eFood
Vol. 1(1); February (2020), pp. 69–76
DOI: https://doi.org/10.2991/efood.k.191004.001; eISSN 2666-3066
https://www.atlantis-press.com/journals/efood

Research Article

Random Forest, Artificial Neural Network, and Support Vector Machine Models for Honey Classification

Cecilia Martinez-Castillo[1,2], Gonzalo Astray[1,*], Juan Carlos Mejuto[1], Jesus Simal-Gandara[2,*]

[1]*Department of Physical Chemistry, Faculty of Food Science and Technology, University of Vigo - Ourense Campus, Ourense E32004, Spain*
[2]*Nutrition and Bromatology Group, Department of Analytical and Food Chemistry, Faculty of Food Science and Technology, University of Vigo - Ourense Campus, Ourense E32004, Spain*

ARTICLE INFO

Article History
Received 14 May 2019
Accepted 29 September 2019

Keywords
Food authenticity
honey
Galician honeys
classification models

ABSTRACT

Different separated protein fractions by the electrophoretic method in polyacrylamide gel were used to classify two different types of honeys, Galician honeys and commercial honeys produced and packaged outside of Galicia. Random forest, artificial neural network, and support vector machine models were tested to differentiate Galician honeys and other commercial honeys produced and packaged outside of Galicia. The results obtained for the best random forest model allowed us to determine the origin of honeys with an accuracy of 95.2%. The random forest model, and the other developed models, could be improved with the inclusion of new data from different commercial honeys.

1. INTRODUCTION

In the past years, the quality of food products has been an important characteristic for consumers [1]. The definition of honey explains that it is a natural product which is produced by *Apis mellifera* (honey bees) from the nectars of different kinds of plants [2–4] or other secretions [3,4] and that has high viscosity, sweetness, and a particular aroma [5]. Honey can be categorized as blossom honey or honeydew honey [4]. Honey is composed of different sugars and other elements such as enzymes, organic acids, vitamins, or aromatic substances, among others [6].

Honey is a product widely consumed due to its health benefits [2]. Honey's composition and properties depend on the botanical origin of the nectar, or secretions, used by the bees during honey production [4,7]. Honey presents in its composition phenolic constituents that give it anticarcinogenic, immune-modulating, and analgesic properties, among others [4]. These properties, together with their anti-inflammatory, antimicrobial, and antioxidant effects, make honey a very valuable food product [4]. Besides this, their specific botanical sources and their geographical origins imply, in a large number of cases, a higher price due to their properties (pharmacoactive or organoleptic) [8]. In this sense, European Union safeguards their foods under many ways [8], such as: (i) protected designations of origin or (ii) protected geographical indications, *inter alia* [9]. Due to its limited availability, and its high price, there is high probability that honey is a product that can be adulterated [6].

Fraud is becoming an increasing phenomenon to report an extra profit, hence the use of trustworthy control methods are necessary to limit, even eliminate, the risks of falsification [10] to ensure the food authenticity. The most common methods to adulterate honey is through addition of cheap sweeteners (corn syrup and maltose syrup, among others) or through use of honeybees that are fed sugar or other types of sucrose [6,11,12]. These two methods are in line with the assertion of Cotte et al. [10] that also reports another method of fraud consisting of misuse of the name of origin by mixing (voluntarily or not) different honeys of diverse varieties.

Different techniques to control and combat the adulteration of honey were reported by Cotte et al. [10]: (i) pollen analysis, (ii) method based on stable carbon isotope ratio analysis, or (iii) method based on site-specific natural isotope fractionation–nuclear magnetic resonance (SNIF–NMR) of ethanol deuterium. Nevertheless, there are important limitations with these two isotopic methods, hence it is necessary to consider new parameters to use in control methods to ensure honey authenticity [10]. In this sense, Cotte et al. [10] proposed a gas and liquid chromatography method combined with principal components analysis to detect the addition of different kinds of sugars. Other methods that do not require the specific compound identification can be interesting as the specific composition depends on the honey category [13]. For this reason, Azevedo et al. [13] extracted the protein to discriminate honeys using principal component analysis. Honey's protein content depends on different factors, including the species of the bee that produced the honey; in fact, honey from *Apis cerana* presents an amount between 0.1 and 3.3%, whereas honey from *A. mellifera* contains proteins in the range of 0.2–1.6% [6,14,15], although the normal content is <0.5% [16]. It is clear that sugars and proteins can be used to categorize the honey type or sugar addition. It is also possible to classify the origin of honey using proteins, aroma compounds, and pollen analysis, among other components [17].

*Corresponding authors. Email: gastray@uvigo.es; jsimal@uvigo.es
Peer review under responsibility of the *International Association of Dietetic Nutrition and Safety*

In this research three kinds of models were developed to differentiate between Galician honeys and commercial honeys: (i) a Random Forest (RF) model, (ii) an Artificial Neural Network (ANN) model, and (iii) a Support Vector Machine (SVM) model.

(i) RF is a method that can be used for classification or regression purposes [18,19], which was introduced in 2001 by Breiman [20]. RF is a powerful prediction method [21] that can be applied in multiple research fields, such as social sciences [22,23] or environmental sciences [24,25], among others.

(ii) ANN models are a computational technique inspired by the biological neural system [26,27] and can be used for different purposes such as prediction, clustering, or pattern recognition [26]. ANNs are formed by artificial units called neurons [26–28]. ANN is formed by input, intermediate, and output layers [29]. ANN models used in this research are based on multilayer perceptron that is a supervised network requiring desired output for each case of study [30]. Artificial neural models are good models for systems with incomplete data, fuzzy information, and complex and ill-defined problems [31]. ANNs are able to find complex relationships between input and output variables [30]. Owing to these advantages, ANNs can be used in different areas, such as environmental sciences [26,27], energy fields [28,32], food technology [30,33,34], or chemistry [35,36], *inter alia*.

(iii) SVM was proposed in 1992 by Boser et al. [37] for classification problem [38]. SVM is a supervised learning method [38,39] that can construct a hyperplane to separate data into many classes [38,39], even a group of hyperplanes, which can be used for different tasks such as regression or classification [18,39]. SVM models present an advantage in comparison with other methods, for example, partial least square-discriminant analysis, to model classification of nonlinear problems [39]. As a result of this advantage, the SVM can be applied in different research areas such as agricultural sciences [38,40], medicine [41,42], or Economics [43,44], *inter alia*.

In this research paper, different separated protein fractions by the electrophoretic method in polyacrylamide gel [45] obtained by Rodríguez-Otero et al. [46] were used to classify two different types of honeys, Galician honeys and commercial honeys produced and packaged outside of Galicia. Therefore, the main aim of this research is to develop a prediction model as a tool for honey authenticity between Galician honeys and the rest of commercial honeys.

2. MATERIALS AND METHODS

2.1. Data Set

The commercial honeys were purchased in Santiago de Compostela and the Galician honeys (belonging to the four Galician provinces) were provided to Rodríguez-Otero et al. [46] by the Regional Centre for Agrarian Extension of Santiago de Compostela. In this research, a total of 104 multifloral honeys have been used. Of these, 82 are Galician honeys and 22 are commercial honeys produced and packaged outside of Galicia.

The gel used to carry out the electrophoretic separation was prepared from four working solutions (two buffer solutions, one gel monomer solution, and one ammonium peroxydisulfate solution). The electrophoresis tubes are loaded with the gel and 25 mg of sample, and an electrical current of 1.5 mA per tube was applied. The intensity of the electrical current is increased up to 3 mA per tube when the penetration of the tracer dye into the tubes is observed. When the bromophenol blue reaches the lower edge of the tube (about 5 mm from the edge), the electrical current is turned off, the gels are removed, and they are stained and excess dye is removed by washing. Afterward, the gels are scanned. For more details consult the complete procedure by Rodríguez-Otero et al. [46].

Twelve different fractions have been found according to their relative mobility; nevertheless, these fractions were not found in all Galician and commercial honeys [46]. The relative mobilities were measured using as reference the distance covered by bromophenol blue [46].

According to the experimental work carried out by Rodríguez-Otero et al. [46], the most frequent band present in the honeys analyzed were: seven, eight, eleven, and twelve, which present relative mobilities between 18.6 and 68.9 for the Galician honeys and between 17.1 and 63.6 for the commercial honeys. These four bands were used to develop different prediction models in this article.

2.2. Methodologies

According to the purpose of this research, it is possible to find in the research literature the prediction models (RFs, ANNs, and SVMs) used in this research but applied to different fields related to the honey.

It is possible to find research papers about RF used with visible/near infrared (VIS-NIR) hyperspectral imaging to classify different honey types based on floral origin and compare its results with other methods such as radial basis function network, principal component analysis, or SVM [47]. It is also possible to find research on neural models to authenticate honey samples using rheological and physicochemical parameters (comparing its performance with other models such as principal component analysis and linear discriminant analysis) [5]. It is also possible to compare ANNs with other techniques such as cluster analysis, principal component analysis, Bayesian method, and partial least-squares regression [17] to differentiate Galician honey from non-Galician honeys. Finally, ANN can be used to authenticate honeys labelled as "Corsica", a European protected designation of origin [48] or to predict the botanical origin of honeys (monofloral or multifloral) using chemical and physical parameters [49]. Finally, SVMs have been used to detect adulterants in honey using near-infrared spectroscopy and then it was compared to other methods such as ANNs or linear discriminant analysis [50]. SVM can be used combined with near infrared spectroscopy to predict the botanical origin of honey samples [51] or using an electronic nose, electronic tongue, and spectral analysis to evaluate raw honey samples [52].

The RF is an algorithm where many decision trees are developed using bootstrap cases from the database used for training [53]. Each decision tree is developed using a subset of independent characteristics while the training phase takes place [47]. Each of these trees represents an individual classifier conforming, together, an ensemble classifier [53]. In comparison with a single decision tree, an RF model achieves better precision values [53]. To find the best RF model must be tested not only the number of trees but also the maximum depth, the criteria for attributes selection (criterion), among other parameters.

Different ANN topologies with different hidden layer configurations were developed using trial and error method to determine the neurons in the output layer [30]. ANN models have two contour layers: (i) a first layer or "input layer" where the experimental values are introduced and (ii) a second layer called output layer where the predicted values are generated. Between these two layers exist one or more layers called hidden layer or layers.

During the neural training phase, different parameters minimize the errors between the input and the predicted variable [54]. The learning process occurs in the intermediate and output layers. To find the best model, it is necessary to use the trial and error approach where different topologies and training cycles are analyzed.

In this research, ANN models were developed using the backpropagation algorithm and the sigmoidal function in the hidden and output neurons. These models consume time and computational resources to optimize the parameters involved in the learning process [55,56].

Finally, the last model was the SVM model. SVM is a powerful technique used for regression and classification [18]. In our case, it was used to classify tasks using C-support vector classification (C-SVC) type for classification tasks. SVM models find an optimal hyperplane to obtain a good separation and maximize the decision surface limit [55]. The learner of the SVM proposed by Chang and Lin, LibSVM [18,57,58], was used in this research. The parameters used were chosen according to the updated guide "*A Practical Guide to Support Vector Classification*" [59].

2.3. Statistical Analysis

The analysis of data reported by Rodríguez-Otero et al. [46] was carried out by means of a Trial/Free version of RapidMiner Studio from RapidMiner Inc. This software was used to develop the different models, and to fit and to plot the results. All models were implemented in an Intel Core i7-8700 processor 3.20 GHz with 16 GB RAM.

3. RESULTS AND DISCUSSION

All models were developed using trial and error method to find the best model configuration. In this sense RF was implemented using: (i) the number of trees (1–100 in 99 steps with linear scale), (ii) the criterion (gain ratio, information gain, Gini index, and accuracy), (iii) maximal depth (−1 to 100 in 101 steps with linear scale), (iv) apply pruning (true or false), and (v) apply pre-pruning (true or false). The best RF model was chosen according to its validation accuracy. ANN was developed with different: (i) topologies (varying the number of hidden neurons between 1 and $2n + 1$, with n being the number of input variables), (ii) training cycles, (iii) learning rates (0.1, 0.2, and 0.3), (iv) momentum (0.1, 0.2, and 0.3), and (v) decay (true or false). The value range of the attributes was automatically normalized between −1 and 1 by the neural net operator. The best neural network model was chosen according to its validation accuracy. SVM was developed with different: (i) type (C-SVC), (ii) gamma values (2^{-15} to 2^{3} in 36 steps with a logarithmic scale), and (iii) C values (2^{-5} to 2^{15} in 40 steps with a logarithmic scale). The SVM model was chosen according to its validation accuracy. Each input variable used for the SVM model was normalized between −1 and 1 for training phase, then this normalization was applied to validation and querying phases. Once the best model of each approach has been chosen, the final model is selected based on its accuracy for validation and training phase. Finally, the chosen model is tested with the querying group.

3.1. Models Implemented with Four Input Variables

To identify honey, the model must be accuracy to reduce material costs and save time. To develop the models with four input variables, honeys with bands seven, eight, eleven and twelve were selected. In this research, a total of 104 honeys have been used. Of these, 22 are commercial honeys and 82 are Galician honeys. These data were divided into three groups, the first group was to train the models (52 Galician honeys and 10 commercial honeys), the second group to validate the models and choose the best model (14 Galician honeys and seven commercial honeys), and finally, the third group to query and check the correct prediction of the selected model (16 Galician honeys and five commercial honeys). The model's predictive power was determined as a function of the accuracy in the validation phase.

Table 1 shows the adjustments of the different models developed with four input variables. The seven, eight, eleven and twelve bands of the electrophoretic gel were selected as input variables to obtain the predictive models. Values are presented as a percentage of the accuracy value for training (T), validation (V), querying (Q) phase, and overall phases (O).

At first, it can be seen how the different values of accuracy are homogeneous among them (Table 1). It can be seen that all models present the same accuracy for the validation phase (100.0%). If we take into account the accuracy provided in the training phase, it can be concluded that the artificial neural network (ANN_1) and the support vector machine (SVM_1) models present the worst accuracy (93.5%).

The prediction errors obtained, during the training phase, by the ANN_1 and SVM_1 models are due to the low classification power of commercial honeys, where only six of 10 commercial honeys are correctly classified (60.0% of accuracy). In the case of Galician honeys, both models predict with total accuracy of 100.0%. The RF model (RF_1) is the best predictive model according to validation (100.0%)

Table 1 | Accuracy (%) for training (subscript T), validation (subscript V), querying (subscript Q), and overall (subscript O) phases for each model developed with four input variables: Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM)

Model	B7	B8	B11	B12	$Accuracy_T$	$Accuracy_V$	$Accuracy_Q$	$Accuracy_O$
RF_1					95.2	100.0	90.5	95.2
ANN_1					93.5	100.0	90.5	94.2
SVM_1					93.5	100.0	90.5	94.2

and training phase [95.2%, where of the 52 Galician honeys, 51 are correctly classified (98.1%) and for the 10 commercial honeys, eight are correctly classified (80.0%)]. All models present an accuracy of 100.0% in the validation phase, therefore, there are no significant differences between them. Related to the querying phase and using the best model selected, RF (RF_1), the querying phase presents an accuracy of 90.5% [where all the Galician honeys are correctly classified and only three of the five commercial honeys (60.0%) are correctly classified].

It can be concluded that the best model chosen according to both accuracies (validation and training) is the RF model, which presents an accuracy of 90.5% in the querying phase. In overall phases, the RF model presents an accuracy of 95.2%.

In Figure 1 we can see the behavior of the best predictive model developed with four input variables, the RF model. In this sense, Figure 1 shows the 21 cases for the validation phase, 14 cases correspond to Galician honeys and 7 to commercial honeys. The predictive model classifies all the honeys perfectly but the behavior of each prediction must be analyzed. It can be seen that most of the Galician honeys (10 samples) have a high confidence level (around 98.5%). Besides these 10 honeys, another four Galician honeys present confidence levels between 63.7% and 68.2%. Despite these relatively low confidence level values, in comparison with the 10 Galician honeys, it can be concluded that the faith in the prediction is good. For the commercial honeys (seven cases), the confidence level varies, for six honeys it is between 65.1% and 97.6%. The remaining commercial honey presents a prediction confidence level close to 51.5%. If the last case is not taken into account then the RF_1 model can predict with high confidence level for both types of honeys.

Figure 2 shows the 21 cases of querying phase of the RF model among which 16 cases are Galician honeys and the remaining five cases are commercial honeys. In this case, not all honeys are classified correctly. All Galician honeys are classified correctly with a confidence level above 98% with three exceptions where the confidence level reaches a maximum around 81.9% as opposed to the other cases where the confidence level reaches a 98.5%. The five commercial honeys present a very contrasted classification (three correct cases and two incorrect). Within the correct cases, two

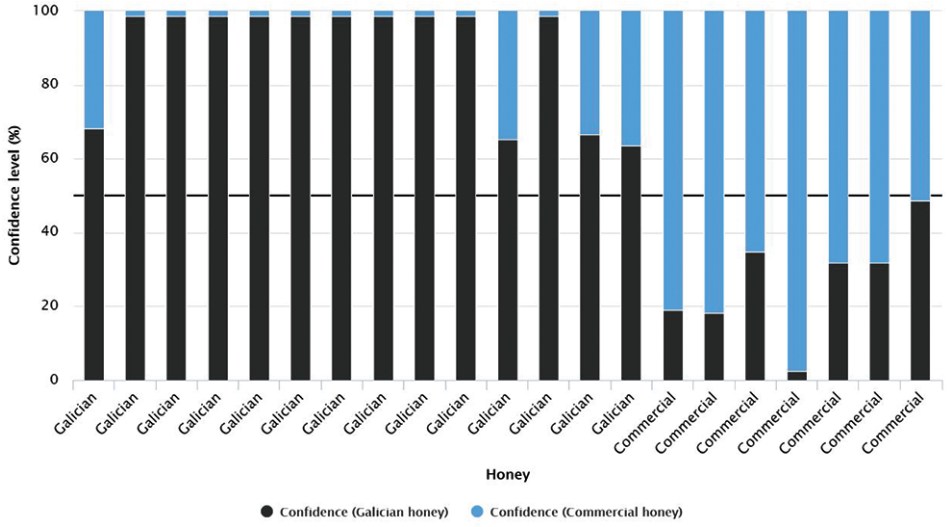

Figure 1 | Bar graph for validation cases according to the confidence value (%) of each prediction for the random forest model with four input variables.

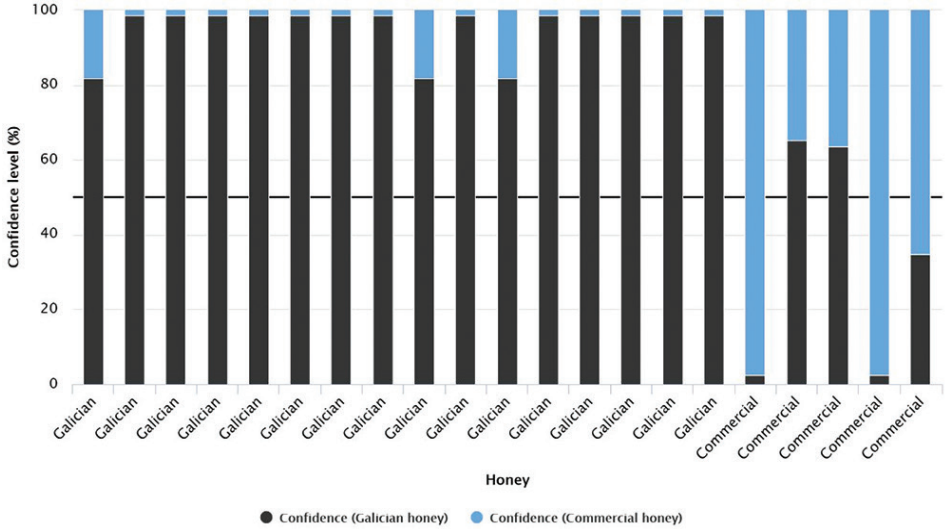

Figure 2 | Bar graph for querying cases according to the confidence value (%) of each prediction for the random forest model with four input variables.

commercial honeys present a confidence level of around 97.6%, whereas the other case presents a confidence of 65.1%. For the misclassified honeys the confidence level varies between 34.8% and 36.3%. According to these results, it can be concluded that the RF model developed with four input variables presents low prediction confidence for commercial honeys.

3.2. Models Implemented with Two Input Variables

It has been decided to simplify the model of four input variables for a simpler model and to compare our results with the model obtained, with the bands eleven and twelve, by Rodríguez-Otero et al. [46].

Table 2 shows the adjustments of the models developed with two input variables. In these models, developed with bands eleven and twelve, the adjustments present greater heterogeneity in comparison to models developed with four bands. In this case, the SVM_2 model has the worst accuracy value in line with the adjustments obtained for the validation phase [90.5%—13 Galician honeys (92.9%) and six commercial honeys (85.7%) were classified correctly]. This model presents a relativity low accuracy for the training phase [90.3%– due to the low classification power for commercial honeys, where only four honeys are correctly predicted (40.0%)]. For the querying phase, the accuracy improves slightly over the accuracy of the training phase, but still remains low (90.5%), which suggests that SVM developed with two bands is not good enough to be used in the food authenticity field. This low predictive power in the querying phase is due to the low classification of commercial honeys, where only three of the five honeys are correctly classified (60.0%).

The second-best model, in agreement with the values obtained for the validation phase, is the model based on ANNs (Table 2). This model, developed with two input variables, provides good accuracy value for the validation phase of around 95.2% (where 13 of the 14 Galician honeys and 100% of the commercial honeys have been correctly classified), but this accuracy descends quickly in the training phase where the value drops to 87.1%. This decrease in the accuracy in training phase is due to the poor classification of the model for Galician honeys (92.3%) and, above all, for commercial honeys (60.0%). For the querying phase, the ANN_2 model provides a relatively high accuracy value of 95.2% (where only one commercial honey is incorrectly classified). These accuracy values for training and validation phases suggest that the neural model developed with two bands is not a good model to food authenticity.

Finally, the model with the best accuracy for the validation phase is the RF developed with bands eleven and twelve. This model obtains the best result for the validation phase, reaching an accuracy value of 100%, which means that all honeys have been correctly classified by the model (Table 2). This fact is reinforced with the adjustment in the training phase where the adjustment is kept very high (96.8%, with only one honey incorrectly classified for each type). These two accuracy values are high and, consequently, we understand that the RF model can be a very useful tool for food authenticity on honeys. In fact, this is confirmed by the high-accuracy value for querying phase where it reaches 95.2% (only one commercial honey is misclassified). To our understanding, the values obtained from the RF model developed with two bands (95.2% for querying phase and 97.1% for overall phase) make this prediction model a good model to be considered for ensuring food authenticity in honeys.

Figure 3 shows the 21 honeys used for validation phase (14 Galician honeys and seven commercial honey). This RF model classifies correctly all honeys with good confidence values for Galician honeys (around 99.4%), although there are three cases that present a lower confidence level, two of them close to 57.0% and 52.5%. On the right side of Figure 3, it can be seen that the commercial honeys are correctly classified but with lower confidence level (within a range of 54.0–78.1%) than the Galician honeys.

Finally, Figure 4 shows the honeys used for querying phase. All the Galician honeys are correctly classified (16 out of 16), thus, honeys from Galicia show a high confidence level, reaching 99.4%, except for two cases where the confidence levels are 62.5% and 93.8%. In the case of commercial honeys (five cases), the confidence level range is 14.9–78.1%. Four honeys are correctly classified with a confidence level between 65.4% and 78.1%; conversely, one case it is erroneously classified as Galician honey (confidence level of 14.9%). Owing to this, the accuracy value of RF_2 for querying phase is 95.2%. With the exception of this last honey, it can be said that the confidence level for the commercial honey predictions is adequate.

Once the best model has been developed, the results can be compared with those from the model developed by Rodríguez-Otero et al. [46] (Table 3). This model was developed using the program BMDP7M to the relative mobility data of bands eleven and twelve of the 82 Galician honeys and 22 commercial honeys to obtain a model based on discriminant analysis [46]. The classification matrix for the different models is shown in Table 3. Regarding the results obtained by Rodríguez-Otero et al. [46], it can be observed that for the 104 total honeys, 71 of the 82 honeys from Galicia were classified correctly, which represents 86.6%, and 19 of the 22 commercial honeys were classified correctly with an accuracy of 86.4%. Considering the RF model selected in this research, it can be seen that for Galician honey the accuracy is higher (98.8%) in comparison with the model developed by Rodríguez-Otero et al. [46] (86.6%). This result can be explained by the fact that 81 Galician honeys were correctly classified and the predictions for the commercial honeys were incorrect for only two honeys (accuracy of 90.9%). Finally, the improvement is remarkable taking into account the general accuracy value presented by the RF model. In this sense, the accuracy value goes from 86.5% to 97.1%.

Therefore, it can be concluded that the use of RF classification model with two input variables can be used to predict the origin of

Table 2 | Accuracy (%) for training (subscript T), validation (subscript V), querying (subscript Q), and overall (subscript O) phases for each model developed with two input variables: random forest (RF), artificial neural network (ANN), and support vector machine (SVM)

Model	B7	B8	B11	B12	$Accuracy_T$	$Accuracy_V$	$Accuracy_Q$	$Accuracy_O$
RF_2					96.8	100.0	95.2	97.1
ANN_2					87.1	95.2	95.2	90.4
SVM_2					90.3	90.5	90.5	90.4

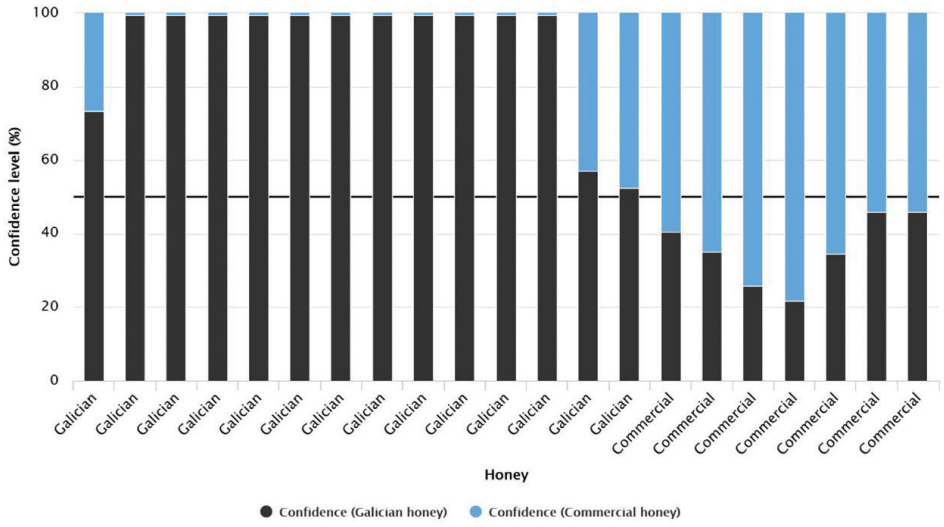

Figure 3 | Bar graph for validation cases according to the confidence value (%) of each prediction for the random forest model with two input variables.

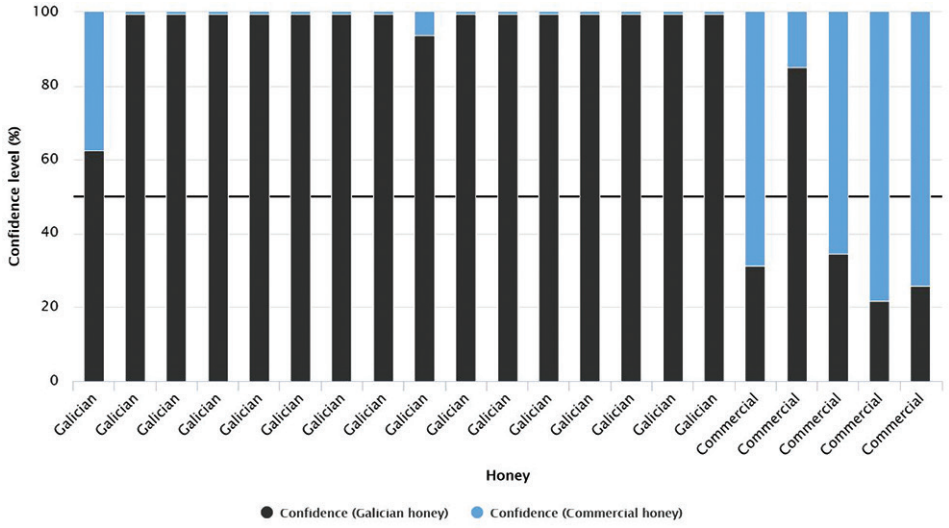

Figure 4 | Bar graph for querying cases according to the confidence value (%) of each prediction for the random forest model with two input variables.

Table 3 | Classification matrix for the models developed by Rodríguez-Otero et al. [46] (top) and the best model (RF_2) developed in this research (bottom)

Model developed by Rodríguez-Otero et al. [46]	**Honey classification**		
	Galician honey	**Commercial honey**	**Correct classification (%)**
Galician honeys (82 samples)	71	11	86.6
Commercial honeys (22 samples)	3	19	86.4
		Accuracy	86.5
Random forest (RF_2) developed in this research	**Honey classification**		
	Galician honey	**Commercial honey**	**Correct classification (%)**
Galician honeys (82 samples)	81	1	98.8
Commercial honeys (22 samples)	2	20	90.9
		Accuracy	97.1

the Galician honey. The possible reason for better results provided by the RF models may be due to the fact that this type of models are specially oriented to classification tasks and that the multiple trees that constitute the RF (RF_1 has six trees and RF_2 has eight trees) are able to offer a weighted value with more precision than an SVM or ANN model. The model could be used within the Galician geographical area to determinate with accuracy the native honeys.

4. CONCLUSION

Honey quality is very important for European consumers. To safeguard this product, the European Union has different geographical indication. Nevertheless, honey is a product that can be easily adulterated with different methods and to ensure its authenticity it is necessary to apply different techniques to control and combat this adulteration.

In this research, RF, ANN, and SVM models were tested to differentiate Galician honeys and other commercial honeys produced and packaged outside of Galicia. In addition to this, our best model was compared with the original model developed by Rodríguez-Otero et al. [46].

The results obtained for the best RF model allowed us to determine the honey's origin with an accuracy of 95.2%. To our understanding, the RF model, and the SVM and ANN models, could be improved with the inclusion of new data from different commercial honeys.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

C.M-C., G.A., C.M. and J.S-G. conceived and designed the models, drafted/revised the different manuscript versions and approved the final version.

ACKNOWLEDGMENTS

Cecilia Martinez-Castillo thanks the University of Vigo and Xunta de Galicia, Consellería de Medio Rural, for her contract supported by FEADER 2018/002B project. Gonzalo Astray thanks the University of Vigo for his contract supported by *"Programa de retención de talento investigador da Universidade de Vigo para o 2018"*. Authors thank RapidMiner Inc. for the Trial and Free License of RapidMiner Studio software.

REFERENCES

[1] Saurina J. Characterization of wines using compositional profiles and chemometrics. Trends Anal Chem 2010;29:234–45.
[2] Maione C, Barbosa F Jr, Barbosa RM. Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: a review. Comput Electron Agric 2019;157:436–46.
[3] Cuevas-Glory LF, Pino JA, Santiago LS, Sauri-Duch E. A review of volatile analytical methods for determining the botanical origin of honey. Food Chem 2007;103:1032–43.
[4] Siddiqui AJ, Musharraf SG, Iqbal Choudhary M, Rahman AU. Application of analytical methods in authentication and adulteration of honey. Food Chem 2017;217:687–98.
[5] Oroian M, Ropciuc S, Paduret S. Honey authentication using rheological and physicochemical properties. J Food Sci Technol 2018;55:4711–18.
[6] da Silva PM, Gauche C, Gonzaga LV, Costa ACO, Fett R. Honey: chemical composition, stability and authenticity. Food Chem 2016;196:309–23.
[7] Bertelli D, Lolli M, Papotti G, Bortolotti L, Serra G, Plessi M. Detection of honey adulteration by sugar syrups using one-dimensional and two-dimensional high-resolution nuclear magnetic resonance. J Agric Food Chem 2010;58:8495–501.
[8] Gallego-Picó A, Garcinuño-Martínez RM, Fernández-Hernando P. Chapter 20 - Honey authenticity and traceability. Compr Anal Chem 2013;60:511–41.
[9] European Commission. Quality schemes explained. https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/quality-schemes-explained_en; 2019 (accessed on August 21, 2019).
[10] Cotte JF, Casabianca H, Chardon S, Lheritier J, Grenier-Loustalot MF. Application of carbohydrate analysis to verify honey authenticity. J Chromatogr A 2003;1021:145–55.
[11] Puscas A, Hosu A, Cimpoiu C. Application of a newly developed and validated high-performance thin-layer chromatographic method to control honey adulteration. J Chromatogr A 2013;1272:132–5.
[12] Guler A, Bakan A, Nisbet C, Yavuz O. Determination of important biochemical properties of honey to discriminate pure and adulterated honey with sucrose (*Saccharum officinarum* L.) syrup. Food Chem 2007;105:1119–25.
[13] Azevedo MS, Valentim-Neto PA, Seraglio SKT, da Luz CFP, Arisi ACM, Costa ACO. Proteome comparison for discrimination between honeydew and floral honeys from botanical species *Mimosa scabrella* Bentham by principal component analysis. J Sci Food Agric 2017;97:4515–9.
[14] Won SR, Li CY, Kim JW, Rhee HL. Immunological characterization of honey major protein and its application. Food Chem 2009;113:1334–8.
[15] Lee DC, Lee SY, Cha SH, Choi YS, Rhee HI. Discrimination of native bee-honey and foreign bee-honey by SDS–PAGE. Korean J Food Sci 1998;30:1–5.
[16] Anklam E. A review of the analytical methods to determine the geographical and botanical origin of honey. Food Chem 1998;63:549–62.
[17] Latorre MJ, Peña R, García S, Herrero C. Authentication of Galician (N.W. Spain) honeys by multivariate techniques based on metal content data. Analyst 2000;125:307–12.
[18] RapidMiner GmbH. RapidMiner Documentation. https://rapidminer.com/; 2018 (accessed on August 21, 2019).
[19] Tian Y, Yan C, Zhang T, Tang H, Li H, Yu J, et al. Classification of wines according to their production regions with the contained trace elements using laser-induced breakdown spectroscopy. Spectrochim Acta B 2017;135:91–101.
[20] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[21] Vigneau E, Courcoux P, Symoneaux R, Guérin L, Villière A. Random forests: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. Food Quality and Preference 2018;68:135–45.
[22] Bauder RA, Khoshgoftaar TM. Medicare fraud detection using random forest with class imbalanced big data. In: Proceedings of the 2018 IEEE 19th International Conference on Information Reuse and Integration (IRI). Salt Lake City, UT, USA: IEEE; 2018. pp. 80–7.
[23] Herzallah W, Faris H, Adwan O. Feature engineering for detecting spammers on Twitter: modelling and analysis. J Inform Sci 2018;44:230–47.
[24] Osawa T, Kohyama K, Mitsuhashi H. Multiple factors drive regional agricultural abandonment. Sci Total Environ 2016;542:478–83.
[25] Vitale M, Proietti C, Cionni I, Fischer R, De Marco A. Random forests analysis: a useful tool for defining the relative importance of environmental conditions on crown defoliation. Water Air Soil Pollut 2014;225:1–17.
[26] Myronidis D, Ioannou K. Forecasting the urban expansion effects on the design storm hydrograph and sediment yield using artificial neural networks. Water (Switzerland) 2019;11:31.
[27] Qaderi F, Babanezhad E. Prediction of the groundwater remediation costs for drinking use based on quality of water resource, using artificial neural network. J Clean Prod 2017;161:840–9.
[28] Fazelpour F, Tarashkar N, Rosen MA. Short-term wind speed forecasting using artificial neural networks for Tehran, Iran. Int J Energy Environ Eng 2016;7:377–90.

[29] Sholahudin S, Han H. Simplified dynamic neural network model to predict heating load of a building using Taguchi method. Energy 2016;115:1672–78.
[30] Azizi A, Abbaspour-Gilandeh Y, Nooshyar M, Afkari-Sayah A. Identifying potato varieties using machine vision and artificial neural networks. Int J Food Prop 2016;19:618–35.
[31] Kalogirou SA. Artificial neural networks in renewable energy systems applications: a review. Renew Sustain Energy Rev 2001;5:373–401.
[32] Dong Q, Xing K, Zhang H. Artificial neural network for assessment of energy consumption and cost for cross laminated timber office building in severe cold regions. Sustainability (Switzerland) 2018;10:1–15.
[33] Astray G, Mejuto JC, Martínez-Martínez V, Nevares I, Alamo-Sanza M, Simal-Gandara J. Prediction models to control aging time in red wine. Molecules 2019;24:pii: E826.
[34] Gonzalez-Fernandez I, Iglesias-Otero MA, Esteki M, Moldes OA, Mejuto JC, Simal-Gandara J. A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. Crit Rev Food Sci Nutr 2019;59:1913–26.
[35] Moldes ÓA, Morales J, Cid A, Astray G, Montoya IA, Mejuto JC. Electrical percolation of AOT-based microemulsions with *n*-alcohols. J Mol Liquids 2016;215:18–23.
[36] Jiang M, Ma C, Xia F, Zhang Y. Application of artificial neural networks to predict the hardness of Ni–TiN nanocoatings fabricated by pulse electrodeposition. Surf Coat Technol 2016;286:191–6.
[37] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT). Pittsburgh, PA, USA: ACM; 1992. pp. 144–52.
[38] Srestasathiern P, Lawawirojwong S, Suwantong R. Support vector regression for rice age estimation using satellite imagery. In: Proceedings of the 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). Chiang Mai, Thailand: IEEE; 2016. pp. 1–5.
[39] Ríos-Reina R, Elcoroaristizabal S, Ocaña-González JA, García-González DL, Amigo JM, Callejón RM. Characterization and authentication of Spanish PDO wine vinegars using multi-dimensional fluorescence and chemometrics. Food Chem 2017;230:108–16.
[40] Wang C, Li Z. Weed recognition using SVM model with fusion height and monocular image features. Trans Chinese Soc Agric Eng 2016;32:165–74.
[41] Qiao Z, Zhang Q, Dong Y, Yang JJ. Application of SVM based on genetic algorithm in classification of cataract fundus images. In: Proceedings of the 2017 IEEE International Conference on Imaging Systems and Techniques (IST). Beijing, China: IEEE; 2017. pp. 1–5.
[42] Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN, Sejnowski TJ. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. IEEE Trans Biomed Eng 2002;49:963–74.
[43] Deng S, Yeh TH. Using least squares support vector machines for the airframe structures manufacturing cost estimation. Int J Prod Econ 2011;131:701–8.
[44] Shen W, Zhang Y, Ma X. Stock return forecast with LS-SVM and particle swarm optimization. In: Proceedings of the 2009 International Conference on Business Intelligence and Financial Engineering (BIFE). Beijing, China: IEEE; 2009. pp. 143–7.
[45] Akroyd P. Acrylamide gel slab electrophoresis in a simple glass cell for improved resolution and comparison of serum proteins. Anal Biochem 1967;19:399–410.
[46] Rodríguez-Otero JL, Paseiro Losada P, Simal-Lozano J, Cepeda Saez A. Intento de caracterización de las mieles de Galicia mediante las fracciones proteicas separadas por electroforesis de disco. Anal Bromatol 1990;XLII-1:83–98.
[47] Minaei S, Shafiee S, Polder G, Moghadam-Charkari N, van Ruth S, Barzegar M, et al. VIS/NIR imaging application for honey floral origin determination. Infrared Phys Technol 2017;86:218–25.
[48] Cajka T, Hajslova J, Pudil F, Riddellova K. Traceability of honey origin based on volatiles pattern processing by artificial neural networks. J Chromatogr A 2009;1216:1458–62.
[49] Anjos O, Iglesias C, Peres F, Martínez J, García Á, Taboada J. Neural networks applied to discriminate botanical origin of honeys. Food Chem 2015;175:128–36.
[50] Zhu X, Li S, Shan Y, Zhang Z, Li G, Su D, et al. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. J Food Eng 2010;101:92–7.
[51] Bisutti V, Merlanti R, Serva L, Lucatello L, Mirisola M, Balzan S, et al. Multivariate and machine learning approaches for honey botanical origin authentication using near infrared spectroscopy. J Near Infrared Spectrosc 2019;27:65–74.
[52] Gan Z, Yang Y, Li J, Wen X, Zhu M, Jiang Y, et al. Using sensor and spectral analysis to classify botanical origin and determine adulteration of raw honey. J Food Eng 2016;178:151–8.
[53] Batista BL, da Silva LRS, Rocha BA, Rodrigues JL, Berretta-Silva AA, Bonates TO, et al. Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Res Int 2012;49:209–15.
[54] Dai X, Shi H, Li Y, Ouyang Z, Huo Z. Artificial neural network models for estimating regional reference evapotranspiration based on climate factors. Hydrol Process 2009;23:442–50.
[55] da Costa NL, Llobodanin LAG, de Lima MD, Castro IA, Barbosa R. Geographical recognition of Syrah wines by combining feature selection with extreme learning machine. Measurement 2018;120:92–9.
[56] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 2006;70:489–501.
[57] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:27.
[58] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/; 2018 (accessed on August 21, 2019).
[59] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf; 2003. pp. 1–16.