RESEARCH ARTICLE

Enhancing Honey Adulteration Detection With Optimal Subspace Wavelength Reduction in Vis-NIR Reflection Spectroscopy

MOKHTAR AL-AWADHI[1,2] **AND RATNADEEP DESHMUKH**[2]

[1]Department of Information Technology, Faculty of Engineering and IT, Taiz University, Taiz, Yemen
[2]Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad 431004, India

Corresponding author: Mokhtar Al-Awadhi (mokhtar.awadhi@taiz.edu.ye)

ABSTRACT The accurate detection of honey adulteration is paramount for maintaining the quality and authenticity of honey products. In this study, we introduce a novel feature selection method, termed Optimal Subspace Wavelength Reduction (OSWR), and integrate it with reflectance Visible-Near Infrared (Vis-NIR) spectroscopy to enhance the discrimination between pure and adulterated honey and predict adulteration levels. OSWR efficiently addresses the dimensionality challenge of large spectral datasets, reducing 2151 wavelengths to a compact and informative set of 39 wavelengths. We comprehensively evaluate machine learning (ML) models, focusing on OSWR as a pivotal component of our methodology. Our results reveal remarkable success in discriminating among pure honey, adulterated honey, and sugar syrup, with an impressive classification accuracy of 96.67% achieved using OSWR, coupled with Standard Normal Variate (SNV) preprocessing, Linear Discriminant Analysis (LDA) feature extraction, and K-Nearest Neighbors (KNN) classification. Furthermore, this study demonstrates the effectiveness of OSWR for predicting adulteration levels, where it achieves an accuracy of as high as 92.67% when coupled with SNV, LDA, and KNN. This work highlights the potential of OSWR as a feature selection method in the context of honey adulteration detection. Through the integration of Vis-NIR spectroscopy and OSWR, our approach offers a tool for enhancing honey products' quality and authenticity assessment, potentially simplifying spectral data analysis.

INDEX TERMS Chemometric analysis, feature selection, honey adulteration, machine learning, optimal subspace wavelength reduction, Vis-NIR spectroscopy.

I. INTRODUCTION

Honey, renowned for its nutritional and therapeutic attributes, has garnered global attention as a natural sweetening agent with remarkable health benefits [1], [2]. Its economic significance is evident, with an annual production exceeding 1.2 million tons [3]. The nutritional and economic value of honey arises from its intricate composition, encompassing saccharides, aqueous content, proteins, organic acids, vitamins, minerals, chromatic constituents, phenolic and volatile compounds, alongside specific particulate matter [4], [5].

Regulations imposed by the European Union emphasize the prohibition of adding or removing constituents from honey, rendering adulteration unlawful [6]. Adulteration involves incorporating substandard honey and synthetic artificial sugars [7]. Due to its distinct health benefits, sensory attributes, and aromatic complexities, honey is more expensive than alternative sweeteners. Consequently, this has led to the susceptibility of honey to adulteration practices aimed at reducing production costs and increasing profits [8], [9], [10].

Methods of honey adulteration encompass feeding bees starch and inverted syrup, adding sugars like high fructose, glucose, and sucrose syrups, and blending inferior honey with premium varieties [9], [11]. Adulteration is a widespread concern across various stages of honey production and

The associate editor coordinating the review of this manuscript and approving it for publication was Norbert Herencsar.

processing. The challenge is exacerbated by the resemblance between fake and pure honey, hampering detection [12]. Conventional quality assessment methods, such as physicochemical analyses, exhibit limitations in accurately identifying adulteration [13], [14], [15], [16]. Consequently, the need for innovative methodologies in honey quality control is evident.

In response, diverse analytical techniques, including chromatography [17], [18] and biosensors [19], [20], [21]. have been proposed for detecting honey adulteration. However, these methods are expensive, time-consuming, and potentially destructive [22]. Leveraging Visible-Near Infrared (Vis-NIR) spectroscopy alongside machine learning (ML) algorithms has demonstrated rapid detection of adulteration in honey from single botanical sources [7], [23], [24], [25], [26], and two botanical sources [27]. However, the application of this technology to detect adulteration across several honey types remains unexplored.

Numerous studies have explored the utility of ML algorithms in detecting honey adulteration. However, the majority of these investigations have predominantly relied on absorbance/transmittance Vis-NIR spectroscopy [7], [23], [24], [25], [26], [27]. A comparatively limited body of work has delved into reflectance Vis-NIR spectroscopy, with only a few notable studies contributing to this branch of research [28], [29]. Consequently, the present study seeks to fill this research gap by focusing on the applicability of reflectance Vis-NIR spectroscopy for detecting adulterated honey and predicting adulteration levels.

Employing a subset of discriminative wavelengths could streamline computation and facilitate the design of cost-effective spectrometers. However, few studies [7], [25] employed feature selection techniques to identify pivotal wavelengths for honey adulteration detection. This paper introduces a novel feature selection approach to identify crucial wavelengths for effective honey adulteration detection and adulteration level prediction. Additionally, ML models are developed for multi-type honey adulteration detection using the selected wavelengths. The key contributions of this study include:

1) Creation of a Vis-NIR reflectance spectral dataset for genuine and adulterated honey across various types. Adulterated samples were prepared by mixing pure honey with different concentrations of three artificial sugar syrups.
2) Introduction of a novel feature selection (FS) method, termed Optimal Subspace Wavelength Reduction (OSWR), to identify critical wavelengths in the Vis-NIR spectral region. The proposed method enhanced the accuracy and speed of ML models for honey adulteration detection and adulteration level prediction. Furthermore, the proposed FS approach enables the development of cost-effective spectrometers for capturing honey spectra.
3) Development of ML models for detecting adulteration with various sugar syrups and predicting adulteration levels across multiple honey types. Model performance was assessed using Leave-One-Out Cross-Validation (LOOCV).

The subsequent sections of this paper are organized as follows: Section II outlines the materials and methods used in this study. Experimental results and findings are presented in Section III. A discussion of the results is given in Section IV. Conclusions and recommendations for future work are provided in Section V.

II. MATERIALS AND METHODS

This section investigates the viability of integrating Vis-NIR spectroscopy with ML techniques for detecting adulteration involving diverse artificial sugars across multiple honey types. Additionally, a pioneering feature selection method is introduced to facilitate the creation of an economical Vis-NIR spectrometer for recording honey spectra. This method aims to identify pivotal spectral wavelengths within the Vis-NIR spectral range, crucial for both honey adulteration detection and determination of adulteration levels. Fig. 1 provides an illustrative block diagram showcasing the methodologies employed in this study.

A. HONEY AND SUGAR SAMPLES

Our study involved a carefully composed sample set totaling 150 samples, designed to encompass a broad range of honey adulteration scenarios. This meticulously structured dataset covers various honey varieties, including nine distinct monofloral botanical sources (Apple, Eucalyptus, Sulai, Ajwain, Litchi, Acacia, Sidr, Jamun, Sulai) and one multifloral honey type. These botanical variations comprehensively represent the floral diversity found in natural honey.

To simulate diverse adulteration scenarios, the sample set includes three types of sugar syrups: glucose syrup, invert syrup, and sucrose syrup. The samples span five adulteration levels, ranging from 0% (pure, unadulterated honey) to 100% (fully adulterated with sugar syrup).

The 150 samples (10 botanical sources × 3 sugar syrups × 5 adulteration levels = 150) were thoughtfully designed to ensure a comprehensive exploration of honey authenticity and quality assessment. They include 30 pure honey samples, 90 adulterated honey samples, and 30 sugar syrup samples. For internal consistency, each of the ten honey botanical sources has three replicates from the same source in the pure honey category. Three sets of 30 adulterated honey samples (10%, 25%, 50%) were prepared. Each sample in the adulterated honey class reflects a unique combination of a specific honey botanical source and one of the three sugar syrups. The sugar syrup group includes ten replicates for each type (glucose syrup, invert syrup, sucrose syrup), allowing for a detailed examination of the distinct spectral characteristics of each sugar syrup type. This carefully designed sample set serves as a fundamental element of our study, offering a realistic depiction of honey diversity and enabling a thorough analysis of authenticity and quality assessment across various honey types and adulteration levels.

B. Vis-NIR SPECTRA ACQUISITION AND DATASET DESCRIPTION

The acquisition of reflectance spectra, a pivotal element of this study, was facilitated by a FieldSpec 4 spectroradiometer,
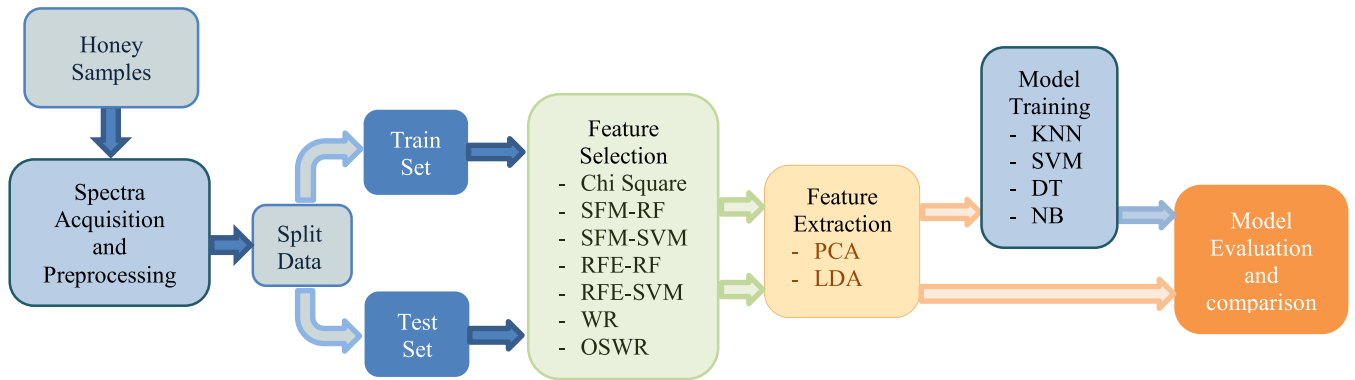
FIGURE 1. Block diagram of the honey adulteration detection method proposed in this study.

an instrument crafted by Analytical Spectral Devices (ASD), Inc., USA (now known as Malvern Panalytical Ltd, Malvern, UK) [30]. The spectroradiometer harnesses graded-index InGaAs photodiode Short-Wave Infrared (SWIR) detectors to capture spectral signatures. Encompassing a broad spectral gamut from 350 nm to 2500 nm, it effectively envelops the Vis-NIR (350 nm to 1000 nm) and the SWIR (1001 nm to 2500 nm) spectral regions. The spectrometer offers resolutions of 3 nm and 8 nm in the Vis-NIR and SWIR domains, respectively. Illumination for spectroscopic data collection is provided by a 75-watt quartz-tungsten-halogen lamp, with the RS3 ASD software managing the acquisition process to generate Vis-NIR spectra for the honey samples.

The acquisition of the honey samples' reflectance spectra occurred in a dark room designed to minimize external interference, as depicted in Fig. 2, emphasizing the importance of data integrity. Each honey sample underwent 40 individual scans to reduce the impact of random fluctuations, and these scans were averaged into a single spectral instance representing the sample's reflective characteristics. To facilitate data manipulation and interpretation, the spectrometer's binary ASD files were transformed into a more accessible Comma-Separated Values (CSV) file format using an asdreader Python script [31].

Diversity in the dataset was achieved by adding various labels to the dataset. These labels, such as sample number, botanical source, sugar adulterant, adulteration level, and type, allowed for a comprehensive examination of the spectra. The sample number served as a distinct identifier, ranging from 1 to 150. The botanical source attribute categorized each sample according to its unique floral context among ten different sources. The sugar adulterant attribute indicated the three different sugar syrups used for adulteration. The adulteration level attribute classified samples into five distinct levels, enabling a detailed analysis of the extent of adulteration. Finally, the type attribute categorized each sample as pure honey, adulterated honey, or a sugar syrup specimen. These thoughtfully assigned labels enhanced the dataset's versatility, enabling a comprehensive exploration of honey samples from various angles.

C. PREPROCESSING WITH STANDARD NORMAL VARIATE
The Standard Normal Variate (SNV) normalization technique was developed in response to the challenges posed by spectral perturbations. SNV normalization is designed to mitigate the multiplicative effects of scattering-induced shifts and address variations in global signal intensities, thereby enhancing spectral consistency [32]. The SNV procedure involves a mathematical transformation applied to each data point in the spectrum. Specifically, it subtracts each data point from the spectrum's mean to isolate it from local influences. Subsequently, the result is divided by the spectrum's standard deviation, ensuring normalization and alignment with the broader spectral context. This process is briefly represented by Equation (1).

$$x_{m,n}^{\text{SNV}} = \frac{\left(x_{m,n} - \bar{x}_m\right)}{\sqrt{\frac{\Sigma_{n=1}^{N}\left(x_{m,n} - \bar{x}_m\right)^2}{N-1}}} \quad (1)$$

where $x_{m,n}^{\text{SNV}}$ is the element of the transformed spectrum and $x_{m,n}$ is the corresponding original element of the spectrum m at variable n, $\bar{x}_m$ is the mean of spectrum m, and N is the number of wavelengths in the spectrum.

D. DIMENSIONALITY REDUCTION
1) FEATURE SELECTION
Feature selection, a crucial technique for dimensionality reduction, involves ranking features based on their significance [33]. This approach confers several notable advantages, encompassing reductions in both classification time and model complexity, improvement in classifier precision, and mitigation of overfitting concerns [34]. Within the domain of spectral data analysis, feature selection assumes an additional important role by enabling the identification of salient wavelengths. This capability further supports the design of cost-effective spectrometers tailored for the precise capture of spectra at specific wavelengths.

In this study, we propose an innovative feature selection method calibrated to identify the most discriminating wavelengths germane to detecting honey adulteration. To ascertain

FIGURE 2. The experimental setup used in this study to acquire the reflectance Vis-NIR spectra of the honey and sugar samples. (1) Honey sample, (2) Sensor lens, (3) Optic fiber probe, (4) ASD Spectroradiometer, (5) Laptop.

the efficacy of our proposition, we undertake a comparative analysis involving four commonly employed feature selection methodologies. These encompass the Chi-square method, the SelectFromModel (SFM) approach, the Recursive Feature Elimination (RFE) procedure, and the Wavelength Reduction (WR) technique.

The Chi-square methodology is firmly rooted in statistical testing, leveraging the statistical significance of the relationship between attributes and the target variable to discern prominent features. Chi-square values are computed to encapsulate the degree of correlation between attributes and target variables, and those features achieving the highest scores are chosen for selection [34]. This approach treats each feature as an independent entity, effectively isolating them from the broader contextual domain. Importantly, it operates as a filter-based procedure dissociated from any particular classifier.

In contrast, the Select-From-Model (SFM) technique is distinguished by its model-centric approach to feature selection [35]. In this methodology, features are chosen based on their significance as assessed by the classifier's feature attribute. The underlying principle involves identifying features that do not meet a pre-defined relevance threshold, deeming them uninformative, and removing them from the feature set. This framework uses two well-established ML models: Support Vector Machines (SVM) and Random Forest (RF). During this iterative process, selecting an optimal feature subset is guided by assessing the classification accuracy obtained through a 10-fold cross-validation procedure.

Parallel to these methodologies, the Recursive Feature Elimination (RFE) mechanism implements a sequential feature selection protocol. It commences by initially ranking features based on their internal coefficients or significance within a model. Subsequently, the RFE iteratively excludes
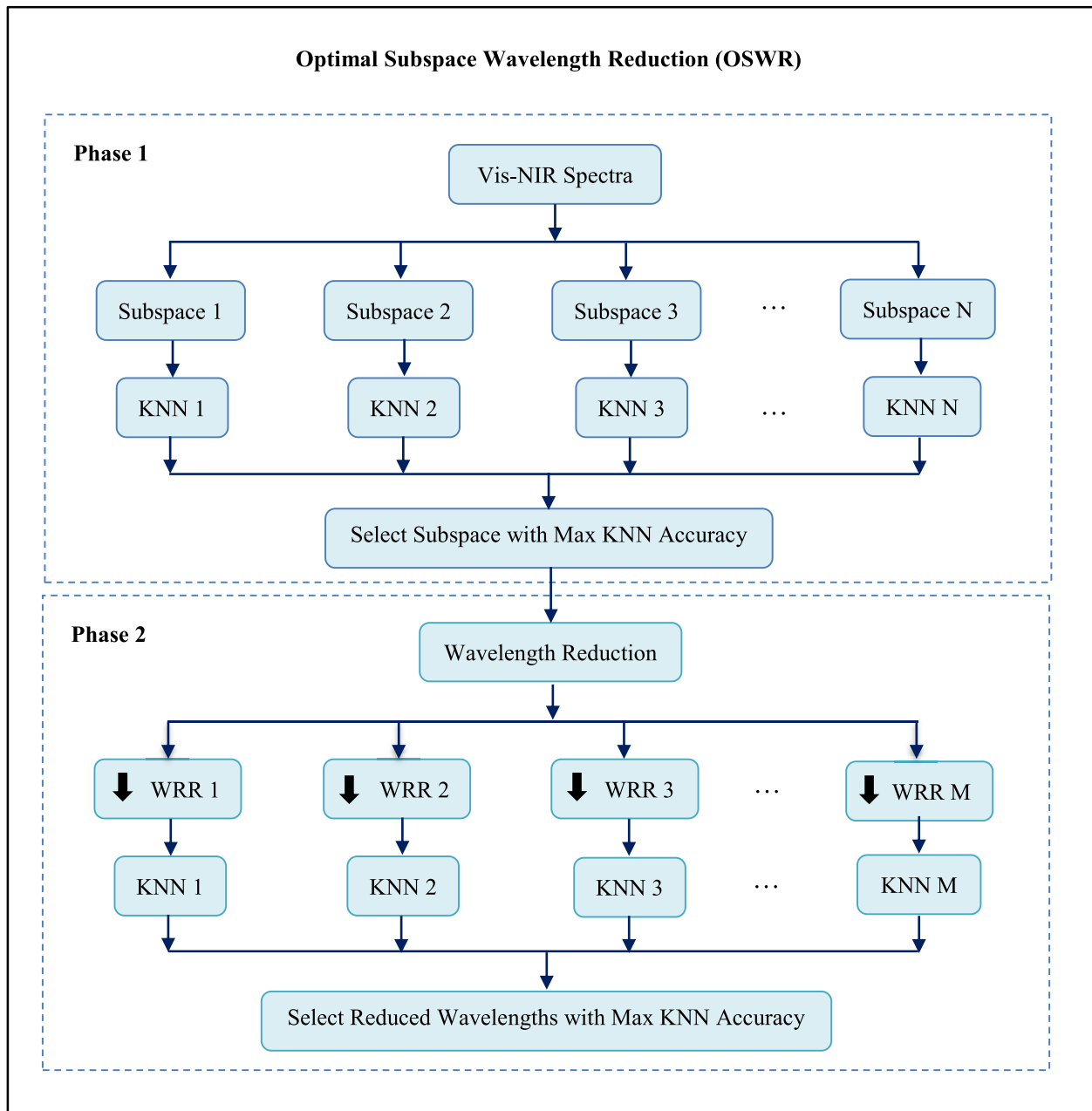
less influential attributes. This iterative process alleviates model collinearity issues by systematically eliminating specific characteristics [36]. Like the SFM technique, RFE interfaces with two external estimators, RF and SVM. The determination of the optimal feature ensemble is guided by evaluating classification accuracy through a 10-fold cross-validation methodology.

The Wavelength Reduction (WR) technique provides a straightforward yet efficacious approach to reducing data dimensionality [37], [38]. This method coordinates a systematic reduction in wavelengths, guided by a predetermined wavelength reduction rate (WRR). The accuracy of ML models is assessed using progressively truncated sets of wavelengths, ultimately leading to the selection of the wavelengths that result in the highest predictive accuracy. While WR is easily implementable, it is important to note that this technique does not provide a quantifiable scoring mechanism for the selected features.

The feature selection method introduced in this study, Optimal Subspace Wavelength Reduction (OSWR), is designed to enhance the effectiveness of ML models in detecting honey adulteration while dealing with a wide spectral range of data, specifically, the 2151 spectral bands spanning from 350 to 2500 nm. Notably, not all these spectral wavelengths are equally informative or discriminative for detecting honey adulteration, which prompted the development of OSWR. The OSWR method, visualized in Fig. 3, is a two-phase process:

a: OPTIMAL SUBSPACE DETERMINATION (PHASE 1)
This phase aims to identify the most effective spectral band interval within the Vis-NIR region (350-2500 nm). This interval contains the wavelengths that offer the highest utility

Optimal Subspace Wavelength Reduction (OSWR)

Phase 1

Vis-NIR Spectra

Subspace 1 | Subspace 2 | Subspace 3 | ··· | Subspace N

KNN 1 | KNN 2 | KNN 3 | ··· | KNN N

Select Subspace with Max KNN Accuracy

Phase 2

Wavelength Reduction

⬇ WRR 1 | ⬇ WRR 2 | ⬇ WRR 3 | ··· | ⬇ WRR M

KNN 1 | KNN 2 | KNN 3 | ··· | KNN M

Select Reduced Wavelengths with Max KNN Accuracy

FIGURE 3. Optimal subspace wavelength reduction feature selection method proposed in this study.

for maximizing an ML model (e.g. KNN) performance in honey adulteration detection. This process unfolds in several steps:

1. The entire spectral range from 350-2500 nm is randomly divided into multiple subspaces with different starting wavelengths and lengths.
2. The performance of the KNN model in detecting honey adulteration is evaluated within these subspaces.
3. The subspace that yields the best KNN performance, thus being the optimal subspace, is selected.

b: WAVELENGTH REDUCTION (PHASE 2)

This phase addresses the high correlation often observed among adjacent Vis-NIR wavelengths. The main objective is

eliminating these correlated wavelengths while retaining the uncorrelated, discriminative ones. The wavelength reduction technique applied in this phase is straightforward and consists of the following steps:

1. Multiple Wavelength Reduction Rates (WRRs) are defined.
2. For each WRR, the following steps are performed:
   a. A reduction of the number of wavelengths in the optimal subspace, as determined in Phase 1, is carried out based on the specific WRR.
   b. The performance of the KNN model in detecting honey adulteration is assessed using the reduced set.
3. The reduced wavelength set that attains the highest KNN performance is selected as the optimal choice.

In summary, OSWR is a structured method that optimizes the spectral wavelength selection process, resulting in a reduced yet highly informative subset of wavelengths for accurately detecting honey adulteration by ML models. This approach ensures that the analysis is not burdened with redundant or irrelevant spectral data, ultimately improving the efficiency and accuracy of the detection process.

2) FEATURE EXTRACTION

Feature extraction is fundamental in dimensionality reduction, transforming an initial raw dataset into more manageable groups for subsequent analysis [39]. Its significance becomes particularly pronounced when dealing with extensive datasets characterized by many variables, which often necessitate substantial computational resources [40]. The core purpose of feature extraction lies in its capacity to alleviate the computational burden by reducing the data volume while retaining essential and pertinent information [40]. Additionally, feature extraction can help minimize the amount of duplicated data, which is especially valuable for specific research inquiries [40]. Furthermore, data reduction accelerates ML algorithms' learning and generalization phases [41].

Within the scope of this study, the application of feature extraction takes on a particular significance. Two distinct methodologies, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), were employed to elucidate common patterns within the dataset. These methods served as tools to discern the inherent attributes that effectively discriminate between different classes [42]. PCA is a transformative technique that restructures data into a space characterized by orthogonal axes or principal components, capturing the most significant variance present in the dataset. This reconfiguration enables the identification of salient features that encapsulate the predominant patterns within the data.

Additionally, LDA complements the analytical process by introducing a classification-centric perspective. LDA seeks to identify discriminative characteristics that enhance class differentiation by maximizing inter-class variance while minimizing intra-class variance [43]. It is essential to highlight the distinction that while PCA primarily focuses on retaining variance, LDA is specifically designed to facilitate the creation of features optimized for discrimination between classes.

Utilizing PCA and LDA feature extraction methodologies, this study positions itself within a streamlined analytical framework characterized by a judicious and well-informed reduction in the dimensionality of the dataset. This approach offers several advantages, including enhanced computational efficiency and an improved ability to disentangle and identify information crucial for detecting honey adulteration and predicting adulteration levels.

E. CLASSIFICATION

This phase of the study encompasses a rigorous assessment of the efficacy of four widely utilized ML classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB). The chosen classifiers span divergent classification categories and have exhibited robust performance in prior investigations [24], [44]. The implementation of these algorithms was carried out through the utilization of the Scikit-Learn Python library [45].

KNN constitutes an instance-based non-parametric learning paradigm, engendering classification based on the attributes of its K nearest neighbors [46]. The selection of K markedly influences the performance of the KNN classifier, as smaller K values can engender overfitting while larger K values render the model more susceptible to noise. Determining the optimal K value involved the experimentation with a range of 20 values (ranging from 1 to 20). A precedent study has effectively harnessed KNN for identifying honey adulteration via Hyperspectral Imaging data [44].

The application of SVM, a supervised ML method, assumes delineating decision boundaries or hyperplanes in high-dimensional space to partition data into distinct classes before classification [47]. SVM employs pivotal data points or vectors to delineate these decision bounds, transforming nonlinear decision boundaries into linear equivalents by incorporating kernel functions. Within this investigation, linear and Radial Basis Function (RBF) kernels were employed, with the latter being substantiated in prior research as outperforming other kernels [48]. The RBF kernel encompasses two pivotal tuning parameters, namely cost, and gamma, with their manipulation potentially affecting classification performance [49]. The exploration of cost and gamma encompassed values of 0.001, 0.01, 0.1, 1, 10, 100, and 1000. SVM has demonstrated commendable proficiency in identifying adulteration in honey through non-imaging Vis-NIR data [24] and imaging Vis-NIR data [44].

DT, a classification approach, organizes instances predicated on arranging their feature values [46]. Each node within a decision tree represents a feature characterizing a classifiable instance, while branches indicate potential values for the node. The root node initiates the categorization process by espousing the feature that optimally divides the training data. A diverse array of techniques, encompassing information gain and the Gini index, may be employed to discern the most productive feature for data division. Their simplicity of interpretation and implementation characterizes DTs.

The NB approach entails the deployment of basic Bayesian networks constructed from directed acyclic graphs with one parent (unobserved node) and numerous children (observed nodes) [46]. Central to NB classifiers is the strong assumption of independence among child nodes concerning their parent. Notably, the computational training time of the NB classifier is expedited by its model structure, which may be translated from a product form into a sum through logarithmic transformation, engendering computational efficiency.

It is important to highlight that all computational processes encompassing preprocessing, feature selection, feature extraction, and classification were conducted through the Anaconda data science platform, employing Python version 3.8.8.

F. PERFORMANCE EVALUATION

The comprehensive evaluation of the ML classifiers' efficacy in detecting honey adulteration and predicting the extent of adulteration was accomplished by using the LOOCV technique. This assessment strategy entails partitioning the dataset into individual samples, iteratively employing one sample for testing and the remaining samples for training. The overall classification accuracy, indicative of the proportion of accurately classified instances, was adopted as the primary performance metric. Notably, the LOOCV method mirrors traditional cross-validation with the number of folds equaling the total sample count, ensuring an accurate and robust testing procedure.

Furthermore, confusion matrices were employed to scrutinize the ML models' performance in discerning each level of adulteration. These matrices concisely depict the model's classification outcomes, facilitating a nuanced analysis of true positive, true negative, false positive, and false negative instances. This evaluation endeavor affords insights into the discriminative capacity of the models across varying degrees of adulteration.

In addition to evaluating the above methodologies, the paired t-test statistic was harnessed to gauge the extent of significant differences among the classification outcomes of the four classifiers [50]. This statistical technique allows for a robust determination of whether the discrepancies in classification results among the classifiers are statistically significant, shedding light on the divergence in their performance.

The assessment of the ML models was conducted across multiple datasets, encompassing raw data, preprocessed data, features reduced through feature selection methods, features reduced via feature extraction methods, and features reduced through feature selection followed by feature extraction. This comprehensive analysis provides a holistic understanding of the impact of different data manipulation strategies on the models' classification performance.

III. RESULTS

This section presents the results of the experiments on discriminating between pure and fake honey and predicting adulteration levels.

A. Vis-NIR REFLECTANCE SPECTRA

The Vis-NIR reflectance spectra, obtained using a spectroradiometer, serve as a foundational basis for deciphering the complex spectral patterns linked to a range of adulteration levels. Fig. 4 (a) presents the averaged reflectance spectra for honey samples, categorized according to varying degrees of impurity, covering a broad wavelength spectrum from 350 nm to 2500 nm. Notably, spectral variations become apparent within the wavelength range below 1500 nm.

Fig. 4 (b) provides a detailed representation of spectral complexities within the Vis-NIR range (350 nm to 1000 nm). Notably, clear and distinct spectral patterns emerge for sugar samples, making them easily distinguishable from honey samples. A discernible trend becomes apparent, wherein increasing degrees of adulteration correspond to higher reflectance levels, while unadulterated honey exhibits the lowest reflectance. Remarkably, the wavelength around 850 nm stands out as the peak reflectance point for the honey samples.

Exploration of the Short-Wave Infrared (SWIR) spectral domain, ranging from 1001 nm to 1800 nm, is presented in Fig. 4 (c), revealing two prominent peaks at 1100 nm and 1300 nm. This spectral range is a critical discriminator, highlighting the divergent reflective behaviors of sugar and honey samples, with sugar demonstrating lower reflectance. Reflection spectra show little reflection in the 1400-1800 nm area because honey's molecular components absorb light strongly. Absorption bands linked with overtone and combination vibrations of chemical bonds make it sensitive to honey's chemical makeup in this wavelength range. Water, a significant component of honey, has strong near-infrared absorption bands, which may explain the low reflectivity. Natural sugars like glucose and fructose in honey have NIR absorption properties. Therefore, their quantities and compositions might affect reflectance in this range. Honey contains proteins, amino acids, and minerals, which may affect its spectral behavior. To fully interpret these spectrum features and comprehend the molecular qualities that cause this phenomenon, in-depth spectroscopic investigations and chemical component identification are needed, frequently using chemometric approaches and reference standards.

Further examination within the SWIR 2 spectral region, from 1801 nm to 2500 nm and illustrated in Fig. 4 (d), emphasizes a significant peak centered at 1850 nm. Sugar samples consistently exhibit lower reflectivity than honey samples across this specific spectral range.

In Fig. 4 (e), we present the mean spectra and standard deviations of pure honey and sugar syrup samples. These spectra offer insight into the inherent variations present within these sample categories. Notably, the variations in the spectra of pure honey are indicative of the diverse honey botanical sources, highlighting the influence of these sources on the spectral signatures. Similarly, the spectra of sugar syrup exhibit variations, underscoring the distinct spectral profiles associated with different sugar types used in the adulteration process.

Despite the variations observed within each class, whether in pure honey or sugar syrup, the discernible disparities between the spectral characteristics of pure honey and sugar syrup remain evident. These distinctions validate the efficacy of the Vis-NIR spectroscopy method in effectively distinguishing between these two fundamental categories.

A recurring pattern becomes evident, where increased reflectance profiles are inversely correlated with diminishing levels of adulteration, except in the case of pure honey samples. This reflectivity pattern underscores the potential importance of spectral signatures as distinguishing markers in detecting honey adulteration.

B. DISCRIMINATION BETWEEN PURE AND ADULTERATED HONEY

1) OVERALL ML MODEL PERFORMANCE

This section assesses the effectiveness of various ML models in distinguishing between unadulterated honey, adulterated
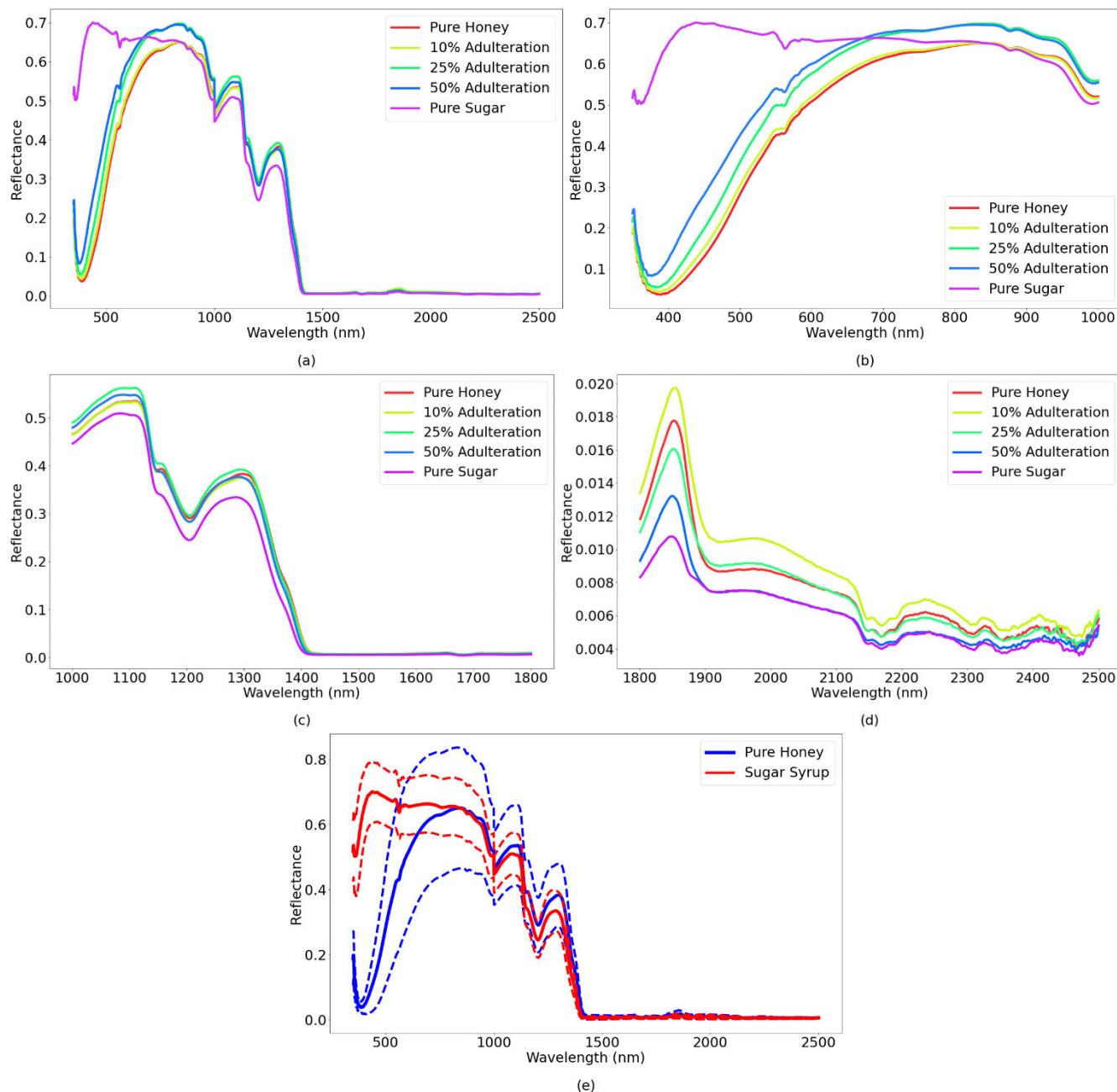
FIGURE 4. The reflectance spectral data of the honey samples averaged according to the adulteration level: (a) Full spectra, (b) Spectra in the Vis-NIR region (350nm-1000nm), (c) Spectra in the SWIR 1 region (1001nm-1800nm), (d) Spectra in the SWIR 2 region ((1801nm-2500nm), and (e) Mean spectra ± standard deviation of pure honey and sugar syrup.

honey, and sugar syrup samples. The classification performances of these models are detailed in Table 1, which presents their accuracy in detecting honey adulteration across various scenarios, including using the full wavelength spectrum and different feature selection and extraction techniques without data preprocessing. Notably, the SVM model stands out with the highest accuracy of 77.33% when utilizing the complete wavelength spectrum and 80% when utilizing a reduced feature set obtained through PCA. In contrast, the DT

model achieves its highest accuracy of 86% through feature reduction facilitated by LDA.

The eighth row of Table 1 holds particular significance, highlighting the effectiveness of the proposed feature selection method when coupled with feature extraction based on LDA. This hybrid approach produces the most promising results, with the SVM achieving the highest classification accuracy of 89.33%. Furthermore, an accuracy rate of 85.33% is observed when combining WR for feature selection, LDA

TABLE 1. Classification accuracy of the ML models for discriminating between genuine and fake honey using full wavelengths, feature selection methods (39 features), and feature extraction methods.

PP [a]	Feature Selection Method	Feature Extraction Method											
		-				PCA				LDA			
		KNN	SVM	DT	NB	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Without Preprocessing	-	70.67	77.33	71.33	61.33	69.33	80.00	69.33	77.33	82.67	84.67	86.00	81.33
	Chi-Square	69.33	80.00	76.00	63.33	69.33	80.00	72.00	72.67	70.00	78.67	66.67	72.00
	SFM-RF	67.33	80.00	71.33	64.67	70.00	80.00	74.00	67.33	73.33	76.00	71.33	74.67
	SFM-SVM	68.00	80.00	74.00	65.33	69.33	80.00	71.33	67.33	68.00	80.00	74.00	71.33
	RFE-RF	71.33	80.00	70.00	64.00	67.33	80.00	74.67	77.33	84.00	84.00	83.33	85.33
	RFE-SVM	71.33	80.00	72.67	67.33	72.67	80.00	68.67	70.67	78.00	82.67	78.67	78.67
	WR	70.00	80.00	68.00	62.67	68.67	80.00	70.00	78.00	82.67	85.33	82.00	85.33
	OSWR	46.67	60.00	47.33	45.33	46.67	60.00	41.33	58.00	86.67	**89.33**	84.00	85.33
SNV	-	71.33	93.33	76.00	66.00	74.00	80.00	76.00	71.33	83.33	83.33	82.67	86.67
	Chi-Square	70.67	80.00	72.00	68.00	71.33	80.00	72.67	71.33	79.33	80.00	74.00	73.33
	SFM-RF	70.67	80.00	70.67	68.67	67.33	80.00	74.00	67.33	76.00	79.33	78.67	81.33
	SFM-SVM	66.67	80.00	72.67	68.67	66.00	80.00	72.00	67.33	73.33	80.00	76.67	74.00
	RFE-RF	73.33	79.33	72.00	69.33	82.00	80.00	78.00	65.33	82.00	81.33	78.67	81.33
	RFE-SVM	80.67	76.00	76.00	68.67	75.33	80.00	74.67	70.00	82.00	87.33	76.67	85.33
	WR	72.67	79.33	76.00	66.67	68.00	80.00	74.67	72.00	88.00	87.33	80.67	84.00
	OSWR	75.33	79.33	80.00	68.00	71.33	80.00	72.00	76.00	**96.67**	95.33	95.33	92.00

[a] PP = Preprocessing Method

TABLE 2. Classification accuracy of the ML models for predicting honey adulteration level using full wavelengths, feature selection methods (39 features), and feature extraction methods.

PP [a]	Feature Selection Method	Feature Extraction Method											
		-				PCA				LDA			
		KNN	SVM	DT	NB	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Without Preprocessing	-	44.67	54.67	49.33	45.33	48.00	42.00	50.00	40.67	64.67	68.00	67.33	66.67
	Chi-Square	46.67	56.67	46.00	56.00	46.00	56.00	47.33	56.00	48.67	50.00	46.00	54.67
	SFM-RF	50.67	58.67	52.67	55.33	48.00	57.33	45.33	54.00	57.33	61.33	48.67	57.33
	SFM-SVM	53.33	57.33	45.33	56.00	52.00	57.33	52.00	53.33	49.33	52.67	47.33	50.00
	RFE-RF	52.67	48.00	52.67	53.33	51.33	54.67	52.00	53.33	61.33	65.33	63.33	63.33
	RFE-SVM	54.67	48.00	54.67	58.00	58.67	49.33	56.00	50.67	64.67	68.67	63.33	69.33
	WR	44.00	38.00	42.67	45.33	48.67	39.33	51.33	38.67	67.33	68.00	66.67	68.67
	OSWR	28.00	18.00	32.67	24.67	22.00	15.33	22.67	16.00	80.00	76.67	75.33	**80.67**
SNV	-	55.33	77.33	57.33	37.33	55.33	50.00	58.00	35.33	70.67	70.67	66.67	73.33
	Chi-Square	50.67	59.33	49.33	58.67	49.33	58.00	50.00	56.67	46.67	50.67	44.67	51.33
	SFM-RF	49.33	57.33	50.00	58.00	50.00	59.33	48.67	52.00	54.67	56.67	59.33	57.33
	SFM-SVM	53.33	60.67	50.67	55.33	54.00	60.67	54.00	52.00	52.00	55.33	56.00	55.33
	RFE-RF	56.67	53.33	58.67	56.00	60.67	58.00	60.67	52.67	58.67	62.67	60.00	62.00
	RFE-SVM	60.00	52.00	62.67	54.00	55.33	53.33	60.67	48.67	69.33	74.67	64.00	73.33
	WR	55.33	53.33	53.33	37.33	54.00	49.33	51.33	36.00	71.33	77.33	69.33	75.33
	OSWR	53.33	45.33	52.67	36.67	43.33	40.00	48.00	34.00	**92.67**	88.67	86.67	90.67

[a] PP = Preprocessing Method

for feature extraction, and employing both SVM and NB for classification.

After applying SNV preprocessing to the data, as outlined in Table 1, notable enhancements in classification accuracies become evident across a range of ML models.

Of particular note is the substantial improvement in the performance of the SVM, which increases from 77.33% to 93.33% when utilizing the entire wavelength spectrum. A particularly noteworthy achievement is observed with the comprehensive approach incorporating SNV preprocessing,

the proposed feature selection method (OSWR), LDA-based feature extraction, and KNN classification, resulting in an impressive classification accuracy of 96.67%.

To assess the variation in classification performance among the ML models, we conducted statistical paired t-tests. The findings indicate a significant difference in performance between KNN, SVM, DT, and NB. No statistically significant distinction is observed among SVM, DT, and NB in classification accuracies.

Table 3 presents an overview of the classification accuracies achieved by KNN, SVM, DT, and NB under various selections of wavelengths. These selections are characterized using RFE-SVM, WR, and the proposed feature selection techniques, all integrated with LDA-based feature extraction. Optimal feature counts, determined through a power-of-2-based WRR (WRR $= 2^x + 2, x = 0 \dots 7$), reveal that KNN achieves its highest accuracy of 96.67% when utilizing 39 optimal wavelengths. Notably, KNN consistently outperforms other classifiers across different configurations of selected wavelengths. As the table highlights, NB effectively identifies adulteration in up to 80% of honey samples using a minimal set of 11 carefully selected wavelengths through the proposed method.

Table 4 outlines the 39 most pivotal wavelengths selected through diverse feature selection methodologies. The distinct selection of wavelengths by different methods is noteworthy, reflecting their varying strategies for optimal wavelength identification. Visible spectral region wavelengths dominate the selections made by the initial three methods, while subsequent approaches predominantly gravitate toward the Visible to Near-Infrared region.

The nuanced classification accuracy of ML models is visualized through the confusion matrices depicted in Fig. 5. Each matrix offers detailed insight into the accuracy of the ML model's classifications for each class. As an illustrative example, Fig. 5 (b) illustrates SVM's classification proficiency in distinguishing pure honey, adulterated honey, and sugar syrup. SVM accurately identifies 26 samples as pure honey while misclassifying four samples as adulterated honey. The second matrix row indicates SVM's accurate identification of 87 adulterated honey samples while misclassifying three samples as pure honey. The third row underscores SVM's accurate classification of all 30 sugar syrup samples.

2) DETECTION OF ADULTERATION BY ADULTERANT TYPE

The evaluation of ML model performance in detecting different types of artificial sugars in honey is presented in Table 5. Across all models, a notable trend emerges wherein adulterated honey is consistently identified, with classification accuracies exceeding 90%. However, the efficiency of the models varies depending on the specific type of artificial sugar. Across all three types of sugar adulterants, the KNN algorithm achieves high accuracy, ranging from 96% to 98%, indicating that KNN effectively captures the patterns and relationships in the data to make accurate predictions. Both SVM and DT also show relatively strong performance across the different types of sugar adulterants, with accuracies ranging

from 94% to 98%, suggesting that these algorithms can handle this classification task effectively. While still achieving respectable accuracies, NB lags slightly behind the other algorithms. It achieves accuracies ranging from 90% to 94%, indicating that the assumption of independence between features in the NB model is not fully met by the data, leading to slightly lower accuracy than other algorithms. It is interesting to note that the performance of the algorithms is relatively consistent across the different types of sugar adulterants, suggesting that the features used for classification might have similar discriminative power for each adulterant type.

3) DETECTION OF ADULTERATION BY HONEY BOTANICAL SOURCE

The ability of classification algorithms to detect adulteration across various botanical sources of honey is examined in Table 6. The results reveal that the models consistently achieve notably high accuracy levels, often reaching 100%, indicating that the selected features hold significant discriminatory power regarding the botanical origin of the honey. While most honey classes attain high accuracy rates, certain classes exhibit slightly lower accuracy values. For instance, the "Rose" category consistently maintains an 86.67% accuracy across all algorithms, implying that distinguishing this particular class may present a more significant challenge. It is noteworthy that, due to the naive feature independence assumption, the NB algorithm occasionally records lower accuracy than other algorithms. It is important to acknowledge the limited number of samples per botanical source in this study, with 12 samples per floral source, three of which are pure honey samples.

Consequently, the results in Table 6 provide initial insights into the performance variations of ML models across distinct botanical sources. In cases where accuracy falls short of 100%, it indicates the models' limitations in discriminating pure samples. For example, Table 6 reveals that the KNN algorithm misclassified five out of 30 pure honey samples. These misclassified pure honey samples originated from various botanical sources such as Acacia (one sample), Eucalyptus (one sample), Rose (two samples), and Sulai (one sample), confirming the results presented in Fig. 6 (a), where KNN accurately classified 25 out of 30 pure honey samples and misclassified the remaining five as 10%-Adulterated honey.

4) TRAINING TIME ANALYSIS

Table 7 sheds light on the training times associated with ML models employing full and selected wavelength sets. Intriguingly, it becomes evident that using selected wavelengths accelerates the training process compared to utilizing the complete wavelength spectrum. Notably swift in both scenarios, KNN emerges as the fastest algorithm for detecting honey adulteration. In stark contrast, DT exhibits the slowest performance in this context. SVM registers slower training times than KNN but surpasses the prolonged training times associated with DT, whether employing full or selected wavelengths.

TABLE 3. Classification accuracy of ML models using SFM-SVM, WR, and OSWR-LDA feature selection-extraction methods with different numbers of wavelengths.

Aim	N [a]	Feature Selection Method											
		RFE-SVM-LDA				WR-LDA				OSWR-LDA			
		KNN	SVM	DT	NB	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Discrimination between Pure and Fake Honey	2	66.67	80.00	67.33	68.00	64.67	80.00	66.67	78.67	72.67	80.00	76.00	79.33
	3	72.00	80.00	67.33	68.00	67.33	80.00	70.00	76.67	77.33	80.00	72.00	77.33
	6	71.33	80.00	77.33	71.33	76.67	80.00	72.67	80.00	76.67	78.67	77.33	78.67
	11	76.67	80.00	76.67	68.00	74.67	78.00	73.33	78.00	77.33	81.33	74.00	80.00
	20	73.33	76.67	72.00	69.33	78.67	81.33	79.33	80.00	91.33	92.00	90.00	92.00
	39	82.00	87.33	76.67	85.33	88.00	87.33	80.67	84.00	**96.67**	95.33	95.33	92.00
	68	88.00	86.67	81.33	85.33	83.33	85.33	86.00	87.33	92.00	91.33	91.33	87.33
	114	85.33	86.00	87.33	85.33	84.00	84.67	80.00	78.67	89.33	88.67	87.33	75.33
Prediction of Honey Adulteration Level	2	52.00	55.33	50.00	52.00	39.33	33.33	40.00	35.33	46.67	39.33	42.00	36.67
	3	60.00	56.00	58.00	54.00	38.67	38.00	48.00	42.67	51.33	40.00	49.33	47.33
	6	52.67	54.00	54.00	53.33	55.33	42.67	44.67	48.67	56.67	54.67	60.67	60.00
	11	60.00	54.00	61.33	51.33	53.33	57.33	50.67	58.00	57.33	64.67	57.33	62.00
	20	54.67	59.33	54.00	55.33	55.33	58.00	58.00	57.33	79.33	80.00	71.33	76.67
	39	69.33	74.67	64.00	73.33	71.33	77.33	69.33	75.33	**92.67**	88.67	86.67	90.67
	68	71.33	74.67	70.00	74.00	74.00	74.67	64.67	72.67	85.33	81.33	75.33	84.67
	114	66.00	68.00	66.00	68.00	67.33	68.00	56.67	66.00	77.33	74.67	67.33	74.67

[a] N = Number of Wavelengths

**TABLE 4.** Optimal wavelengths selected by various feature selection methods.

Method	Selected Wavelengths												
Chi-Square	391	390	396	395	392	393	389	394	399	398	397	388	400
	387	385	386	401	402	384	403	404	383	405	382	406	407
	408	409	381	410	379	380	411	378	412	413	377	414	415
SFM-RF	446	408	442	391	388	420	390	438	418	395	409	413	407
	357	374	457	447	424	369	364	416	421	426	397	1406	444
	402	417	382	392	429	399	406	398	449	387	430	378	400
SFM-SVM	419	420	422	421	423	418	424	417	426	425	416	427	415
	428	414	429	413	430	431	412	411	432	433	410	434	435
	409	436	437	408	438	407	439	406	440	405	441	442	404
RFE-RF	403	400	439	394	405	1406	402	2465	395	2486	409	399	387
	985	1400	1404	407	2442	398	393	2480	2483	1409	611	416	1958
	406	397	2453	2464	425	2482	401	1401	1035	449	1121	386	1677
RFE-SVM	422	421	425	424	420	419	423	418	417	416	415	414	426
	413	541	427	412	411	428	410	409	408	429	430	539	1385
	407	406	431	1386	1384	1000	1387	432	1388	540	1383	999	433
WR	350	406	462	518	574	630	686	742	798	854	910	966	1022
	1078	1134	1190	1246	1302	1358	1414	1470	1526	1582	1638	1694	1750
	1806	1862	1918	1974	2030	2086	2142	2198	2254	2310	2366	2422	2478
Proposed OSWR Method	750	768	786	804	822	840	858	876	894	912	930	948	966
	984	1002	1020	1038	1056	1074	1092	1110	1128	1146	1164	1182	1200
	1218	1236	1254	1272	1290	1308	1326	1344	1362	1380	1398	1416	1434

C. PREDICTION OF HONEY ADULTERATION LEVEL

1) OVERALL ML MODEL PERFORMANCE

This section delves into the performance evaluation of ML models in predicting the levels of honey adulteration. Table 2 offers an insight into the classification accuracies achieved by the ML models in predicting adulteration levels. Both full wavelength data and variations involving feature selection and extraction methods were assessed on the raw data without preprocessing. SVM exhibited the highest classification accuracy of 54.67% and 68%, utilizing the full spectral range and LDA-reduced features. In contrast, DT achieved the highest accuracy with PCA-reduced, yielding an accuracy of 50%. The concluding row of Table 2 emphasizes that improved performance was obtained by employing the proposed feature
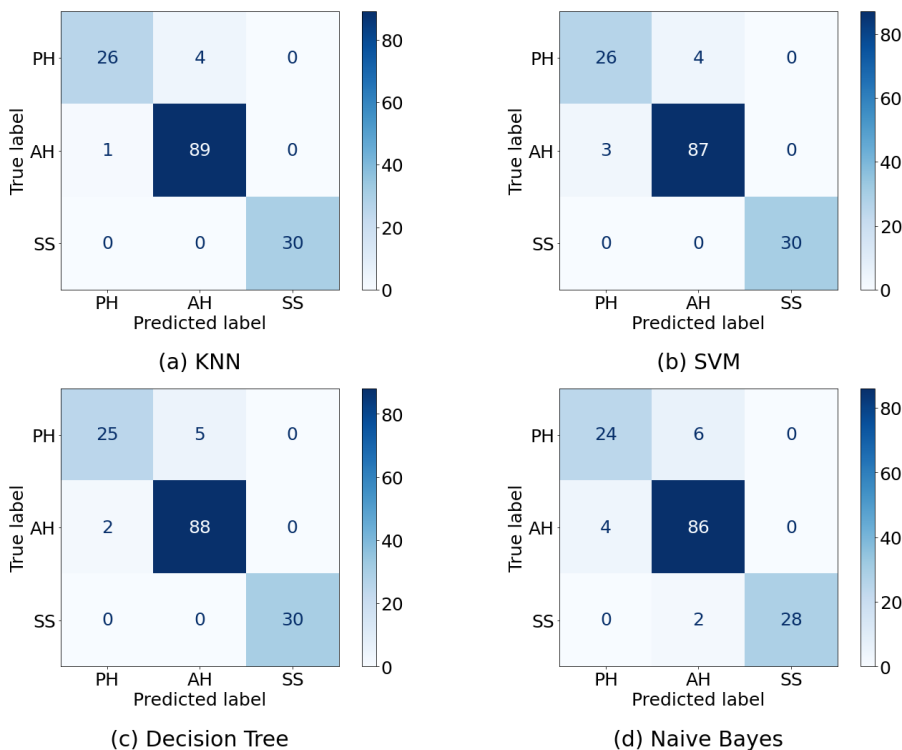
FIGURE 5. Confusion matrices of ML models for discriminating between genuine and fake honey using the proposed method. (a) K-Nearest Neighbors, (b) Support Vector Machine, (c) Decision Tree, (d) Naïve Bayes. PH = Pure Honey, AH = Adulterated Honey, SS = Sugar Syrup.

TABLE 5. Adulterant-based classification accuracies of ML models for discriminating between pure and fake honey and predicting honey adulteration level using the proposed method.

Sugar Adulterant	Discrimination between Pure and Fake Honey				Prediction of Honey Adulteration Level			
	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Glucose Syrup	96.00	94.00	96.00	90.00	90.00	88.00	88.00	90.00
Invert Syrup	96.00	96.00	96.00	94.00	94.00	90.00	84.00	90.00
Sucrose Syrup	98.00	96.00	94.00	92.00	94.00	88.00	88.00	92.00

TABLE 6. Classification accuracies of the ML models for discriminating between pure and fake honey and predicting the adulteration level in honey from various botanical origins using the proposed method.

Honey Botanical Origin	Discrimination between Pure and Fake Honey				Prediction of Honey Adulteration Level			
	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Acacia	93.33	86.67	93.33	93.33	86.67	80.00	86.67	93.33
Ajwain	100.00	100.00	100.00	100.00	93.33	80.00	86.67	93.33
Apple	100.00	100.00	100.00	93.33	93.33	86.67	93.33	93.33
Eucalyptus	93.33	93.33	93.33	93.33	86.67	86.67	80.00	86.67
Jamun	100.00	100.00	93.33	86.67	93.33	100.00	86.67	93.33
Litchi	100.00	100.00	100.00	100.00	100.00	100.00	93.33	93.33
Multiflora	100.00	100.00	100.00	86.67	100.00	100.00	100.00	100.00
Rose	86.67	86.67	86.67	86.67	80.00	73.33	73.33	80.00
Sidr	100.00	93.33	93.33	93.33	100.00	93.33	80.00	86.67
Sulai	93.33	93.33	93.33	86.67	93.33	86.67	86.67	86.67

selection method (OSWR) followed by LDA-based feature extraction. This approach led to KNN and NB achieving the highest classification accuracies of 80% and 80.67%, respectively.
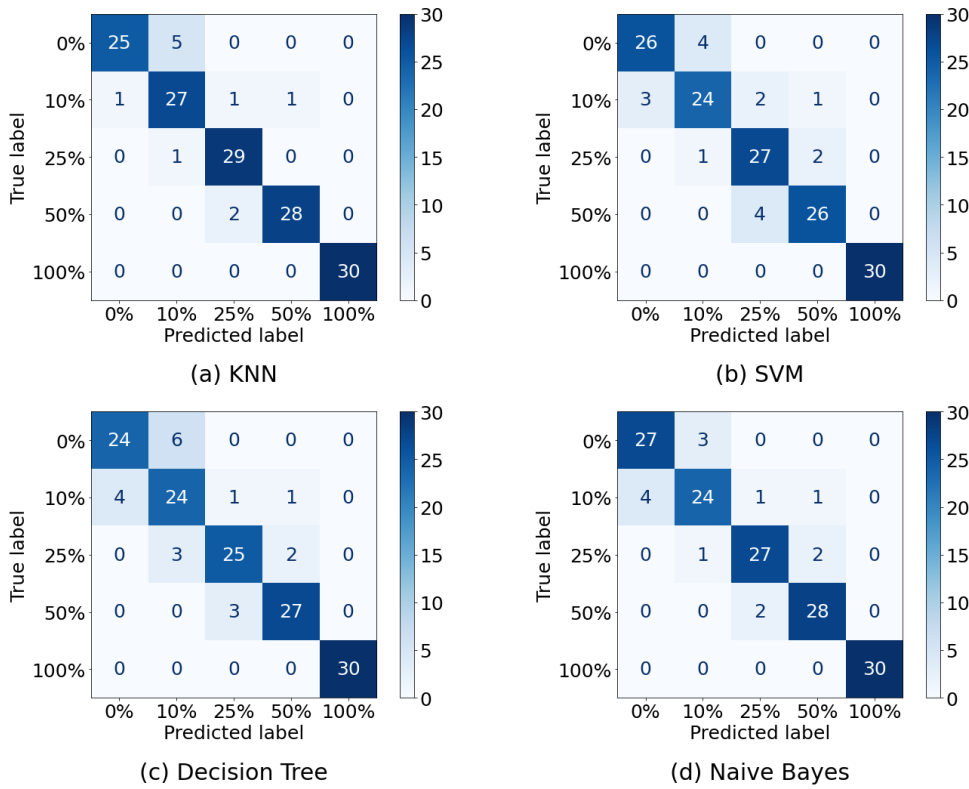
FIGURE 6. Confusion matrices of ML models for predicting the adulteration level using the proposed method.
(a) K-Nearest Neighbors, (b) Support Vector Machine, (c) Decision Tree, (d) Naïve Bayes.

TABLE 7. Training time (in milliseconds) of ML models for discriminating between pure and fake honey and detecting honey adulteration using full and selected wavelengths.

Features	Discrimination between pure and fake honey				Prediction of Honey Adulteration Level			
	KNN	SVM	DT	NB	KNN	SVM	DT	NB
Full Wavelengths	0.72	14.02	96.60	4.74	0.79	41.95	148.46	4.12
Selected Wavelengths	0.23	1.10	1.90	0.52	0.22	2.27	2.75	0.63

The complexities within the dataset, encompassing a combination of artificial sugar samples, pure honey samples, and various fake honey samples, contributed to the challenges faced in predicting honey adulteration levels. The diverse composition rendered the accurate grouping of spectral data according to adulteration levels difficult for the ML models.

Table 3 further illuminates the classification accuracies in predicting adulteration levels using data preprocessed with the SNV method. Overall, SNV preprocessing yielded performance improvements in predicting honey adulteration levels. Notably, the fusion of SNV, the proposed feature selection method (OSWR), LDA-based feature extraction, and KNN classification yielded the highest accuracy of 92.67%, surpassing other combinations. Statistical paired t-tests highlighted that the classification accuracy of the KNN model exhibited significant differences compared to SVM, DT, and NB. However, no significant distinctions emerged among SVM, DT, and NB classification accuracies.

The classification accuracies of varying numbers of wavelengths using different methods are presented in Table 3.

Results reveal that KNN achieves a peak accuracy of 92.67% utilizing 39 optimal wavelengths selected by the proposed feature selection method.

Illustrative confusion matrices, as depicted in Fig. 6, offer a deeper understanding of the ML models' effectiveness in predicting honey adulteration levels. The matrices represent five adulteration levels, ranging from pure honey (0% adulteration) to pure sugar syrup (100% adulteration). For instance, the confusion matrix of SVM in Fig. 6 (b) displays a satisfactory classification of pure honey (26 samples from 30 samples). Sugar syrup samples were more consistently classified correctly.

2) PREDICTION OF ADULTERATION LEVEL BY ADULTERANT TYPE

The performances of ML models in predicting honey adulteration levels based on the adulterant type are shown in Table 5. The accuracy values for different sugar adulterants and algorithms vary, suggesting that the performance

of the classifiers is influenced by both the characteristics of the adulterant type and the choice of the classification algorithm.

Invert Syrup has varying accuracy across different algorithms, ranging from 84.00% (DT) to 94.00% (KNN), indicating that some algorithms better capture the features distinguishing Invert Syrup from others. For Glucose Syrup, KNN, SVM, and NB achieve similar accuracy (90.00% or 88.00%), suggesting that these algorithms might effectively capture the distinguishing features of Glucose Syrup. At the same time, the DT might struggle with this classification task.

In several cases, NB achieves competitive accuracy, indicating that the naive feature independence assumption does not significantly impact the classification performance.

Sucrose Syrup achieves consistent accuracy across KNN, SVM, and DT (88.00%). NB achieves slightly higher accuracy (92.00%), suggesting that the features of Sucrose Syrup might have clear patterns that multiple algorithms can capture effectively.

3) PREDICTION OF ADULTERATION LEVEL BY HONEY BOTANICAL SOURCE

Table 6 offers a comprehensive perspective by presenting the classification performance of models in predicting adulteration levels across different honey types using the proposed methodology. As observed in previous scenarios, the accuracy values exhibit variation depending on the botanical origins and algorithms employed. Certain botanical origins consistently yield high accuracy, while others display variability in performance. For instance, Multiflora consistently achieves a perfect accuracy rate of 100% across most algorithms, indicating that it possesses highly distinctive features effectively captured by these models. Litchi also consistently attains high accuracy, suggesting well-separated features from other classes.

Conversely, 'Eucalyptus' exhibits slightly lower accuracy consistently, indicating a potentially more challenging classification task for this class. On the other hand, Rose demonstrates relatively lower accuracy across all algorithms, implying that distinguishing it based on the available features may pose challenges.

4) TRAINING TIME ANALYSIS

Lastly, Table 7 offers insights into the training times of the ML models for predicting honey adulteration levels, comparing full and selected wavelengths. The results underscore that utilizing selected wavelengths led to quicker model training times. KNN emerged as the fastest algorithm, demonstrating superior efficiency in full and optimal wavelength scenarios. Conversely, DT exhibited the longest training times, while SVM exhibited intermediate performance in terms of training time.

In summary, the investigation into predicting honey adulteration levels revealed the complexity of the dataset's composition and its challenges. Despite these complexities, the presented ML models displayed promising capabilities in achieving accurate predictions. The experimental results, derived from various configurations and preprocessing techniques, offer valuable insights into refining techniques to predict honey adulteration levels accurately.

IV. DISCUSSION

The presented results illuminate several noteworthy aspects in the context of honey adulteration detection and prediction of adulteration levels. The analyses encompassed a comprehensive evaluation of ML models, employing diverse feature selection, extraction methods, and preprocessing techniques. The following observations emerge from the discussion of the results:

A. DISCRIMINATION AND DETECTION OF ADULTERATION

The study underscores the effectiveness of ML models in distinguishing between pure honey, adulterated honey, and sugar syrup. The observed classification accuracies demonstrate the feasibility of employing spectral data to discern these distinct categories. KNN emerges as a robust classifier, consistently achieving notable accuracy and proving its potential for the classification task. The utilization of diverse feature selection and extraction methodologies further enhances the discrimination capability of the models.

B. IMPACT OF PREPROCESSING WITH SNV

The impact of preprocessing using SNV on detecting honey adulteration is notably positive in this study. SNV preprocessing is a crucial step in enhancing the performance of ML models when applied to Vis-NIR spectroscopy data.

SNV is primarily employed to correct baseline shifts and reduce the multiplicative effects of light scattering and path-length variations. In the context of honey adulteration detection, SNV plays a pivotal role in enhancing the quality of spectral data by mitigating background noise and variability, thereby improving classification accuracy.

Our work shows that preprocessing Vis-NIR spectral data using SNV enhances ML model performance in honey adulteration detection and prediction. Specifically, SVM achieved a classification accuracy of 93.33% using full wavelengths with SNV preprocessing, compared to 77.33% without preprocessing. This substantial increase in accuracy indicates SNV's effectiveness in reducing data noise and enhancing the discrimination between genuine and fake honey samples.

Furthermore, SNV preprocessing with the proposed feature selection algorithm, OSWR, LDA-based feature extraction, and KNN classification achieved the highest classification accuracy of 96.67% in detecting honey adulteration, outperforming other combinations, highlighting the crucial role of SNV in facilitating the success of advanced data processing techniques, such as feature selection and extraction, ultimately leading to improved detection of honey adulteration. In summary, preprocessing with SNV significantly enhances the accuracy and reliability of ML models for honey adulteration detection by effectively addressing data noise and variability, thus improving the overall quality of spectral data.

C. FEATURE SELECTION AND EXTRACTION

The proposed feature selection method, OSWR, outperforms other feature selection and feature extraction methods in this study for several compelling reasons. OSWR stands out by efficiently tackling the dimensional challenge posed by the large dataset, reducing it from 2151 wavelengths to a mere 39 while retaining crucial information. It excels in selecting high-relevance wavelengths and distinguishing between pure and fake honey. This adaptability is crucial as OSWR's data-driven approach tailors feature selection to the specific dataset, enhancing classification performance by capturing spectral attributes indicative of adulteration. When coupled with LDA for feature extraction, OSWR amplifies class separability, vital for differentiating various honey types effectively.

Furthermore, its sequential nature ensures consistent and stable feature selection, mitigating the risk of overfitting and promoting model generalization. This adaptability, consistency, and ability to deliver a compact yet informative feature set enhance classification accuracy and make OSWR a practical choice for real-world applications, facilitating the design of cost-effective spectrometers optimized for these specific wavelengths. In summary, OSWR's addictiveness, complementarity with LDA, and stability make it the ideal choice for addressing dataset complexities and achieving exceptional performance in discriminating between genuine and fake honey samples.

D. SUPERIORITY OF KNN OVER OTHER ML MODELS

KNN emerges as this study's top-performing ML model for several compelling reasons. First, KNN excels at capturing local patterns and relationships within the spectral data, making it especially adept at detecting subtle variations that signal honey adulteration. Its non-parametric nature allows it to handle the complex and nonlinear relationships that often characterize spectral datasets. Furthermore, KNN's simplicity and robustness are advantageous when dealing with high-dimensional data, like spectral readings, as it requires minimal parameter tuning and makes no assumptions about data distribution. Applying feature selection techniques in this study further enhances KNN's performance, allowing it to focus on the most informative aspects of the data.

Additionally, KNN's principle of assigning a data point to the majority class among its nearest neighbors aligns well with the homogeneity expected in honey adulteration classification. It thrives in situations where adulteration levels form clusters or localized patterns. Lastly, KNN's flexibility, adapting to different datasets and characteristics, adds to its appeal. However, the choice of distance metric and the selection of the number of neighbors (K) demand careful consideration, highlighting the importance of optimization and validation for realizing KNN's full potential.

The limited capacity of ML models to achieve perfect accuracy in this study can be attributed to several factors and inherent challenges. Firstly, the dataset's complexity plays a significant role, encompassing various types of honey, differing adulteration levels, and diverse artificial sugars. Such diversity introduces considerable variability that ML models may struggle to address comprehensively. Secondly, the overlapping spectral signatures of pure honey and adulterated samples, particularly at lower adulteration levels, can confound the models, resulting in misclassifications. Thirdly, spectral data's sensitivity to noise, variations in data collection conditions, or instrument limitations can introduce unpredictability that impacts accuracy.

Moreover, some spectral regions or wavelengths may lack the discriminative power necessary for precise classification. Each ML model has limitations, such as KNN's preference for local patterns over global relationships. Imbalanced class distribution, data preprocessing, feature engineering, and the selection of optimal hyperparameters all contribute to the challenge of achieving perfect accuracy. In essence, while ML models excel at identifying patterns, real-world datasets are inherently complex and often contain nuances and uncertainties that prevent the attainment of 100% accuracy. The goal is to develop models that offer practical utility and insights, even if perfection remains elusive.

E. PREDICTION OF ADULTERATION LEVELS

Predicting honey adulteration levels remains a complex task due to the intricate nature of the dataset. Despite the inherent challenges posed by varying honey types and varying types and concentrations of artificial sugar, ML models showcase commendable predictive capabilities. The notable influence of preprocessing, feature selection, and extraction on prediction accuracies underscores the importance of employing comprehensive methodologies to address this complex problem.

F. COMPARISON WITH PREVIOUS WORK

Our study's results align with previous research in several aspects, demonstrating the reliability of our findings. Specifically, our work substantiates prior studies that employed spectral data for honey quality assessment, showcasing that Vis-NIR spectroscopy is an effective tool for detecting honey adulteration.

Table 8 provides a comparative analysis of our approach with methods proposed in prior research. Most previous works, such as those referenced in [7], [23], [24], [25], [26], [28], and [29], primarily focused on evaluating the performance of ML models for detecting adulteration in a single type of honey. In contrast, our study broadens the scope by assessing the effectiveness of the models in detecting adulteration across various botanical sources.

Two distinctive features set our work apart from previous endeavors. First, introducing the OSWR feature selection method offers two notable advantages: A) Enhanced performance of ML models regarding prediction accuracy and reduced time complexity (as fewer wavelengths lead to simpler models and faster training times). B) The potential to utilize cost-effective, portable spectrometers designed for measuring spectral data at a reduced set of wavelengths. Secondly, our study rigorously evaluated the performance of ML models for detecting honey adulteration across ten distinct honey types. While some prior models achieved higher classification accuracies, they were predominantly tailored to detect adulteration in a single honey type.

TABLE 8. Comparative analysis of our study and previous research.

Author	Technology/ Range (nm)	Operation Mode	Honey Types	Adulterant	AD [a] Levels	Preprocessing	Feature Selection	Feature Extraction	ML Models	Performance
Aliaño-González et al. [7]	Vis-NIR 400-2500	Abs.	1	IS, RS, BS, FS	2 8	-	LDA	-	LDA PLSR	Accuracy = 100% R^2 =0.96
Raypah et al. [23]	Vis-NIR 400-1100	Abs.	1	DW, ACV, HFS	10	Moving Average Smoothing	-	PCA	PLSR	R^2= 0.96, RMSE = 5.88
Azmi et al. [24]	Vis-NIR 450-969	Trans.	1	SS	5	-	-	PCA	LDA QDA SVM	Accuracy = 99.33% Accuracy = 99.33% Accuracy = 99.33%
Ferreiro-González et al. [25]	Vis-NIR 400-2500	Abs.	1	HFCS	11	Moving Average Smoothing	Stepwise LDA	PCA	LDA PLSR	Accuracy = 100% R^2 = 0.9855
Calle et al. [26]	Vis-NIR 400-2500	Abs.	1	RS, HFCS, FS, IS, BS	8	-	-	-	SVM	Accuracy = 93.98%
Calle et al. [27]	Vis-NIR 400-2500	Abs.	2	Low-cost honey	2 2 10	First Derivative, SGF	-	-	SVM RF SVR	Accuracy = 100% Accuracy = 100% R^2 = 0.99, RMSE =1.89
Woeng et al. [28]	NIR 900-1700	Refl.	1	GS	11	SGF, Cut, Smoothing	-	PCA	KNN RF	Accuracy = 87% Accuracy = 90.2%
Tan et al. [29]	NIR 900-1700	Refl.	1	HFCS	11	Cut, EMSC, Smoothing, Normalization	-	PCA	LR	Accuracy = 98.2%
Proposed Method	**Vis-NIR 350-2500**	**Refl.**	**10**	**GS, IS, SS**	**3 5**	**SNV**	**OSWR**	**LDA**	**KNN**	**Accuracy = 96.67% Accuracy = 92.67%**

[a] AD = Adulteration, Abs. = Absorbance, Trans. = Transmittance, Refl. = Reflectance, IS = Invert Sugar, RS = Rice Syrup, BS = Brown Cane Sugar, FS = Fructose Syrup, PLSR = Partial Least Squares Regression, R^2 = Coefficient of Determination, DW = Distilled Water, ACV = Apple Cider Vinegar, HFS = High Fructose Syrup, RMSE = Root Mean Squared Error, SS = Sucrose Syrup, QDA = Quadratic Discriminant Analysis, HFCS = High Fructose Corn Syrup, SVR = Support Vector Regression, GS = Glucose Syrup, SGF = Savitzky-Golay filter, EMSC = Extended Multiplicative Signal Correction, LR = Logistic Regression.

Nevertheless, it is worth noting that while our study provides compelling evidence of the efficacy of the proposed feature selection method, OSWR, other research may favor different feature selection or extraction approaches. Variations in honey botanical sources, sugar adulterants, adulteration levels, preprocessing methods, and classification algorithms can lead to divergent findings in different studies. Therefore, while our results are consistent with the broader theme of using Vis-NIR spectroscopy for honey quality evaluation, some variations might exist when making detailed comparisons with specific methodologies or datasets from other research.

In summary, our study's findings generally agree with prior research that highlights the potential of spectral analysis in honey quality assessment while also showcasing the effectiveness of our proposed feature selection and extraction method for consistently detecting fake honey samples across various botanical sources and sugar adulterants.

G. IMPLICATIONS OF THE STUDY
This research holds significant implications for various stakeholders. First, it helps the honey business verify product authenticity and quality. This may help honey producers and suppliers fight fraud, protect their brands, and preserve consumer confidence. The discoveries may also enhance regulatory and food safety, improving counterfeit honey monitoring and enforcement. Furthermore, the study suggests the potential for cost-effective spectrometry techniques to transform on-site honey quality assessment through the utilization of portable spectrometers. Ultimately, consumers stand to benefit the most, as they can make more informed choices and have increased confidence in the authenticity and purity of the honey products they purchase. In summary, this research can make a substantial positive impact on the honey industry by promoting transparency, integrity, and consumer trust.

H. LIMITATIONS OF THE STUDY
While offering valuable insights into honey adulteration detection, this study has certain limitations that warrant consideration in future research. Future studies should prioritize the inclusion of more diverse datasets, encompassing variations in geographic regions, floral sources, and processing methods to enhance the reliability and real-world applicability of detection models. Cross-spectrometer validation can ensure broader hardware compatibility while addressing the impact of environmental factors such as temperature and humidity on spectral data should be explored. Large-scale experiments with extensive sample sizes can improve the representation of honey types and adulteration levels. Additionally, future research should investigate alternative ML algorithms, including deep learning approaches, for more accurate and interpretable models. Developing cost-effective spectrometry solutions and exploring emerging sensor technologies, such as portable and smartphone-based spectroscopy, can democratize honey quality assessment. Collaborative efforts with experts from various fields, including chemistry and food science, can further deepen our understanding of honey composition and variations. Addressing these recommendations will contribute to more reliable

and accessible honey adulteration detection methods, benefiting the food industry and consumer protection.

V. CONCLUSION

In conclusion, integrating reflectance Vis-NIR spectroscopy with advanced techniques such as feature selection, feature extraction, and classification methods has demonstrated substantial efficacy in discriminating pure and fake honey samples and in precisely determining adulteration levels. A compelling solution emerged through systematically exploring various combinations of these techniques: the proposed feature selection algorithm coupled with LDA and KNN classification. This amalgamation yielded impressive outcomes, achieving a remarkable classification accuracy of 96.67% for honey adulteration detection and 92.67% accuracy in discerning the level of adulteration.

Importantly, the developed model showcases its prowess in detecting adulteration across a spectrum of ten distinct honey types. By leveraging the proposed feature selection methodology, a significant reduction in data dimensionality was realized, downsizing the original 2151 wavelengths to a mere 39 wavelengths. This reduction offers substantial benefits for practical implementation, including the potential for designing and deploying cost-effective spectrometers, enabling rapid and precise detection of honey adulteration.

The research underscores the promising prospect of leveraging cutting-edge techniques to safeguard the integrity and quality of honey products. As this study primarily focused on detection, there is the potential for further exploration into quantifying the extent of adulteration and extending the methodology to accommodate variations in honey sources and adulterant types. The findings herein contribute to enhancing food authenticity and quality assessment through the innovative fusion of spectroscopic technologies and advanced data analysis methods.

ACKNOWLEDGMENT

REFERENCES

[1] L. Castro-Vázquez, M. C. Díaz-Maroto, C. de Torres, and M. S. Pérez-Coello, "Effect of geographical origin on the chemical and sensory characteristics of chestnut honeys," *Food Res. Int.*, vol. 43, no. 10, pp. 2335–2340, Dec. 2010.

[2] P. M. Kuś, I. Jerković, Z. Marijanović, M. Kranjac, and C. I. G. Tuberoso, "Unlocking phacelia tanacetifolia Benth. Honey characterization through melissopalynological analysis, color determination and volatiles chemical profiling," *Food Res. Int.*, vol. 106, pp. 243–253, Apr. 2018.

[3] E. Alissandrakis, P. A. Tarantilis, C. Pappas, P. C. Harizanis, and M. Polissiou, "Investigation of organic extractives from unifloral chestnut (Castanea sativa L.) and eucalyptus (Eucalyptus globulus Labill.) honeys and flowers to identification of botanical marker compounds," *LWT-Food Sci. Technol.*, vol. 44, no. 4, pp. 1042–1051, May 2011.

[4] P. M. da Silva, C. Gauche, L. V. Gonzaga, A. C. O. Costa, and R. Fett, "Honey: Chemical composition, stability and authenticity," *Food Chem.*, vol. 196, pp. 309–323, Apr. 2016.

[5] A. J. Siddiqui, S. G. Musharraf, and M. I. Choudhary, "Application of analytical methods in authentication and adulteration of honey," *Food Chem.*, vol. 217, pp. 687–698, Feb. 2017.

[6] European Commission, "Council directive 2001/110/EC of 20 December 2001 relating to honey," *Off. J. Eur. Communities L*, vol. 10, pp. 47–52, Dec. 2002.

[7] M. J. Aliaño-González, M. Ferreiro-González, E. Espada-Bellido, M. Palma, and G. F. Barbero, "A screening method based on visible-NIR spectroscopy for the identification and quantification of different adulterants in high-quality honey," *Talanta*, vol. 203, pp. 235–241, Oct. 2019, doi: 10.1016/j.talanta.2019.05.067.

[8] A. Devi, J. Jangir, and K. A. Anu-Appaiah, "Chemical characterization complemented with chemometrics for the botanical origin identification of unifloral and multifloral honeys from India," *Food Res. Int.*, vol. 107, pp. 216–226, May 2018, doi: 10.1016/j.foodres.2018.02.017.

[9] M. J. Aliaño-González, M. Ferreiro-González, E. Espada-Bellido, G. F. Barbero, and M. Palma, "Novel method based on ion mobility spectroscopy for the quantification of adulterants in honeys," *Food Control*, vol. 114, Aug. 2020, Art. no. 107236, doi: 10.1016/j.foodcont.2020.107236.

[10] F. Huang, H. Song, L. Guo, P. Guang, X. Yang, L. Li, H. Zhao, and M. Yang, "Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion," *Spectrochimica Acta A, Mol. Biomolecular Spectrosc.*, vol. 235, Jul. 2020, Art. no. 118297, doi: 10.1016/j.saa.2020.118297.

[11] N. Arroyo-Manzanares, M. García-Nicolás, A. Castell, N. Campillo, P. Viñas, I. López-García, and M. Hernández-Córdoba, "Untargeted headspace gas chromatography—Ion mobility spectrometry analysis for detection of adulterated honey," *Talanta*, vol. 205, Dec. 2019, Art. no. 120123, doi: 10.1016/j.talanta.2019.120123.

[12] M. Esteki, J. Simal-Gandara, Z. Shahsavari, S. Zandbaaf, E. Dashtaki, and Y. Vander Heyden, "A review on the application of chromatographic methods, coupled to chemometrics, for food authentication," *Food Control*, vol. 93, pp. 165–182, Nov. 2018, doi: 10.1016/j.foodcont.2018.06.015.

[13] M. Al-Mahasneh, M. Al-U'datt, T. Rababah, M. Al-Widyan, A. A. Kaeed, A. J. Al-Mahasneh, and N. Abu-Khalaf, "Classification and prediction of bee honey indirect adulteration using physiochemical properties coupled with K-Means clustering and simulated annealing-artificial neural networks (SA-ANNs)," *J. Food Qual.*, vol. 2021, pp. 1–9, Apr. 2021, doi: 10.1155/2021/6634598.

[14] M. E. Crăciun, O. C. Pârvulescu, A. C. Donise, T. Dobre, and D. R. Stanciu, "Characterization and classification of Romanian acacia honey based on its physicochemical parameters and chemometrics," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, Nov. 2020, doi: 10.1038/s41598-020-77685-9.

[15] S. Amiry, M. Esmaiili, and M. Alizadeh, "Classification of adulterated honeys by multivariate analysis," *Food Chem.*, vol. 224, pp. 390–397, Jun. 2017, doi: 10.1016/j.foodchem.2016.12.025.

[16] N. S. Sotiropoulou, M. Xagoraris, P. K. Revelou, E. Kaparakou, C. Kanakis, C. Pappas, and P. Tarantilis, "The use of SPME-GC-MS IR and Raman techniques for botanical and geographical authentication and detection of adulteration of honey," *Foods*, vol. 10, no. 7, p. 1671, Jul. 2021, doi: 10.3390/foods10071671.

[17] M. K. Islam, T. Sostaric, L. Y. Lim, K. Hammer, and C. Locher, "Sugar profiling of honeys for authentication and detection of adulterants using high-performance thin layer chromatography," *Molecules*, vol. 25, no. 22, p. 5289, Nov. 2020, doi: 10.3390/molecules25225289.

[18] M. J. Aliaño-González, M. Ferreiro-González, E. Espada-Bellido, M. Palma, and G. F. Barbero, "A screening method based on headspace-ion mobility spectrometry to identify adulterated honey," *Sensors*, vol. 19, no. 7, p. 1621, Apr. 2019, doi: 10.3390/s19071621.

[19] N. Irawati, N. M. Isa, A. F. Mohamed, H. A. Rahman, S. W. Harun, and H. Ahmad, "Optical microfiber sensing of adulterated honey," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5510–5514, Sep. 2017, doi: 10.1109/JSEN.2017.2725910.

[20] Z. Gan, Y. Yang, J. Li, X. Wen, M. Zhu, Y. Jiang, and Y. Ni, "Using sensor and spectral analysis to classify botanical origin and determine adulteration of raw honey," *J. Food Eng.*, vol. 178, pp. 151–158, Jun. 2016, doi: 10.1016/j.jfoodeng.2016.01.016.

[21] A. Zakaria, A. Y. M. Shakaff, M. J. Masnan, M. N. Ahmad, A. H. Adom, M. N. Jaafar, S. A. Ghani, A. H. Abdullah, A. H. A. Aziz, L. M. Kamarudin, N. Subari, and N. A. Fikri, "A biomimetic sensor for the classification of honeys of different floral origin and the detection of adulteration," *Sensors*, vol. 11, no. 8, pp. 7799–7822, Aug. 2011, doi: 10.3390/s110807799.

[22] M. A. Al-Awadhi and R. R. Deshmukh, "A review on modern analytical methods for detecting and quantifying adulteration in honey," in *Proc. Int. Conf. Mod. Trends Inf. Commun. Technol. Ind. (MTICTI)*, Dec. 2021, pp. 1–6, doi: 10.1109/MTICTI53925.2021.9664767.

[23] M. E. Raypah, A. F. Omar, J. Muncan, M. Zulkurnain, and A. R. Abdul Najib, "Identification of stingless bee honey adulteration using visible-near infrared spectroscopy combined with aquaphotomics," *Molecules*, vol. 27, no. 7, p. 2324, Apr. 2022, doi: 10.3390/molecules27072324.

[24] M. F. I. Azmi, D. Jamaludin, S. A. Aziz, Y. A. Yusof, and A. M. Mustafah, "Adulterated stingless bee honey identification using VIS-NIR spectroscopy technique," *Food Res.*, vol. 5, pp. 85–93, Jan. 2021, doi: 10.26656/fr.2017.5(S1).035.

[25] M. Ferreiro-González, E. Espada-Bellido, L. Guillén-Cueto, M. Palma, C. G. Barroso, and G. F. Barbero, "Rapid quantification of honey adulteration by visible-near infrared spectroscopy combined with chemometrics," *Talanta*, vol. 188, pp. 288–292, Oct. 2018, doi: 10.1016/j.talanta.2018.05.095.

[26] J. L. P. Calle, M. Ferreiro-González, M. J. Aliaño-González, E. Espada-Bellido, M. Palma, and G. F. Barbero, "Quality control of honey by visible and near infrared spectroscopy (Vis-NIRS) and headspace-ion mobility spectrometry (HS-IMS)," in *Proc. 3rd Int. Electron. Conf. Environ. Res. Public Health*. Basel, Switzerland: MDPI, Jan. 2021, doi: 10.3390/ECERPH-3-09079.

[27] J. L. P. Calle, I. Punta-Sánchez, A. V. González-de-Peredo, A. Ruiz-Rodríguez, M. Ferreiro-González, and M. Palma, "Rapid and automated method for detecting and quantifying adulterations in high-quality honey using vis-NIRs in combination with machine learning," *Foods*, vol. 12, no. 13, p. 2491, Jun. 2023, doi: 10.3390/foods12132491.

[28] V. Woeng, L. Y. Lim, L. A. K. Saleena, M. I. Solihin, and L. P. Pui, "Physicochemical properties and detection of glucose syrup adulterated Kelulut (*Heterotrigona itama*) honey using near-infrared spectroscopy," *J. Food Process. Preservation*, vol. 46, no. 7, p. e16686, Jul. 2022, doi: 10.1111/jfpp.16686.

[29] S. H. Tan, L. P. Pui, M. I. Solihin, K. S. Keat, W. H. Lim, and C. K. Ang, "Physicochemical analysis and adulteration detection in Malaysia stingless bee honey using a handheld near-infrared spectrometer," *J. Food Process. Preservation*, vol. 45, no. 7, p. e15576, Jul. 2021, doi: 10.1111/jfpp.15576.

[30] *ASD FieldSpec 4 Hi-Res: High Resolution Spectroradiometer*. Accessed: May 20, 2023. [Online]. Available: https://www.malvernpanalytical.com/en/products/product-range/asd-range/fieldspec-range/fieldspec4-hi-res-high-resolution-spectroradiometer

[31] A. Taggart, "Python ASD parser," Univ. Sheffield, Jun. 2017, doi: 10.15131/shef.data.5164816.v1.

[32] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Correction to the description of standard normal variate (SNV) and de-trend (DT) transformations in practical spectroscopy with applications in food and beverage analysis—2nd edition," *J. Near Infr. Spectrosc.*, vol. 1, no. 3, pp. 185–186, Jun. 1993, doi: 10.1255/jnirs.21.

[33] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019, doi: 10.1109/MGRS.2019.2911100.

[34] A. K. Verma and S. Pal, "Prediction of skin disease with three different feature selection techniques using stacking ensemble method," *Appl. Biochemistry Biotechnol.*, vol. 191, no. 2, pp. 637–656, Jun. 2020, doi: 10.1007/s12010-019-03222-8.

[35] X. Wei, M. A. Johnson, D. B. Langston, H. L. Mehl, and S. Li, "Identifying optimal wavelengths as disease signatures using hyperspectral sensor and machine learning," *Remote Sens.*, vol. 13, no. 14, p. 2833, Jul. 2021, doi: 10.3390/rs13142833.

[36] N. Zhang, M. Chen, F. Yang, C. Yang, P. Yang, Y. Gao, Y. Shang, and D. Peng, "Forest height mapping using feature selection and machine learning by integrating multi-source satellite data in Baoding City, North China," *Remote Sens.*, vol. 14, no. 18, p. 4434, Sep. 2022, doi: 10.3390/rs14184434.

[37] D. Tsankova and S. Lekova, "Botanical origin-based honey discrimination using Vis-Nir spectroscopy and statistical cluster analysis," *J. Chem. Technol. Metall.*, vol. 50, no. 5, pp. 638–642, 2015.

[38] Y. Li and H. Yang, "Honey discrimination using visible and near-infrared spectroscopy," *ISRN Spectrosc.*, vol. 2012, pp. 1–4, Nov. 2012, doi: 10.5402/2012/487040.

[39] I. Jolliffe, "Principal component analysis," in *Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ, USA: Wiley, 2014.

[40] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[41] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[42] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Process. Soc. Workshop*, Aug. 1999, pp. 41–48.

[43] R. O. Duda and P. E. Hart, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2006.

[44] M. A. Al-Awadhi and R. R. Deshmukh, "Honey adulteration detection using hyperspectral imaging and machine learning," in *Proc. 2nd Int. Conf. Artif. Intell. Signal Process. (AISP)*, Feb. 2022, pp. 1–5, doi: 10.1109/AISP53593.2022.9760585.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[46] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, Oct. 2007.

[47] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of Complexity: Festschrift for Alexey Chervonenkis*. Cham, Switzerland: Springer, 2015, pp. 11–30.

[48] M. A. Al-Awadhi and R. R. Deshmukh, "Detection of adulteration in coconut milk using infrared spectroscopy and machine learning," in *Proc. Int. Conf. Modern Trends Inf. Commun. Technol. Ind. (MTICTI)*, Dec. 2021, pp. 1–4, doi: 10.1109/MTICTI53925.2021.9664764.

[49] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998, doi: 10.1023/A:1009715923555.

[50] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

MOKHTAR AL-AWADHI received the B.Sc. degree in information technology from Cairo University, Cairo, Egypt, and the M.Sc. degree in computer science and communications engineering from the University of Duisburg–Essen, Duisburg, Germany. He is currently pursuing the Ph.D. degree in computer science and information technology with Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India. He is a Lecturer with the Department of Information Technology, Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen. His research interests include machine learning, computer vision, pattern recognition, image processing, hyperspectral imaging and non-imaging, data analytics, and natural language processing.

RATNADEEP DESHMUKH received the M.Sc., M.E., and Ph.D. degrees in computer science and engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India. He is currently a Professor and the former Head of the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University. Previously, he was an Associate Professor with the Department of Computer Science and Information Technology and held roles as the Director with VIASMSC, Aurangabad. He has an extensive professional track record, including coordinating the DST-FIST Program and the GIAN Program at Dr. Babasaheb Ambedkar Marathwada University. His research interests include human–computer interaction, data mining, image processing, and artificial intelligence. He is the Chairperson of the IETE Aurangabad Center and was elected as the Sectional President of the ICT Section of the Indian Science Congress Association (ISCA), in 2019.

• • •