IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2022, accepted 19 July 2022, date of publication 16 August 2022, date of current version 25 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3199021

RESEARCH ARTICLE

Prediction of Pakistani Honey Authenticity Through Machine Learning

NOUREEN FATIMA[1], GHULAM MUJTABA[1], ADNAN AKHUNZADA[2], (Senior Member, IEEE), KHURSHED ALI[3], ZAHID HUSSAIN SHAIKH[4], BABY RABIA[5], (Senior Member, IEEE), AND JAVED AHMED SHAHANI[1]

[1]Center of Excellence for Robotics, Artificial Intelligence, and Block Chain, Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan
[2]Faculty of Computing and Informatics, University Malaysia Sabah, Kota Kinabalu 88400, Malaysia
[3]Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan
[4]Department of Mathematics, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan
[5]Department of Education, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan

Corresponding authors: Noureen Fatima (noureen.mscss19@iba-suk.edu.pk) and Adnan Akhunzada (adnan.akhunzada@ums.edu.my)

This work was supported by the Sind Research Support Program (SRSP) initiated by Sindh Higher Education Commission under Project SHEC/SRSP/ITCS-3/25/2020-21, and in part by the University Malaysia Sabah (UMS). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

ABSTRACT Honey is a high-demand product in many countries because it is high in nutritional value and rich in antioxidants. Thus, the demand for honey is increased. However, the productivity of honey is naturally lower than its demand. Therefore, honey has often become a target for adulteration. Adulteration of honey is a critical issue because the nutritional value of pure honey is reduced by adding cheap and easily available sweeteners, affecting the consumers' health. Thus, investigating honey authenticity is popular among regulatory bodies, the food industry, retail sellers, and consumers. Several works have been done to predict the authenticity of honey using various physicochemical features. Few other works have also classified honey on the basis of geographical or botanical origin. However, previous studies have three major limitations. First, the existing studies used the imbalanced datasets, and the performance of these studies further needs attention. Second, as far as we know, no researcher has attempted to use machine learning approaches in investigating the adulteration of Pakistani honey. Finally, the dataset for predicting the authenticity of Pakistani honey is lacking. Therefore, this study proposes a novel classification model to address the aforementioned weaknesses by classifying the authenticity of Pakistani honey using machine learning algorithms and several physicochemical features. This work also presents three classification models systematically to classify the Pakistani honey into three levels. The first level classifies whether the honey is original or branded. The second level classifies the geographical origin. The botanical origin of honey is classified in the third level. Our experimental results show that the proposed features coupled with machine learning algorithms can predict the authenticity of Pakistani honey with outstanding results. We believe that our proposed work will be proved beneficial in reducing the adulteration of Pakistani honey.

INDEX TERMS Machine learning, botanical origin, geographical origin, physicochemical properties, Pakistani honey.

I. INTRODUCTION

Honey is organic food mainly composed of minerals, sugar, glucose, fructose 95% – 99%, fructose-oligosaccharides 4%-5% of dry honey mater, aromatic substances, and water [1]. Honey's composition, aroma, and flavor depend on the flower honeybees eat and the geographical and climate region they belong [2]. Therefore, the physicochemical properties of honey may vary in various regions and countries [3], [4], [5]. Moreover, honey production may vary depending on the species of bees, weather conditions, and the process of formulation, manipulation, packing, and storage

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

time [6], [7]. Honey is widely consumed as medicine and sweetener worldwide. Honey is also considered the best energy source. According to nutritional standards, honey is not a complete food. However, it is identified as a potential dietary supplement. Honey is an easy-to-digest sweetener; compared with saccharose, it can be used in various food products [8]. Thus, honey has been widely consumed in various commercially available products as a sweetener. In medicine, it is used as an ointment, prebiotic, skincare product, food preservative, and treatment of cough and eye ailments [9]. Honey can provide many health benefits; honey with excessive trace minerals can harm humans [10]. Honey contains the high amounts of natural sweetener and provides many therapeutic effects [11].

The standard regularization of honey is presented in the European Council (EC) directive in 2001/110, further amended by the 2014/63/European Union (EU) directive [12]. The consumption criteria of honey in the market and the human consumption of honey as medicine are provided in Annex II of study [13]. The conditions of other legislation on the adulteration of honey may vary at the national level [14]. However, the existing standard regularization do not specify the chemical composition of nectar honey. These standards provide the limits in the physicochemical properties (moisture content, sugar content, or electrical conductivity) of honey based on a few botanical origins. Pakistan produces an abundant amount of unique-tasting and quality honey in the Middle East and exports around 4000 tons of honey for export to Arab countries, amounting to approximately 23 million dollars [15]. Pakistan produces high-quality honey from different floras or colonies, including *Apis dorsata* (*Ziziphus, eucalyptus, Sheesham, sunflower, kalonji, and Robinia*) and *Apis Florea*, and other species in various ecological areas. However, most of honey from *Apis dorsata* and *Apis Florea* are used for personal consumption or sold locally [15]. Both colonies are found in the hilly and mountainous areas in Pakistan. Moreover, Honey is more produced in three provinces of the Pakistan namely, Punjab, Khyber Pakhtunkhwa (KPK), and Sindh [15].

Adulteration has two major aspects: the origin of the honey and production mode. For the origin of honey, knowing the geographical and botanical origin of honey is vital. The manipulation and adulteration of honey can be identified by its physicochemical properties, such as moisture, ash, pH acidity, and geographical and botanical origin. In addition, the authenticity of honey must be ensured by analyzing honey samples thoroughly according to their chemical formula [16].

Multivariate data analysis is used in evaluating the set of random variables statistically [17]. The major concepts of multivariate data analysis have been recently fused into computer science and the artificial intelligence field. The concept of algorithm has been also incorporated to determine the hidden pattern of the complex datasets. Multivariate data analysis provided feasibility to machine learning algorithms for both predictive and exploratory data analysis. Moreover, machine learning allows the systems to learn and improve from the experience without being explicitly programmed. Machine learning algorithms are often categorized as either supervised or unsupervised. In supervised machine learning algorithms, datasets are labeled by domain experts. Meanwhile, for unsupervised machine learning algorithms datasets are not labeled.

Several researchers have worked on machine learning and multivariate data analysis to find the most important and irrelevant features of the classification model's performance [18]. Many researchers have worked on the adulteration of the honey by using the geographical [19], [20], [21] and botanical origin of honey [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. However, the existing works have three main limitations. First, the existing studies used the imbalanced datasets and the performance of these studies needs further attention. Second, as far as we know, no researcher has attempted machine learning approaches in predicting the authenticity of Pakistani honey. Finally, the dataset for predicting the authenticity of Pakistani honey is lacking. Therefore, this study aims to predict the authenticity of Pakistani honey by using eight informative physicochemical features (namely, pH, moisture, electrical conductivity, protein, ash content, sugar, sucrose, and acidity) from the collected Pakistani honey samples. Then, these features were fed as input to seven machine learning algorithms namely, Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Extreme Gradient Boosted Tree (XGBoost), and Linear Discriminant Analysis (LDA). The physicochemical properties of Pakistani honey are assumed to provide information on the botanical and geographical origin of the original or branded honey. With the classifying pattern of Pakistani honey according to the physicochemical properties, the machine learning algorithms can successfully predict information for new and unknown samples. We also developed the three-level classification model. The top level predicts whether the collected honey sample is original or branded. The second level predicts the geographical location (namely, Sindh, Punjab, and KPK) of a given honey sample. Finally, the third level predicts the botanical origin (namely, Dorsata, Florea, and others) of the given honey sample. The experimental results showed that our proposed method for predicting the authenticity of honey has an the overall accuracy of 100%. The core contributions of this work are as follows:

1) In this study, a new Pakistani honey dataset consisting of 140 honey samples was developed. The dataset contains three distinct botanical origins (namely, Dorsata, Florea, and others origin) belonging to three distinct regions of Pakistan (namely, Punjab, Sindh, and KPK).
2) This study is the first to attempt to work on the authenticity of Pakistani honey by employing several machine learning algorithms with the fusion of multivariate data analysis. An accuracy of 100% is achieved by employing the XGBoost machine learning algorithm.

3) Our collected dataset is naturally imbalanced. Thus, we employed the latest balancing approach named feature construction and smote-based imbalance handling (FCSMI) to balance the dataset and to improve the overall classification accuracy of XGBoost.

The rest of this study is arranged as follows. Section II describes the existing works on honey authenticity. SectionIII presents material and methods used in this study. Section IV presents the experimental settings and results. Section V presents the theoretical analysis of our obtained results in the form of a discussion. Finally, Section VI concludes our research.

II. RELATED WORK

The manipulation of honey and adulteration can be identified by tracing the geographical and botanical origin of honey. Therefore, honey samples must be comprehensively analyzed to ensure the authenticity of the honey [36]. Several researchers have worked to determine the adulteration of honey in various countries, such as India [7], [24], China [26], [31], [35], [37], Saudi Arabia [19], Italy [29], Brazil [33], Poland [38], Argentina [39], Spain [32], Finland [40], Turkey [27], New Zealand [25] and Uruguay [22]. These studies have investigated honey in various regions. Then, the honey is classified based on its geographical origin or botanical origin. For instance, Anjos *et al.* [41] classified the honey based on botanical origin from the six regions in Uruguay. Similarly, Sun *et al.* [31] classified the honey based on botanical origin from different regions in China. Wei and Wang [34] collected honey samples from 22 countries. They classified honey based on the botanical region. The details for each study are provided in Table 1.

Many researchers have investigated the authenticity of honey by using different methods, such as, Melissopalynology (analyzing honey by its pollen grain). However, this method has two limitations. First, this method is quite time consuming method. Secondly, highly specialized personnel is required [42] to evaluate it for authenticity. Many researchers have worked on the geographical and botanical origin profile of honey to overcome the limitations of above-mentioned method to recognize honey adulteration. In both geographical and botanical origin of the honey, chemical profile of honey was analyzed using various methods such as, atomic spectroscopy (determining the elements of the honey by using its mass spectrum), inductively coupled plasma mass spectrometry (ICP-MS), nuclear magnetic resonance spectroscopy (NMR) and many other. Moreover, many researchers have worked on physicochemical properties of honey. This is a well-known process in which both physical and chemical properties of honey are determined. These properties include pH, moisture, electrical conductivity, water content, and color. Physicochemical properties have been used in the recent decade to determine the botanical and geographical origin of honey [17]. However, this approach requires standard means, time, and expertise. On contrary, using the physicochemical properties with machine learning and multivariate data analysis has an advantage. In particular, machine learning algorithms can find the hidden pattern and multivariate data analysis provides the statistical and intelligent method to uncover the important information that an expert cannot sometimes determine. Many researchers have also worked on sensory data through electronic tongues and electronic noses to extract the physicochemical properties. The details are provided in Table 1.

The sample size of each studies varies from 20 to 300 instances. Two studies did not mention the size of the sample explicitly [32], [35]. Fifteen out of twenty-five studies worked with a sample size of 100, as shown in Table 1. These studies have good performances and are substantial. However, these studies cannot find the hidden pattern of the data because machine learning or multivariate analysis need much information in terms of data to learn the pattern [43]. For instance, Kortesniemi *et al.* [40] utilized a sample size of 20 to classify honey by its botanical origin through nine regions in Finland. The nine botanical origin classes were differentiated using 20 samples. Therefore, learning the pattern from such a small sample is really difficult for multivariate data analysis or principle component analysis (PCA) [44]. A similar kind of work was conducted by Anjos *et al.* [41]. They classified 39 samples of honey into seven botanical classes. In this case, two or three samples per botanical class were used. Studies with a large sample size utilized spectrography, melissopalynology, chromatography, NMR spectroscopy, spectrometry, and sensorial techniques to analyze the geographical or botanical origin of honey. However, these techniques are time-consuming and require highly specialized personnel [42].

Another major drawback found in the aforementioned studies is the imbalanced class problem (Table 1). Almost none of the studies have worked on balancing the datasets. Therefore, the imbalanced datasets produced biased classification results toward the majority class [45]. Thus, Maione *et al.* [43] suggested to deal with the imbalanced class problem first and then report the classification results.

The present study is different from the existing studies in three aspects. First, as far as we know, this study is the first to attempt to work on Pakistani honey classification based on geographical and botanical origin. Second, this work prepared and compiled Pakistani honey dataset. This dataset contains 140 samples of honey belonging to three geographical origins and three botanical origins of Pakistan. Finally, this study employed a state-of-the-art data balancing technique, to balance our Pakistani honey dataset and produce robust classifiers. We believe that this work can bring new information to Pakistani researchers working in the field of food technology.

III. MATERIAL AND METHODS

This section provides the detailed methodology for predicting the Pakistani honey authenticity through machine learning algorithms. The detailed research methodology is

TABLE 1. Detailed literature on honey classification.

Aim of study is to classify	Property analyzed	Sample Size	Technique	Remarks
based on Italian Geographical origin from 7 botanical origins [46]	Chemical Profile	39	PCA and DA	Small number of sample and class imbalanced problem
based on the geographical origins of 4 regions from New Zealand [34]	Chemical Profile	83	DA	Small number of sample and class imbalanced problem
based on the botanical origin of the 8 regions of Turkish and Anatolian [27]	Molecular composition and transform spectrography	120	Clustering and PCA	not dealt with the imbalanced problem
based on the botanical origin of different countries of Italy, east Europe, and Spain [23]	Melissopalynology and Spectrograph	184	PCA	not dealt with the imbalanced problem
based on the botanical origin 8 regions of Spain [32]	Mineral composition from spectrometry	Not mention	Clustering and variance analysis	not dealt with the imbalanced problem
based on the botanical origin of 4 regions of China [35]	Sensorial data from PE-tongue VE-tongue	Not mention	PCA	not dealt with the imbalanced problem
based on the Geographical origin of 4 regions of Argentina [39]	Physicochemical property	141	PCA and LDA	not dealt with the imbalanced problem
based on the botanical origin of 9 regions of Finland [40]	Data obtained from nuclear magnetic resonance metabolomics	20	PCA and DA	Small number of sample and class imbalanced problem
based on the botanical origin Geographical of 22 countries [21]	Ultra-performed liquid chromatography	49	Artificial neural network	Small number of sample and class imbalanced problem
based on the botanical origin of 11 regions of Spanish and Portugal [25]	Physicochemical property	49	PCA and LDA	Small number of sample and class imbalanced problem
based on the botanical origin 6 regions of Uruguayan [41]	Physicochemical property	30	PCA, PFA, and DA	Small number of sample and class imbalanced problem
based on the botanical origin of 5 regions of China [22]	Spectrography	250	LDA and PCA	not dealt with the imbalanced problem
based on the botanical origin of 3 regions of China [30]	Electrical (Tongue nose) and physicochemical	154	soft modeling and PCA	not dealt with the imbalanced problem
based on the botanical origin of 2 regions of Andalusia [46]	Physicochemical property	29	PCA and LDA	Small number of sample and class imbalanced problem
based on the geographical origin of a region of Argentinean [26]	Color histogram and image processing	210	Success projection algorithms and LDA	not dealt with the imbalanced problem
based on and rheology Geographical origin of Ethiopian [20]	Melissopalynology	320	PCA and Analysis of variance	not dealt with the imbalanced problem
based on the botanical origin of 5 regions of Poland [28]	Physicochemical property	72	Regression Tree and Cart	Small number of sample and class imbalanced problem

TABLE 1. ***(Continued.)*** **Detailed literature on honey classification.**

based on the botanical origin of 6 regions of Poland [38]	Spectrography and Chromatography	62	PCA and KNN	Small number of sample and class imbalanced problem
based on the botanical origin of Oceania(New Zealand and Australia) [47]	Nuclear Magnetic Resonance	264	PCA	not dealt with the imbalanced problem
based on the Geographical origin of Brazilian [33]	Chemical Profile	57	Machine Learning	Small number of sample and class imbalanced problem
based on the Botanical origin of 6 regions of Saudi Arabia [19]	Spectrography	18	Hierarchical clustering and PCA	Small number of sample and class imbalanced problem
based on the Botanical origin of Chines [31]	Chromatography	75	PCA, clustering , and DA	Small number of sample and class imbalanced problem
based on the Botanical origin of Indian [24]	Phenolic and Volatile Compounds of honey	30	PCA	Small number of sample and class imbalanced problem
based on the Botanical origin of Indian [48]	Antioxidant and Macro Mineral	24	PCA and LDA	Small number of sample and class imbalanced problem
based on the Botanical origin of Spain [49]	Physicochemical property	100	PCA, LDA, and Machine Learning	Small number of sample and class imbalanced problem
based on the Geographical origin of the region of Palestine [50]	Physicochemical property	33	PCA	Small number of sample and class imbalanced problem
based on the Botanical Origin of 6 regions of Italy [29]	Spectrography	206	PCA	Not dealing with the imbalanced problem
(In this studies) based on Botanical and Geographical origin of Pakistani Honey	Physicochemical property	140	Machine Learning	**Deal with imbalanced class problem**

also depicted in Figure 1. First, we collected and prepared the Pakistani honey dataset. Then, we identified discriminative physicochemical features from the collected honey samples and handled the class imbalance issue. Finally, we constructed a prediction model and evaluated the performance of proposed model. Each step is shown in Figure 1 and further described in the subsequent subsections.

A. COLLECTION OF PAKISTANI HONEY DATA AND FEATURE EXTRACTION

For this research, we collected 140 honey samples as secondary data from well-known and credible academic articles [37], [51], [52], [53], [54], [55], [56], [57], [58], [59]. The secondary data are selected, primarily because no researcher has attempted to use machine learning approaches in recognizing adulteration in Pakistani honey. Therefore, no publicly available dataset of Pakistani honey is present. Moreover, these papers are published in well-known national and international peered-reviewed and credible journals. Finally, various laboratory experiments were performed in these studies [37], [51], [52], [53], [54], [55], [56], [57], [58], [59] to identify several physicochemical properties and their quantities from different types of Pakistani honey samples. These experiments were performed following the standards of well-known food regulatory authorities including, European Union Council [13] and Association of Official Agricultural Chemists [60].

These 140 samples were collected from either original or branded Pakistani honey. Branded honey refers to the processed laboratory honey; during the process, the producers mix certain nanomaterials, such as hydrogen pre-oxide to preserve the honey [61], [62]. On the other hand, original honey is chemical free. These samples belong to three distinct geographical origins in Pakistan namely, Sindh, Punjab, and KPK. Furthermore, each sample belongs to either of three botanical origin namely, Dorsata, Florea, and other species. We collected these samples because they are widely consumed in the Pakistan. Table 2 shows the eight discriminative and informative physicochemical features from each of

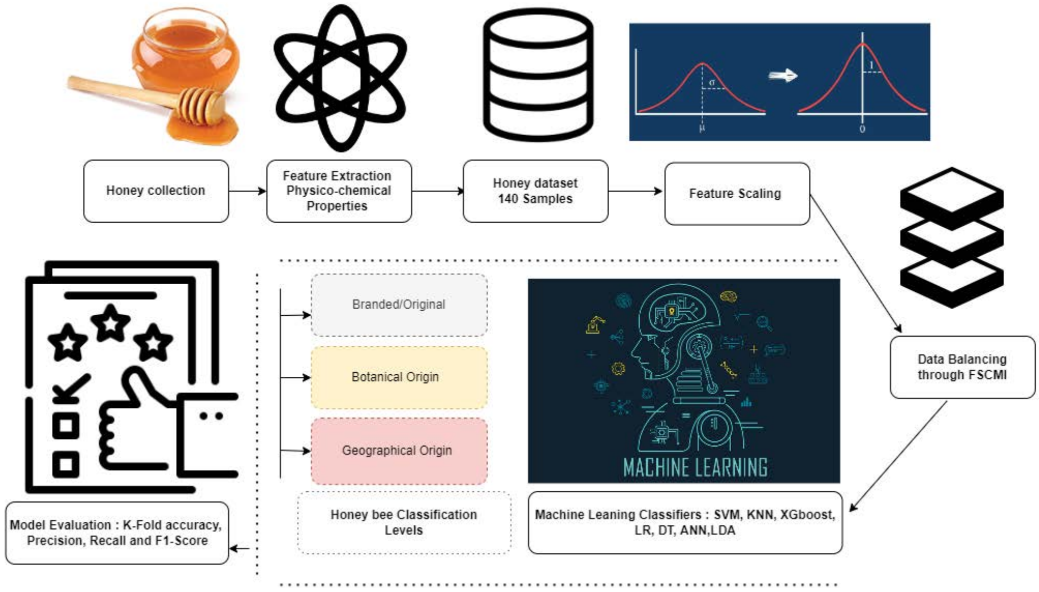

Honey collection
Feature Extraction Physico-chemical Properties
Honey dataset 140 Samples
Feature Scaling
Data Balancing through FSCMI
Machine Leaning Classifiers : SVM, KNN, XGboost, LR, DT, ANN,LDA
MACHINE LEARNING
Branded/Original
Botanical Origin
Geographical Origin
Honey bee Classification Levels
Model Evaluation : K-Fold accuracy, Precision, Recall and F1-Score

FIGURE 1. Detailed diagram for the Pakistani honey authenticity classification model.

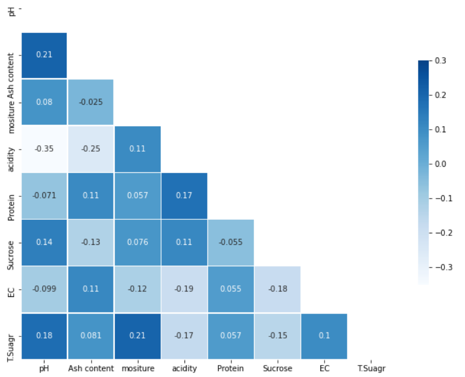


FIGURE 2. Correlation diagram for eight physicochemical features.

TABLE 2. Physicochemical properties extracted from honey.

physicochemical	Method	Units
pH Measurement	pH electrode or pH-meter [63]	per 100 g
Moisture	Digital Moisture meter	%
Ash Content	AOAC Official Method [64]	%
Free Acidity	Titrimetric method [65]	meq/kg
Protein	Spectrophotometrically [66]	per 100 g
Sugar	Spectrophotometrically [66]	per 100 g
Sucrose	Spectrophotometrically [66]	per 100 g
Electricity Conductivity	Milwaukee-301 Meter [65]	mS/cm

the collected honey sample. These features include, electrical conductivity, pH, protein, moisture, free acidity, sugar, sucrose, and ash content. The master feature set contains the eight physicochemical features, type (original vs branded), geographical origin (Punjab, Sindh, and KPK), and botanical origin (Dorsata, Florea, and others species). The correlation between all these eight features is as shown in Figure 2.

B. CONSTRUCTION OF MASTER FEATURE VECTORS FROM PAKISTANI HONEY DATASET

A master feature vector contains all the honey samples, feature values, and class labels. We constructed three master feature vectors from our collected dataset: type master feature vector (TMFV), geographical master feature vector (GMFV), and botanical master feature vector (BMFV). We constructed these three master feature vectors to address the unique classification problem in the present study. The TMFV contains 140 rows and nine columns. First eight columns represent the eight physicochemical features and the ninth column represents the class label (either original or branded). Similarly, the GMFV contains 140 rows and nine columns. Each row represents the unique honey sample, whereas first eight columns represent the eight physicochemical features and the ninth column represents the class label (either Punjab, Sindh, or KPK). Finally, the BMFV contains 140 rows and nine columns. Each row represents the unique honey sample whereas first eight columns represent the eight physicochemical features and the ninth column represents the class label (either Dorsata, Florea, or others).

C. DATA BALANCING TECHNIQUES

Class Imbalance problem is found in many classification models [67]. Class imbalance problem occur when the number of instances from one or more classes is considerably greater than that from another class. Class imbalance problem causes the biased classification results because the majority classes are overwhelmed, and the minority classes are ignored. Sometimes, they also result in poor accuracy in minority classes because machine learning algorithms work best when an approximately equal number of instances exist in each class [68]. Our collected dataset is also naturally imbalance, particularly in GMFV and BMFV, as shown in Figure 3. Therefore, we used FCSMI, the latest class balancing technique, to address the class imbalance issue. FCSMI was introduced by Mishra *et al.* [69]. The performance of FCSMI method is better than that of the prevailing state-of-the-art sampling methods [70]. The idea behind FCSMI involves applying synthetic minority oversampling technique (SMOTE) to balance the ratio between minority and majority instances.

D. CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Many classification techniques exist, but we chose the most widely used classification models for numerical data. According to the ''no free lunch'' theorem of Wolpert and Macready [71], a single machine learning algorithm that performs best in all application areas does not exist. Hence, various machine learning algorithms should be tested. Therefore, Fernández-Delgado *et al.* [72] evaluated the performance of 179 machine learning classifiers on 121 different datasets. The experimental results showed that boosting algorithm, SVM, random forest (RF), ANN, and LR perform well on most of the datasets. Thus, in this study we employed seven machine learning algorithms (LR, SVM, KNN, RF, LDA, ANN and XGBoost) to evaluate which one is suitable for predicting the authenticity of Pakistani honey. In the subsequent subsections we have provided the brief description of each of these machine learning algorithms.

1) LR

Linear regression is widely used in the regression problems of supervised machine learning. It predicts the dependent variable (y), i.e. output, based on the independent variable, (x) i.e. input. This technique finds the linear relation between x and y; hence, it is called linear regression. It is mostly concerned with minimizing errors and predicting the best possible results [73].

2) DT

The DT is used for classification and regression problems in supervised machine learning. However, it is frequently chosen for classification problems. It classifies a tree structure comprising nodes, branches, and leaf nodes. The features of datasets are represented by nodes, decision rules are represented by branches, and the outcome is represented by the leaf node. The treelike structure is easy to understand [74].

3) KNN

KNN is used in classification and regression problems. It works by calculating the k nearest and k closest training examples in the dataset. For classification problems, the output of KNN is based on the most frequent vote of the neighbors. However, for the regression problems, the output is the average of the values of KNNs. Euclidean distance is commonly used for calculating the distance between neighbors. K value is defined by the user in the classification problems for choosing a good value of k in various heuristic techniques. Noisy and irrelevant features can degrade the performance of KNN [75].

4) SVM

It is a supervised machine learning algorithm that is extensively used for linear classification problems [76]. It provides the best accuracy results for various problems in the field of text classification, image processing, and bioinformatics problems. It classifies data by constructing the hyperplanes in high- or infinite-dimensional space. Hyperplane separates the training examples by the maximal margin.

5) ANN

ANN is a form of machine learning. It has been in the research domain for decades. It works on the notion of self-learning from the training examples, which are usually labeled in advance. The basic architecture consists of connected neurons. Neurons i and j have a link with weight $w_{i,j}$. The weight decides the strength of the information propagated to unit.

6) XGBoost

XGBoost is the advanced version of DT. It is capable of gradient boosting; it bags DTs to improve the overall performance and avoid overfitting and underfitting [77].

7) LDA

LDA is like PCA; however, it focuses on maximizing the separability among classes, and it reduces dimensionality from an original number of features to K features, where K is the number of classes [78]. With the aim to classify objects into one of two or more groups based on some sets of parameters that describes objects, LDA has come up with specific functions and applications.

E. EVALUATION OF MACHINE LEARNING MODEL

The dataset was split into two parts (90%-10%) training and testing set respectively. In this step, we evaluated the performance of seven constructed machine learning models (namely, LR, DT, KNN, SVM, ANN, XGBoost, and LDA) by using the test set. Moreover, we used four performance metrics, namely, precision, recall, F1-score, and accuracy to measure the performance of the constructed classification models. In this step, we ensured that our testing accuracy is less than or at least equal to training set accuracy to avoid model overfitting and model variance. The definitions of each performance metrics are listed below:

1) **Precision** is the ratio of the correctly predicted labels for the specific class concerning all the predicted labels of the classes. It evaluates the performance of the proposed models to detect the actual class sample correctly.

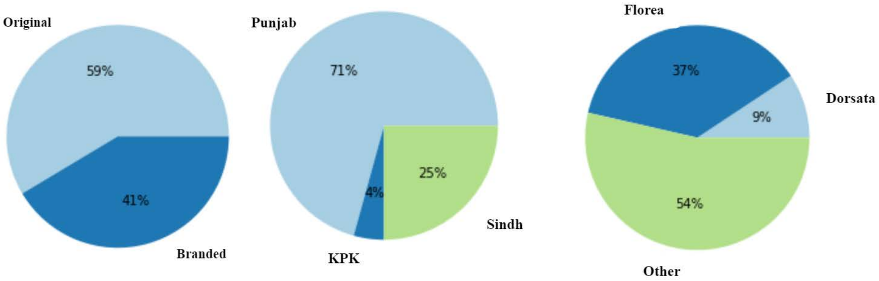


FIGURE 3. Data distribution for each level.

2) **Recall** is the ratio of all predicted labels for the specific class to the actual labels of the class. It calculates the number of accurately detected instances as positive instances.
3) **F1-Score** computes the weighted harmonic mean/ balanced ratio of recall and precision.
4) **Accuracy** computes the frequency of the accurate detection of each class of honey from the total number of instances.
5) **Confusion matrix** computes the number of false positive, true positive, false negative, and true negative for each sample class, the detection was measured with the labels.

IV. EXPERIMENTAL SETTINGS AND RESULTS

This section describes the detailed experimental settings and the results. In this work, we employed five experimental settings on three distinct master feature vectors (namely, TMFV, GMFV, BMFV). In the first setting, we fed TMFV as an input to seven machine learning algorithms (mentioned in section III-D) to classify Pakistani honey samples into either original or branded class. In the second experimental setting, we fed GMFV as an input to seven machine learning algorithms (mentioned in section III-D) to classify Pakistani honey samples into respective geographical region (either Punjab, Sindh, or KPK). In the third experimental setting, we fed the balanced GMFV as an input to seven machine learning algorithms (mentioned in section III-D) to classify Pakistani honey samples into either Punjab, Sindh, or KPK class. In the fourth experimental setting, we fed BMFV as an input to seven machine learning algorithms (mentioned in section III-D) to classify Pakistani honey samples into either Dorsata, Florea, or other origin class. In the fifth experimental setting, we fed the balanced BMFV as an input to seven machine learning algorithms (mentioned in section III-D) to classify Pakistani honey samples into either Dorsata, Florea, or other origin class. In each experimental settings, we evaluated the classification results in terms of precision, recall, F1-score, and accuracy. We also provided the confusion matrix of the machine learning classifier, which yielded the best results in each experimental setting. All the experiments were performed on Windows 10 with GPU (GeForce MX130) using Python language.

TABLE 3. Classification results of Pakistani honey according to type.

ML Algorithm	Accuracy	Precision	Recall	F1_score
LR	100.0	100.0	100.0	100.0
DT	92.3	93.3	92.8	92.8
KNN	78.5	84.4	78.5	76.6
SVM	71.4	80.9	71.4	67.1
ANN	71.4	80.0	71.4	67.1
XGBoost	100.0	100.0	100.0	100.0
LDA	71.5	80.0	71.2	67.4

A. HONEY CLASSIFICATION ACCORDING TO TYPE

This section presents the results of experimental setting-I. In this setting, we provided TMFV as an input to LR, DT, KNN, SVM, ANN, XGBoost, and LDA machine learning algorithms and evaluated its precision, recall, accuracy, and F1-score. In this setting, 90% of the samples were used as a training set to construct the classification model, whereas 10% of the samples were used as a test set to evaluate the performance of the constructed classification model.

The experimental results of the test set of each machine learning algorithm are presented in Table 3. The highest results (100% accuracy and 100% F1-score) were observed using LR and XGBoost algorithms followed by DT (92.3% accuracy and 92.8% F1-Score). The lowest results were observed in SVM, ANN, and LDA algorithms (approx. 71% accuracy and approx. 67% F1-Score) followed by KNN (78.5% accuracy and 76.5% F1-Score).

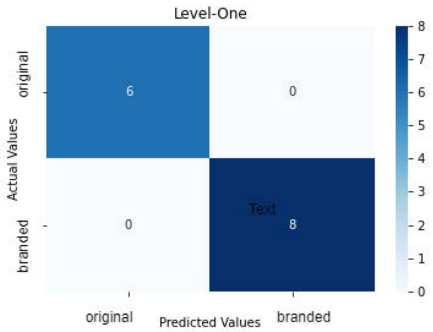


FIGURE 4. Confusion matrix of XGBoost model for setting-I.

Figure 4 also shows the confusion matrix of experiment analysis with XGBoost machine learning algorithm in setting-I. Figure 4 shows that all the instances were correctly classified into their respective classes.

B. HONEY CLASSIFICATION ACCORDING TO GEOGRAPHICAL ORIGIN

This section presents the results of experimental setting-II. In this setting, we provided GMFV as an input to LR, DT, KNN, SVM, ANN, XGBoost, and LDA machine learning algorithms and evaluated its precision, recall, accuracy, and F1-score. In this setting, 90% of the samples were used as a training set to construct the classification model, whereas 10% of the samples were used as a test set to evaluate the performance of the constructed classification model.

The experimental results of the test set of each machine learning algorithms are given in Table 4. The highest results (92.8% accuracy and 90% F1-score) were observed using XGBoost algorithm followed by DT, LR, KNN, and LDA (85.5% accuracy and 82.7% F1-Score). The lowest results were observed in ANN algorithm (approx. 71% accuracy and approx. 67% F1-score) followed by SVM (78.5% accuracy and 68.1% F1-score).

TABLE 4. Classification results of Pakistani honey according to geographical origin.

ML Algorithm	Accuracy	Precision	Recall	F1_score
LR	85.5	80.9	85.7	82.7
DT	85.7	85.7	85.7	84.3
KNN	85.7	80.9	85.7	82.8
SVM	78.5	61.9	71.4	68.1
ANN	71.4	64.0	71.4	67.1
XGBoost	92.8	88.9	92.8	90.0
LDA	85.5	80.9	85.7	82.7

We also reported the confusion matrix of the experiment analysis with XGBoost machine learning algorithm in setting-II (shown in Figure 5). Almost 92% of the instances were classified correctly into their respective classes, whereas 8% were misclassified. This misclassification is possibly caused by the class imbalance issue. Therefore, we employed FCSMI data balancing technique (described in Section III-C) in experimental setting-III to further reduce the misclassification rate. Using the FCSMI technique, we oversample the minority classes and balanced all the classes. After data balancing, the balanced dataset contains 297 honey samples. Each class has 99 samples. Moreover, 90% of the samples from this dataset were used for training set, whereas 10% were used for the test set.

The experimental results on the test set of each machine learning algorithm are presented in Table 5. The highest results (96.6% accuracy and 96.6% F1-score) were observed using XGBoost, ANN, and LR algorithms followed by SVM (86.6% accuracy and 86.6% F1-score). The lowest results were observed in the LDA algorithm (approx. 73% accuracy and approx. 73% F1-Score).

Compared with experimental setting-II, experimental setting-III further improved the classification performance by showing an increase in accuracy and F1-score by 8% to 25%.

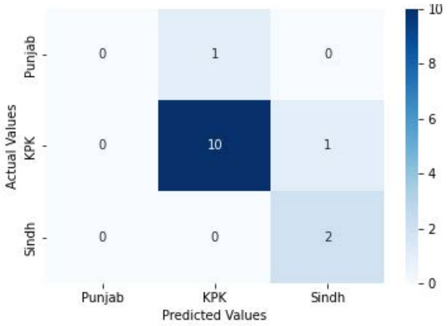


FIGURE 5. Confusion matrix of XGBoost model for setting-II.

TABLE 5. Classification results of Pakistani honey according to geographical origin after balancing the dataset.

ML Algorithm	Accuracy	Precision	Recall	F1_score
LR	96.6	96.6	96.6	96.6
DT	93.3	93.3	93.3	93.3
KNN	96.6	96.6	96.6	96.6
SVM	86.6	86.6	86.6	86.6
ANN	96.6	96.6	96.6	96.6
XGBoost	96.6	96.6	96.6	96.6
LDA	73.3	76.3	73.3	73.3

Figure 6 also shows the confusion matrix of the experiment analysis with XGBoost machine learning algorithm in setting-III. Figure 6 shows that almost all the instances were correctly classified into their respective classes except for one instance of KPK class.

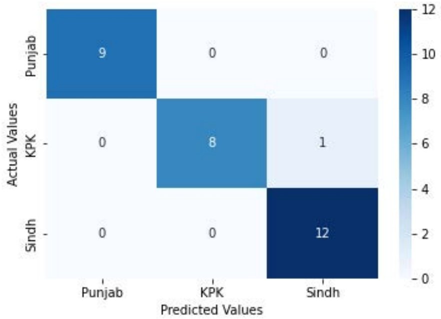


FIGURE 6. Confusion matrix of XGBoost model for setting-III.

C. HONEY CLASSIFICATION ACCORDING TO BOTANICAL ORIGIN

This section presents the results of experimental setting-IV. In this setting, we provided BMFV as an input to LR, DT, KNN, SVM, ANN, XGBoost, and LDA machine learning algorithms and evaluated its precision, recall, accuracy, and F1-score. In this setting, 90% of the samples were used as a training set to construct the classification model, whereas 10% of the samples were used as a test set to evaluate the performance of the constructed classification model.

The experimental results of the test set of each machine learning algorithms are presented in Table 6. The highest results (64.2% accuracy and 64.2% F1-Score) were observed using XGBoos, LR, DT, and ANN algorithms followed by KNN and LDA (57.1% accuracy and 57.1% F1-score). The lowest results were observed in SVM algorithm (approx. 35% accuracy and approx. 32% F1-score).

TABLE 6. Classification results of Pakistani honey according to botanical origin.

ML Algorithm	Accuracy	Precision	Recall	F1_score
LR	64.2	84.1	64.2	64.2
DT	64.2	84.1	64.2	64.2
KNN	57.1	80.5	57.1	54.4
SVM	35.7	41.6	35.7	32.9
ANN	64.2	69.8	64.2	63.6
XGBoost	64.2	84.1	64.2	64.2
LDA	57.1	78.9	57.1	58.7

We also reported the confusion matrix of experiment analysis with XGBoost machine learning algorithm in setting-IV. Figure 7 shows that almost 64% of the instances were classified correctly into their respective classes, and 36% were misclassified. This misclassification is possibly caused by the class imbalance issue. Therefore, we employed FCSMI data balancing technique (described in section III-C) in experimental setting-V to further reduce the misclassification rate. Using the FCSMI technique, we oversampled the minority classes and balanced all the classes. After data balancing, the balanced dataset contains 228 honey samples, and each class has 76 samples. Moreover, 90% of the samples from this dataset were used for training set, and 10% were used for the test set.

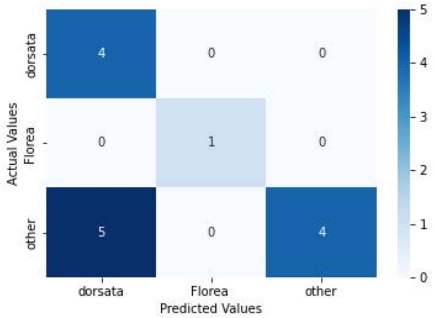


FIGURE 7. Confusion matrix of XGBoost model for setting-IV.

The experimental results on test set of each machine learning algorithms are presented in Table 7. The highest results (100% accuracy and 96.6% F1-score) were observed using XGBoost and LR algorithms followed by DT (95.6% accuracy and 95.6% F1-score). The lowest results were observed in ANN algorithm (approx. 73% accuracy and approx. 73% F1-score).

Compared with experimental setting-IV, this experimental setting V further improved the classification performance by showing an increase in accuracy and F1-score by 36% to 47%.

Figure 8 also shows the confusion matrix of the experimental analysis with XGBoost machine learning algorithm in setting-V. Figure 8 shows that almost all the instances were correctly classified into their respective classes.

TABLE 7. Classification results of Pakistani honey according to botanical origin after balancing the dataset.

ML Algorithm	Accuracy	Precision	Recall	F1_score
LR	100.0	100.0	100.0	100.0
DT	95.6	96.0	95.6	95.6
KNN	86.9	90.5	86.9	86.8
SVM	82.6	85.2	82.6	82.5
ANN	73.9	82.9	73.9	72.5
XGBoost	100.0	100.0	100.0	100.0
LDA	82.6	85.2	82.6	82.5

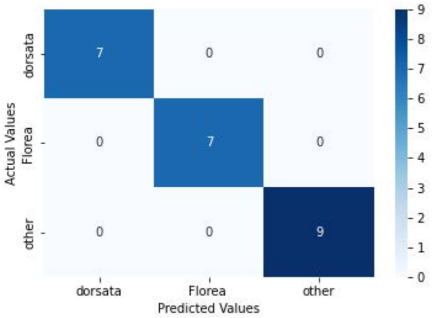


FIGURE 8. Confusion matrix of XGBoost model for setting-V.

V. DISCUSSION

This section provides the theoretical analysis of proposed three level classification models for classifying the Pakistani honey samples by using eight physicochemical features and employing seven machine learning algorithms. This section aims to analyze critically the obtained results and justify why the classification model classified the honey accurately. Moreover, the error analyses of misclassification instance during different experimental settings are also provided.

The results of experimental setting-I (section IV-A) showed that the classification of Pakistani honey samples according to type as original or branded can classify (71.4% to 100%) correctly. However, the error rate of setting-I is between 4% and 12%. This error is possibly due to the inability of the features to produce the discriminative and representative patterns for the original and branded honey sample. In our future work, we will investigate the features that best classify the honey and minimize the misclassification rate.

Similarly, the results of experimental setting-II (section IV-B) showed that the classification of Pakistani honey samples according to geographical origin as Punjab, Sindh, or KPK can classify (78.5% to 92.8%) correctly. However, the error rate of setting-II is between 7.2% and 21.5%. This misclassification is possibly caused by the class imbalance issue. Therefore, we should not rely on imbalanced dataset results. Several studies have recently proved that balancing the dataset yields promising results [79]. Most of the researchers suggested utilizing the FCSMI data balancing technique (described in section III-C) to obtain robust results [70]. After the dataset in experimental setting-III was balanced, the overall performance of honey classification further improved by showing an increase in accuracy and F1-score by 8% to 25%, compared with experimental setting-II results.

Likewise, the results of experimental setting-IV (section IV-C) showed that the classification of Pakistani honey samples according to botanical origin as Dorsata, Florea, or others can classify 35.2% to 64.2% correctly. The error rate is between 35.8% and 64.4%. This huge misclassification rate is possibly caused by the class imbalance issue. Therefore, we applied the FCSMI data balancing technique (described in section III-C) to obtain robust results, as described in result setting-V. The results obtained after applying the FCSMI data balancing technique were quite surprising. The considerable change in setting-V, is 36% to 47% in terms of accuracy and F1-score.

VI. CONCLUSION

In this study, three level classification model is proposed to classify the Pakistani honey samples. The first level classifies whether the honey is original or branded. The second level classifies the honey according to geographical origin. The final level classifies the honey according to botanical origin. The experimental results showed that our proposed work performed excellently by achieving an accuracy of 100%. Furthermore, this work developed a new Pakistani honey dataset consisting of 140 honey samples. The nature of the dataset was highly imbalanced. Thus, the latest data balancing technique named FCSMI was employed to balance the dataset. The experimental results showed good results on the balanced dataset. In the future work, more honey samples will be collected, and other features or properties (such as, mineral profiles) of honey will be extracted to enhance the robustness of our dataset. Moreover, new botanical origins and geographical origins will be added to expand our dataset further. We believe that our proposed work will serve as a baseline in recognizing adulteration of Pakistani honey. In addition, this work will be proved beneficial in reducing the adulteration of Pakistani honey.

CONFLICTS OF INTEREST:

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. Sharif, M. Iftikhar, A. Hussain, M. U. Rehman, S. F. Zaidi, M. Akram, M. Daniyal, and K. Usmanghani, ‘‘Evaluation of Physio-chemical properties of honey collected from local markets of Lahore, Pakistan,’’ *Pakistan J. Med. Biol. Sci.*, vol. 2, no. 1, pp. 1–6, 2018.

[2] O. Escuredo, I. Dobre, M. Fernandez-Gonzalez, and M. C. Seijo, ‘‘Contribution of botanical origin and sugar composition of honeys on the crystallization phenomenon,’’ *Food Chem.*, vol. 149, pp. 84–90, Apr. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24295680

[3] M. Thakur, N. Gupta, H. K. Sharma, and S. Devi, ‘‘Physicochemical characteristics and mineral status of honey from different agro-climatic zones of Himachal Pradesh, India,’’ *Brit. Food J.*, vol. 123, no. 11, pp. 3789–3804, Oct. 2021.

[4] M. Ismail, E. M. Abdallah, and E. R. Elsharkawy, ‘‘Physico-chemical properties, antioxidant, and antimicrobial activity of five varieties of honey from Saudi Arabia,’’ *Asia Pacific J. Mol. Biol. Biotechnol.*, vol. 2021, pp. 27–34, Oct. 2021.

[5] A. Abselami, A. Tahani, M. Sindic, M.-L. Fauconnier, E. Bruneau, and A. Elbachiri, ‘‘Physicochemical properties of some honeys produced from different flora of Eastern Morocco,’’ *J. Mater. Environ. Sci.*, vol. 9, no. 3, pp. 879–886, 2018.

[6] S. M. Kadri, R. Zaluski, and R. D. O. Orsi, ‘‘Nutritional and mineral contents of honey extracted by centrifugation and pressed processes,’’ *Food Chem.*, vol. 218, pp. 237–241, Mar. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27719904

[7] M. Oroian, S. Ropciuc, S. Paduret, and E. T. Sanduleac, ‘‘Authentication of Romanian honeys based on physicochemical properties, texture and chemometric,’’ *J. Food Sci. Technol.*, vol. 54, no. 13, pp. 4240–4250, Dec. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29184230

[8] E. Mendes, E. B. Proença, I. Ferreira, and M. Ferreira, ‘‘Quality evaluation of Portuguese honey,’’ *Carbohydrate Polym.*, vol. 37, no. 3, pp. 219–223, 1998.

[9] D. C. Abell, H. Friebe, C. Schweger, A. S. K. Kwok, and P. Sporns, ‘‘Comparison of processed unifloral clover and canola honey,’’ *Apidologie*, vol. 27, no. 6, pp. 451–460, 1996.

[10] M. F. Lanjwani and F. A. Channa, ‘‘Minerals content in different types of local and branded honey in Sindh, Pakistan,’’ *Heliyon*, vol. 5, no. 7, 2019, Art. no. e02042.

[11] F. A. Zulkhairi Amin, S. Sabri, S. M. Mohammad, M. Ismail, K. W. Chan, N. Ismail, M. E. Norhaizan, and N. Zawawi, ‘‘Therapeutic properties of stingless bee honey in comparison with European bee honey,’’ *Adv. Pharmacol. Sci.*, vol. 2018, pp. 1–12, Dec. 2018.

[12] E. Baglio, *Chemistry and Technology of Honey Production*. Springer, 2017.

[13] A. Thrasyvoulou, C. Tananaki, G. Goras, E. Karazafiris, M. Dimou, V. Liolios, D. Kanelis, and S. Gounari, ‘‘Legislation of honey criteria and standards,’’ *J. Apicultural Res.*, vol. 57, no. 1, pp. 88–96, Jan. 2018.

[14] C. Alimentarius, ‘‘Standard for honey (CXS 12-1981),’’ Tech. Rep., 1981.

[15] K. Khan, ‘‘Beekeeping in Pakistan (history, potential, and current status),’’ 2020.

[16] A. Blouch, R. Mahmood, K. Rafique, F. A. Shaheen, M. Munir, A. Qayyum, and R. Ali, ‘‘Comparative analysis of physicochemical properties of honey from ecological zones and branded honey of Pakistan,’’ *Journal*, vol. 4, pp. 40–49, 2016.

[17] A. J. Izenman, ‘‘Modern multivariate statistical techniques,’’ in *Regression, Classification and Manifold Learning*, vol. 10. Springer, 2008, p. 978.

[18] G. Chandrashekar and F. Sahin, ‘‘A survey on feature selection methods,’’ *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[19] M. J. Ansari, A. Al-Ghamdi, K. A. Khan, N. Adgaba, S. H. El-Ahmady, H. A. Gad, A. Roshan, S. A. Meo, and S. Kolyali, ‘‘Validation of botanical origins and geographical sources of some Saudi honeys using ultraviolet spectroscopy and chemometric analysis,’’ *Saudi J. Biol. Sci.*, vol. 25, no. 2, pp. 377–382, Feb. 2018.

[20] M. A. Domínguez, P. H. G. D. Diniz, M. S. Di Nezio, M. C. U. de Araújo, and M. C. U. de Araújo, ‘‘Geographical origin classification of Argentinean honeys using a digital image-based flow-batch system,’’ *Microchem. J.*, vol. 112, pp. 104–108, Jan. 2014.

[21] G. Di Bella, V. Lo Turco, A. G. Potortì, G. D. Bua, M. R. Fede, and G. Dugo, ‘‘Geographical discrimination of Italian honey by multi-element analysis with a chemometric approach,’’ *J. Food Composition Anal.*, vol. 44, pp. 25–35, Dec. 2015.

[22] E. Corbella and D. Cozzolino, ‘‘Classification of the floral origin of uruguayan honeys by chemical and physical characteristics combined with chemometrics,’’ *LWT-Food Sci. Technol.*, vol. 39, no. 5, pp. 534–539, Jun. 2006.

[23] F. Corvucci, L. Nobili, D. Melucci, and F.-V. Grillenzoni, ‘‘The discrimination of honey origin using melissopalynology and Raman spectroscopy techniques coupled with multivariate analysis,’’ *Food Chem.*, vol. 169, pp. 297–304, Feb. 2015.

[24] 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.

[25] Z. Jandrić, R. D. Frew, L. N. Fernandez-Cedi, and A. Cannavan, ‘‘An investigative study on discrimination of honey of various floral and geographical origins using UPLC-QToF MS and multivariate data analysis,’’ *Food Control*, vol. 72, pp. 189–197, Feb. 2017.

[26] 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.

[27] S. Gok, M. Severcan, E. Goormaghtigh, I. Kandemir, and F. Severcan, ‘‘Differentiation of Anatolian honey samples from different botanical origins by ATR-FTIR spectroscopy using multivariate analysis,’’ *Food Chem.*, vol. 170, pp. 234–240, Mar. 2015.

[28] S. Popek, M. Halagarda, and K. Kursa, ‘‘A new model to identify botanical origin of Polish honeys based on the physicochemical parameters and chemometric analysis,’’ *LWT*, vol. 77, pp. 482–487, Apr. 2017.

[29] E. Schuhfried, J. Sánchez del Pulgar, M. Bobba, R. Piro, L. Cappellin, T. D. Märk, and F. Biasioli, ‘‘Classification of 7 monofloral honey varieties by PTR-ToF-MS direct headspace analysis and chemometrics,’’ *Talanta*, vol. 147, pp. 213–219, Jan. 2016.

[30] S. Serrano, M. Villarejo, R. Espejo, and M. Jodral, ‘‘Chemical and physical parameters of andalusian honey: Classification of citrus and eucalyptus honeys by discriminant analysis,’’ *Food Chem.*, vol. 87, no. 4, pp. 619–625, Oct. 2004. [Online]. Available: https://www.infona.pl/resource/bwmeta1.element.elsevier-269d82aa-b79d-3b2f-b84e-a8bc92b3b797/tab/summary

[31] Z. Sun, L. Zhao, N. Cheng, X. Xue, L. Wu, J. Zheng, and W. Cao, ‘‘Identification of botanical origin of Chinese unifloral honeys by free amino acid profiles and chemometric methods,’’ *J. Pharmaceutical Anal.*, vol. 7, no. 5, pp. 317–323, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29404055 and [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5790708/

[32] C. de Alda-Garcilope, A. Gallego-Picó, J. C. Bravo-Yagüe, R. M. Garcinuño-Martínez, and P. Fernández-Hernando, ‘‘Characterization of Spanish honeys with protected designation of origin ‘Miel de Granada’ according to their mineral content,’’ *Food Chem.*, vol. 135, no. 3, pp. 1785–1788, 2012.

[33] B. L. Batista, L. R. S. da Silva, B. A. Rocha, J. L. Rodrigues, A. A. Berretta-Silva, T. O. Bonates, V. S. D. Gomes, R. M. Barbosa, and F. Barbosa, ‘‘Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques,’’ *Food Res. Int.*, vol. 49, no. 1, pp. 209–215, Nov. 2012.

[34] Z. Wei and J. Wang, ‘‘Tracing floral and geographical origins of honeys by potentiometric and voltammetric electronic tongue,’’ *Comput. Electron. Agricult.*, vol. 108, pp. 112–122, Oct. 2014.

[35] Z. Wei and J. Wang, ‘‘Tracing floral and geographical origins of honeys by potentiometric and voltammetric electronic tongue,’’ *Comput. Electron. Agricult.*, vol. 108, pp. 112–122, Oct. 2014.

[36] E. Anklam, ‘‘A review of the analytical methods to determine the geographical and botanical origin of honey,’’ *Food Chem.*, vol. 63, no. 4, pp. 549–562, 1998.

[37] H. Fahim, J. I. Dasti, I. Ali, S. Ahmed, and M. Nadeem, ‘‘Physico-chemical analysis and antimicrobial potential of A pis dorsata, A pis mellifera and Z iziphus jujube honey samples from Pakistan,’’ *Asian Pacific J. Tropical Biomed.*, vol. 4, no. 8, pp. 633–641, 2014. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84922638414&doi=10.12980%2fAPJTB.4.2014APJTB-2014-0095&partnerID=40&md5=8c5cbe772d757fa5eea25d97b69388ae

[38] P. M. Kuś and S. van Ruth, ‘‘Discrimination of Polish unifloral honeys using overall PTR-MS and HPLC fingerprints combined with chemometrics,’’ *LWT-Food Sci. Technol.*, vol. 62, no. 1, pp. 69–75, Jun. 2015.

[39] D. C. Fechner, A. L. Moresi, J. D. Ruiz Díaz, R. G. Pellerano, and F. A. Vazquez, ‘‘Multivariate classification of honeys from corrientes (Argentina) according to geographical origin based on physicochemical properties,’’ *Food Bioscience*, vol. 15, pp. 49–54, Sep. 2016.

[40] M. Kortesniemi, C. M. Slupsky, T. Ollikka, L. Kauko, A. R. Spevacek, O. Sjövall, B. Yang, and H. Kallio, ‘‘NMR profiling clarifies the characterization of Finnish honeys of different botanical origins,’’ *Food Res. Int.*, vol. 86, pp. 83–92, Aug. 2016.

[41] O. Anjos, C. Iglesias, F. Peres, J. Martínez, Á. García, and J. Taboada, ‘‘Neural networks applied to discriminate botanical origin of honeys,’’ *Food Chem.*, vol. 175, pp. 128–136, May 2015.

[42] D. Milojković-Opsenica, D. Lušić, and L. Tešić, ‘‘Modern analytical techniques in the assessment of the authenticity of Serbian honey,’’ *Arhiv za Higijenu Rada i Toksikologiju*, vol. 66, no. 4, pp. 233–241, 2015.

[43] C. Maione, F. Barbosa, and R. M. Barbosa, ‘‘Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review,’’ *Comput. Electron. Agricult.*, vol. 157, pp. 436–446, Feb. 2019.

[44] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, ‘‘Sample size selection in optimization methods for machine learning,’’ *Math. Program.*, vol. 134, no. 1, pp. 127–155, 2012.

[45] J. Burez and D. Van den Poel, ‘‘Handling class imbalance in customer churn prediction,’’ *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009.

[46] L. Lin, F. Lei, D.-W. Sun, Y. Dong, B. Yang, and M. Zhao, ‘‘Thermal inactivation kinetics of rabdosia serra (Maxim.) hara leaf peroxidase and polyphenol oxidase and comparative evaluation of drying methods on leaf phenolic profile and bioactivities,’’ *Food Chem.*, vol. 134, no. 4, pp. 2021–2029, Oct. 2012.

[47] M. Spiteri, K. M. Rogers, E. Jamin, F. Thomas, S. Guyader, M. Lees, and D. N. Rutledge, ‘‘Combination of 1H NMR and chemometrics to discriminate manuka honey from other floral honey types from oceania,’’ *Food Chem.*, vol. 217, pp. 766–772, Feb. 2017.

[48] X. Song, H. Hu, and B. Zhang, ‘‘Drying characteristics of Chinese Yam *(Dioscorea opposita Thunb.)* by far-infrared radiation and heat pump,’’ *J. Saudi Soc. Agricult. Sci.*, vol. 17, no. 3, pp. 290–296, 2018.

[49] F. Mateo, A. Tarazona, and E. M. Mateo, ‘‘Comparative study of several machine learning algorithms for classification of unifloral honeys,’’ *Foods*, vol. 10, no. 7, p. 1543, Jul. 2021.

[50] H. Imtara, Y. Elamine, and B. Lyoussi, ‘‘Physicochemical characterization and antioxidant activity of palestinian honey samples,’’ *Food Sci. Nutrition*, vol. 6, no. 8, pp. 2056–2065, Nov. 2018.

[51] H. Ashkani, K. Badinij, A. Bulfati, U. Chutani, T. Dareshak, and D. Darzada, ‘‘Assessment of physico-chemical and antimicrobial of honey of Apis dorsata from different locations of Pakistan,’’ *Global Sci. Res. J.*, vol. 2, no. 6, pp. 186–191, 2014.

[52] M. Gulfraz, F. Iftikhar, S. Raja, S. Asif, S. Mehmood, Z. Anwar, and G. Kaukob, ‘‘Quality assessment and antimicrobial activity of various honey types of Pakistan,’’ *Afr. J. Biotechnol.*, vol. 9, no. 41, pp. 6902–6906, 2010. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-78049458893&partnerID=40&md5=6417bcb70c70405ada28eb4ad71a8951

[53] M. Gulfraz, F. Iftikhar, M. Imran, A. Zeenat, S. Asif, and I. Shah, ‘‘Compositional analysis and antimicrobial activity of various honey types of Pakistan,’’ *Int. J. Food Sci. Technol.*, vol. 46, no. 2, pp. 263–267, Feb. 2011.

[54] M. Farooque Lanjwani and F. Ahmed channa, ‘‘Investigation of antioxidant activity and physicochemical properties of local and branded honeys in Sindh, Pakistan,’’ *Sustain. Chem. Eng.*, pp. 43–50, Apr. 2020.

[55] M. F. Lanjwani and F. A. Channa, ‘‘Minerals content in different types of local and branded honey in sindh, Pakistan,’’ *Heliyon*, vol. 5, no. 7, Jul. 2019, Art. no. e02042. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069568312&doi=10.1016%2fj.heliyon.2019.e02042&partnerID=40&md5=4fd537b1e063c1eede2015f864d302b7

[56] M. Sajid, M. Yamin, F. Asad, S. Yaqub, S. Ahmad, M. A. M. S. Mubarik, B. Ahmad, W. Ahmad, and S. Qamer, ‘‘Comparative study of physiochemical analysis of fresh and branded honeys from Pakistan,’’ *Saudi J. Biol. Sci.*, vol. 27, no. 1, pp. 173–176, Jan. 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/31889832

[57] K. A. Khan, A. A. Al-Ghamdi, and M. J. Ansari, ‘‘The characterization of blossom honeys from two provinces of Pakistan,’’ *Italian J. Food Sci.*, vol. 28, no. 4, pp. 625–638, 2016. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994525360&partnerID=40&md5=5aa323ea89a35ab9931a8f4c0f93202a

[58] R. Kousar, ‘‘Physciochemical variations in the honey produced by Apis dorsata from Punjab, Pakistan,’’ *Pure Appl. Biol.*, vol. 6, no. 2, pp. 733–739, Jun. 2017.

[59] M. Nasiruddin Khan, M. Qaiser, S. M. Raza, and M. Rehman, ‘‘Physicochemical properties and pollen spectrum of imported and local samples of blossom honey from the Pakistani market,’’ *Int. J. Food Sci. Technol.*, vol. 41, no. 7, pp. 775–781, Aug. 2006.

[60] W. Horowitz and G. W. Latimer, *Official Methods of Analysis of AOAC International*, vol. 18. Gaithersburg, MD, USA: AOAC International, 2006.

[61] M. Shafiq, S. Anjum, C. Hano, I. Anjum, and B. H. Abbasi, ‘‘An overview of the applications of nanomaterials and nanodevices in the food industry,’’ *Foods*, vol. 9, no. 2, p. 148, Feb. 2020.

[62] D. A. Marrez, A. Shaker, M. A. Ali, and H. M. Fathy, ‘‘Food preservation: Comprehensive overview of techniques, applications and hazards,’’ *Egyptian J. Chem.*, 2022.

[63] C.-W. Pan, J.-C. Chou, T.-P. Sun, and S.-K. Hsiung, "Development of the tin oxide pH electrode by the sputtering method," *Sens. Actuators B, Chem.*, vol. 108, nos. 1–2, pp. 863–869, Jul. 2005.

[64] A. Pascual-Mate, S. M. Oses, M. A. Fernandez-Muino, and M. T. Sancho, "Methods of analysis of honey," *J. Apicultural Res.*, vol. 57, no. 1, pp. 38–74, 2018.

[65] A. Pascual-Mate, S. M. Oses, M. A. Fernandez-Muino, and M. T. Sancho, "Methods of analysis of honey," *J. Apicultural Res.*, vol. 57, no. 1, pp. 38–74, 2018.

[66] M. M. Bradford, "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding," *Anal. Biochem.*, vol. 72, nos. 1–2, pp. 248–254, May 1976.

[67] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1388–1400, Oct. 2010.

[68] T. O. Ayodele, "Types of machine learning algorithms," in *New Advances in Machine Learning*, vol. 3. Portsmouth, U.K.: Univ. Portsmouth, 2010, pp. 19–48.

[69] N. K. Mishra and P. K. Singh, "Feature construction and smote-based imbalance handling for multi-label learning," *Inf. Sci.*, vol. 563, pp. 342–357, Jul. 2021.

[70] R. Hou, Z. Chen, J. Chen, S. He, and Z. Zhou, "Imbalanced fault identification via embedding-augmented Gaussian prototype network with meta-learning perspective," *Meas. Sci. Technol.*, vol. 33, no. 5, 2022, Art. no. 055102.

[71] D. H. Wolper and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[72] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, Jan. 2014.

[73] D. Muchlinski, D. Siroky, J. He, and M. Kocher, "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data," *Political Anal.*, vol. 24, no. 1, pp. 87–103, 2016.

[74] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Adv. Space Res.*, vol. 41, no. 12, pp. 1955–1959, 2008.

[75] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[76] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

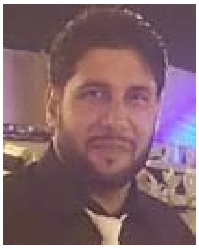
[77] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[78] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. Springer, 2013, pp. 237–280.

[79] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.

NOUREEN FATIMA received the master's degree in computer science from Sukkur IBA University, Sukkur, Pakistan, in 2021. She is currently a Research Assistant with the Center of Excellence for Robotics, Artificial Intelligence, and Blockchain, Department of Computer Science, Sukkur IBA University. She served as a Reviewer for IEEE Access. She is the author of several articles published in international journals. Her research interests include applied research in the field of artificial intelligence and its application to signal processing, and natural language processing.

GHULAM MUJTABA received the master's degree (Hons.) in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has received the gold medal for his master's degree. He has been an Associate Professor with Sukkur IBA University, Sukkur, Pakistan, since 2006. Prior to joining Sukkur IBA University, he was with a well-known soft ware house in Karachi for four years. He has vast experience in teaching and research. He has also published several articles in academic journals indexed in well-reputed databases, such as ISI and Scopus. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval.



ADNAN AKHUNZADA (Senior Member, IEEE) is currently working as an Associate Professor with the Faculty of Computing and Informatics, University Malaysia Sabah, Malaysia. His experience as an Educator and a Researcher is diverse that includes as an Assistant Professor with COMSATS University Islamabad (CUI), a Senior Researcher at RISE SICs Vasteras AB, Sweden, a Research Fellow and the Scientific Lead at DTU Compute, The Technical University of Denmark (DTU), the Course Director of Ethical Hacking at The Knowledge Hub Universities (TKH), Coventry University, U.K., a Visiting Professor having mentorship of graduate students, and a Supervision of academic and research and development projects both at UG and PG levels. He has also been involved in international accreditation, such as Accreditation Board for Engineering and Technology (ABET), and curriculum development according to the guidelines of ACM/IEEE. He is a PI of national, and a co-PI of several Swedish and Horizon 2020 EU funded projects. He has a proven track record of high impact published research and commercial products. His research interests include cyber security, secure future internet, artificial intelligence, such as machine learning, deep learning, and reinforcement learning, large scale distributed systems, such as edge, fog, cloud, SDNs, the IoT, industry 4.0, and the internet of everything (IoE). He is also a member of technical program committee of varied reputable conferences, journals, and editorial boards. He is also a Professional Member of ACM with extensive 13 years of research and development (R&D) experience both in ICT industry and academia.

KHURSHED ALI received the B.E. degree in software engineering from the Mehran University of Engineering and Technology (MUET), Pakistan, in 2006, the M.E. degree in software engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2013, and the joint Ph.D. degree in computer science from the National Tsing Hua University, Hsinchu, in collaboration with Academia Sinica, Taipei, Taiwan. He is currently working as an Assistant Professor with Sukkur IBA University. His research interests include reinforcement learning, social networks, data mining, and machine learning.

ZAHID HUSSAIN SHAIKH received the M.Sc. degree in mathematics from Shah Abdul Latif University Khairpur, in 2005, the M.Phil. degree in econometrics and statistics from the Pakistan Institute of Development Economics, Islamabad, in 2014, and the Ph.D. degree in economics from the City University of Hong Kong, in 2019. He has been working as an Assistant Professor with the Department of Mathematics and Social Sciences, Sukkur-IBA University Sukkur, since 2015. He has research and teaching experience of diversified fields. His research interests include applied econometrics and statistics, mathematical economics, applied mathematics, and optimization.

BABY RABIA (Senior Member, IEEE) is currently working as an Associate Professor with the Faculty of Computing and Informatics, University Malaysia Sabah, Malaysia. Her experience as an Educator and a Researcher is diverse that includes as an Assistant Professor with COMSATS University Islamabad (CUI), a Senior Researcher at RISE SICs Vasteras AB, Sweden, a Research Fellow and the Scientific Lead at DTU Compute, The Technical University of Denmark (DTU), the Course Director of Ethical Hacking at The Knowledge Hub Universities (TKH), Coventry University, U.K., a Visiting Professor having mentorship of graduate students, and a Supervision of academic and research and development projects both at UG and PG levels. She has also been involved in international accreditation, such as Accreditation Board for Engineering and Technology (ABET), and curriculum development according to the guidelines of ACM/IEEE. She is a PI of national, and a co-PI of several Swedish and Horizon 2020 EU funded projects. She has a proven track record of high impact published research and commercial products. Her research interests include cyber security, secure future internet, artificial intelligence, such as machine learning, deep learning, and reinforcement learning, large scale distributed systems, such as edge, fog, cloud, SDNs, the IoT, industry 4.0, and the internet of everything (IoE). She is also a member of technical program committee of varied reputable conferences, journals, and editorial boards. She is also a Professional Member of ACM with extensive 13 years of research and development (R&D) experience both in ICT industry and academia.

JAVED AHMED SHAHANI received the M.S. degree in computer science from the National University of Emerging Science, Karachi, in 2007, and the dual Ph.D. degree, in 2017, on an Erasmus Mundus Scholarship from the following University, such as the University of Bologna, the University of Turin, the Autonoma de Barcelona Universitat, and the University of Luxembourg. He completed a postdoctoral research with the Norwegian University of Science and Technology, in 2020. He was awarded an ERCIM Fellowship for conducting postdoctoral research. Prior to this, he was awarded a EURECA Scholarship from European Commission to conduct short-term research at the University of Paderborn, Germany, in 2009. He has vast teaching and research experience. He has been associated with Sukkur IBA University, since 2002. He also taught at IBA Karachi as a Visiting Faculty, from 2006 to 2009. He has more than 30 research publications. His research interests include blockchain, GDPR, data protection and privacy, and online social networks. He also has the status of being a HEC Approved Supervisor.