Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 2, pp. 215–233
DOI : 10.47974/JIOS-1546

Efficient machine learning models for the detection of coconut milk adulteration

Ashwini Niteen Yeole [†]
M. S. Guru Prasad *
Santosh Kumar [§]
Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun
Uttrakhand
India

Abstract

Coconut milk adulteration detection involves the use of analytical methods, including machine learning, to detect the presence of impurities or additives in coconut milk. Adulteration can occur when substances such as water or other cheaper ingredients are added to coconut milk, compromising its quality, nutritional value, and authenticity. Identifying adulteration is crucial for ensuring consumer safety, maintaining product quality, and upholding industry standards. The aim of this study was to propose a machine learning model to detect adulteration in coconut milk. To implement the proposed work, a coconut milk adulteration dataset is collected from a standard source. The amount of available data is limited; hence, synthetic data is generated by applying the CTGAN algorithm. The proposed framework employs three different Feature Extraction (FE) strategies, i.e., Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Autoencoder (AE). Then, the feature-extracted dataset is classified through four effective machine learning algorithms, such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). The machine learning algorithms outcomes are evaluated using four performance metrics, i.e., Accuracy, Precision, Recall, and F1-score. RF provides the highest accuracy of 99.17%, precision value of 97.29%, recall value of 96.71%, and F1 Score of 97% for the LDA techniques.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Food adulteration, Coconut milk, Machine learning, Infrared spectra, Feature extraction.*

[†] *E-mail:* **anyeole@gmail.com**
* *E-mail:* **guru0927@gmail.com (Corresponding Author)**
[§] *E-mail:* **drsantosh.cse@geu.ac.in**

1. Introduction

At present, the safety of food and beverage has become a concern for public health on a worldwide scale, particularly in terms of food cleanliness and quality [1]. Food is a worldwide essential that must be consumed on a daily basis for the purposes of obtaining energy and development, maintaining health, and preventing sickness. As a result, it is essential to make certain that only healthy and appropriately certified food items are made available to customers. But the fact of the matter is that customers are having a tough time selecting legal food goods owing to difficulties [2].

Human health is directly affected by food adulteration, making it one of the most serious consumer concerns in low-income nations. Because of rising consumer demand and falling manufacturing costs, food is sometimes tampered with. Because food that has been contaminated has had other chemicals added to it, either legally or illegally, it is difficult for people to identify it with their bare senses and requires laboratory equipment [3]. Therefore, creating food and beverage evaluations is crucial and should be prioritised to protect the public's health. Numerous technological applications and scientific processes have been created for evaluating foods and beverage [4].

A common liquid item that may be easily tampered with is coconut milk. Because it has a high concentration of saturated fat—particularly in the form of easily absorbed medium-chain fatty acids—coconut milk has several advantages over other sources of saturated fat [5]. The danger of mortality and disease is increased when coconut milk is altered in any way. For this reason, testing coconut milk for contaminants that compromise its quality is essential [6].

Coconut milk adulteration can be detected using chemical techniques; however, these approaches are damaging and time-consuming. These days, it's easy and quick to tell if coconut milk has been tampered with. Among these methods are infrared spectroscopy and artificial intelligence. Molecular absorption of infrared light is measured using infrared spectroscopy. Only a small number of studies have combined IR spectroscopy with ML methods for detecting adulteration in coconut milk.

The proposed Machine learning based classification model is used to determine the level of adulteration in the samples of coconut milk. This samples are comprised of a newly collected data set of samples of contaminated coconut milk using an infrared spectroscopic instrument, as well as a machine learning model that was trained on this data. Machine learning techniques are used to train a classifier on this data set, which can

then be used to calculate the amount of water that is included in coconut milk. The concept that the model should be applicable to any variety of coconut milk is an important one that possesses the potential to be used all over the world for a variety of coconut milk preparations.

In this work, a benchmark coconut milk adulteration dataset is collected, and then that dataset is first pre-processed. After that, three different FE approaches, such as PCA, LDA, and AE, are utilised to translate the datasets into a suitable format for subsequent evaluations. Following this step, the feature extracted datasets are categorised using four different effective classification algorithms (LR, SVM, DT, and RF), and the most effective classification models are determined. In the meanwhile, the experimental outcomes of the modified dataset are analysed in order to investigate the importance of the FE approaches that were used to the dataset.

2. Literature Survey

Dhritiman Saha et al., [1] expressed that over time, the role of non-destructive testing methods in tracking food quality has grown. Because of the spatial and spectral information, it gives, hyperspectral imaging is a significant non-destructive quality assessment technology. The potential for this method to be used in food applications has increased as machine learning methods have advanced to allow for faster analysis and greater classification accuracy. This research covers the use of several machine learning methods for analysing hyperspectral pictures to establish food quality. The article discusses the hyperspectral imaging theory, the benefits and drawbacks of various machine learning methods, and their applications. The hyperspectral pictures of food goods were quickly and accurately analysed using machine learning techniques, allowing for reliable classification and regression models to be built. In order to decrease the computing burden and time, selecting appropriate wavelengths using the hyperspectral data is crucial. This opens up more opportunities for real-time applications. Because of its feature learning nature, deep learning has emerged as one of several promising and effective methodologies for real-time applications. The entire promise of deep learning, however, cannot be realised until the field matures. same, continual machine learning provides the route for contemporary HSI solutions but requires future investigation to include the seasonal fluctuations in food quality. They also look ahead to the potential of

machine learning techniques in hyperspectral image processing and discuss their current limits.

Suhaili Othman et. al., [2] discussed that in the food and agriculture production, in particular, AI approaches have developed into practical, rapid, and successful methods when coupled with detecting equipment for quality evaluation, in particular for adulteration and deficiency identification. In this research, we looked at how current developments in artificial intelligence approaches may be used in tandem with different types of sensors to identify food adulteration and flaws in agricultural products. This study proposes research directions and guidelines for applying case-specific AI techniques to enhance the knowing of the gathered high-dimensional dataset, in addition to classification and prediction, with the aim to offer references and guidelines for academic researchers and companies operating in the field of food and agriculture quality assessment. The benefits and drawbacks of AI techniques were laid bare. It is also important to research and test prospective innovations in the form of innovative sensors and algorithms. It is reasonable to expect that, in the not-too-distant future, AI will be able to identify food adulteration and agricultural product faults, especially with the integration of multiple sensors and the use of deep learning algorithms. Possible areas of focus include real-time monitoring and predictive modelling, both of which have the potential to forewarn of quality issues, reduce the likelihood of fraud, and guarantee goods of high value.

Kavitha Rachineni et. al., [3] illustrated Nuclear magnetic resonance (NMR) is a highly effective analytical technology that may be used to verify the authenticity of honey down to the level of its chemical constituents, through the determination of their identities and the measurement of their spectrum patterns. However, it remains difficult since it could require a person-centric review or a time-consuming method to analyse a large number of honey samples in a short amount of time. In light of this, automating NMR spectrum analysis of honey with the use of supervised machine learning models speeds up the analysis process, which is especially useful for food chemistry researchers or the food business that lacks NMR specialists. In this paper, the scientists effectively proved their approach by taking into account three of the most common sugar adulterants in honey over a range of concentrations. An appropriate supervised machine learning classification analysis is carried out to put the planned work into effect.

Bezuayehu Gutema Asefa et al., [4] demonstrated that for the purpose of detecting milk that has been adulterated with water, an efficient

approach that is based on digital image analysis and a technology known as machine learning has been developed. When many different machine learning algorithms were examined, the results showed that SVM performed the best overall with an accuracy of 89.48% and a precision of 95.10%. It was shown that the performance of the categorization system improved in the extreme classes. It was possible to get a better quantitative determination of the water that was superfluous. The method that has been described may be utilised to screen raw milk on the basis of the amount of added extraneous water without the requirement of any additional reagent being utilised.

Qianqian Li et al., [5] illustrated about the the detection of adulteration with rice syrup in honey of varying botanical origin in food products. It is vital, regardless of the honey's initial categorization, to build a generic model for the detection of adulteration because the process is both time-consuming and fraught with a high risk of incorrect diagnosis. In their research, they used infrared (IR) spectra in conjunction with four different supervised pattern recognition techniques to develop a generic model for the detection of rice syrup adulteration in acacia, linden, and jujube honey samples all at the same time. This was done concurrently. In addition, the Monte Carlo sampling technique was utilised so that the models could be evaluated. A remarkable performance was delivered by the first Der-LS-SVM, which included increased accuracy (97.09%), increased sensitivity (96.64%), and increased specificity (97.58%). In addition to that, this study takes further attempts to regulate the quality of the honey product that is available on the market about the adulteration of rice syrup.

3. Methodology

In this work, data is collected from a standard source. But the collected data is limited for training machine learning models; hence, a sufficient amount of synthetic data is generated through real data. The three different FE approaches are utilised to translate the datasets into a suitable format for subsequent evaluations. Then, the feature-extracted dataset is classified through four effective machine learning algorithms. The performance of the proposed machine learning classifiers was analysed to find out which FE method is more useful to detect coconut milk adulteration. Figure 1 depicts the proposed framework.

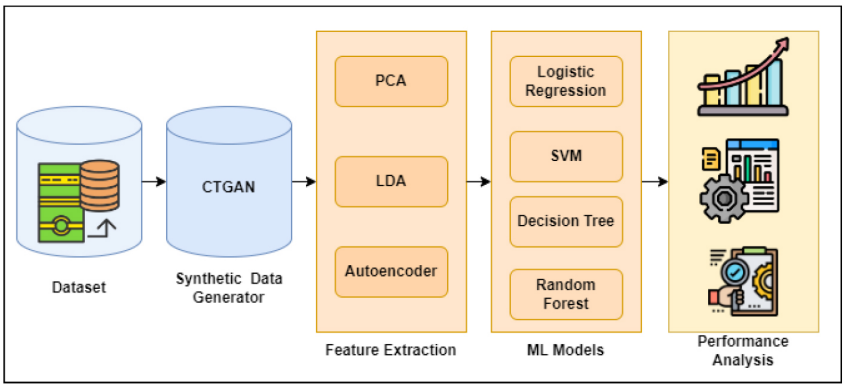

Figure 1

Proposed framework for coconut milk adulteration detection

3.1 *Dataset*

Infrared spectral data of coconut milk was collected form a standard source [11]. Absorbance spectral data were acquired and recorded in wavelength range from 2500 to 4000 nm for a total of 42 coconut milk samples. Coconut milk ware prepared in three forms of adulteration and the dataset is balanced. Coconut milk comes from traditional markets and instant coconut milk in Indonesia. Table 1 shows the description of the dataset.

Table 1

Dataset description

Sl. No.	Coconut Milk Type	No. of samples
	Without Adulteration (0%)	14
	10% Adulterated with water	14
	20% Adulterated with water	14

3.2 *CTGAN*

Conditional Tabular Generative Adversarial Network (CTGAN) is a generative model designed to create synthetic tabular data that resembles real-world data. It can generate data samples while considering the conditional dependencies present in the input dataset. This is particularly valuable when there is structured data with relationships between

Table 2
Synthetic data description

Sl. No.	Coconut Milk Type	No. of samples
	Without Adulteration (0%)	250
	10% Adulterated with water	250
	20% Adulterated with water	250

features. CTGAN is particularly useful when the amount of real-world data is limited or sensitive [7].

The CTGAN design is based on a generative adversarial network (GAN), which trains two neural networks against one another. Both the discriminator and the generator undergo extensive training to enable them to distinguish between true and fabricated information. As time goes on, the generator becomes better at generating synthetic data that is harder for the discriminator to tell apart from the genuine thing.

The amount of data collected was limited for training machine learning models; hence, CTGAN is utilized to generate a sufficient amount of synthetic data using real data. Table 2 shows the description of the synthetic data.

3.3 *Feature Extraction*

A more concise and useful representation of raw data is achieved by feature extraction, which involves selecting or transforming essential information or features.

3.3.1 Principal Component Analysis

PCA is useful for reducing the number of dimensions in a dataset without losing any of the information of interest. What PCA does is find the linear combinations of the original characteristics (the principal components) and then represent those [8]. The working of PCA technique is as follows:

a. To begin PCA, the data must be centred, which is done by eliminating the mean of each feature.
b. The centred data is then used to generate a covariance matrix. How the characteristics in the dataset are connected to one another is defined by the covariance matrix.

c. The covariance matrix must next be subjected to an eigenvalue decomposition, The eigenvalues and eigenvectors are obtained from this decomposition.

d. Each eigenvector's eigenvalue indicates how much of the data's variation it explains. When using PCA, eigenvectors are ranked by their associated eigenvalues. The primary components are the names given to these sets of eigenvectors. In order to minimise the dimensionality of the data, it is common practise to select the top K eigenvectors (principal components), where K original number of features.

e. The chosen PCs can be used to transform high-dimensional data into a more manageable area. With this new form, much of the data's inherent variability is preserved despite a reduction in dimensionality.

3.3.2 Linear Discriminant Analysis

LDA is a useful method for feature extraction and dimensionality reduction. It seeks to discover a linear combination of characteristics that best separates two or more classes in a given dataset. LDA's primary goal is to lessen the amount of variation between classes while increasing the gap between them. The working of LDA technique is as follows:

a. Mean feature values should be determined for each dataset class. The centres of these groups.

b. The within-class scatter matrix provides a numerical representation of the variation in data across categories. Covariance matrices for each group are added together and the results are weighted by the number of observations in each group to determine the overall covariance.

c. The dispersion between classes is measured by the between-class scatter matrix. It is computed by factoring in the number of data points in each class alongside the variations in class means.

d. Use the scatter matrices from both within and between classes to solve the generalised eigenvalue issue. This results in a collection of eigenvalues and eigenvectors.

e. Order the eigenvalues from lowest to highest. The primary components are chosen from the eigenvectors that map to the most

significant eigenvalues. These are the dimensions we're down to now.

f. A lower-dimensional representation of the data can be obtained by projecting the original data onto the new axes specified by the chosen primary components.

3.3.3 Autoencoder

Dimensionality reduction may be accomplished with the help of an autoencoder, a form of artificial neural network. An effective representation of data is learned by this feedforward neural network type, which is then used for feature extraction. The working of autoencoders is as follows:

a. The autoencoder's initial component is an encoder, which transforms raw data into a more compact form. The encoder, made up of one or more layers of neurons, learns to extract the most crucial properties from the input data, thereby decreasing the data's dimensionality.

b. The autoencoder's compressed representation is created at the bottleneck layer, which is also called the latent space or encoding layer. This layer captures the most important aspects of the data and has a substantially smaller dimensionality than the original data.

c. The decoder, the autoencoder's second component, decompresses the representation and reconstructs the original input data. It uses one or more layers to replicate the input as accurately as feasible.

d. Autoencoders are trained by minimising a loss function that quantifies the degree to which the reconstructed data deviates from the original.

3.4 *Machine Learning*

3.4.1 Logistic Regression

In the realm of machine learning, logistic regression serves as a common statistical technique and classification tool. It estimates the probability of two possible outcomes based on the values of a dependent variable that is either a binary or a multinomial [9]. The key characteristics of Logistic Regression are as follows:

a. Sigmoid Function: The sigmoid function is used to represent the connection between the independent factors and the binary dependent variable in Logistic Regression. Any real number

may be transformed into a value between zero and one using the logistic function. This is incredibly important when trying to depict probability.

b. Probability Estimation: The likelihood that a given input example is of a certain class is determined using Logistic Regression. Decisions involving more than two categories can also be made using these probabilities.

c. Linear Combination: Logistic Regression models the log-odds of the probability as a linear combination of the input features.

d. Training: Logistic Regression minimises a cost function during training, which is often the cross-entropy between the estimated probabilities and the actual class labels in the training data. Optimisation methods such as gradient descent are commonly used for this purpose.

3.4.2 Support Vector Machine

Support Vector Machine (SVM) is a sophisticated and adaptable supervised machine learning technique that may be utilised for both classification and regression applications. SVM is particularly well-suited for situations where the aim is to discover a distinct border (hyperplane) that best divides various classes or predicts numeric values, with an emphasis on maximising the margin or distance between data points and the decision boundary [10]. The key characteristics of SVM are as follows:

a. Linear Separability: The goal of support vector machines (SVMs) is to locate a hyperplane that cleanly divides data into distinct categories. Finding the hyperplane with the largest difference between the two classes is the goal of SVM analysis while doing binary classification. Cases that lie along this hyperplane are being supported by these vectors.

b. Margin: The margin is the average separation of each class's data points (support vectors) from the hyperplane. SVM aims to maximise this margin since doing so typically leads to improved generalisation and reduced overfitting.

c. Kernel Trick: By projecting the original feature space onto a higher-dimensional space, SVM is able to handle data that is not linearly separable. In order to translate the data to a space where linear

separation is achievable, kernel functions like polynomial kernels or radial basis function (RBF) kernels are used.

d. C Parameter: In SVM, the regularisation parameter "C" determines how much weight is given to optimising margin vs reducing classification mistakes. A bigger C results in a tighter classification, whereas a smaller C encourages a broader margin but may permit some misclassified points.

e. Multi-class Classification: SVM was originally developed for use in binary classification, but it has since been expanded to handle multi-class issues with the help of methods like one-versus-all (OvA) and one-versus-one (OvO) classification.

f. Regression: SVMs may also be utilised for regression tasks, in which the objective is to predict a continuous numerical value while minimising the margin violations.

g. Solving the Dual Problem: The dual problem of SVM optimisation concerns the constrained maximisation of a function. When using this formulation, a sparse collection of support vectors is typically the result.

3.4.3 Decision Tree

For both classification and regression, many researchers turn to Decision Trees, one of the most well-known machine learning algorithms. It's a basic model, yet it can capture intricate data linkages and decision-making constraints. The Key characteristics of Decision Trees are as follows:

a. Tree Structure: A Decision Tree is a hierarchical structure made of nodes. The top node is referred to as the "root," and the sub-nodes that it gives rise to are termed "internal nodes." Leaves, also known as terminal nodes, are the very last nodes on a branch.

b. Decision Nodes: Internal nodes in the tree represent judgements or tests on feature values. Based on a characteristic and a threshold, the data is partitioned at these nodes.

c. Leaf Nodes: Leaf nodes reflect the final predictions or outcomes for a given input. Each leaf node represents a class label in a classification challenge, and a numerical prediction in a regression task.

d. Splitting Criteria: Each node in a Decision Tree is used by an algorithm to make a split in the data based on some combination

of attributes and thresholds. Gini impurity (for classification) and mean squared error (for regression) are two examples of metrics that may be used to evaluate a split's quality.

e. Pruning: Decision Overfitting the training data is possible when trees get too complicated. Pruning is a method for streamlining a tree by deleting its less-important branches in order to boost its prediction accuracy.

f. Entropy and Information Gain: Decision Trees are commonly used in classification tasks, with entropy and information gain being used to determine the optimal split. A dataset's entropy is a measure of its impureness or disorder, and a split's information gain is a quantification of the entropy reduction accomplished by that split.

g. Gini Impurity: The Gini impurity index is another useful indicator for sorting. It is a statistical measure of how likely it is that a randomly selected element would be incorrectly labelled based on the node's class distribution. The purpose of splitting is to get the lowest possible Gini impurity.

h. Categorical and Numerical Features: Both category and numeric information may be processed using decision trees, albeit various splitting algorithms may be employed for each kind of feature.

3.4.4 Random Forest

In machine learning, Random Forest is a popular ensemble method for both classification and regression. It relies on pooling the information from several separate decision trees to boost predictive power while decreasing the likelihood of overfitting. Random Forests are well-known for their sturdiness and their capacity to process data with a large number of dimensions. The key characteristics of Random Forest are as follows:

a. Ensemble of Decision Trees: A Random Forest consists of a collection of independent decision trees. These "weak learners" are built from scratch and are referred to as "base learners."

b. Bagging (Bootstrap Aggregating): Random Forest utilises bootstrap samples to train each decision tree. The original dataset is sampled at random with replacement to produce a bootstrap sample. The training data for each tree benefits from the diversity that is introduced by this sampling method.

c. Random Feature Selection: Random Forests employ random feature selection in addition to bootstrapping the data. Usually, the square root of the total number of features is chosen at random as the collection of features to examine at each decision node in a decision tree. This feature selection adds extra variation among the trees.

d. Voting or Averaging: For classification tasks, the final prediction of the Random Forest is determined by a majority vote among the individual trees. In regression tasks, the final prediction is typically the average of the predictions from all the trees.

e. Out-of-Bag (OOB) Error: Each tree is trained on a unique subset of the data, which allows for an assessment of the model's performance to be made using the remaining (out-of-bag) data points without the need for a separate validation set. OOB error is a helpful predictor of the Random Forest's future success on unseen data.

f. Randomization for Robustness: The randomization incorporated in Random Forests helps prevent overfitting, as each tree focuses on different subsets of data and characteristics. Because of its stability, Random Forests are less affected by data noise and outliers.

g. Parallelization: Random Forest is a scalable and computationally economical approach since it allows for concurrent training of numerous decision trees.

4. Results and Discussion

In order to successfully carry out the experiment, the scikit-learn package of the Python programming language is utilised to successfully carry out the data preparation, feature extraction, and classification tasks. In this study, a 5-fold cross-validation approach is utilised to develop a proposed model on coconut milk adulteration datasets. The goal of this work is to detect adulteration in coconut milk. During the process of creating the model, four of the folds are utilised for training, while the fifth is reserved for testing. As a consequence, this process is carried out five times, and then the final step is to take the average of the findings. In this situation, since there are not enough samples in the datasets, 5-fold cross-validation is used to prevent the model from overfitting, reduce the variance while the model is being built, and generalise the model using a little quantity of data. The different statistical assessment metrics such as accuracy, precision, recall, and F1-score are taken into consideration in evaluating the findings of an experiment.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$F1 = \frac{(2*\text{Precission}*\text{Recall})}{(\text{Precision}+\text{Recall})} \tag{4}$$

4.1 *Analysis of Accuracy*

Accuracy measures how well a classifier really makes predictions. The greater the value of accuracy suggests better prediction and lower the miss-classification. Table 3 shows the accuracy results of ML classifiers using a number of distinct feature extraction methods. The accuracy of the proposed ML classifiers for the PCA technique is LR: 96.23%, SVM: 93.45%, DT: 92.67%, and RF: 97.23%. For the LDA technique, the accuracy obtained by the proposed ML classifiers is LR: 97.17%, SVM: 95.32%, DT: 94.28%, and RF: 99.17%. The accuracy of the proposed ML classifiers for the autoencoder is LR: 96.77%, SVM: 94.78%, DT: 93.84%, and RF: 98.45%.

Table 3

Accuracy of different ML classifiers on Coconut Milk Adulteration dataset

Feature Extraction	LR	SVM	DT	RF
PCA	96.23	93.45	92.67	97.23
LDA	97.17	95.32	94.28	99.17
Autoencoder	96.77	94.78	93.84	98.45

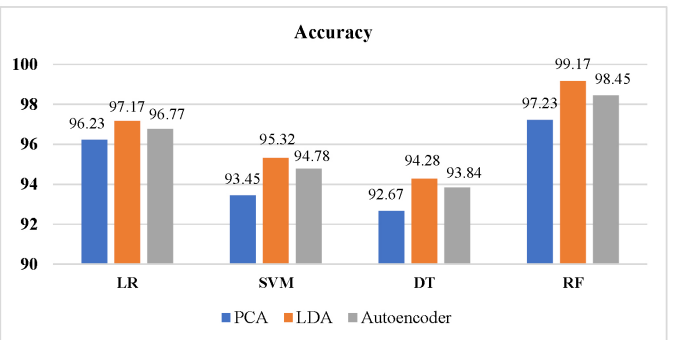

Figure 2

Analysis of Accuracy

RF provides the highest accuracy of 99.17% for the LDA technique. Figure 2 depicts the accuracy analysis of the proposed ML classifiers for feature-extracted dataset.

4.2 *Analysis of Precision*

An increase in true positive value and a decrease in false positive value correspond to an increase in precision. Table 4 shows the precision values of proposed ML classifiers for different feature extraction methods. The precision score of the proposed ML classifiers for the PCA technique is LR: 95.73%, SVM: 92.59%, DT: 91.49%, and RF: 96.71%. For the LDA technique, the precision score obtained by the proposed ML classifiers is LR: 96.85%, SVM: 94.18%, DT: 93.81%, and RF: 97.29%. The precision score of the proposed ML classifiers for the autoencoder is LR: 96.18%, SVM: 93.78%, DT: 91.96%, and RF: 97.18%. RF provides the highest precision value of 97.29% for the LDA technique. Figure 3 depicts the precision analysis of the proposed ML classifiers for feature-extracted dataset.

Table 4

Precision of different ML classifiers on Coconut Milk Adulteration dataset

Feature Extraction	LR	SVM	DT	RF
PCA	95.73	92.59	91.49	96.71
LDA	96.85	94.18	93.81	97.29
Autoencoder	96.18	93.78	91.96	97.18

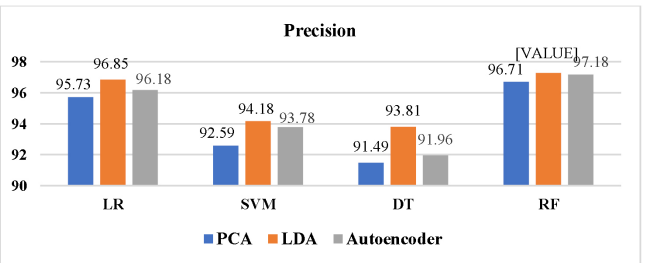

Figure 3

Analysis of precision

4.3 *Analysis of Recall*

In statistics, recall measures the proportion of correct predictions made, therefore a high recall indicates few false negatives. When the true

positive is high and the false negative is low it suggests better prediction. Table 5 shows the recall values of proposed ML classifiers for different feature extraction methods. The recall score of the proposed ML classifiers for the PCA technique is LR: 94.92%, SVM: 91.39%, DT: 90.92%, and RF: 95.83%. For the LDA technique, the recall score obtained by the proposed ML classifiers is LR: 95.74%, SVM: 93.74%, DT: 92.59%, and RF: 96.71%. The recall score of the proposed ML classifiers for the autoencoder is LR: 95.88%, SVM: 92.69%, DT: 91.17%, and RF: 96.29%. RF provides the highest recall value of 96.71% for the LDA technique. Figure 4 depicts the recall analysis of the proposed ML classifiers for feature-extracted dataset.

Table 5

Recall of different ML classifiers on Coconut Milk Adulteration dataset

Feature Extraction	LR	SVM	DT	RF
PCA	94.92	91.39	90.92	95.83
LDA	95.74	93.74	92.59	96.71
Autoencoder	95.88	92.69	91.17	96.29

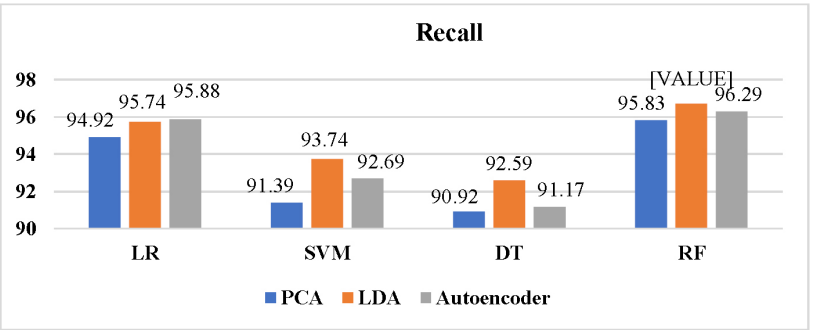

Figure 4

Analysis of Recall

4.4 *Analysis of F1 Score*

A higher F1-score suggests more accurate predictions since it is the harmonic mean of the recall and accuracy values. Table 6 shows the F1 Score of proposed ML classifiers for different feature extraction methods. The F1 Score of the proposed ML classifiers for the PCA technique is LR: 95.32%, SVM: 91.98%, DT: 91.20%, and RF: 96.27%. For the LDA technique, the F1 Score obtained by the proposed ML classifiers is LR: 96.29%, SVM:

93.96%, DT: 93.20%, and RF: 97%. The F1 Score of the proposed ML classifiers for the autoencoder technique is LR: 96.03%, SVM: 93.23%, DT: 91.56%, and RF: 96.88%. RF provides the highest F1 Score of 97% for the LDA technique. Figure 5 depicts the F1 score analysis of the proposed ML classifiers for feature-extracted dataset.

Table 6

F1 Score of different ML classifiers on Coconut Milk Adulteration dataset

Feature Extraction	LR	SVM	DT	RF
PCA	95.32	91.98	91.20	96.27
LDA	96.29	93.96	93.20	97
Autoencoder	96.03	93.23	91.56	96.88

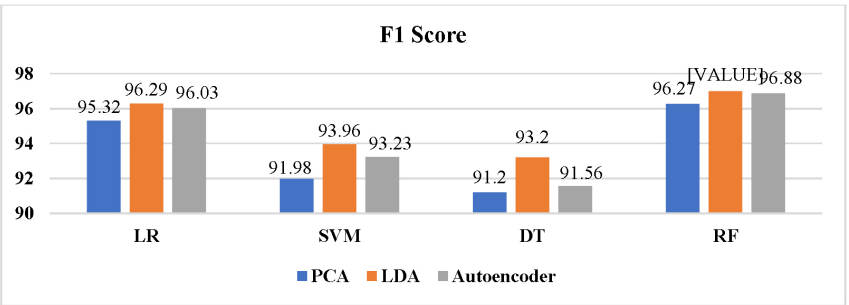

Figure 5

Analysis of F1 Score

5. Conclusion

In this work, we propose a machine-learning framework for coconut milk adulteration detection. We show that predictive models based on ML techniques are useful tools for this task. After completing the initial data processing, the three-feature extraction (PCA, LDA, and AE) techniques are applied to extract features from the dataset. The feature-extracted dataset is classified through four effective machine learning (LR, SVM, DT, and RF) algorithms. The proposed machine learning model's performance is analysed, and the best-performing feature extraction and classification approaches are identified. The experimental findings of the proposed work are justified through different statistical evaluation measures (accuracy, precision, recall, and F1 score). Tables 3–6 illustrate a

performance analysis of the proposed model. The Random Forest algorithm on the LDA feature-extracted dataset outperformed with the highest accuracy of 99.17%, precision of 97.29%, recall of 96.71%, and F1 score of 97%. Our work demonstrates the efficiency of machine learning models in detecting food adulteration. However, a limitation of our research is the insufficient training data for the proposed model. In the future, we intend to collect more data related to food adulteration and construct a robust model for detecting food adulteration.

References

[1] Saha, D., & Manickavasagan, A., Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. *Current Research in Food Science*, 4, 28-44 (2021).

[2] Othman, S., Mavani, N. R., Hussain, M. A., Abd Rahman, N., & Ali, J. M., Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review. *Journal of Agriculture and Food Research*, 100590 (2023).

[3] Rachineni, K., Kakita, V. M. R., Awasthi, N. P., Shirke, V. S., Hosur, R. V., & Shukla, S. C., Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification. *Current Research in Food Science*, 5, 272-277 (2022).

[4] Asefa, B. G., Hagos, L., Kore, T., & Emire, S. A., Computer vision based detection and quantification of extraneous water in raw milk (2021).

[5] Q. Li et al., Low risk of category misdiagnosis of rice syrup adulteration in three botanical origin honey by ATR-FTIR and general model, vol. 332. Elsevier Ltd. (2020). doi: 10.1016/j.foodchem.2020.127356.

[6] Noviyanto and W. H. Abdulla, "Honey Dataset Standard Using Hyperspectral Imaging for Machine Learning Problems," pp. 503–507 (2017).

[7] M. Hussain Khan, Z. Saleem, M. Ahmad, A. Sohaib, H. Ayaz, and M. Mazzara, "Hyperspectral imaging for color adulteration detection in red chili," *Appl. Sci.*, vol. 10, no. 17 (2020), doi: 10.3390/app10175955.

[8] S. P. S. Brighty, G. S. Harini, and N. Vishal, "Detection of Adulteration in Fruits Using Machine Learning," pp. 2021–2024 (2021).

[9] M. Pérez-Calabuig, S. Pradana-López, A. Ramayo-Muñoz, J. C. Cancilla, and J. S. Torrecilla, "Deep quantification of a refined adulterant blended into pure avocado oil," *Food Chem.*, vol. 404, (no. September 2022, 2023), doi: 10.1016/j.foodchem.2022.134474.

[10] P. kumar Goyal, Kashish, "Advances in Machine Learning and Hyperspectral Imaging in the Food Supply Chain," *Food Eng. Rev.*, pp. 596–616 (2022), doi: 10.1007/s12393-022-09322-2.

[11] Sitorus, A., Muslih, M., Cebro, I. S., & Bulan, R., Dataset of adulteration with water in coconut milk using FTIR spectroscopy. *Data in Brief*, 36, 107058 (2021).