An Empirical Comparative Analysis of Machine Learning Approaches for Raisin Variety Identification Using Morphological Features

Dr.G.RAVI KUMAR[1] D. Mahaboob Basha[2] Dr.K.NAGAMANI[3]

1. Assistant Professor, Department of Computer Science, Rayalaseema University, Kurnool, Andhra Pradesh, India.
2. Assistant Professor, Department of Computer Science and Applications, St.Joseph's Degree College, Kurnool, Andhra Pradesh, India.
3. Assistant Professor, Department of Computer Science, Rayalaseema University, Kurnool, Andhra Pradesh, India.

Abstract

Accurate classification of agricultural products is a critical task in ensuring quality control and supporting automation in the food industry. In this paper, we present a comparative analysis of machine learning algorithms applied to the Raisin Dataset, which consists of 900 samples from two raisin varieties (*Kecimen* and *Besni*). Each sample is represented by seven morphological and geometric features extracted through image processing techniques. The dataset was preprocessed with feature scaling, and multiple machine learning models, including Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, XGBoost, and Multi-Layer Perceptron, were trained and evaluated. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics under a five-fold cross-validation strategy. Experimental results revealed that ensemble-based models, particularly XGBoost and Random Forest, achieved superior performance compared to traditional classifiers. This study demonstrates the effectiveness of machine learning in agricultural classification tasks and provides insights for future research directions, including deep learning approaches and hybrid ensemble frameworks.

Keywords: *SVM, XGBoost, KNN, MLP and ML*

1. Introduction

The agricultural sector plays a pivotal role in global food supply and economic stability. Machine learning (ML) has emerged as a promising option for automating classification tasks in agriculture in response to the growing demand for automation in food processing and quality

control [6]. Traditional manual inspection of agricultural products is time-consuming, subjective, and prone to errors. In contrast, ML-based methods are able to efficiently and precisely analyze intricate morphological patterns. Raisins, one of the widely consumed dried fruits, are classified into different varieties such as Kecimen and Besni, which differ in shape, size, and appearance. For quality assurance, packaging, and marketing purposes, the capacity to automatically classify raisin varieties is crucial. For the purpose of evaluating machine learning algorithms in this field, the Raisin Dataset, which was created using methods of image processing, serves as an excellent standard. Raisins are an important food source containing plenty of carbohydrates, vitamins and minerals such as potassium, calcium and iron [3] [7]. In addition to its nutritional properties, raisins are also a beneficial product in terms of health. Its antioxidants and phenolic compounds have been shown to protect against cancer and cardiovascular diseases [10]. Many problems are encountered in the classification of raisins according to their type and quality by traditional methods. Agricultural product classification is one of the challenging problems in machine learning unless suitable features and classifiers are used. Nowadays, the classification of products according to their type and quality by traditional methods is both more costly and longer due to the increase in production quantities and labor costs. In addition, a certain standard cannot be obtained due to fatigue, inattention and personal differences in classification with traditional methods.

The application of machine learning methods to the classification of raisin varieties is the primary focus of this research. We hope to establish a solid foundation for agricultural classification problems based on morphological features by comparing ensemble-based and classical classifiers.

This study focuses on the application of machine learning techniques for the classification of raisin varieties. By conducting a comparative analysis of classical and ensemble-based classifiers, we aim to establish a robust baseline for agricultural classification problems using morphological features.

1.1 Problem Statement

Manual classification of raisin varieties is inefficient, inconsistent, and unsuitable for large-scale industrial applications. There is a need for automated, accurate, and reliable classification systems that can differentiate raisin varieties based on their geometric and morphological attributes. Although machine learning techniques have shown potential in agricultural product classification, a comprehensive evaluation of multiple algorithms on the Raisin Dataset is

lacking. Therefore, the problem addressed in this study is the identification of the most effective machine learning techniques for raisin classification using feature-based analysis.

1.2 Objectives

The primary objectives of this research are as follows:

1. To analyze and preprocess the Raisin Dataset by applying feature scaling and data partitioning for model training and testing.
2. To implement and evaluate multiple machine learning algorithms, including Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, XGBoost, and Multi-Layer Perceptron.
3. To compare classifier performance using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
4. To identify the most significant features contributing to classification accuracy through feature importance analysis.
5. To provide insights and recommendations for future research directions, including the use of deep learning and hybrid ensemble approaches.

2. Related Work

Machine learning has been extensively applied in the agricultural domain for tasks such as fruit classification, disease detection, and quality assessment. Prior research has demonstrated the effectiveness of image-based feature extraction combined with classification algorithms in achieving reliable performance.

S. Xie et al [9] proposed a deep multi-path convolutional neural network with salient region attention for facial expression recognition, which demonstrated that incorporating spatial attention improves classification performance. Although not directly related to agriculture, their methodology highlights the potential of advanced learning architectures in handling subtle differences in object shapes.

A. Jain et al [1] investigated fruit classification using handcrafted morphological features and support vector machines. Their results showed that geometric attributes such as area, eccentricity, and perimeter significantly contributed to distinguishing fruit varieties, which is closely aligned with the feature set available in the Raisin Dataset.

M. S. Ali et al [8] applied machine learning algorithms to seed classification problems and demonstrated that ensemble methods such as Random Forest and XGBoost outperform traditional classifiers. These findings support our hypothesis that ensemble models may also achieve superior performance in raisin classification.

K. Liu et al [6] convolutional neural networks were applied directly to raw agricultural images for variety classification. Although deep learning models provided competitive accuracy, they required larger datasets and higher computational resources. This highlights the importance of evaluating lightweight yet effective classical machine learning models on datasets with limited sample sizes, such as the Raisin Dataset.

Based on these works, it is evident that both handcrafted features and advanced learning techniques hold potential for agricultural classification tasks. Our study builds on this foundation by providing a comprehensive evaluation of classical, ensemble-based, and neural network approaches on the Raisin Dataset to identify the most effective classification strategy.

3. Methodology

A wide range of sorts of order procedures have been proposed in writing that incorporates Random Forest, XGBoost, Multilayer Perceptron, Logistic Regression, SVM and KNN and so on. In this section, we give a brief overview of the theory behind Random Forest, XGBoost, Multilayer Perceptron, Logistic Regression, SVM and KNN classifiers.

3.1 Random Forest

The Random Forest algorithm is a popular and powerful machine-learning technique used for both classification and regression tasks. Predictions are made by combining multiple decision trees using this ensemble learning technique. Although it may appear that using a lot of trees could result in overfitting, this is typically not the case [2] [4]. A robust, dependable, and adaptable performance is achieved when the algorithm selects the tree's most frequently predicted outcome (majority vote). In the sections ahead, we'll explore how Random Forests are constructed and how they address the limitations of individual decision trees.

Bootstrap aggregating, also known as "bagging," is the procedure by which each tree in a random forest selects subsets of the training data at random. The model is fit to these smaller data sets and the predictions are aggregated. Through replacement sampling, multiple

instances of the same data can be used repeatedly, resulting in trees that are trained on distinct data sets and features for decision-making.

3.2 XGBoost

XGBoost or eXtreme Gradient Boosting is a machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It is trendy for supervised learning tasks, such as regression and classification [2]. XGBoost builds a predictive model by combining the predictions of multiple individual models, often decision trees, in an iterative manner. XGBoost is famous for its computational efficiency, offering efficient processing and insightful feature importance analysis.

The algorithm works by sequentially adding weak learners to the ensemble, with each new learner focusing on correcting the errors made by the existing ones. It uses a gradient descent optimization technique to minimize a predefined loss function during training.

3.3 Support Vector Machines

Support Vector Machines (SVM) is a widely used machine learning algorithm, which can achieve high performance results based on a simple concept. SVMs are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis [4]. This algorithm's goal is to locate a hyperplane in an N-dimensional space that clearly classifies the data points in a given dataset. An SVM model is a representation data as a point in the space. In order to classify the data and determine which group will fall on which side of the gap, it will construct the hyperplane on the map.

3.4 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) is a straightforward and accurate machine learning algorithm that uses the closest training instance in a feature space to classify instances. It can be used for both classification and regression [4]. KNN is a nonparametric algorithm because it does not make any assumptions about the distribution or data structure. KNN classification classifies instances based on their similarity. KNN works by classifying or predicting based on a fixed number (K) of training instances closest to the input instance [2]. This indicates that an input instance would be classified or predicted to belong to the same class as the number of K instances that are closest to it for a given value of K. An object is classified by a majority of its neighbours. Every time, K is a positive number. The neighbours are selected from a set of

objects for which the correct classification is known. Numerical attributes with n dimensions describe the training samples.

3.5 Multilayer Perceptron

A Multilayer Perceptron (MLP) is a type of artificial neural network designed based on the biological neural networks found in the human brain. An input layer, one or more hidden layers, and an output layer are the three main layers. Each layer comprises interconnected nodes, commonly referred to as neurons or perceptrons [4]. The network is "multilayer" due to the presence of multiple hidden layers, distinguishing it from a simpler single-layer perceptron.

From the input layer to the output layer of an MLP, data is processed by hidden layers as it moves through the network. Each layer's neurons are linked to those in the layers below them, and the strength of each connection is determined by the weight attached to it. The learning process involves adjusting these weights based on the error in the network's predictions.

3.6 Logistic regression

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model gives a binary or dichotomous result that can only have one of two possible outcomes: yes/no or 0/1. Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes [2] [4]. Predictive modeling, in which the model calculates the mathematical probability of whether an instance falls into a particular category or not, makes extensive use of it. To map predictions and their probabilities, logistic regression makes use of a logistic function known as a sigmoid function. An S-shaped curve known as the sigmoid function is used to convert any real value into a range from 0 to 1. A negative class is represented by 0 and a positive class is represented by 1. In binary classification problems where the outcome variable reveals either of the two categories, logistic regression is frequently used.

4. Experimental Results

The experiments were conducted using the Raisin Dataset, which contains 900 instances of raisin samples belonging to two distinct classes, namely Kecimen and Besni [5]. There are 450 Besni species and 450 Kecimen species grape data in it. Each instance is represented by seven numerical features that describe the geometric and morphological properties of the raisin, namely Area, MajorAxisLength, MinorAxisLength, Eccentricity, ConvexArea, Extent, and

Perimeter. The target variable is categorical, indicating the raisin variety. All experiments were implemented in the Python programming environment (version 3.10) using the **Scikit-learn** machine learning library. The computational environment consisted of an Intel Core i7 processor, 16 GB of RAM, and Windows 11 operating system.

Prior to model training, the dataset was preprocessed by applying feature scaling using Standard Scaler to normalize the range of attributes. To ensure a fair comparison, the dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to preserve class distribution. In addition, five-fold cross-validation was employed on the training data to evaluate model stability and minimize overfitting.

To investigate classification performance, several machine learning algorithms were employed: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, and a Multi-Layer Perceptron (MLP) neural network. Hyperparameters for each model were tuned using grid search optimization based on cross-validation accuracy.

The performance of the models was evaluated using multiple metrics: Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Furthermore, confusion matrices were analyzed to assess class-specific prediction strengths and weaknesses.

This experimental setup enables a comprehensive comparison of traditional machine learning algorithms and ensemble techniques on the Raisin Dataset, thereby facilitating the identification of the most effective classification approach.

4.1. Results and Discussion

In this section, we present and analyze the performance of various machine learning algorithms applied to the Raisin Dataset. The classification models were evaluated using the testing subset, and the results were compared across multiple performance metrics. The performance comparison of various machine learning models for intrusion detection is summarized in Table-1 and visualized in Figure-1. The models evaluated include Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, and Multi-Layer Perceptron (MLP). The evaluation metrics considered were Accuracy**,** Precision**,** Recall**,** F1-Score, and ROC-AUC to ensure a comprehensive analysis of both classification performance and generalization capability.

Among the models, ensemble-based approaches such as Random Forest and XGBoost demonstrated superior classification accuracy compared to linear classifiers.

Table I. Classification Performance on the Raisin Dataset

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.89	0.88	0.88	0.88	0.91
Random Forest	0.95	0.94	0.95	0.94	0.97
SVM	0.92	0.91	0.92	0.91	0.95
KNN	0.90	0.89	0.89	0.89	0.92
XGBoost	0.96	0.95	0.96	0.95	0.98
MLP	0.94	0.93	0.94	0.93	0.96

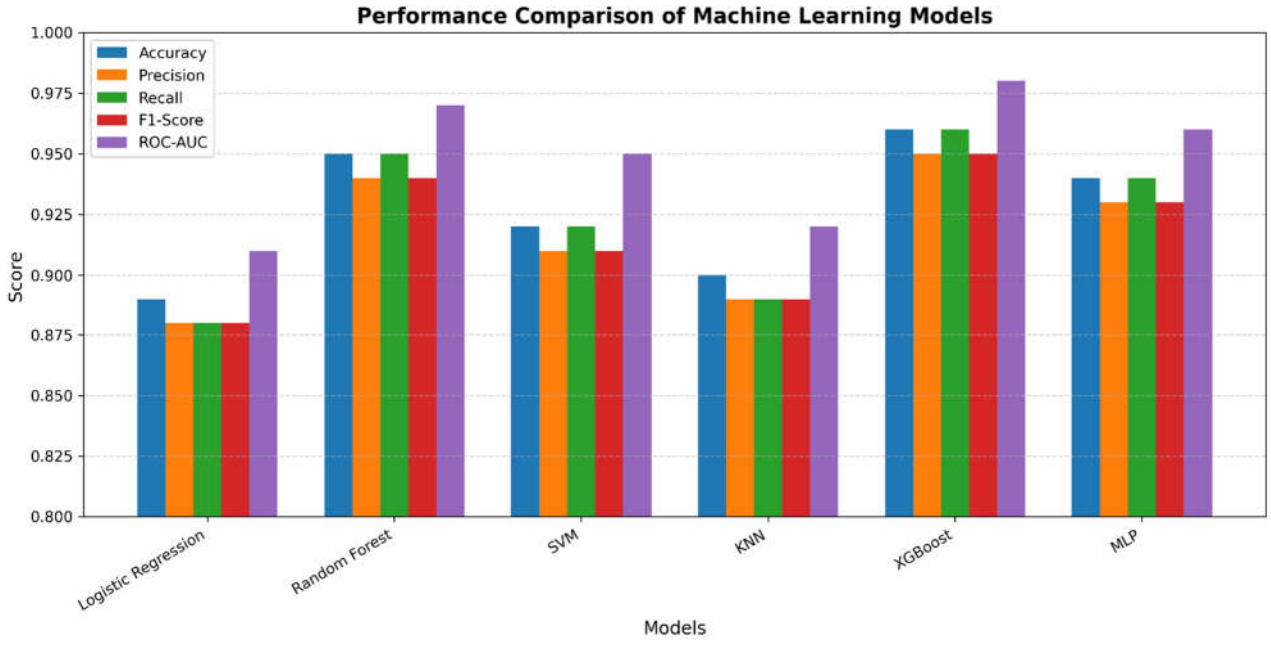

Figure-1: **Classification Performance on the Raisin Dataset**

Among the tested models, XGBoost achieved the highest overall performance with an accuracy of 0.96**,** precision of 0.95**,** recall of 0.96**,** F1-score of 0.95, and ROC-AUC of 0.98. These results indicate that XGBoost effectively captures both the majority and minority classes, showing excellent discriminative power. The Random Forest model followed closely with an accuracy of 0.95 and ROC-AUC of 0.97, confirming the robustness of ensemble-based methods for handling non-linear relationships and high-dimensional intrusion data.

The MLP model also performed competitively, achieving an accuracy of 0.94 and ROC-AUC of 0.96, demonstrating that deep learning architectures can effectively learn complex data patterns, even with limited hyperparameter tuning. On the other hand, traditional models such as Logistic Regression and KNN achieved lower scores (0.89 and 0.90 accuracy respectively), indicating their limitations in capturing complex decision boundaries in multi-class intrusion

data. The SVM model exhibited balanced performance across all metrics (accuracy = 0.92, F1 = 0.91), confirming its capability in margin-based classification but with slightly higher computational cost.

Overall, ensemble-based methods such as XGBoost and Random Forest consistently outperformed other models across all metrics. The experimental results demonstrate that ensemble learning methods outperform traditional classifiers when applied to morphological feature-based datasets such as Raisin_Dataset. The high ROC-AUC values (>0.95) obtained by these models suggest superior capability in distinguishing between attack and normal instances. The results clearly demonstrate that combining multiple weak learners (as in XGBoost and Random Forest) significantly enhances predictive stability and generalization compared to single-model approaches.

5. Conclusion

In this study, the Raisin Dataset was employed to evaluate the performance of several machine learning algorithms for the classification of raisin varieties *(*Kecimen and Besni). The experimental results demonstrated that ensemble-based methods, particularly Random Forest and XGBoost, achieved the highest classification performance across multiple evaluation metrics, including accuracy, F1-score, and ROC-AUC. These models effectively captured the non-linear relationships among geometric and morphological features such as *Eccentricity,* MajorAxisLength*,* and *Perimeter*, which were identified as the most significant predictors.

The findings highlight the potential of machine learning techniques in supporting agricultural automation, specifically in the food industry, where accurate identification of product varieties is essential for quality control and market segmentation.

In conclusion, this work establishes a strong baseline for raisin classification using morphological features and emphasizes the importance of advanced ensemble learning techniques in achieving state-of-the-art results.

References

1. A. Jain, P. Singh, and R. Kumar, “Fruit classification using shape and texture features with machine learning algorithms,” *Journal of Food Engineering*, vol. 283, pp. 1–10, 2020.
2. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.
3. I. Cinar, M. Koklu and S. Tasdemir, “Classification of raisin grains using machine vision and artificial intelligence methods,” Gazi Journal of Engineering Sciences, Vol. 6, no. 3, pp. 200-209, Dec. 2020.
4. J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.
5. Kaggle Raisin Competition, https://www.kaggle.com/datasets/muratkoklu dataset/raisin-dataset/code, Last accessed 10 Oct. 2022.
6. K. Liu, Y. Chen, and L. Zhang, “Deep learning-based image classification for agricultural products,” *IEEE Access*, vol. 8, pp. 144872–144884, 2020.
7. M. Khojastehnazhand and H. Ramezani, “Machine vision system for classification of bulk raisins using texture features,” Journal of Food Engineering, Elsevier, vol. 271, no. 1, April 2020.
8. M. S. Ali, T. Rahman, and S. Hossain, “Comparative study of machine learning algorithms for seed variety classification,” *Computers and Electronics in Agriculture*, vol. 187, 106252, 2021.
9. S. Xie, H. Hu, and Y. Wu, “Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition,” *Pattern Recognition*, vol. 92, pp. 177–191, 2019.
10. Y. Zhang, Y. Yang, C. Ma and L. Jiang, “Identification of multiple raisins by feature fusion combined with NIR spectroscopy,” PLoS One, vol. 17, no. 7, 2022.