

This is an electronic reprint of the original article.

This reprint *may differ* from the original in pagination and typographic detail.

Author(s): Pekka Jounela, Jouni Karhu, Saila Karhu, Anna-Liisa Välimaa

Title: Authenticity markers of cultivated blueberries (*Vaccinium* sp.) and European wild blueberry (*V. myrtillus*) assessed by hyperspectral imaging and machine learning algorithms

Year: 2024

Version: Published version

Copyright: The Author(s) 2024

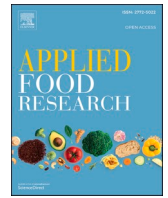
Rights: CC BY 4.0

Rights url: <https://creativecommons.org/licenses/by/4.0/>

Please cite the original version:

Pekka Jounela, Jouni Karhu, Saila Karhu, Anna-Liisa Välimaa, Authenticity markers of cultivated blueberries (*Vaccinium* sp.) and European wild blueberry (*V. myrtillus*) assessed by hyperspectral imaging and machine learning algorithms, Applied Food Research, Volume 4, Issue 2, 2024, 100597, ISSN 2772-5022, <https://doi.org/10.1016/j.afres.2024.100597> .

All material supplied via *Jukuri* is protected by copyright and other intellectual property rights. Duplication or sale, in electronic or print form, of any part of the repository collections is prohibited. Making electronic or print copies of the material is permitted only for your own personal use or for educational purposes. For other purposes, this article may be used in accordance with the publisher's terms. There may be differences between this version and the publisher's version. You are advised to cite the publisher's version.



Authenticity markers of cultivated blueberries (*Vaccinium* sp.) and European wild blueberry (*V. myrtillus*) assessed by hyperspectral imaging and machine learning algorithms

Pekka Jounela^{a,*}, Jouni Karhu^b, Saila Karhu^a, Anna-Liisa Välimaa^b

^a Natural Resources Institute Finland (Luke), Itäinen Pitkätie 4 A, FI-20520, Turku, Finland

^b Natural Resources Institute Finland (Luke), Paavo Havaksen tie 3, FI-90570, Oulu, Finland

ARTICLE INFO

Keywords:

Authenticity
Bilberry
Computer vision
Feature selection
Food fraud

ABSTRACT

The present study was undertaken to tackle food fraud in the food chain by developing a novel panel of analytical fingerprints to distinguish the fruit of bilberry (*Vaccinium myrtillus*), a wild European blueberry grown in Finland, from their commercial cultivated counterparts based on *V. corymbosum* or other originally American *Vaccinium* species. The analytical fingerprints were derived using hyperspectral imaging (HSI). The cultivated domestic blueberry samples from Finland ($N = 188$), those imported ($N = 338$), and domestic Finnish wild bilberry fruit ($N = 247$) were imaged using Specim IQ and FX17 HSI cameras. The fruit HSI data were then classified using seven machine learning model types. The supervised evolutionary feature selection algorithm was used to recognize the most separating subset of wavelengths (nm predictors, colors) between the fruit class labels, which were 1) wild bilberry fruit vs. cultivated blueberry fruit and 2) domestic cultivated blueberry fruit vs. imported cultivated blueberry fruit. Moreover, we ran 190,822 individual models and 54,137 predictor combinations to determine whether subsets or all wavelengths better distinguished fruit class labels. Overall, the best model type, the rbf kernel support vector machine (rbf-SVM), showed excellent prediction performance when classifying wild bilberries vs. cultivated blueberries with HSI data obtained using both camera types (4.1% and 3.2% holdout set errors). Also the rbf-SVM performed well when classifying domestic cultivated blueberries vs. those imported with HSI data obtained using an IQ camera (2.3% holdout set error). Our results showed that hyperspectral imaging could predict food fraud attempts concerning the *Vaccinium* fruit of different origin and species.

1. Introduction

The origin of food is among the most important criteria in European consumers' purchasing decisions (Efsa, 2022). Food adulteration is occurring increasingly often today, affecting a wide range of food products. Recently, an initiative of the European Parliament (EP) identified several food products as the target of fraudulent activities such as olive oil, fish, honey, dairy products, and meat (Ec, 2023). This initiative calls for the development of technologies and reliable methods to detect food fraud. To maintain trust, consumers should urgently be assured that the products they buy comply with the label. Food adulteration may cause significant consequences for consumers and food companies concerning health and food safety issues, economic damage and reputational damage to businesses, and a loss of consumer trust in official controls, governments, and the industry (Robson et al., 2021; Visciano &

Schirone, 2021; Winkler et al., 2023).

Both the business operators of food supply chains and organized crime groups actively commit food fraud to gain economic benefit and the associated profits (Efsa, 2022). Food fraud categories are diverse, including adulteration and product tampering by substitution, dilution, removal, unapproved/undeclared enhancement and concealment, and unapproved/undeclared treatment, process, or product concealment, counterfeit and misdescription, mislabeling, and misbranding (Rizzuti, 2022). In the EU, there is also a significant increase of fraud cases involving fruit and vegetables.

The European blueberry or bilberry (*V. myrtillus*) grows wild in boreal forests and highland areas as an understory dwarf shrub. The cultivated blueberry cultivars most commonly belong to the species *V. corymbosum* or are hybrids of it, and *V. darrowii* and *V. ashei* are also progenitors of several cultivars, for example. The fruit of stands of the

* Corresponding author at: Natural Resources Institute Finland, Itäinen Pitkätie 4 A, FI-20520, Turku, Finland.

E-mail addresses: pekka.jounela@luke.fi (P. Jounela), jouni.karhu@luke.fi (J. Karhu), saila.karhu@luke.fi (S. Karhu), anna-liisa.valimaa@luke.fi (A.-L. Välimaa).

lowbush species *V. angustifolium* and *V. myrtilloides* are also machine harvested and traded globally (Hancock & Draper, 1989). Commercially cultivated blueberries thus represent not only a large number of cultivars but also several species in their genetic background.

The cultivated blueberry market is much larger than the bilberry market, with worldwide production ca. 1113.260 tn. and exports of 459,544 tn. in 2021 (Fao, 2023). The cultivated blueberry fruit and bilberry fruit resemble each other in appearance and taste to a considerable extent. Intentional mislabeling has occurred when wild bilberries have been replaced with cultivated blueberries (Lee, 2016). Sometimes, unintentional mislabeling is caused when species names are confused – for example, the names “blueberry” and “bilberry” (Salo et al., 2021).

Various analytical techniques are available for detecting food fraud. They can be divided into targeted and nontargeted approaches. Targeted methods detect only one or a small number of potential analytes (adulterants) at a time, whereas nontargeted methods such as spectroscopy assess whether the profile of an unknown sample matches the reference profile of a specific product (Ballin & Laursen, 2019; Everstine et al., 2021). Recently, nontargeted analytical methods have been increasingly used to detect food adulteration. They are considered rapid cost-effective analytical techniques and well suited to the screening of large numbers of samples.

Food authenticity testing should ideally be carried out in real time *on site* at the food production plant or retail outlet with no or minimal sample preparation. Traditional analytical methods such as mass spectrometry are beyond those requirements. Hyperspectral imaging (HSI) is a nontargeted and nondestructive technique that integrates conventional imaging and spectroscopy, collecting spectral and spatial information in parallel. In addition, HSI is a powerful machine vision tool for various real-time applications such as in-field estimation of strawberry ripeness with deep learning (Gao et al., 2020). Furthermore, it is chemical-free and can make rapid measurements with limited sample preparation (Müller-Maatsch & van Ruth, 2021). Hyperspectral cameras process hundreds of spectral bands instead of only assigning the three primary RGB (red, green, blue) colors for each pixel. Moreover, hyperspectral cameras extend far beyond the visible wavelengths undetectable by RGB cameras and unseen by the human eye. Furthermore, HSI techniques provide a detailed vision of the physical and geometrical features of the objects through spectral analysis (Nicolai et al., 2007). In other words, HSI generates a unique fingerprint (characteristic spectra, nm subsets, common denominators among samples) of the object, unlike RGB cameras or the human eye.

Pre-processing hyperspectral images presents computational challenges in the numerical analysis stage. At first, the hyperspectral camera transforms continuous spectral bands into numbers, which are analyzed using a multivariate method that is frequently employed in statistics, computer science, and applied mathematics. The better-known subfields and analytical tasks of these multivariate computer vision techniques are object detection, pattern recognition, image classification, discrimination, segmentation, and anomaly (outlier) detection. The aim of all these multivariate methods is to be both reliable and fast in testing the authenticity of the bilberry fruit.

The present study aimed to tackle food fraud in the food chain by developing a novel panel of analytical fingerprints to distinguish wild bilberries and cultivated Finnish blueberries from commercial blueberries grown abroad. We assessed the statistical performance of an array of modern machine learning techniques and recognized the best ones that could be employed in the detection of food fraud.

2. Material and methods

2.1. Fruit samples

The fruit sample set, 773 in total, consisted of individual intact fresh fruit. The number of berries (sample size) was 247 for wild blueberries, 188 for cultivated Finnish (domestic) blueberries, and 338 for cultivated

blueberries imported into Finland. The wild bilberry fruit were purchased at local Finnish markets or directly from the pickers between July and August in 2021 and 2022. The area's geographical location, where bilberries (FIN) were collected, ranged between latitudes (63° 31' – 66° 18' N) and longitudes (26° 15' – 29° 10' E), consisting of 37 berry batches. The major part (38 varieties) of the cultivated domestic blueberry fruit (FIN) was harvested at Luke's Piikkiö Experimental Site in Kaarina in South-West Finland (60° 23' N, 22° 33' E). The minor part (5 batches) of Finnish cultivated blueberry fruit was purchased from the berry farmers in Central Finland (62° 53' – 64° 07' N, 25° 22' – 27° 40' E). The imported cultivated blueberry fruit, consisting of 56 berry batches, were purchased in different shops in the Oulu region in 2021 and 2022. The fruit originated from Germany (GER), Poland (POL), Ukraine (UKR), Spain (ESP), Portugal (PRT), Morocco (MAR), Zimbabwe (ZWE), South Africa (ZAF), Peru (PER), and Argentina (ARG). The blueberry and bilberry samples were stored at 4 °C and allowed to reach room temperature before HSI.

2.2. HSI

Imaging of the bilberry and blueberry fruit samples was performed at room temperature. Initially, the fresh intact berries were spread evenly over a black 96-well plate (Greiner Bio-One GmbH, Austria) and were then imaged using the Specim IQ and Specim FX17 hyperspectral cameras (Spectral Imaging Ltd., Oulu, Finland). The Specim IQ camera operates in a spectrum wavelength range of 400–1000 nm and the spectral resolution is 7 nm with 204 spectral bands. The Specim FX17 camera operates in a spectral wavelength range of 900–1700 nm and the spectral resolution is 8 nm with 224 spectral bands. Both cameras were mounted on the LabScanner 40 × 20 (Spectral Imaging Ltd., Oulu, Finland) rack with a moving tray. The tray was locked in a fixed position for the IQ camera. The rack had direct-current halogen illumination in two rows, each with three 20-watt halogen light bulbs (OSRAM DECOSTAR 51 ALU 20 W 12V 36 deg GU5.3 halogen). The image acquisition was made through Specim IQ Studio software for the IQ camera and LumoScanner software for the FX17 camera.

2.3. Pre-processing of the HSI image data

The sample images were pre-processed by retaining (segmenting) only the fruit. The background removal from the sample images was based on the spectral differences between the background and fruit. The fruit mean spectral reflectance was calculated over the remaining pixels. The number of individual fruit (sample size) of cultivated domestic blueberry was 188 and of those imported 338 and the sample size of domestic wild-grown bilberry was 247 (773 in total; Fig. 1).

The results of Specim IQ and FX17 based imaging and image pre-processing were the HSI data consisting of 204 and 224 wavelength predictors (columns) in 3 nm intervals, between 397 and 1003 nm and 935–1720 nm respectively (Fig. 2).

2.4. Statistical methods

The computational challenge was to search for the best model type and the most separating subset of nm predictors (columns) to distinguish the fruit class labels, which were 1) wild bilberry vs. cultivated blueberry fruit and 2) cultivated domestic vs. imported blueberry fruit (Fig. 3).

2.4.1. Feature selection

The statistical performance of a subset of HSI predictors (wavelengths, colors) that best separated the class labels was tested a) with the trial subsets of predictors and b) with all predictors. The supervised evolutionary feature selection algorithm (Guyon & Elisseeff, 2003) was used to recognize the most separating subset of predictors (columns) between the class labels that were 1) wild bilberry fruit vs. cultivated

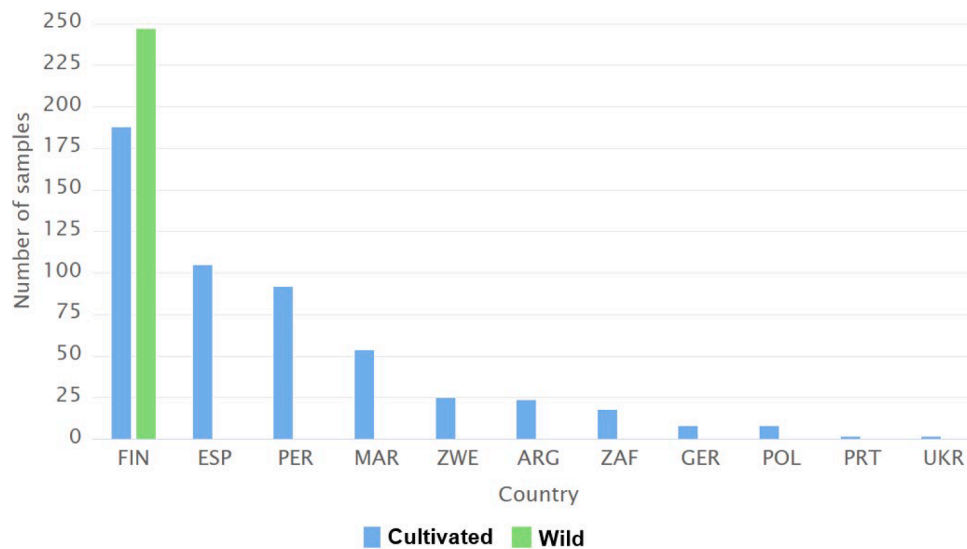


Fig. 1. The numbers of wild-grown bilberry fruit samples and cultivated domestic (FIN) and imported blueberry fruit samples by country.

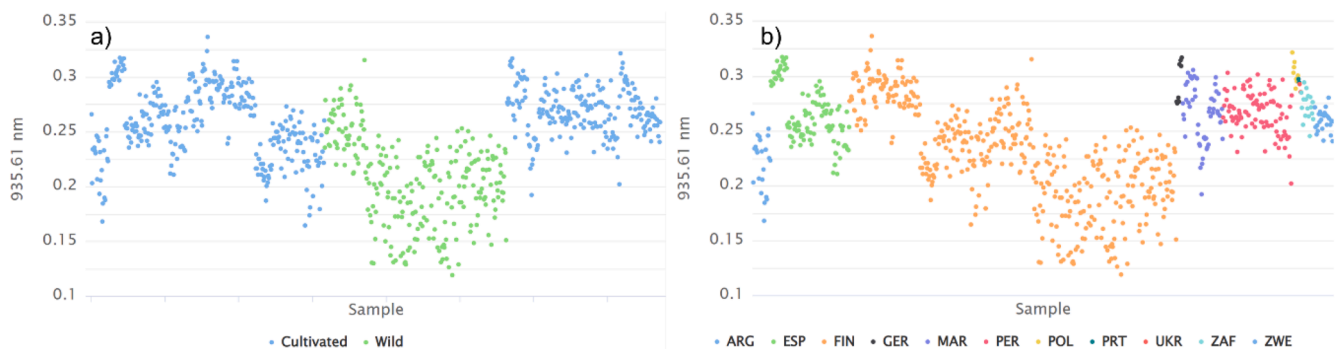


Fig. 2. An example of the HSI wavelength distributions (y-axis, only 935 nm) of 773 samples (x-axis) a) by species group (cultivated, wild) and b) by country, when imaged using an FX17 camera. The total HSI data (i.e., all nm wavelength distributions) consisted of $N = 204$ (IQ camera) and $N = 224$ (FX17 camera) predictors in 3 nm intervals, between 397 and 1003 nm and 935–1720 nm respectively.

blueberry fruit and 2) domestic cultivated blueberry fruit vs. those imported. Specifically, in predictor selection, we selected existing predictors “as is,” but we did not simultaneously derive the new predictor combinations (functions).

2.4.2. Classification models (learners)

Seven classification model types were used to analyze HSI data: support vector machine (rbf-SVM; type: C-SVC; kernel: rbf; library: LibSVM, Chang & Lin, 2011; Cortes & Vapnik, 1995), Fast Large Margin linear SVM (lin-SVM, Fan et al., 2008), deep learning (DL) neural network (library: H2O, Candel & LeDell, 2020), gradient boosted machine (GBM, library: H2O), random forests (RF, Breiman, 2001), Naive Bayes (NB, Hand & Yu, 2001), and decision tree (DT, Quinlan, 1983) models. The goal of using many model types was to identify the best for each classification task and all tasks on average. We limited our study to account for the prediction potential of various models and the transparency of the methods used. For the explainability of those spectral bands (colors) that distinguish the fruit classes, hundreds of approaches such as local explanations, explanations by example, explanations by simplification, and feature relevance explanation techniques (Arrieta et al., 2020; Linardatos et al., 2021) nevertheless remain.

2.4.3. Parameter optimization, training, validation, and holdout testing

The parameters of each trial model were optimized using a 60% subset of data rows and n -fold ($n = 3$ –10) cross-validation (Kohavi,

1995). The performance of each trial model and trial predictor (column) subset was then tested using the 40% holdout set, i.e., the subset of samples (rows) that were not used to train the model (out-of-sample). This was done to check whether the model is really capable of generalizing outside the training data used to optimize the model. Stratified sampling was used to ensure that the class distribution in the subsets was the same as in the whole dataset. The 40% holdout-set was further split into seven batches (also referred to as “repeated holdout testing”). The best and worst performing holdout set batches were then removed, and the average performance of the retained remaining five hold-out batches provided a more robust performance estimate than a single split hold-out set validation would have provided. This repeated hold-out testing provided shorter runtimes in parameter optimization and feature selection than that using nested cross-validation.

Cost-sensitive learning (Chawla et al., 2004; Turney, 2000) was used to mitigate the possible problem of unequal sample sizes between the class labels using the unequal misclassification cost matrix. That is, for example, misclassifying a half less frequently occurring class was penalized twice more than misclassifying a twice more frequently occurring class. The search time for the best model, parameter, and predictor combination was limited to three hours per analysis. The statistical analyses were performed using RapidMiner software (<https://rapidminer.com>, version Studio Enterprise 10.1.001, Mierswa et al., 2006).

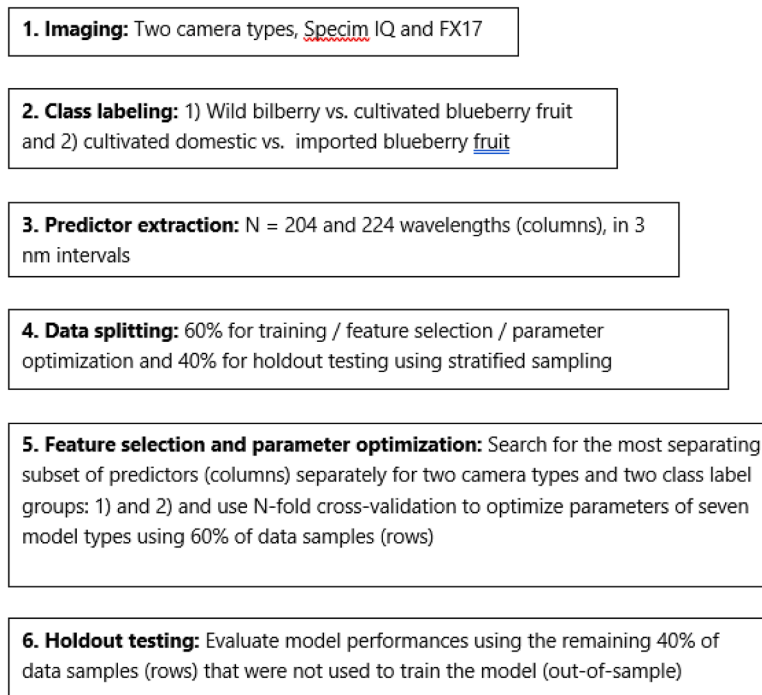


Fig. 3. The sequence of experimental implementation process and computational analysis used to detect food origin.

3. Results

The search for the best model and predictor subset included the testing of 190,822 individual models and 54,137 predictor subset combinations. In general, Fast Large Margin SVM and DL were the best model types when classifying wild-grown bilberries vs. cultivated blueberries with an FX17 camera (2.7% holdout set errors). With both camera types (FX17 and IQ), classifying wild-grown bilberries vs. cultivated blueberries resulted in smaller errors (6.6% and 3.6%, respectively) than classifying cultivated domestic vs. imported blueberries (20.5% and 16.1% respectively; Table 1) on average. Further, the rbf-SVM model performed best when classifying domestic cultivated blueberries vs. those imported with an IQ camera (4% holdout set error). Overall, the rbf-SVM model was also the best model type, with a 6.3% holdout set error on average (Table 1). The best model type in terms of the fewest predictors on average was NB, whereas the best four rbf-SVM

Table 1

The classification error percentages (%) of the class labels 1) wild-grown bilberry vs. cultivated blueberry fruit and 2) domestic vs. imported blueberry fruit, obtained using 7 different model types with an evolutionary feature selection algorithm or without it (marked by *), tested separately with HSI data using Specim FX17 and IQ hyperspectral cameras. The classification error percentages refer to 40% holdout set error i.e., a 40% subset of data rows that were not used to optimize model parameters using cross-validation.

Model	FX17 camera		IQ camera		Average
	Wild vs. Cultivated	Domestic vs. Imported	Wild vs. Cultivated	Domestic vs. Imported	
rbf-SVM	4.1*	14*	3.2*	4.0*	6.3
lin-SVM	2.7*	20	2.7*	16.7	10.5
DL	2.7	20.7*	2.7	8.0*	8.5
GBM	6.8	22*	2.7	11.4	10.1
RF	8.6	20	2.7*	19.5	12.7
DT	10.4*	28	8.1*	30	19.1
NB	10.9	18.7	5.9	23.3	14.7
Average	6.6	20.5	3.6	16.1	11.7

* Holdout set classification error obtained using all (wavelength, nm) predictors.

models consisted of full predictor sets (Table 2).

4. Discussion

The present study shows statistical performances of seven machine learning methods with and without feature selection. The aim of feature selection was to mitigate the computational cost and to improve the model performance outside the data used to optimize the model. By far, the smallest computational costs in terms of only a few (~7) wavelength predictors were obtained using the NB model on average. With a larger number of predictors, the performance of the NB model was severely degraded because NB assumes conditional independence between features and therefore only needs a small amount of data to estimate the parameters (Sharma & Manavjeet, 2013). This is exponentially fewer than with SVMs, DL, and ensembles models like GBM and RF. However, the latter four model types are extensively used in image pattern recognition due to their tolerance of a larger number of predictors, which in its part contributes smaller prediction errors.

In practice, the events (sample sizes) per variable and especially the

Table 2

The estimated number of predictors in each model-analysis-camera combination and those on average used to distinguish the class labels 1) wild-grown bilberry vs. cultivated blueberry fruit and 2) domestic vs. imported blueberry fruit.

Model	FX17 camera		IQ camera		Average
	Wild vs. Cultivated	Domestic vs. Imported	Wild vs. Cultivated	Domestic vs. Imported	
Rbf-SVM	204*	204*	224*	224*	214
Lin-SVM	204*	105	224*	117	163
DL	60	204*	108	224*	118
GBM	27	204*	108	89	85
RF	21	3	224*	16	66
DT	204*	116	224*	9	138
NB	5	6	6	10	7
Average	87	90	160	98	113

* Consists of all (wavelength, nm) predictors.

required holdout set (out-of-sample) performances may vary considerably between cases, and providing clearcut rules therefore regarding the superiority of the used methods is not simple. In general, SVMs have been shown to perform well in high dimensionality cases with small and medium-sized hyperspectral data (Gualtieri & Chettri, 2000; Mercier & Lennon, 2003; Pal & Mather, 2004), whereas deep neural networks are mostly used with larger sample sizes (Rosebrock, 2017), and as our study suggested, NB showed a reasonable performance, with very small predictor numbers. However, it is difficult to analytically derive exact estimates for minimum or optimal sample sizes a priori. This is because the variable numbers (columns, multidimensionality) and sample size (rows) required for a reliable nondestructive food fraud analysis depends on factors such as class imbalance, class separability, the performance measure used, model type, and the quality of features.

The statistical performance of a single prediction model also varies based on the dimensionality of the data, which is known as the peaking phenomenon (Hughes, 1968). With a fixed number of training samples, the error first decreases as the number of dimensions increases, but at some point, it starts to increase instead of decreasing steadily (Trunk, 1979). In other words, with a small sample size, a better performance is usually obtained by keeping the number of features to a relatively small number. With a larger sample size, a larger number of features can be selected that may yield better performance. Perhaps a typical rough rule of thumb is that there should be at least 2–10 training examples for each dimension in the representation (Koutroumbas & Theodoridis, 2009). Kim and Landgrebe (1990) suggest higher values, ~ 6–10 times larger sample size compared to the dimensions. Our HSI data consisted of 2.5–2.8 samples for each dimension.

The existing plethora of feature selection methods and selection among them will potentially have a major impact on the statistical performance of a single model. This is because a feature by itself may have little correlation with the target concept, but when it is combined with (an)other feature(s), they can be strongly correlated with the target concept (Zhao & Liu, 2007). The unintentional removal of these “hidden” features without deeper computational evaluation will potentially result in a poor classification performance. In our case, the search for the best feature subset took 3 h, and the diversity of feature subsets could easily be increased by simply increasing the search time.

A possible way to improve the presented methods and the detection rate of food fraud in general would be to combine many models by stacking (stacked generalization, Wolpert, 1992). In general, stacking exploits testing data performance to combine the models rather than to choose among them, thereby typically obtaining a better performance outside the data used to optimize the models than any single model of the trained ones. Instead of using a winner-takes-all approach, stacking combines the base learners. Using a variety of strong learning algorithms in an ensemble yields significant diversity among the models that tends to reduce problems related to the overfitting of the training data. Stacking has been successfully used in both supervised learning tasks (e.g., classification and regression) and unsupervised learning (e.g., density estimation). We did not apply stacking because the method requires the same predictors in each model type to be used.

The economic incentives to collect small sample HSI data may unintentionally advocate using models that tolerate limited data such as SVMs. In our study, the rbf-SVM model resulted in the best average prediction performance among seven alternative supervised classification model types with and without evolutionary feature selection. Previous food fraud and adulteration detection studies have also used an array of supervised classification methods (Callao & Ruisánchez, 2018) such as artificial neural networks (Mu et al., 2016), density-based methods (Oliveri et al., 2014), and PCA guided SVMs (Majcher et al., 2015). That is, the PCA factor loadings can be used as an alternative to feature selection algorithms when distinguishing HSI data with a supervised classification method.

5. Conclusion

Our study showed that HSI, and particularly an IQ camera, is a highly accurate research tool to classify blueberry labels of different species and even of different origin. Both camera types could be used on site in real time for industrial applications - for example, mounted on a process line above a conveyor belt. Although both SVM and a DL models showed excellent holdout performance, more samples are needed for the establishment of a representative reference database.

Ethical statement - Studies in humans and animals

This study did not involve human or animal subjects.

Funding statement

The research was facilitated by Natural Resources Institute Finland (Luke).

CRediT authorship contribution statement

Pekka Jounela: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jouni Karhu:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis. **Saila Karhu:** Writing – review & editing, Supervision. **Anna-Liisa Välimaa:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

All authors declare there are no competing interests.

Acknowledgment

The authors thank Tuuli Haikonen and Eila Järvenpää for providing Finnish blueberry fruit and Tiina Väyrynen and Annika Uimonen for assistance in hyperspectral imaging.

Data availability

Data will be made available on request.

References

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bannetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science and Technology*, 86, 537–543. <https://doi.org/10.1016/j.tifs.2018.09.025>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Callao, M. P., & Ruisánchez, I. (2018). An overview of multivariate qualitative methods for food fraud detection. *Food Control*, 86, 283–293. <https://doi.org/10.1016/j.foodcont.2017.11.034>
- Candel, A., & LeDell, E. (2020). Deep learning with H2O. <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf>
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–27. <https://doi.org/10.1145/1961189.1961199>
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 6(1), 1–6. <https://doi.org/10.1145/1007730.1007733>
- Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Ec (2023). https://joint-research-centre.ec.europa.eu/scientific-activities-z/food-authenticity-and-quality_en.

- European food safety authority (Efsa), (2022). Special Eurobarometer Wave EB97.2 Food safety in the EU. https://www.efsa.europa.eu/sites/default/files/2022-09/EB97_2-food-safety-in-the-EU_report.pdf, <https://doi.org/10.2805/729388>.
- Everstine, K., Popping, B., & Gendel, S. M. (2021). Chapter 3 - Food fraud mitigation: Strategic approaches and tools. In R. S. Hellberg, K. Everstine, & S. A. Sklare (Eds.), *Food fraud, a global threat with public health and economic consequences* (pp. 23–43). Academic Press. <https://doi.org/10.1016/B978-0-12-817242-1.00015-4>.
- Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research : JMLR*, 9, 1871–1874. <https://doi.org/10.1145/1390681.1442794>
- Fao (2023). FAOSTAT statistical database. <https://www.fao.org/faostat/en/#home>.
- Gao, Z., Shao, Y., Xuan, G., Wang, Y., Liu, Y., & Han, X. (2020). Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. *Artificial Intelligence in Agriculture*, 4, 31–38. <https://doi.org/10.1016/j.aiia.2020.04.003>
- Gualtieri, J. A., & Chettri, S. (2000). Support vector machines for classification of hyperspectral data. *IGARSS 2000. In 2. IEEE 2000 International geoscience and remote sensing symposium. Taking the pulse of the planet: the role of remote sensing in managing the environment. proceedings (Cat. No. 00CH37120)* (pp. 813–815). <https://doi.org/10.1109/IGARSS.2000.861712>. Honolulu, HI, USA.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research : JMLR*, 3, 1157–1182. <https://doi.org/10.1162/153244303322753616>
- Hancock, J. F., & Draper, A. D. (1989). Blueberry culture in North America. *HortScience : A Publication of the American Society for Horticultural Science*, 24(4), 551–556. <https://doi.org/10.21273/HORTSCI.24.4.551>
- Hand, D. J., & Yu, K. (2001). Idiot's Bayes — Not so stupid after all? *International Statistical Review = Revue internationale de statistique*, 69(3), 385–399. <https://doi.org/10.2307/1403452>
- Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1), 55–63. <https://doi.org/10.1109/TIT.1968.1054102>. S2CID 206729491
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellis (Ed.), *Proceedings of the 14th international joint conference on artificial intelligence, August 20–25 1995* (pp. 1137–1143). Montreal, Quebec, Canada: Morgan Kaufmann Publishers. San Mateo, CA <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>.
- Koutroumbas, K., & Theodoridis, S. (2009). *Pattern recognition* (4th ed.). Burlington, USA: Academic Press.
- Kim, B., & Landgrebe, D. A. (1990). Prediction of optimal number of features. In *10th Annual international symposium on geoscience and remote sensing* (pp. 2393–2396). <https://doi.org/10.1109/IGARSS.1990.689021>. College Park, MD, USA.
- Lee, J. (2016). Anthocyanin analyses of Vaccinium fruit dietary supplements. *Food Science & Nutrition*, 4(5), 742–752. <https://doi.org/10.1002/fsn3.339>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 8. <https://doi.org/10.3390/e23010018>
- Majcher, M. A., Kaczmarek, A., Klensporf-Pawlik, D., Pikul, J., & Jelen, H. H. (2015). SPME-MS-based electronic nose as a tool for determination of authenticity of PDO cheese, Oscypek. *Food Analytical Methods*, 8, 2211–2217. <https://doi.org/10.1007/s12161-015-0114-x>
- Mercier, G., & Lennon, M. (2003). Support vector machines for hyperspectral image classification with spectral-based kernels. In *IGARSS 2003. 2003 IEEE International geoscience and remote sensing symposium. Proceedings (IEEE Cat. No. 03CH37477)* (pp. 288–290). Toulouse, France: IEEE. <https://doi.org/10.1109/IGARSS.2003.1293752>.
- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06)* (pp. 935–940). New York, NY: ACM. <https://doi.org/10.1145/1150402.1150531>. <https://rapidminer.com>.
- Mu, T., Chen, S., Zhang, Y., Chen, H., Guo, P., & Meng, F. (2016). Portable detection and quantification of olive oil adulteration by 473-nm laser-induced fluorescence. *Food Analytical Methods*, 9, 275–279. <https://doi.org/10.1007/s12161-015-0199-2>
- Müller-Maatsch, J., & van Ruth, S. M. (2021). Handheld devices for food authentication and their applications: A review. *Foods (Basel, Switzerland)*, 10, 2901. <https://doi.org/10.3390/foods10122901>
- Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, 46(2), 99–118. <https://doi.org/10.1016/j.postharvbio.2007.06.024>
- Oliveri, P., López, M. L., Casolino, M. C., Ruisánchez, I., Callao, M. P., Medini, L., & Lanteri, S. (2014). Partial least squares density modeling (PLS-DM)—A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. *Analytica chimica acta*, 851, 30–36.
- Pal, M., & Mather, P. M. (2004). Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generations Computer Systems : FGCS*, 20(7), 1215–1225. <https://doi.org/10.1016/j.future.2003.11.011>
- Quinlan, R. (1983). Learning efficient classification procedures. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning an artificial intelligence approach* (pp. 463–482). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-12405-5_15.
- Rizzuti, A. (2022). Organised food crime: an analysis of the involvements of organised crime groups in the food sector in England and Italy. *Crime, Law and Social Change*, 78(5), 463–482.
- Robson, K., Dean, M., Haughey, S., & Elliott, C. (2021). A comprehensive review of food fraud terminologies and food fraud mitigation guides. *Food Control*, 120, Article 107516. <https://doi.org/10.1016/j.foodcont.2020.107516>
- Rosebrock, A. (2017). *Deep learning for computer vision with python, imagenet bundle*. Pyimagesearch. <https://www.pyimagesearch.com/deep-learning-computer-vision-pythonbook/>.
- Salo, H. M., Nguyen, N., Alakärppä, E., Klavins, L., Hykkerud, A. L., Karppinen, K., Jaakola, L., Klavins, M., & Häggman, H. (2021). Authentication of berries and berry-based food products. *Comprehensive Reviews in Food Science and Food Safety*, 20(5), 5197–5225. <https://doi.org/10.1111/1541-4337.12811>
- Sharma, P., & Manavjeet, K. (2013). Classification in pattern recognition: A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(4), 298–306.
- Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3), 306–307. <https://doi.org/10.1109/TPAMI.1979.4766926>
- Turney, P. (2000). Types of cost in inductive concept learning. In *Proceedings of the ICML'2000 workshop on cost-sensitive learning* (pp. 15–21). <https://doi.org/10.48550/arXiv.cs/0212034>
- Visciano, P., & Schirone, M. (2021). Food frauds: Global incidents and misleading situations. *Trends in Food Science and Technology*, 114, 424–442. <https://doi.org/10.1016/j.tifs.2021.06.010>
- Winkler, B., Maquet, A., Reeves-Way, E., Siegener, E., Cassidy, T., Valinhas De Oliveira, T., Verluyten, J., Jelic, M., & Muznik, A. (2023). Fighting fraudulent and deceptive practices in the agri-food chain. *Technical report "Implementation of article 9 (2) of regulation (EU) 2017/625"*. Luxembourg: Publications Office of the European Union. <https://doi.org/10.2760/31366>. JRC131525.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks : The Official Journal of the International Neural Network Society*, 5(2), 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- Zhao, Z., & Liu, H. (2007). Searching for interacting features. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI '07)* (pp. 1156–1161). Hyderabad, India.