

An efficient machine learning framework for optimizing hyperspectral data analysis in detecting adulterated honey

Ashwini N. Yeole, Guru Prasad M. S., Santosh Kumar

Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, India

Article Info

Article history:

Received Sep 22, 2024

Revised Apr 6, 2025

Accepted Jul 2, 2025

Keywords:

Feature selection
Food adulteration
Honey
Machine learning
Spectroscopy

ABSTRACT

Honey adulteration detection involves employing spectral data, often utilizing machine learning (ML) techniques, to identify the presence of impurities or additives in honey. This study aims to explore ML models through the collection of a hyperspectral honey dataset with limited samples and 128 features. Three distinct feature selection (FS) methods i.e., Boruta, repeated incremental pruning to produce error reduction (RIPPER), and gain ratio attribute evaluator (GRAE) are applied to extract important features for decision-making. Then, the feature-selected dataset is classified through four effective ML algorithms, such as support vector machine (SVM), random forest (RF), logistic regression (LR), and decision tree (DT). Accuracy, F1-score, Kappa Statistics, and Matthews correlation coefficient (MCC) are the performance metrics used to assess the results of ML algorithms. RIPPER FS technique gave the best results by improving its accuracy values from 79.05% (primary data) to 91.89% (augmented data) for the RF classifier model and 74.93% (primary data) to 91.89% (augmented data) for the DT classifier model. These detailed examinations of the experiments demonstrate that proper finetuning of the ML methods can play a vital role in optimizing hyperspectral data analysis for detecting adulteration levels in honey samples.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Guru Prasad M. S.

Department of Computer Science and Engineering, Graphic Era (Deemed to be University)
Dehradun, India

Email: guru0927@gmail.com

1. INTRODUCTION

The domain of food safety presents a complicated and diverse situation characterized by a plethora of risks and impurities, which have an impact on various regions worldwide. The need for food products of exceptional quality and safety is a universally anticipated requirement, thus making food security a matter of foremost importance for the well-being of the general public. The immediate health concerns caused by the intake of contaminated food lead to long-lasting harmful effects such as dietary deficiency and the spread of diseases. The frets related to external pollutants, microbial contamination, and the presence of chemicals and harmful toxins are incorporated into the vast extent of food safety. The issue impacts both developing and developed nations and has worldwide ramifications. Understanding the severity of this problem, it becomes essential to effectively tackle these challenges and implement strategies to raise the bar for food safety [1].

Honey a renewable sugar substitute prominent for its nutritive and therapeutic uses is in popular demand and is experiencing trouble sustaining its authenticity and purity. A surge in demand has stimulated the honey adulteration business placing this valuable product's integrity and quality at risk. Common adulterants like corn syrup, sugar syrup, beet syrup, rice syrup, or water used in tampering with honey, reduce its purity causing serious health issues. The detection and elimination of these harmful additives is of

utmost importance for the producers, authorities, and consumers. A simple strategy to cope with this problem is using efficient machine learning (ML) models [2]. These models have the potential to explore complex designs and slight modifications in the honey structure to differentiate genuine honey from its fake counterparts. This approach determines a proactive structure to guarantee quality honey market by enhancing the speed and accuracy of spotting tainted honey. The goal of this project is to investigate and apply ML models that are specially designed for this particular situation of identifying adulteration in honey. Employing data analytics skills, the conventional monitoring process within the honey industry can be changed by using reliable and scalable ML models to protect the authenticity of this precious element. Through the utilization of constructive ML techniques, we anticipate a significant advancement in preserving the honeys virginity, establishing consumer trust, and maintaining the reputation of this traditional natural sweetener [3].

Honey adulteration being a major risk in the food industry calls for the use of labor-intensive and perhaps dangerous chemical screening procedures. However, recent advancements have facilitated the quick and easy identification of adulterated honey. Artificial intelligence and infrared spectroscopy are two of these cutting-edge techniques that stand out. Infrared spectroscopy essentially measures the molecular absorption of infrared light, thereby furnishing invaluable insights into the composition of honey instances. Although presently in its beginning stages, the amalgamation of infrared spectroscopy with ML algorithms evinces great potential in terms of detecting adulteration in honey. Using spatial data in addition to spectrum data, hyperspectral imaging builds on spectroscopy and is a promising method for ensuring the quality of food. In terms of acquiring morphological and chemical data from food and food items, hyperspectral imaging is a significant technological advancement. Hyperspectral imaging (HSI) combined with methods for ML can be used to detect honey adulteration. By investing in these innovative approaches, the culinary industry stands to bolster the authenticity and quality assurance of honey products, thereby ensuring consumer confidence and safety. The study aims to predict characteristic features in honey. The research focuses on the detection of tainted honey using ML and HSI techniques [4].

The proposed ML-based classification model is used to determine the degree of adultery in the samples of honey. These samples are comprised of a typical dataset of samples of contaminated as well as pure honey using HSI. Techniques for ML are employed to train a classifier on this dataset, that subsequently is employable to calculate the amount of sugar syrup that is included in honey samples. The concept that the model should apply to any variety of honey samples is a crucial one that holds the potential to be used all over the world for a wide variety of honey types. This study gathers a honey dataset from a standard source and four distinct but effective ML algorithms (support vector machine (SVM), random forest (RF), logistic regression (LR), and decision tree (DT)) were applied to this unstructured data. In the next step, three FS techniques (Boruta, RIPPER, and GRAE) are implemented on the unstructured honey dataset to extract the most informative subset of features, upon which the above four ML models were applied. As it was observed that the ML models did not perform well on these fewer samples of data, synthetic data was generated using the CTGAN algorithm. To check the competency of the above classification model it was again applied to the synthetic data generated to first obtain the best feature-selection technique and then determine the most effective classification model among them. The work contributions are as follows:

- Integrated synthetic data generated via the CTGAN algorithm with a standard dataset to address sample size limitations, enhancing model robustness.
- Implemented advanced FS techniques (Boruta, RIPPER, GRAE) to isolate key features, optimizing classifier performance effectively.
- Evaluated and enhanced the performance of SVM, RF, LR, and DT models through tailored FS, achieving perfect scores in all key metrics with RIPPER for RF and DT models.

The remaining portion of the document is organized subsequently. Section II demonstrates the examination of the existing literature conducted within the scope of this research, Section III presents an exposition of the proposed research methodology and depicts the materials employed in the study. Section IV analyses detailed experimental outcomes while Section V brings the observations and conclusions to a close.

2. LITERATURE REVIEW.

Rachineni *et al.* [5] discussed that food adulteration presents serious problems for customers influencing the food industry, especially when it comes to honey. Honey is often diluted with an assortment of adulterants. Traditional strategies are costly, time-consuming and frequently operator-dependent. Nuclear magnetic resonance (NMR) spectroscopy is effective but hinder scalability for numerous honey samples. The combination of supervised ML and NMR spectroscopy shows promise in automating the identification of adulterated honey, benefiting food chemistry researchers and the food industry. Moreover, the efficacy of this methodology establishes the groundwork for its utilization in verifying the authenticity of other food

products, including tea, oils, spices, and more, indicating its potential for wider uses in ensuring food quality and safety.

Hao *et al.* [6] shows how various fluorescence spectroscopic detection techniques were used to determine the authenticity of acacia honey, or *Robinia pseudo acacia* L. honey. While techniques like DNA technology, NMR spectroscopy, high-performance liquid chromatography (HPLC), and intelligent sensing can be expensive and time-consuming, fluorescence spectroscopy provides a rapid, precise, and effective means for verifying the authenticity of honey. The study revealed notable discrepancies in fluorescence durations, peak values, and intensities among syrups, concentrated acacia honey, and acacia honey, underscoring the sensitivity and reliability of fluorescence spectrometry in detecting honey adulteration and upholding the genuineness of honey products.

Al-Awadhi and Deshmukh [7] illustrates that the quality of food products like honey is altered due to food adulteration causing an impact on the health and economy. Honey is often targeted due to its value and the ease of adding cheap sugar syrups without affecting its taste or look, making it prone to adulteration. HSI technology merges spatial and spectral data for a 3D image capturing sample properties. HSI has shown exceptional performance in identifying the botanical origins of honey and identifying adulteration when used with methods of ML as the K-nearest neighbors (KNN) and SVM. While previous research has mostly concentrated on the classification of botanical origins or the detection of adulteration independently, this study has developed an ML-based approach that combines the two tasks, using HSI data to simultaneously classify the botanical sources of honey and spot adulteration. This study's dataset, which includes spectrum examples from hyperspectral photographs of samples of pure and contaminated honey, made it easier to create and assess the recommended approach.

Noviyanto and Abdullah [8] assert that prior studies focused on using optical spectroscopy for examining honey, neglecting the potential of HSI. Optical spectroscopy is extensively studied for its rapid and non-invasive analysis, but HSI is not fully utilized despite its attractive features. Through obtaining spatially resolved spectrum information, HSI provides unique advantages over standard spectroscopy. This allows for the development of hypercubes that encompass both spectral and spatial dimensions. With the ability to anticipate distinct ingredients and classify various kinds, including authenticity, adulteration, brand identification, and geographical and botanical origin, this special capability offers a novel way to advance honey analysis. A thorough framework for data collecting and handling, as well as a standardized dataset for honey hypercubes, are conspicuously absent from the literature, despite the potential advantages of HSI in honey analysis. This work aims to address this gap by suggesting a methodical process to create the first common honey hypercube dataset in addition to a flexible and scalable dataset module that makes it easier to work with ML tools. This study seeks to promote further developments in HSI-based honey analysis by creating such standards and tools.

Phillips and Abdulla [9] expressed that as honey is the third most contaminated food commodity, honey fraud, especially the mixing of honey and sugar, is a serious worldwide problem. The combination of ML approaches and recent developments in HSI presents a potential approach to honey adulteration detection. This research proposes a novel approach that uses HSI and ML to detect contaminated honey samples with excellent accuracy-above 95% for both binary adulteration detection and multi-class classification over a range of adulterant concentrations. Prior research has investigated a range of techniques for evaluating the quality of honey, including chemical analyses, assessments of taste and smell, and spectroscopy methods including visible and near-infrared (VNIR) and fourier transform infrared (FTIR) spectroscopy. These techniques may not have the same spatial resolution as HSI, but they have demonstrated promise in identifying adulteration, especially in premium honey varieties like Manuka honey. One benefit of HSI is that it can record both spatial and spectral data, which enables it to analyze honey samples in great detail. Although HSI has been applied to classify honey according to its botanical origins or qualities in earlier research, its use in identifying contaminated honey has been restricted. The main contributions are the production of a publicly accessible dataset of tampered honey samples detected by HSI and the invention of a novel feature smoothing method to improve classification accuracy. To accurately detect the sugar concentration in honey, the suggested approach entails training an ML classifier on the dataset of contaminated honey samples. Large datasets may be captured with HSI, though, which makes ML models more reliable and improves adulteration detection accuracy. Finally, employing HSI and ML, this research proposes a new and robust approach to honey adulteration detection. Future goals for the research include broadening the dataset to encompass a more diverse range of adulterants and honey kinds, as well as assessing and enhancing the ML models generalization to new honey varieties.

3. METHOD

The methodology section illustrates a comprehensive framework, as shown in Figure 1, which outlines our approach. Initially, a detailed exposition of the dataset used for honey adulteration detection is presented, outlining its composition and characteristics. Following this, our discussion expands to the techniques employed for data analysis to enhance the size of the datasets. Moreover, a detailed description of three distinct FS methods is provided, which include Boruta, RIPPER, and GRAE. Lastly, we debate the incorporation of four ML models - SVM, RF, LR, and DT, each meticulously chosen for their distinct strengths and appropriateness for the categorization assignment.

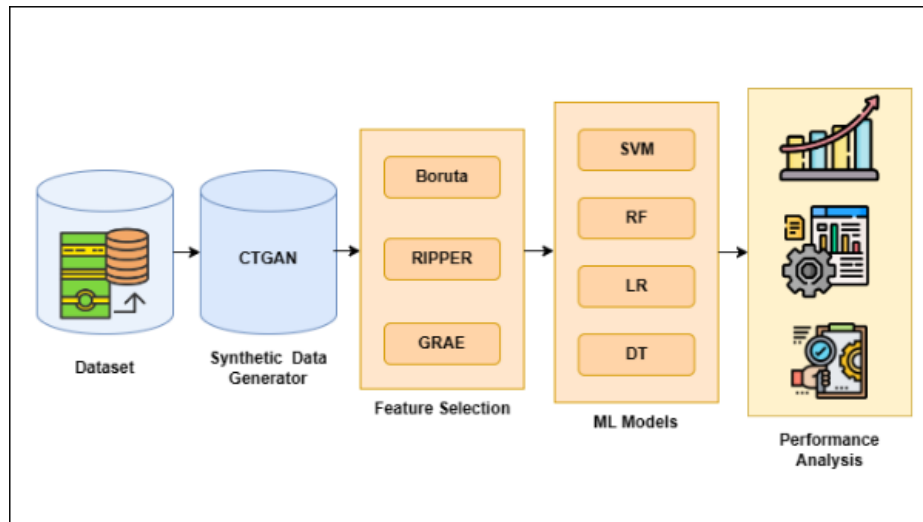


Figure 1. Proposed framework for honey adulteration detection

3.1. Dataset description

In this work, data is collected from a standard source [9]. The collected data was hyperspectral. The complete dataset contains a total of 8,675 instances and each instance is characterized by 128 features. By eliminating entries with missing values, we cleaned the dataset to make the model simpler and improve the classification accuracy. We then eliminated a few redundant and insignificant features (i.e., those unrelated to the honey datasets spectral characteristics) such as the brand of the datasets honey samples, their relative concentration, and the class name belonging to the particular brand respectively. The spectral values between 400 nm to 1,000 nm were used for the classification of adulterated honey.

3.2. Synthetic data generation/ data augmentation

Augmented data accurately reflects real-world events based on mathematical and statistical principles [10], [11]. Generative models are the most ideal methods to be employed for creating augmented data. CTGAN proves to be exceptionally beneficial when the availability of actual real-world data is restricted or when it is of a sensitive nature [12], [13]. The CTGAN design is founded on a GAN, a generative adversarial network, where two neural networks (i.e., discriminator and generator) are pitted against each other in training. Gradually, the generator becomes more proficient in generating artificial data that the discriminator cant differentiate from real data [14]. Table 1 shows the description of the primary as well as augmented data.

Table 1. Primary and augmented dataset description

| Sl. No. | Honey type | Primary data samples | Augmented data samples |
|---------|----------------------------------|----------------------|------------------------|
| 1. | 0% Without adulteration | 1,200 | 2,300 |
| 2. | 5% Adulterated with sugar syrup | 1,825 | 4,850 |
| 3. | 10% Adulterated with sugar syrup | 1,850 | 4,150 |
| 4. | 25% Adulterated with sugar syrup | 1,875 | 3,000 |
| 5. | 50% Adulterated with sugar syrup | 1,925 | 6,200 |

3.3. Feature selection techniques

Feature selection (FS) is a strategy for picking a useful subset of features from a larger set. The goal is to eliminate redundant and irrelevant features that either hinder or do not contribute to the learning process [15], [16]. Three options are accessible for recognizing the finest feature subset: filter, wrapper, and hybrid techniques. Filter methods use information theory steps like interclass distance to find relevant features. Wrapper methods focus on classifier performance for FS. Hybrid methods combine filtering and wrapping benefits [17]. The honey dataset utilized Boruta, RIPPER, and GRAE FS methods summarised below to extract relevant features.

3.3.1. Boruta

Developed as an extension of the RF algorithm, Boruta employs a novel and robust approach to identify the most relevant features in a dataset. By iteratively refining the feature set based on this rigorous comparison process, Boruta effectively identifies a feature subset that exhibits a genuine predictive power while mitigating the risk of overfitting. The methodical approach of Boruta makes it a significant tool in various domains of research. Renowned for its ability to handle high-dimensional datasets and complex feature interactions, Boruta stands as a powerful tool for enhancing the performance, interpretability, and generalization capabilities of ML models [18].

3.3.2. Repeated incremental pruning to produce error reduction (RIPPER)

Prominent and frequently employed rule-based classification algorithm RIPPER operates by generating decision rules iteratively using the training data and applying the heuristic pruning method to improve and minimize the set of rules. Subsequently, using the pruning mechanism RIPPER eliminates the redundant or less informative rules thus giving a more enhanced rule set and reduced model complexity as included in (1).

$$Gain(R_0, R_1) = c \left[\log \left(\frac{T_1}{T_1 + F_1} \right) - \log \left(\frac{T_0}{T_0 + F_0} \right) \right] \quad (1)$$

Where, R_0 = initial rule, R_1 = rule following the conjunct, c = total actual instances of R_0 and R_1 , T_0 = number of true instances by R_0 , T_1 = number of true instances by R_1 , F_0 = number of false instances by R_0 , F_1 = number of false instances by R_1 [19], [20]. This iterative cycle of rule generation, addition, and pruning continues till a predefined stopping criterion is attained.

3.3.3. Gain ratio attribute evaluator (GRAE)

Determining the most informative features from high-dimensional datasets, GRAE stands as a proficient and effective FS method. This ensemble-based approach integrates the evolutionary search capabilities of genetic algorithms, the discerning power of rough sets, and the adaptability of ensemble learning principles. Additionally, rough set theory is employed to assess the discernibility of features within each subset, ensuring the retention of non-redundant and informative features [21]. Each attributes gain ratio value to the class variable is determined in mathematical reference (2) below, where, GR is the gain ratio. Renowned for its systematic approach, adaptability, and effectiveness in handling high-dimensional data, GRAE represents a valuable tool for FS within the field of ML.

$$GR(C, A) = \frac{H(C) - H(C|A)}{H(A)} \quad (2)$$

3.4. Machine learning algorithms

Various ML algorithms have been useful in overcoming the issues of spectral data to make miniature spectroscopy a widely accepted technology for effective and efficient food quality checks. Feature subsets were generated using the above FS techniques. Four classifiers i.e. SVM, RF, LR, and DT were applied to each of these subsets and their performance was compared to detect the best classifier. A brief description of all the classifiers is given:

3.4.1. Support vector machine

Useful for both regression and classification applications, SVM is a sophisticated approach to supervised ML. Maximizing the space or margin between data points and decision border being the main focus, this technique is referred to as a maximum margin classifier. Different SVM types are used for different types of data i.e. when dividing data that cannot be divided by a straight line, non-linear SVM is employed, and linear SVM is used for data that can be divided directly [22]. The function in (3) defines the linear kernel, where w is the hyperplanes normal vector, and b is its distance from the origin. SVM works

well in high-dimensional instances and consumes less memory because the decision function only utilizes a portion of training data referred to as support vectors.

$$K(w, b) = w^T x + b \quad (3)$$

3.4.2. Random forest

Random forest (RF) is a widely utilized ensemble method that exhibits high effectiveness in combining information from multiple DTs to enhance predictive power and mitigate overfitting risk in both classification and regression tasks. It functions in two distinct phases i.e., the process of creating a forest by combining a specified quantity of DTs and subsequently making predictions for each tree. RF randomly selects data points from the training set constructing DTs to forecast results for fresh data points either by averaging or by majority voting [19]. Mathematical reference (4) is the prediction \hat{y} , where N denotes the number of forest trees and $f_i(x)$ is forecast for i th DT for input features x . Features of the RF method are its flexibility while processing large amounts of data and good accuracy even in the presence of a substantial portion of missing data.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N f_i(x) \quad (4)$$

3.4.3. Logistic regression

Logistic regression (LR) stands as a supervised ML algorithm implemented in predictive analysis and is founded on the concept of probability. The sigmoid function (S-shaped curve) converts every real integer to a value between 0 and 1. Its specific design caters to scenarios where the data exhibits linear separability and the outcome is binary or dichotomous in nature. Imbalanced class representation generally leads to a skewed class distribution which hampers the LR performance [23]. A sigmoid function in (5) converts predictions into probabilities and gives the LR hypothesis as in (6), where, e represents the natural logarithm base, X the input features (predictors), θ the hypothesis function, and β gives the model coefficients (parameters).

$$f(x) = \frac{1}{1+e^{-(x)}} \quad (5)$$

$$h\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1X)}} \quad (6)$$

3.4.4. Metrics for evaluating classifier performance

Performance measurement plays a pivotal role in the assessment of a classification models capability to accurately forecast instances and attain a desired target [19]. The confusion matrix offers a more thorough summary of a predictive models performance and is classified into one of four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [24], [25]. In this study, evaluation metrics consist of the following metrics to assess the effectiveness of the four classifiers. Here, P_o is observed agreement between raters, and P_e is expected agreement between raters for random guessing.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (7)$$

$$F1 = \frac{(2*Precision*Recall)}{(Recall+Precision)} \quad (8)$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (9)$$

$$MCC = \frac{(TP*TN - FP*FN)}{\{(TP+FP)(TP+FN)(TN+FP)(TN+FN)\}^{1/2}} \quad (10)$$

4. EXPERIMENTAL RESULTS ANALYSIS

A comprehensive evaluation of honey adulteration detection is conducted as shown in Figure 2 below. The performance of ML models and FS methods are assessed across two datasets: primary and augmented. Our analysis aims to elucidate the effectiveness of ML models and FS methods in improving Honey adulteration detection accuracy and contributing to advancements in the field of food quality assurance and consumer protection.

4.1. Experimental setup

The research is conducted using Kaggle Notebooks, an open-source cloud-based technology offered by Kaggle. The scikit-learn package (1.3.0) and libraries like pandas and matplotlib for the Python programming language are used to accomplish the duties of FS, classification, artificial data generation, data preprocessing, and plotting graphs efficiently to effectively complete the experiment. In this study, a prediction model on honey adulteration datasets is developed using a 5-fold cross-validation approach [24], [25]. The goal of this work is to detect sugar syrup adulteration in honey. Four of the folds are employed in training during the model-building process and a fifth fold is retained for testing. The procedure is repeated five times, with the average of the results being the last step. 5-fold cross-validation, commonly used for moderate-sized datasets like the honey dataset provides a balance between computational efficiency and robust estimation of model performance. It helps in reducing the variance in performance estimation compared to using only one train-test split thus helping model generalization to unseen data. To support the experimental results few statistical evaluation metrics with mathematical references (7-10), such as accuracy, F1-score, Kappa Statistic, and MCC are taken into account, and a comparative analysis for each metric is done.

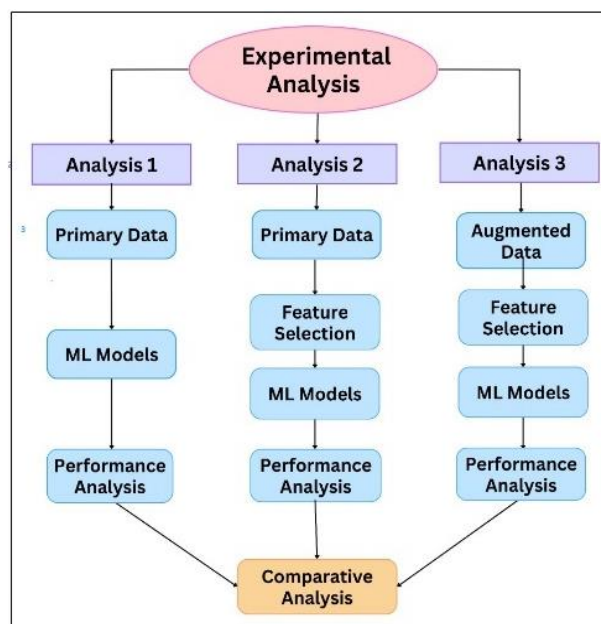


Figure 2. Experimental analysis of honey adulteration detection

4.2. Analysis of accuracy

The accuracy metric refers to the percentage of correctly identified examples in a dataset, which represents the overall number of instances. It serves as an indicator of the general accuracy of a models predictions. The accuracy results of various ML classifiers on honey dataset, using distinct FS methods, are presented in Table 2. Analysing the honey data accuracy values, it is observed that RIPPER achieves 91.89% accuracy for RF and DT while SVM and LR follow with 90.70% and 88.49% accuracy, respectively. Notably, RF exhibits the best average outcome among all the four classifiers.

Table 2. Accuracy analysis of various classifiers with the honey dataset

| Dataset | Sample size | Feature selection | ML models | | | |
|----------------|-------------|-------------------|-----------|-------|-------|-------|
| | | | SVM | RF | LR | DT |
| Primary data | 8,675 | | 46 | 79.05 | 49 | 74.93 |
| Primary data | 8,675 | Boruta | 54 | 87.02 | 54.09 | 85 |
| | | RIPPER | 82.56 | 84.54 | 68.58 | 84.54 |
| | | GRAE | 54.06 | 86.76 | 51.04 | 87.46 |
| | | | | | | |
| Augmented data | 20,500 | Boruta | 78.68 | 91.82 | 72.21 | 90.66 |
| | | RIPPER | 90.70 | 91.89 | 88.49 | 91.89 |
| | | GRAE | 75.02 | 90.80 | 70.21 | 90.56 |
| | | | | | | |

4.3. Analysis of F1-score

The F1-score serves as an evaluation metric for a models precision and recall, especially beneficial in cases of imbalanced class distribution. Table 3. showcases the F1-score results of different ML classifiers on feature-selected honey datasets. Analysis of the feature-selected honey dataset reveals that RF and DT yield a top F1-score value of 92.75% and 92.56% respectively, trailed by SVM and LR attaining 90.25% and 88.73% respectively, using the RIPPER method. Similarly, RF and DT demonstrate strong performance with F1-scores above 90% for all three FS methods.

Table 3. F1-score analysis of various classifiers with honey dataset

| Dataset | Sample size | Feature selection | ML models | | | |
|----------------|-------------|-------------------|-----------|-------|-------|-------|
| | | | SVM | RF | LR | DT |
| Primary data | 8,675 | | 53 | 78 | 49 | 77 |
| Primary data | 8,675 | Boruta | 66.72 | 85.82 | 56.70 | 86.72 |
| | | RIPPER | 72.05 | 87.26 | 70.53 | 88.77 |
| | | GRAE | 54.66 | 85.76 | 57.49 | 85.46 |
| Augmented data | 20,500 | Boruta | 87.60 | 90.82 | 79.03 | 90.66 |
| | | RIPPER | 90.25 | 92.75 | 88.49 | 92.56 |
| | | GRAE | 81.60 | 91.80 | 73.12 | 90.56 |

4.4. Analysis of Kappa

Kappa statistics are commonly used to evaluate the agreement between the predicted labels and the true labels. The increased value of kappa signifies improved predictive performance, reflecting a heightened level of similarity between observed and forecasted values. The kappa values of different ML classifiers applied to various feature-selected honey datasets are displayed in Table 4. Examining the feature-selected honey datasets kappa results, it is observed that DT yields a top Kappa value of 92.05%, while RF gives 91%, trailed by SVM and LR attaining 90.06% and 88.45% respectively, using the RIPPER method. Similarly, RF demonstrates strong performance with Kappa above 90% for all three FS methods.

Table 4. Kappa analysis of various classifiers with honey dataset

| Dataset | Sample size | Feature selection | ML models | | | |
|----------------|-------------|-------------------|-----------|-------|-------|-------|
| | | | SVM | RF | LR | DT |
| Primary data | 8,675 | | 58 | 79 | 51 | 77 |
| Primary data | 8,675 | Boruta | 64.72 | 85.82 | 57.76 | 86.72 |
| | | RIPPER | 72.05 | 87.26 | 70.53 | 86.77 |
| | | GRAE | 54.66 | 85.76 | 57.49 | 85.46 |
| Augmented data | 20,500 | Boruta | 80.61 | 90.75 | 74.01 | 86.23 |
| | | RIPPER | 90.06 | 91.00 | 88.45 | 92.05 |
| | | GRAE | 74.61 | 90.07 | 71.44 | 90.20 |

4.5. Analysis of MCC

MCC takes into account true positives, true negatives, false positives, and false negatives, providing a measure of the quality of binary classifications. The higher value of MCC represents better prediction and a strong correlation between actual and predicted class. The MCC values of various ML classifiers on distinct feature-selected honey datasets are given in Table 5. Examination of the feature-selected honey dataset reveals that RF and DT demonstrate the highest MCC rate of 92% for the RIPPER approach while SVM and LR give 90.07% and 91.02% MCC values respectively. Furthermore, RF exhibits superior average results compared to other classifiers.

Table 5. MCC analysis of various classifiers with honey dataset

| Dataset | Sample size | Feature selection | ML models | | | |
|----------------|-------------|-------------------|-----------|-------|-------|-------|
| | | | SVM | RF | LR | DT |
| Primary data | 8,675 | | 60 | 76 | 55 | 76 |
| Primary data | 8,675 | Boruta | 64.72 | 81.84 | 67.06 | 38.72 |
| | | RIPPER | 72.05 | 84.89 | 72.52 | 89.07 |
| | | GRAE | 58.66 | 83.28 | 59.09 | 82.95 |
| Augmented data | 20,500 | Boruta | 84.61 | 90.31 | 70.03 | 40.02 |
| | | RIPPER | 90.07 | 92.00 | 91.02 | 92.00 |
| | | GRAE | 78.61 | 90.75 | 76.51 | 91.20 |

5. CONCLUSION

In this study, we proposed an ML-based framework for the detection of honey adulteration. We demonstrate that ML-based predictive models are practical instruments for this purpose. After completing the initial data preprocessing and augmentation techniques, the three FS (Boruta, RIPPER, and GRAE) techniques are applied to extract features from the datasets. The four efficient ML algorithms (SVM, RF, LR, and DT) are applied to the feature-selected datasets for classification. The performance analysis of the proposed ML model identifies the most effective FS and classification strategies. Experimental findings of the proposed work are justified through different statistical evaluation measures (accuracy, F1-score, Kappa, and MCC). It is observed that the effectiveness of the model trained using a feature-selected augmented dataset is far better than the feature-selected primary dataset. RIPPER FS technique outperformed with 91.89% accuracy value, 92.75% f1-score value, 91% Kappa value, and 92% MCC value for the RF classifier model and trained on the feature-selected datasets. This work demonstrates synthetic data generation using the CTGAN model, detects the most efficient FS technique to optimize classifier performance effectively, and evaluates the efficiency of ML models in detecting food adulteration. Our study was limited by the fact that there was not enough data to create a generalized model for honey of all types and different origins. In the future, we plan to gather more data connected to honey adulteration and build a prediction model that is more universal for any type of honey sample and any origin of honey samples to detect sugar syrup or any other type of adulterant in the honey.

ACKNOWLEDGMENTS

The authors gratefully acknowledge for this project under an internal research grant. The authors are thankful to the administration of the University for facilitating laboratory access to complete the work.

FUNDING INFORMATION

This research was financially supported by Graphic Era (Deemed to be University), Dehradun.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Ashwini N. Yeole | ✓ | ✓ | | | ✓ | | | | ✓ | | | | | ✓ |
| Guru Prasad M.S. | | ✓ | ✓ | | | ✓ | | | | ✓ | | ✓ | | |
| Santosh Kumar | | | | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ | | ✓ | |

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in [Kaggle repository] at <https://www.kaggle.com/datasets/ashwiniteenyeole/honey-data1>, (Phillips & Abdulla, 2023) [9].




REFERENCES

- [1] D. P. Aykas, "Determination of possible adulteration and quality assessment in commercial honey," *Foods*, vol. 12, no. 3, p. 523, Jan. 2023, doi: 10.3390/foods12030523.
- [2] G. Machuca *et al.*, "Hyperspectral microscopy technology to detect syrups adulteration of endemic guindo santo and quillay honey using machine-learning tools," *Foods*, vol. 11, no. 23, p. 3868, Nov. 2022, doi: 10.3390/foods11233868.




- [3] S. Sneha, S. Surjith, and S. M. Alex Raj, "A review on food adulteration detection techniques: methodologies, applications, and challenges," in *2023 International Conference on Control, Communication and Computing, ICC3 2023*, May 2023, pp. 1–5, doi: 10.1109/ICC357789.2023.10165065.
- [4] G. Machuca *et al.*, "Hyperspectral Microscopy Technology to Detect Syrups Adulteration of Endemic Guindo Santo and Quillay Honey Using Machine-Learning Tools," *Foods*, vol. 11, no. 23, 2022, doi: 10.3390/foods11233868.
- [5] K. Rachineni, V. M. R. Kakita, N. P. Awasthi, V. S. Shirke, R. V. Hosur, and S. C. Shukla, "Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification," *Current Research in Food Science*, vol. 5, pp. 272–277, 2022, doi: 10.1016/j.crfs.2022.01.008.
- [6] S. Hao, J. Yuan, Q. Wu, X. Liu, J. Cui, and H. Xuan, "Rapid identification of corn sugar syrup adulteration in wolfberry honey based on fluorescence spectroscopy coupled with chemometrics," *Foods*, vol. 12, no. 12, p. 2309, Jun. 2023, doi: 10.3390/foods12122309.
- [7] M. A. Al-Awadhi and R. R. Deshmukh, "Honey adulteration detection using hyperspectral imaging and machine learning," in *2022 2nd International Conference on Artificial Intelligence and Signal Processing, AISP 2022*, Feb. 2022, pp. 1–5, doi: 10.1109/AISP53593.2022.9760585.
- [8] A. Noviyanto and W. H. Abdullah, "Honey dataset standard using hyperspectral imaging for machine learning problems," in *25th European Signal Processing Conference (EUSIPCO)*, Aug. 2017, vol. 2017, pp. 473–477, doi: 10.23919/EUSIPCO.2017.8081252.
- [9] T. Phillips and W. Abdulla, "A new honey adulteration detection approach using hyperspectral imaging and machine learning," *European Food Research and Technology*, vol. 249, no. 2, pp. 259–272, Feb. 2023, doi: 10.1007/s00217-022-04113-9.
- [10] R. Goyal, P. Singha, and S. K. Singh, "Spectroscopic food adulteration detection using machine learning: current challenges and future prospects," *Trends in Food Science and Technology*, vol. 146, p. 104377, Apr. 2024, doi: 10.1016/j.tifs.2024.104377.
- [11] L. Tan, "Generating synthetic tabular data," *Towards Data Science*, 1384. <https://towardsdatascience.com/generating-synthetic-tabular-data-503fe823f377>.
- [12] M. H. Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Hyperspectral imaging for color adulteration detection in red chili," *Applied Sciences (Switzerland)*, vol. 10, no. 17, p. 5955, Aug. 2020, doi: 10.3390/app10175955.
- [13] A. Pathare, R. Mangrulkar, K. Suvarna, A. Parekh, G. Thakur, and A. Gawade, "Comparison of tabular synthetic data generation techniques using propensity and cluster log metric," *International Journal of Information Management Data Insights*, vol. 3, no. 2, p. 100177, Nov. 2023, doi: 10.1016/j.ijime.2023.100177.
- [14] C. Dilmegani, "Cem dilmegani, synthetic data generation in 2024: techniques & best practices, 2024," 2024. <https://research.aimultiple.com/synthetic-data-generation/>.
- [15] A. Cutler, D. R. Cutler, and J. R. Stevens, *Ensemble Machine Learning*. New York, NY: Springer New York, 2012.
- [16] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022, doi: 10.3389/fbinf.2022.927312.
- [17] U. Stańczyk and L. C. Jain, "Feature selection for data and pattern recognition: an introduction," in *Studies in Computational Intelligence*, vol. 584, 2015, pp. 1–7.
- [18] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, "Efficient machine learning models for early stage detection of autism spectrum disorder," *Algorithms*, vol. 15, no. 5, p. 166, May 2022, doi: 10.3390/a15050166.
- [19] S. M. Mahedy Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, pp. 15038–15057, 2023, doi: 10.1109/ACCESS.2022.3232490.
- [20] V. H. Khang, C. T. Anh, and N. D. Thuan, "Detecting fraud transaction using ripper algorithm combines with ensemble learning model," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 336–345, 2023, doi: 10.14569/IJACSA.2023.0140438.
- [21] Z. Kang, Y. Zhao, L. Chen, Y. Guo, Q. Mu, and S. Wang, "Advances in machine learning and hyperspectral imaging in the food supply chain," *Food Engineering Reviews*, vol. 14, no. 4, pp. 596–616, Dec. 2022, doi: 10.1007/s12393-022-09322-2.
- [22] M. A. C. Lengua and E. A. P. Quiroz, "A systematic literature review on support vector machines applied to classification," in *Proceedings of the 2020 IEEE Engineering International Research Conference, EIRCON 2020*, Oct. 2020, pp. 1–4, doi: 10.1109/EIRCON51178.2020.9254028.
- [23] A. Pant, "Introduction to logistic regression," *Medium*. <https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>.
- [24] A. Satria, O. S. Sitompul, and H. Mawengkang, "5-Fold cross validation on supporting k-nearest neighbour accuracy of making consimilar symptoms disease classification," in *Proceedings - 2nd International Conference on Computer Science and Engineering: The Effects of the Digital World After Pandemic (EDWAP), IC2SE 2021*, Nov. 2021, pp. 1–5, doi: 10.1109/IC2SE52832.2021.9792094.
- [25] T. Fontanari, T. C. Fróes, and M. Recamonde-Mendoza, "Cross-validation strategies for balanced and imbalanced datasets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2022, vol. 13653 LNAI, pp. 626–640, doi: 10.1007/978-3-031-21686-2_43.

BIOGRAPHIES OF AUTHORS






Ashwini N. Yeole    is a research scholar in the Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, Uttarakhand. She was awarded M.Tech. (IT) Degree from Graphic Era (Deemed to be University), Dehradun, and B.E. (E&TC) Degree from SSGMCE, Shegaon, affiliated to Amravati University, Amravati, Maharashtra. Her research areas are machine learning, deep learning, and data science. She can be contacted at email: anyeole@gmail.com.



Dr. Guru Prasad M. S.    is working as a professor in the Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, Uttarakhand. He got his Ph.D. in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Karnataka. He has given more than 25 technical talks and lectures at reputed conferences, universities, colleges, and events. He has authored more than 50 research papers. He has filed 5 Indian patents. He was invited as a keynote speaker and session chair at reputed international conferences. His areas of interest are machine learning, deep learning, and data science. He can be contacted at email: guru0927@gmail.com.



Dr. Santosh Kumar    received Ph.D. from Indian Institute of Technology, Roorkee (India) in 2012, M.Tech. (CSE) from Aligarh Muslim University, Aligarh (India) in 2007 and B.E. (IT) from C.C.S. University, Meerut (India) in 2003. He has more than 18 years of experience in teaching/research of UG (B.Tech.) and PG (M.Tech./Ph.D.) level courses as a Lecturer/Assistant Professor/Associate Professor/Professor in various academic /research organizations. He has supervised 05 Ph.D. Thesis and 28 M.Tech. dissertation and presently mentoring 07 Ph.D. students and 02 M.Tech. students. He has also completed a consultancy project titled “MANET Architecture Design for Tactical Radios” of DRDO, Dehradun in between 2009-2011. He is an active reviewer board member in various national/International Journals and Conferences. He has memberships of ACM (senior member), IAENG, ACEEE, ISOC (USA) and contributed more than 100 research papers in National and International Journals/conferences in the field of wireless networks, wireless sensor networks, IoT, grid computing, AI and machine learning, software engineering. Currently holding position of Professor and Chairman DRC in the Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun (India). His research interest includes wireless networks, wireless sensor networks, IoT, AI and machine learning, software engineering. He can be contacted at email: drsantosh.cse@geu.ac.in.