, Ba, V, and Mn, which also have relatively high F-scores. It was in good agreement with the VIP scores generated by PLS-DA. Ions such as P, As, and B with very low F-scores have relatively no discriminating power and were less important for training classification models. Hence, the metals Ni, Ca, Ba, V, Mn, Fe, and Cu were the variable features used in training and testing the RF machine learning classification model.

The model development was built using an open-source Colaboratory by Google Research running in Python 3.10.12 programming language built on Jupyter Notebook. The library for this implementation was scikit-learn or sklearn libraries for machine learning classification together with other Python libraries such as NumPy and Pandas for statistical data analysis and, matplotlib and seaborn for data visualization.

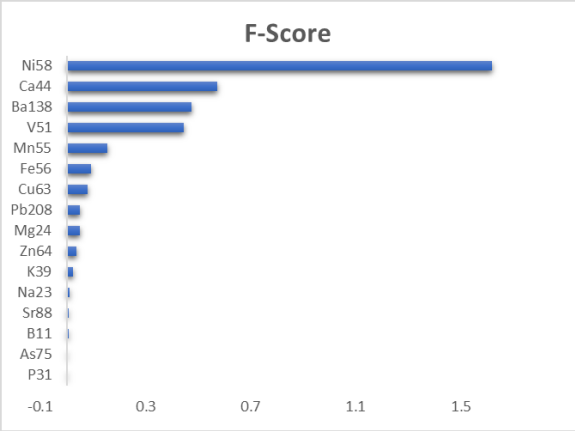

Figure 1 Calculated F-scores for relative variable importance

The dataset was split into training-validation set and test set that were used to train-validate and evaluate the model, respectively. Preprocessing of both sets were done using MinMax Scaler and Standard Scaler normalization to reduce the magnitude of the data. A 10-fold cross validation was implemented during the training phase. The variables used as a feature in this model were the top seven metal ions (Ni, Ba, Ca, Mn, V, Fe, and Cu) based on the F-scores. The class that was used to train and test RF model was Random Forest Classifier from sklearn.ensemble with minimum sample leafs = 5, max depth = 5, n estimators = 10. During the training-validation phase, the model was trained with 100 observations and was able to predict the validation set with 95.0% accuracy and out-of-bag score of 96.25%. Using the test dataset (data points that were not seen by the model during training phase) with 20 observations for both locations, it was able to predict classification with 87.5% accuracy and precision of 84.6% for Guimaras class and 90.0% for Zambales class. A plot tree of how the trained model came up with its prediction is outlined in Figure 2. The "tree" contains three layers of node with one root node, two middle nodes and four-leaf nodes at the last layer that give its decision based on its threshold value. Gini impurity values per node indicate the probability of choosing a metal ion times the probability of being misclassified. An average of these trees comprised the decisions made by the RF model.

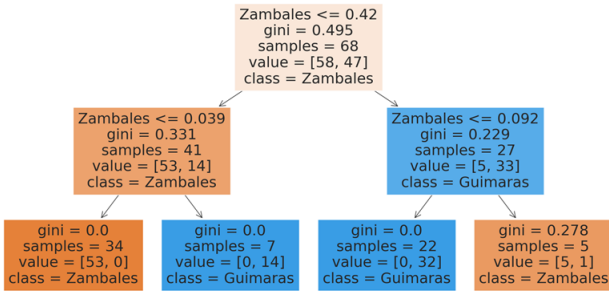

Figure 2 Plot tree of the RF model trained from the mango metal dataset

The RF model's predictive performance was evaluated using confusion matrix and computed for accuracy, precision, recall, and f1-score. During the training-validation phase of model building, the RF model achieved 95% accuracy, 100% precision, and 83% sensitivity. To assess the model, an independent test set was used that contained an equal number of Guimaras and Zambales datapoints (12 observations each), where the RF model performed fairly well in the confusion matrix (Figure 3).

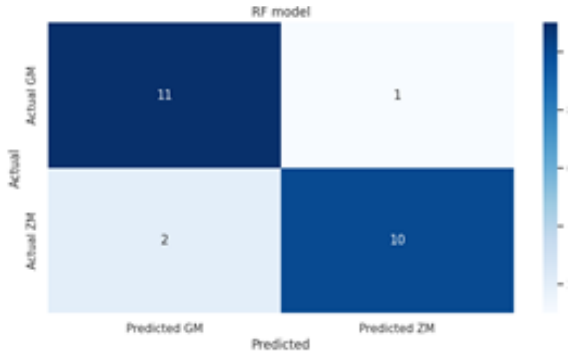

Figure 3 Confusion matrix from the testing phase of RF model

The performance of RF in discriminating mango samples between Guimaras and Zambales based on their metal composition was further investigated by plotting the model's receiver operating characteristic (ROC) and the area under the curve (AUC). ROC-AUC is often calculated to measure the overall performance of the machine learning model. ROC is a plot of true positive rate (TPN) vs false positive rate (FPN). In this study, the true positive rate and false positive rate refer to the rate of predicting true Guimaras and false

Guimaras, respectively. The AUC values of RF model (Figure 4) for predicting Guimaras and Zambales was 0.91. The closer the AUC value to 1, the better the model in discriminating between two classes. Based on the plot, RF model performed well in predicting mango samples from Guimaras and Zambales.

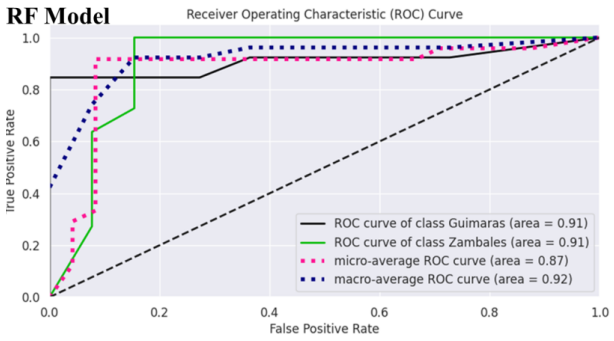

Figure 4 Plot of the receiver operating characteristic (ROC) and area under the curve (AUC) of the trained RF machine learning model

The RF machine learning classification model was also trained and evaluated using subsets of the important metal features as well as subsets containing minerals only and trace elements only (Table 2). From the seven important metal ion variables where the models were initially trained and tested, each of the following subsets contained one less variable and used to train and test using the same train and test datasets. The RF model performance diminished significantly as variables were removed at each iteration of the subset. It was also observed that train accuracy is always better than test accuracy at each subset. This indicates that the RF model is overfitting during training phase. The same is true for mineral and trace elements subsets.

The performance of the model, however, decreased significantly when Ni was the only variable feature of the model. One of the highlights of the study is the authentication of the mango samples based on geographical origin. This can supplement ways in the identification and certification of mangoes in the Geographical Indication (GI) system. Guimaras mangoes, being one of the first registered GI mangoes in the Philippines, can benefit from this result supplementing their claim by metallomics data.

Table 2 Evaluation of model's accuracy using different subsets of metal ions

	Models	RF	
Subset	Variables	Train	Test
1	Ni, Ca, Ba, V, Mn, Fe, Cu	95.0%	87.5%
2	Ni, Ca, Ba, V, Mn, Fe	95.0%	79.2%
3	Ni, Ca, Ba, V, Mn	100.0%	83.3%
4	Ni, Ca, Ba, V	95.0%	83.3%
5	Ni, Ca, Ba	90.0%	66.7%
6	Ni, Ca	95.0%	83.3%
7	Ni	70.0%	50.0%
Mineral	Ca, K, Mg, Na, P	85.0%	70.8%
Trace	As, B, Ni, Pb, Zn	100.0%	75.0%

4. CONCLUSIONS

Using a validated method of ICP-MS, 17 trace and mineral metals in mango samples from Guimaras and Zambales farms were determined with different degrees of correlation among the locations. The seven characteristic ionomes screened by F-Score calculations were Ni, Ba, Ca, Mn, V, Fe, and Cu. The PLS-DA model attests to these characteristic ionomes, and the feasibility of using metal fingerprinting for the traceability of mango origin from the two locations. In addition, Random Forest machine learning classification algorithms was developed to classify mango samples based on their metal compositions.

Further studies are recommended utilizing the mango fruit metal composition data using other machine learning models such as Support Vector Machine and Artificial Neural Network/Multilayer Perceptron, among others. Moreover, correlations with soil nutrition and farming practices are recommended. New models that are physiologically sound, coherent, and statistically unbiased are recommended to be explored to generate practical applications in agronomic farming of mangoes.

5. ACKNOWLEDGMENTS

CD Laurio would like to acknowledge the DOST-Science Education Institute (DOST-SEI) through the Accelerated Science and Technology Human Resource Development (ASTHRDP) program for the graduate scholarship grant, and to the De La Salle University-Central Instrumentation Facility (DLSU-CIF) for the Graduate Thesis Residency Program. The DLSU-CIF is acknowledged for the use of the instruments. We wish to acknowledge Dr. Anna Karen Laserna, Ms. Alona Intac, Mr. Jude Rolan Dela Cruz, and Ms. Jessica Monterde for their assistance in conducting the experiments. We wish to acknowledge Mr. Gerald Dicen, Mr. Kurt Louis Solis, Ms. Mica Binongo, and the rest of the Mango Project team and staff of the Agriculture Research Section of the Department of Science and Technology – Philippine Nuclear Research Institute (DOST-PNRI) for sharing Zambales mango samples. To Mr. Marvin Ayende for the mango samples collected in Masinloc, Zambales. We also acknowledge the Guimaras Mango Growers and Producers Development Cooperative and its chairperson, Mr. Felipe Z. Gamarcha, for the sample collection in Guimaras Island.

6. REFERENCES

David, D. (2020). Random forest classifier tutorial: how to use tree-based algorithms for machine learning. freeCodeCamp. Retrieved from https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/amp/

Kingston, H. M., & Kingston, M. L. (1994). Nomenclature in laboratory robotics and automation (IUPAC Recommendations 1994). Pure and Applied Chemistry, 66, 609–630. https://doi.org/10.1155/S1463924694000040.

Maione, C., Batista B.L., Campiglia, A.D., Barbosa Jr., F., Barbosa, R.M. (2016). Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Comput. Electron. Agric. 121, 101-107. https://doi.org/10.1016/j.compag.2015.11.009

Shah, T. (2017, December 7). About Train, Validation and Test Sets in Machine Learning [Blog Post]. Retrieved from https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7.

Smith, M. (2018). Food Fraud. In: Smith, M. (Ed.), Food Safety and Inspection, An Introduction, 1st Ed. Taylor & Francis Group, London. pp.76-93. Retrieved from https://www.researchgate.net/publication/347883258_Food_fraud