# Predicting Impurities in Mānuka Honey Using Machine Learning

**Volodymyr Zaslavskyi*, Ihor Volokhovych*, Ibraim Didmanidze****

*Taras Shevchenko National University of Kyiv

**Batumi Shota Rustaveli State University

zas@unicyb.kiev.ua, ibraim.didmanidze@bsu.edu.ge

**Abstract**

Mānuka honey is a high-value monofloral honey known for its unique medicinal properties, but its premium status makes it a prime target for adulteration and contamination. This article explores how machine learning (ML) techniques can predict and detect impurities in Mānuka honey, including adulterants (e.g., added sugars or cheaper honeys) and chemical contaminants. We provide an overview of recent research, highlighting both theoretical foundations and practical implementations of ML in honey authenticity testing. Various data sources – from spectral signatures (NIR, FTIR, Raman, NMR) to hyperspectral imaging and microscopic pollen analysis – are leveraged in combination with algorithms such as support vector machines (SVM), random forests, artificial neural networks (ANN), and deep learning models. The methodologies for data preprocessing, feature selection, model training, and validation are detailed. We discuss performance metrics reported in the literature, with many ML models achieving high accuracy in classifying authentic vs. adulterated honey and strong predictive power in quantifying adulterant levels.

**Keywords**:Manuka honey; Machine learning; Food authenticity; Spectroscopy; Hyperspectral imaging; Support vector machine; Neural networks;

## 1. Introduction.

Mānuka honey (derived from *Leptospermum scoparium* nectar in New Zealand) commands a high price due to its unique antibacterial components (e.g., methylglyoxal) and limited supply. This premium value has led to frequent fraud and impurities in marketed “Mānuka” honey. Adulteration can occur by diluting genuine Mānuka honey with cheaper honeys or sugar syrups, or by mislabeling other floral honeys as Mānuka [2][5]. Such fraudulent practices undermine consumer trust and pose economic and health risks. Chemical contaminants (like pesticides, antibiotics, or heavy metals) can also appear in honey via environmental exposure or improper beekeeping, though the primary focus in authenticity is on adulterants. Ensuring honey purity is therefore both a regulatory requirement and a market necessity. Traditional methods for honey authentication include melissopalynology (microscopic pollen analysis) and chemical marker tests. For example, authenticating Mānuka honey often relies on detecting four signature chemical markers (such as leptosperin) and the DNA of Mānuka pollen [5]. However, these conventional methods have limitations: melissopalynology is labor-intensive and may fail to catch sugar adulteration (since sugar syrups contain no pollen) [1], and specialized chemical tests are costly and specific to Mānuka, not applicable to other honey types [5]. Increasingly, researchers are turning to machine learning to address these challenges in a more automated and robust way.

This article provides a comprehensive review of how modern machine learning techniques are being applied to predict and detect impurities in Mānuka honey. We survey related work in the field, then detail common methodologies (data collection, model building, validation) and discuss the results and implications of these approaches. By focusing on Mānuka honey as a case study, we illustrate general principles of food authenticity testing with ML, along with specific considerations for this

unique honey. In doing so, we aim to inform both practitioners in food quality control and IT professionals interested in the intersection of machine learning and food authenticity.

## 2. Related Work.

Related work spans from classical analytical chemistry to modern AI-driven methods. The most successful recent approaches use some form of spectroscopy or imaging combined with machine learning to capture complex signals of authenticity.

- *Calle et al. (2023)* – Vis/NIR spectroscopy of honey blends, SVM classifier (100% accuracy) and SVR regression ($R^2 \approx 0.99$) for adulteration quantification [2].
- *Yang et al. (2020)* – NIR aquaphotomics for Mānuka syrup adulteration, multivariate regression to detect 10–50% syrup adulterants [4].
- *Ahmed (2024)* – Hyperspectral imaging, multiple ML algorithms (ANN, SVM, etc.), >98% accuracy on multi-origin adulteration classification [1].
- *Wu et al. (2023)* – Raman spectroscopy with 2D-COS and deep CNN (DRSN), RMSE ~3% in predicting adulterant levels in Mānuka honey [6].
- *He et al. (2019)* – Microscopy images of pollen, deep learning segmentation + classification, effective botanical authentication (with limitations on non-pollen additives) [5]

These studies collectively demonstrate the breadth of ML applications in honey quality control, setting the stage for a deeper look into the methodologies and results for the specific case of Mānuka honey.

## 3. Methodology.

While each study has unique aspects, we can outline a general methodology for predicting impurities in Mānuka honey using machine learning. Raw data typically require preprocessing to enhance signal quality and remove noise or irrelevant variation. For spectral data, this can include smoothing (e.g., Savitzky–Golay filter), baseline correction, normalization, and derivative transforms to resolve overlapping peaks [2]. Dimensionality reduction is also important, given spectra can have hundreds of wavelength features, many of which are correlated. Techniques such as principal component analysis or even domain-specific methods like selecting key wavelengths (features) by genetic algorithms or Boruta algorithm have been applied [2]. In one study, Ahmed (2024) used an autoencoder (a neural network for unsupervised feature learning) to reduce the dimensionality of hyperspectral data, as well as feature ranking methods to pick the most informative bands [1]. For image data, preprocessing might involve contrast enhancement, resizing/cropping, or color channel selection. Pollen images, for example, were processed and then a segmentation model was trained to isolate pollen grains from the background [5].

With prepared features, ML models can be trained. To simply detect if a honey sample is pure or adulterated (a binary classification) or to identify the type of impurity (multi-class classification), algorithms like SVM, decision trees, random forests, k-nearest neighbors, or neural networks are used. SVMs and RFs have been popular for high-dimensional spectral data; indeed, SVM and RF classifiers achieved 100% accuracy in detecting any adulteration in a Vis-NIR study [2]. Neural networks (including deep CNNs) are powerful especially for image-based tasks or when the relationship between features and labels is highly nonlinear. In cases where the goal is to predict the quantity of impurity (e.g., percentage of adulterant), regression techniques are needed. Traditional approaches use PLS

regression on spectral data to predict concentration. More recently, support vector regression (SVR) and even deep learning models have been applied. For example, using Vis-NIR data, Calle et al. (2023) trained an SVR model that could predict the adulteration percentage in honey blends with an $R^2$ of 0.991 and RMSE ~1.89% [2] – meaning the model's predictions were within about ±2% of the true adulterant levels, an excellent result. A specialized scenario is when we only have reliable data for authentic honey, and we want to detect any anomaly as a potential impurity. One-class classifiers (like one-class SVMs or autoencoder-based detectors) can be trained on genuine Mānuka honey data to model the "normal" class. Cheng et al. (2024) implemented a GANomaly model (which uses a generative adversarial network to learn the distribution of authentic data) on HSI images of Mānuka honey.

During training, techniques like cross-validation are employed to tune hyperparameters and avoid overfitting, especially given the relatively limited number of honey samples typically available. Data augmentation may be applied in some cases (for instance, augmenting spectral data by adding noise or augmenting image data by rotations/flips) to increase the robustness of the model.

After training, the models are evaluated on unseen test data or via cross-validation. For classification: Accuracy, precision, recall, F1-score, and confusion matrices are reported. High accuracies have been common; e.g., 98–100% accuracy for detecting adulteration in controlled studies [1] [2].. In multi-class scenarios (e.g., classifying which adulterant is present), performance might vary by class; authors often report if certain adulterant types are harder to identify. In Ahmed's HSI study, most misclassifications occurred between specific botanical classes of honey (like confusion between clover and other honey types) [1], highlighting the importance of a diverse training set. For regression: $R^2$, RMSE (root mean squared error), and MAE (mean absolute error) are used to judge how closely predicted impurity levels match true values. An $R^2$ near 1.0 and low RMSE (on the order of a few percent for concentration) indicate successful calibration [2]. For one-class models: Metrics like false positive rate (flagging pure honey as adulterated) and false negative rate (missing a fraudulent sample) are relevant. Since one-class models don't output a class label per se, one often sets an anomaly score threshold and reports detection rate. Cheng et al. (2024) would have evaluated how well their system catches known adulterated samples as anomalies without false alarming on authentic samples – though specific numbers would depend on the threshold set and were not reported in the snippet available.

Each methodology step is underpinned by domain expertise and theoretical considerations. Feature selection and preprocessing leverage knowledge of spectroscopy (e.g., knowing which spectral regions relate to sugars vs. unique Mānuka compounds). Model selection is guided by understanding algorithm capabilities (e.g., SVMs handle high-dimension small-sample problems well, CNNs excel at image pattern recognition). Throughout, the approach remains interdisciplinary: it combines apiculture and chemistry knowledge with data science techniques to yield a practical solution for honey impurity prediction.

## 4. Discussion.

The successful application of machine learning to Mānuka honey impurity detection demonstrates several important points, but also raises practical considerations. The studies reviewed consistently show that ML models can achieve very high accuracy in detecting adulteration. When provided with quality data, models like SVM, RF, and ANN have little trouble separating pure vs. impure honey, as adulteration often introduces detectable spectral or compositional changes [2] [1]. The fact that SVM and RF reached 100% accuracy in one case [2] suggests that, under controlled conditions, the chemical

differences between authentic and adulterated honey are stark enough for algorithms to latch onto. Similarly, extremely high $R^2$ in regression models [2]. implies that not only classification but also quantification of adulterant levels is feasible with fine precision. Deep learning approaches further offer the ability to discover complex feature relationships; Wu et al.'s DRSN model capturing nonlinear patterns in Raman 2D spectra is one example of leveraging deep networks for subtle signal detection [6].

Implementing these ML solutions outside the lab involves considerations of cost, speed, and user expertise. One advantage of many ML-compatible techniques (like NIR spectroscopy or imaging) is that they can be fast and require minimal sample prep, which is crucial for industry adoption [2]. Handheld NIR spectrometers now exist that could, in principle, be loaded with a pretrained model to test honey on-site (e.g., at a beekeeping operation or import checkpoint). Hyperspectral imaging was historically expensive and confined to labs, but portable HSI devices are emerging.

The 2019 study by He et al. even emphasized building a low-cost microscopy setup for pollen, using basic lab microscopes and open-source ML, aiming for something accessible to honey producers or testing labs without big budgets [5] A barrier to adoption is the need for a robust reference database – models need to be trained on authentic and adulterated samples that reflect real-world variations. Creating a reference library of authentic Mānuka honeys (capturing the natural variation due to region, season, etc.) is important so that models don't mistake normal variation for anomalies. Likewise, continuously updating models with new types of fraud (e.g., a new syrup formulation) will be necessary. This is analogous to how malware detection or credit card fraud models in IT need regular updates as adversaries change tactics.

From a theoretical ML perspective, these applications highlight interesting challenges. Honey datasets are relatively small (often only tens to a few hundred samples in studies) yet very high dimensional (spectra with hundreds of features, images with thousands of pixels). This is a regime where careful feature reduction and regularization are paramount to avoid overfitting [2]. It also explains why simpler models like SVM have excelled – kernel methods can work well on small data if the classes are separable in some high-dimensional space. Deep learning models, which usually require large datasets, have been successfully applied (like CNN on 2D-COS or pollen images) by leveraging domain-driven augmentation and architectures that embed some prior (e.g., 2D-COS imposes a structure to exploit, and DRSN includes built-in feature filtering).

The success of deep residual shrinkage networks in Wu et al.'s work [6] underscores that integrating signal denoising into the model architecture can yield performance gains in noisy, high-dimensional spectral data. As ML theory and tools progress, we may see more customized networks for spectroscopic data, or the use of transfer learning – for instance, pretraining on a large dataset of food spectra or images and fine-tuning for honey.

In the specific context of Mānuka honey, one future direction is the development of a standardized, ML-based authentication platform. With New Zealand's UMF (Unique Mānuka Factor) grading system and government-defined criteria, there is an opportunity to incorporate ML models to cross-verify those criteria rapidly. For example, a portable device that checks the chemical marker profile (via spectroscopy) and the pollen content (via imaging) with ML could provide a field-deployable authenticity score. This would augment the current chemical tests and could add a layer of defense against sophisticated fraud (like blending genuine Mānuka with just enough markers to pass tests – an ML model considering full spectral data might catch inconsistencies that a handful of markers miss).

## 5. Results.

The convergence of machine learning with advanced sensing techniques has yielded impressive results in impurity prediction for Mānuka honey and honey in general.

Supervised ML models can detect adulterated honey with very high accuracy. In a comparative evaluation using Vis-NIR spectroscopy data from honey samples, both SVM and Random Forest classifiers achieved **100% accuracy** in distinguishing pure honey from honey adulterated with up to 50% cheap honey [2]. Similarly, Ahmed (2024) reported that various ML algorithms (including ANN and SVM) on hyperspectral data all exceeded **98% classification accuracy** for identifying adulterated samples [1].

Taken together, the results across various investigations reinforce that machine learning methods are not only theoretically sound for detecting honey impurities, but in practice they deliver high accuracy and actionable insight. ML models can effectively serve as "virtual chemical sensors," picking up on the multi-dimensional spectral or image cues of authenticity. They can provide a level of assurance (e.g., 98–100% confident classification) that rivals or exceeds traditional tests, and do so with greater speed.

## 6. Conclusion.

In conclusion, machine learning has proven to be a "sweet" solution for honey authenticity – particularly for Mānuka honey, which has a lot riding on its purity. It provides tools that are fast, accurate, and adaptable, embodying the modern approach to an age-old problem of food fraud. By continuing to refine these methods and address practical deployment challenges, stakeholders can ensure that the extraordinary honey labeled as "Mānuka" is truly worthy of the name. The success of this endeavor not only protects consumers and honest producers but also showcases how interdisciplinary innovation (melding IT, biology, and chemistry) can enhance food safety and quality assurance in the 21st century.

## References:


1. Ahmed, E. (2024). Detection of honey adulteration using machine learning. PLOS Digital Health, 3(6), e0000536. https://doi.org/10.1371/journal.pdig.0000536
2. Calle, J. L. P., Punta-Sánchez, I., González-de-Peredo, A. V., Ruiz-Rodríguez, A., Ferreiro-González, M., & Palma, M. (2023). Rapid and automated method for detecting and quantifying adulterations in high-quality honey using Vis-NIRs in combination with machine learning. Foods, 12(13), 2491. https://doi.org/10.3390/foods12132491
3. He, C., Gkantiragas, A., & Glowacki, G. (2019). Honey authentication with machine learning augmented bright-field microscopy. arXiv:1901.00516 [cs.LG]. https://doi.org/10.48550/arXiv.1901.00516
4. Yang, X., Guang, P., Xu, G., Zhu, S., Chen, Z., & Huang, F. (2020). Manuka honey adulteration detection based on near-infrared spectroscopy combined with aquaphotomics. LWT – Food Science and Technology, 132, 109837. https://doi.org/10.1016/j.lwt.2020.109837
5. Fadelli, I. (2019, January 21). Researchers develop a machine learning method to identify fake honey. TechXplore/Phys.org. https://phys.org/news/2019-01-machine-method-fake-honey.html
6. Zhao, M., Zhong, S., Fu, X., Tang, B., & Pecht, M. (2020). Deep residual shrinkage networks for fault diagnosis. IEEE Transactions on Industrial Informatics, 16(7), 4681–4690. https://doi.org/10.1109/TII.2019.2943898