Volume 16, Issue 4, 162526 (1-12)

International Journal of Nano Dimension (IJND)

https://dx.doi.org/10.57647/j.ijnd.2025.1604.26

Feature-driven fruit classification using leveraging machine learning algorithms and nano semiconductor sensors

Venkateswarlu Gundu[1,*], Krishna Kumba[2], Sishaj P Simon[3], Parusharamulu Buduma[4], Mithu Sarkar[2]

[1]*Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur, AP, India.*
[2]*Vellore Institute of Technology Chennai, Vandalur, Tamil Nadu, India.*
[3]*National Institute of Technology Tiruchirappalli, Tamil Nadu, India.*
[4]*Lendi Institute of Engineering and Technology, Vizianagaram, AP, India.*

[*]**Corresponding author:** psv2482109@gmail.com

Original Research

Received:
30 January 2025
Revised:
3 April 2025
Accepted:
2 May 2025
Published online:
10 May 2025
Published in issue:
17 May 2025

Abstract:

The proposed work compares machine learning algorithms for fruit classification using apples, mandarins, oranges, and lemons. The goal is to identify the most accurate and precision-score algorithm. To assemble a robust dataset, we purchased several dozen oranges, lemons, and apples of various subtypes and meticulously recorded their mass, width, height, and color score using nano semiconductor sensors. Using this recorded data statistical analysis is conducted for identifying the accurate fruit classification machine learning method. The accurate prediction rate is determined by subtracting the actual and anticipated values. K-Nearest Neighbors (KNN) shows superior performance, achieving accuracies of 0.989, 0981 and 0.979 on training, validation and testing sets. The performance of the KNN algorithm combined with the W-H-CS feature combination technique is highly dependent on the choice of k and relevance of the selected features.

Keywords: Classification algorithms; Nano-scale feature detection; Machine learning algorithms and statistical analysis; Nano-enabled image sensors; Nano semiconductor sensors

1. Introduction

Fruit classification is critical for digital supermarkets and their management. Fruits obtained from various retailers provide an inexpensive way to characterize and classify the fruit mix. Initial investigations into fruit categorization primarily relied on datasets that contained only one feature. This method had its limitations in terms of precision and adaptability, as it failed to accommodate the extensive diversity in the visual characteristics of fruits [1, 2]. Traditional methods relying on single-feature data are prone to confusion when distinguishing similar ground objects due to the lack of comprehensive dimensional information [3, 4]. However, with the rapid evolution of digital techniques, researchers have increasingly turned to exploring multi-feature data to achieve enhanced classification accuracy [5–7]. The advent of advanced agricultural robots has led to a significant reduction in manual labour in the fields. These robots are now handling tasks like weeding [8, 9], pesticide spraying [10], monitoring plant growth [11], estimating crop yields [12, 13], and even harvesting fruits [14], and so on. All these advanced yielding techniques require an effective classification technique.

The traditional method of fruit classification entails collecting a variety of fruits from various trees and manually sorting them with labour assistance for sale in physical markets. However, this approach is inefficient and not scalable. In the current scenario, advanced robots allow for the re-

mote and convenient collection of fruit images, which are then processed using computer vision techniques. However, images taken within orchards frequently pose challenges: 1) They may have varying levels of illumination due to natural lighting conditions, and 2) the camera is typically positioned 1-2 meters from the nearest foliage, resulting in smaller fruits taking up a small portion of the image's pixels. These inherent characteristics present significant challenges to computer vision-based fruit detection. Recent advances in machine learning techniques have revealed [15].

Many efforts have been made to develop an automated fruit classification model, but the practical application of such a system has yet to be demonstrated [16–19]. Previous efforts have typically used a variety of sensors and machine learning techniques to detect features such as shape, colour, texture, and size in produce items for classification. Fruits, with their irregular shapes, sizes, and colours, present a significant challenge for identification, necessitating the investigation of nearly every fruit aspect as a classification feature. Early classification methods relied on global features such as shape and colour, whereas more advanced techniques delved into local features like texture. A variety of sensors have been used to capture fruit characteristics, ranging from basic black-and-white cameras to sophisticated hyperspectral cameras [20–22].

The classification, recognition, and detection of fruit and vegetables has proven to be challenging in the subfield of object recognition. By using techniques from similar domains, such as leaf classification for vegetable classification, the categorization of fruits and vegetables has also advanced [23]. The majority of research in this domain are focused on fusing machine learning algorithms for recognition or classification with image analysis for feature description [24–26]. The goal of these efforts is to represent physical characteristics using machine vision-based representations called feature descriptions. In order to obtain a qualitative outcome, these features are then passed into a classification algorithm. Many methods have been researched for feature description and categorization; however, substantial redesigns and enhancements are necessary for successful classification. The implementation of a fruit and vegetable classification system requires a thorough rethinking of all associated attributes, sensors, and classification algorithms. In order to offer a comprehensive analysis of the fruit classification endeavors, the complete procedure has been segmented into smaller steps, as illustrated in figure 1. The constituent processes are presented in this study in order, starting with a basic summary before going on to detail the particular variants that are used to classify fruits.

These machine learning-based detectors have been demonstrated to be effective for fruit detection and classification [14]. Though, the reasons why this state of arts are unable to handle the classification with high accuracy includes the following 1) single feature data set 2) usage of pre-trained models and fine-tuning on the trained data set only 3) non-visual sensor data. To overcome the above limitations this article incorporates a multi-feature classification technique to improve the prediction accuracy by conducting statistical analysis using different fruit features and machine learning algorithms. As a result of this article, the following contributions have been made:

- Proposed a novel machine learning classifier model for fruit classification, utilizing multilevel features to train the classifiers.
- Provided a comparative analysis of different machine learning models.

2. Data acquisition & experimental design

Data collection is a critical step in the development of any AI system. It entails gathering and preparing data for use in training and testing machine learning models. Depending on the specific problem and the type of data required, there are several methods for acquiring data for AI. Out of all these the following are the most common methods:

Open-source data collection: There is a massive amount of data available online that can be used to train machine learning models. For example, social media platforms, open data repositories, and government websites can all be valuable data sources.

Crowdsourcing: It is the process of gathering data samples from a large number of people via online surveys, questionnaires, or other means. This method is frequently used to collect data for natural language processing tasks like sentiment analysis or language translation.

Synthetic data: It is the information that is created fraudulently with the intent of imitating certifiable information. This approach is useful when there is a limited amount of certifiable information accessible or when the information contains sensitive data that cannot be shared.

Building custom data assortment pipelines: It is necessary to assemble custom information assortment pipelines from time to time in order to obtain information that is well-defined for the issue being addressed. Whatever method is used, it is critical to ensure that the data is accurate, relevant, and of high quality. This entails thoroughly cleaning and pre-processing the data to remove any errors, inconsistencies, or biases that may affect the AI system's performance. In this work building custom data assortment pipelines is used to acquire the data for experimental design and the same is described as follows. The fruit data acquisition process is shown in figure 1. To assemble a robust dataset, we purchased several dozen(sample size of 3000,) oranges, lemons, and apples of various subtypes and meticulously recorded their mass, width, height, and color score using nano semiconductor sensors, as illustrated in figure 2. Each entry in the dataset corresponds to a single piece of fruit, encompassing several features outlined in the figure.

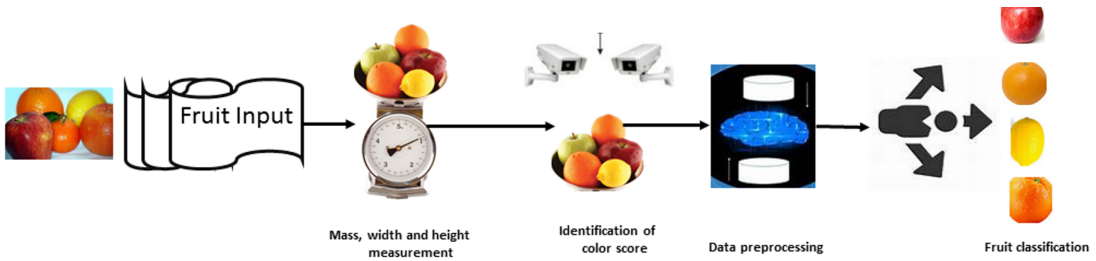

Figure 1. Data preprocessing and classification.

2.1 Data pre-processing

Because of noise, missing values, and unusable formats, a machine learning model cannot be directly applied to real-world data. Apart from sanitizing and organizing data for machine learning models, data pre-processing enhances the precision and effectiveness of those models. Therefore, it is necessary to deal with the dataset's missing values. The most popular techniques for handling missing data are as follows:

- Deleting a missing data row
- Calculating the mean

As per the literature, deleting the missing data may lead to a loss of information, which will not give an accurate output. As a result, the proposed method employs the calculating mean technique for efficient data pre-processing. In this method, for accurate fruit classification, the missing row/column is replaced with the mean of that particular row/column. We used the Scikit-learn library for data pr-processing, which contains various libraries for building machine learning models. The Imputer class from the Sklearn preprocessing library is used here. Categorical data is data that has multiple categories; for example, in our data set, there are four categorical variables: apple, mandarin, orange, and lemon. Despite the fact that machine learning works completely on mathematics and numbers, if our dataset contains categorical variables, then it may be difficult to build a model for it. Therefore, categorical variables must be encoded into numbers. For the effective encoding of categorical data, we used the Label Encoder () class from the pre-processing library. Finally, feature scaling is conducted to standardize the independent variables of the dataset in a specific range. In this work, feature scaling is done using the normalization technique as given in equation (1).

$$\text{New value} = \frac{\text{Original value} - \min(\text{original value})}{\max(\text{original value}) - \min(\text{original value})} \quad (1)$$

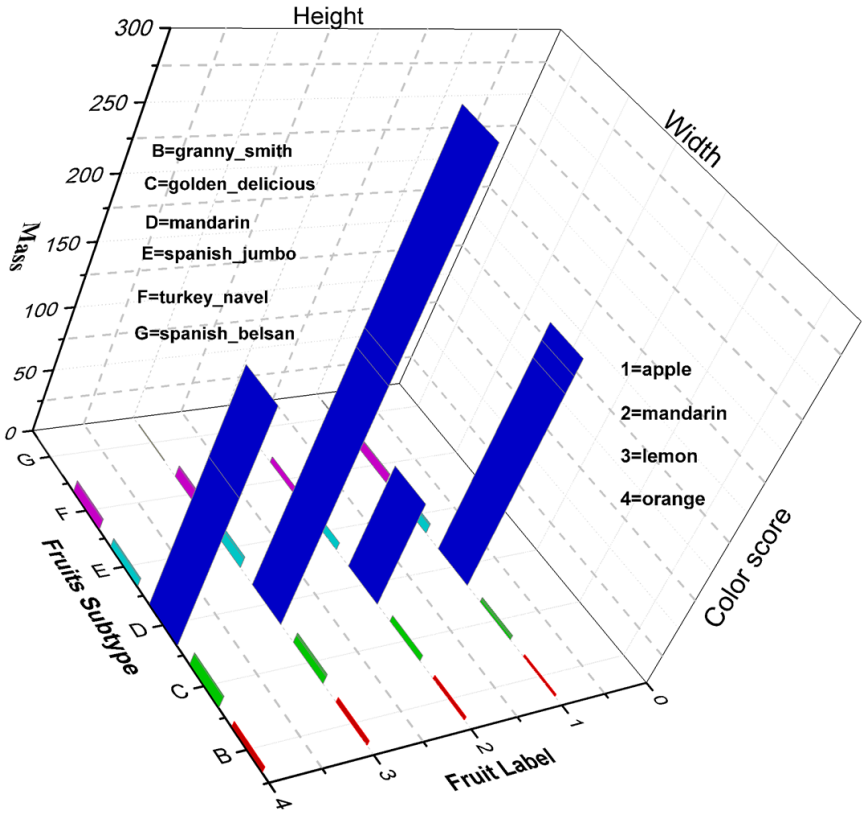

Figure 2. Optimal feature selection.

3. Feature selection

The research explores ways to increase model prediction capabilities, expedite data collection, and facilitate data interpretation by addressing input data gathering in the network model. The input data samples of fruit data (Mass(M), Width(W), Height(H), and Colour Score (CS)) are identified to be nonhomogeneous in nature. Because all of the fruit data is nonhomogeneous, the correlation between any combination of the four waveforms should be investigated for accurate classification. In order to select the appropriate feature data for effective classification, statistical analysis is performed, which can be accomplished by considering various possible combinations with different classification techniques, which are described in Table 1:

Table 1. Optimal fruit feature selection.

Fruit features	Classification techniques
M	Logistic regression Decision tree K-nearest neighbors Linear discriminant analysis Gaussian naive bayes Support vector machine
W	
H	
CS	
M-W	
M-H	
M-CS	
W-H	
W-CS	
H-CS	
M-W-H	
M-W-CS	
W-H-CS	
H-M-CS	
M-W-H-CS	

Multi classification building models

In machine learning, an observation or instance is classified by determining which category (subpopulation) it belongs to using a training set that contains observations (or instances) whose category membership is known. This paper describes various feature types, as shown in figure 1. Using this feature type, various fruit classification algorithms are proposed in order to build an effective model to classify the fruit types. The proposed models are described below:

A. Logistic regression

A logistic regression model, along with a collection of input variables, is used to forecast the probability of an event taking place, given the input data. Logistic regression model gives an outcome's probability, which goes from 0 to 1. The logistic function serves as the foundation for the logistic regression model, which is constructed around it using the sigmoid function, which transforms any real number into a value between 0 and 1. The following is the logistic function:

$$h(x) = \frac{1}{1+e^{-\varnothing x}} \tag{2}$$

Equation (2) uses x as the input data sample and $\varnothing$ as the learning parameter to optimize. The prediction value, which is nearer to 1, is the result. This suggests that the case (output = 1) has a higher probability of being a positive sample. The instance is more likely to be a negative sample (output = 0) if the value is close to zero. The log-likelihood loss function, which is the objective function for this fruit categorization, is found in equation (3).

$$J(\varnothing) = -\frac{1}{m}\sum_{i=1}^{m}(op^i \log(p^i) + (1-op^i)\log(1-p^i)) \tag{3}$$

where $J(\varnothing)$ represents the cost function, m is the number of samples, p^i represents the true label for the i^{th} sample

B. Decision tree

A decision tree is a graphical representation of various options and their potential outcomes. Decision trees are built using an algorithmic approach that identifies alternative ways to segment a data set based on certain conditions. To design decision trees, an algorithmic technique is utilized that identifies multiple ways to segment a data set depending on various conditions. It is one of the most popular and practical supervised learning algorithms. Decision trees are a non-parametric supervised learning method that can be used for classification as well as regression. In the decision tree classification technique, entropy is used for effective attribute selection. There is a lot of difficulty involved in deciding which attributes should be placed at the root or at different levels of the tree as internal nodes when there are N attributes in the dataset. Choosing any node as the root at random won't fix the problem. In order to solve attribute selection issues, most of the researchers worked and derived some solutions. Entropy is one of the most effective techniques for attribute selection. In this technique, a high entropy makes it harder to draw any conclusions about that information. A mathematical representation of entropy for multiple attributes is as follows:

$$E(t,x) = \sum_{c\in x} p(c)E(c) \tag{4}$$

$E(t,x)$ indicated the expected value of function between t and x.

x is a set, and c is the elements within the set.

$p(c)$ is the weighting factor associated with each element c.

$E(c)$ is the expected value of component c.

C. K-nearest neighbors

K-nearest neighbours is a powerful supervised machine learning algorithm that is utilized on open AI platforms for both classification and regression issues. As a result of the characteristics (labelled data) of the training data, KNN makes predictions on the test dataset. Using distance calculations between test and training data, these predictions are made by assuming the data points possess similar characteristics or attributes. k nearest neighbour's algorithm to classify fruit data $h(x)$ can be described as follows:

$$h(x) = \text{mode}[\{y'' : (x'', y'') \in s_x\}] \tag{5}$$

$h(x)$ gives the label of highest occurrence in the K-nearest neighbours. k nearest neighbors of x as s_x. formally s_x is

defined $s_x \in \text{dist}(x,y)$

$$\text{dist}(x,y) = (\sum_{r=1}^{d}\{|x_r - y_r|^p)\}^{\frac{1}{p}} \quad (6)$$

$h(x)$ is the predicted value for input x, $p=1$ Manthattan distance & $p=2$ Euclidean distance, $\{y'' : (x'',y'') \in s_x\}$ represents the set of output values y'' corresponding to input-output pairs (x'',y'') in the subset s_x.

D. Linear discriminant analysis

This technique reduces dimensionality through linear discriminant analysis. Pattern classification and machine learning applications use it as a pre-processing step. In order to sidestep the challenges posed by high-dimensional data, LDA aims to transform features from a higher-dimensional space to a lower-dimensional one. The primary objective of LDA is the projection of high dimensional features into a lower-dimensional space, which not only helps alleviate the curse of dimensionality but also results in resource and cost savings. It can be achieved with class variance, as shown below:

$$b_c = \sum_{i=1}^{k} ni(\bar{x}_i - \bar{x})\{\bar{x}_i - \bar{x}\}^T \quad (7)$$

The distance between mean and sample of each class is calculated using the equation given below:

$$b_w = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_{ij} - \bar{x}_i)\{\bar{x}_{ij} - \bar{x}_i\}^T \quad (8)$$

In the final step, construct the lower-dimensional space that minimizes the variance within a class but maximizes the variance between classes. Fisher's criterion is considered to be a lower-dimensional space projection.

$$p_{lda} = \text{argmax}\frac{|p^T b_c p|}{p^T b_w p} \quad (9)$$

where b_c represents the between-class scatter matrix, b_w represents the within-class scatter matrix and p is the direction.

E. Gaussian naive bayes

A naive Bayesian analysis assumes an independent relationship between the features. The covariance matrix for the classes is still class-specific, so we still assume the class-specific matrix, However, diagonal matrices are used to model covariance. In this case, the features are assumed to be independent. (Gaussian) Naive Bayes assumes a normally distributed distribution of class-conditional densities for a dataset with n input variables x with corresponding target variables t.

$$p(x|t = v, \mu_v, \Sigma_v. = n(x|\mu_v, \Sigma_v.)\} \quad (10)$$

In this example, μ is the class-specific mean vector, and $\Sigma.$ is the class-specific covariance matrix. Determine the class of x using Bayes theorem as follows:

$$h(x) = \{\text{argmx}\}vp(t = v|x,\}\mu v, \Sigma\{v\} \quad (11)$$

where $p(x)$ and $h(x)$ are the conditional probability and probability density of x at t.

F. Support vector machine

In Supervised Learning, an SVM, or Support Vector Machine, is used to solve classification and regression issues. Its primary application in machine learning is problem classification. The SVM algorithm makes it easier to assign additional data points to the relevant category in the future by generating the optimal feasible line or decision boundary that can divide an n-dimensional space into classes. An optimal decision boundary is referred as a hyperplane to maximize the minimum distance as shown in the figure 3. From figure it is observed that the goal of SVM is to minimize the distance and identifying a hyperplane with the maximum margin using the following equation (12).

$$d_h(x_0) = \text{argmax}\frac{|w^T(Q(x_o) + b|}{||w||_2} \quad (12)$$

where $d_h(x_0)$ = optimal direction, $w^T(Q(x_o) + b)$ = absolute value $Q(x_o)$, $Q(x_o)$ = transformation applied at x_o, $||w||_2$ = euclidean norm.

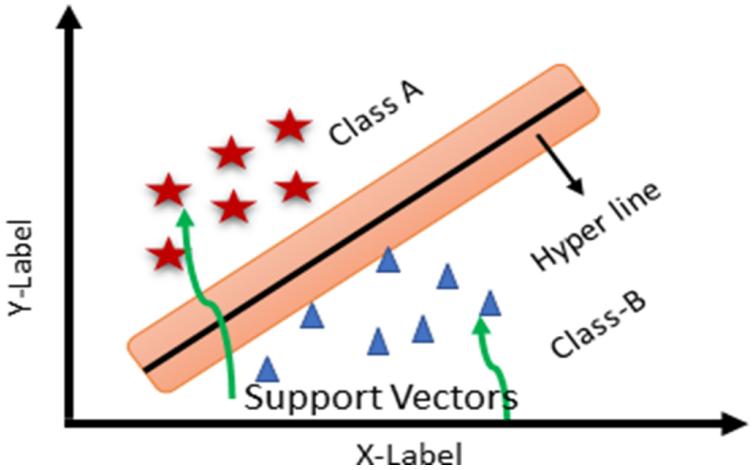

Figure 3. SVM hyper line graph.

4. Result and discussions

The proposed methodology is implemented using collected data from a rural area fruit market in Ippatam in September 2022, it is a certain remote region in Andhra Pradesh, as described in figure 2. The gathered information encompasses a sample size of 3000, featuring an assortment of apples, oranges, mandarins, and lemons, each exhibiting various subtypes alongside a spectrum of mass, width, height, and color ratings, as illustrated in figure 2. In this work, 70%, 20%, and 10% of the samples are used for training, validation, and testing, respectively. The scatter matrix for each input variable is shown in figure 4. As discussed in Table 1, the proposed model's accuracy can be improved by conducting statistical analysis. Here, statistical analysis is performed for various fruit feature combinations, and all possible combinations are evaluated using various classification techniques as detailed in section 5. The resulting possible combinations (15 combinations) are listed below.

Case:1

In this case {LR}-{4} indicates the logistic regression with four labels such as apples, oranges, mandarins, and lemons. Using {LR}-{4} with M, W, H and CS features there are total 15 possible combinations available which are listed below and are evaluated using logistic regression with a set of regularization = 1.0. The corresponding training and testing classifier accuracy is tabulated in the Table 2.

{M}-{LR}-{4},{W}-{LR}-{4},{H}-{LR}-{4},{CS}-{LR}-{4}{M-W}-{LR}-{4},{M-H}-{LR}-{4},{M-CS}-{LR}-{4},{W-H}-{LR}-{4},{W-CS}-{LR}-{4},{H-CS}-{LR}-{4}{M-W-H}-{LR}-{4},},{M-W-CS}-{LR}-{4},},{W-H-CS}-{LR}-{4},},{H-M-CS}-{LR}-{4},},{M-W-H-CS}-{LR}-{4}

where M = mass,W = width,H = height,CS = color score and

Case:2

In this case {DT}-{4} indicates the decision tree with four labels such as apples, oranges, mandarins, and lemons. Using {DR}-{4} with M, W, H and CS features there are total 15 possible combinations available which are listed below and are evaluated using decision tree tuned between 1 and 20 by setting optimal value = 5. The corresponding training and testing classifier accuracy is tabulated in the Table 3.

{M}-{DT}-{4},{W}-{DT}-{4},{H}-{DT}-{4},{CS}-{DT}-{4},{M-W}-{DT}-{4},{M-H}-{DT}-{4},{M-CS}-{DT}-{4},{W-H}-{DT}-{4},{W-CS}-{DT}-

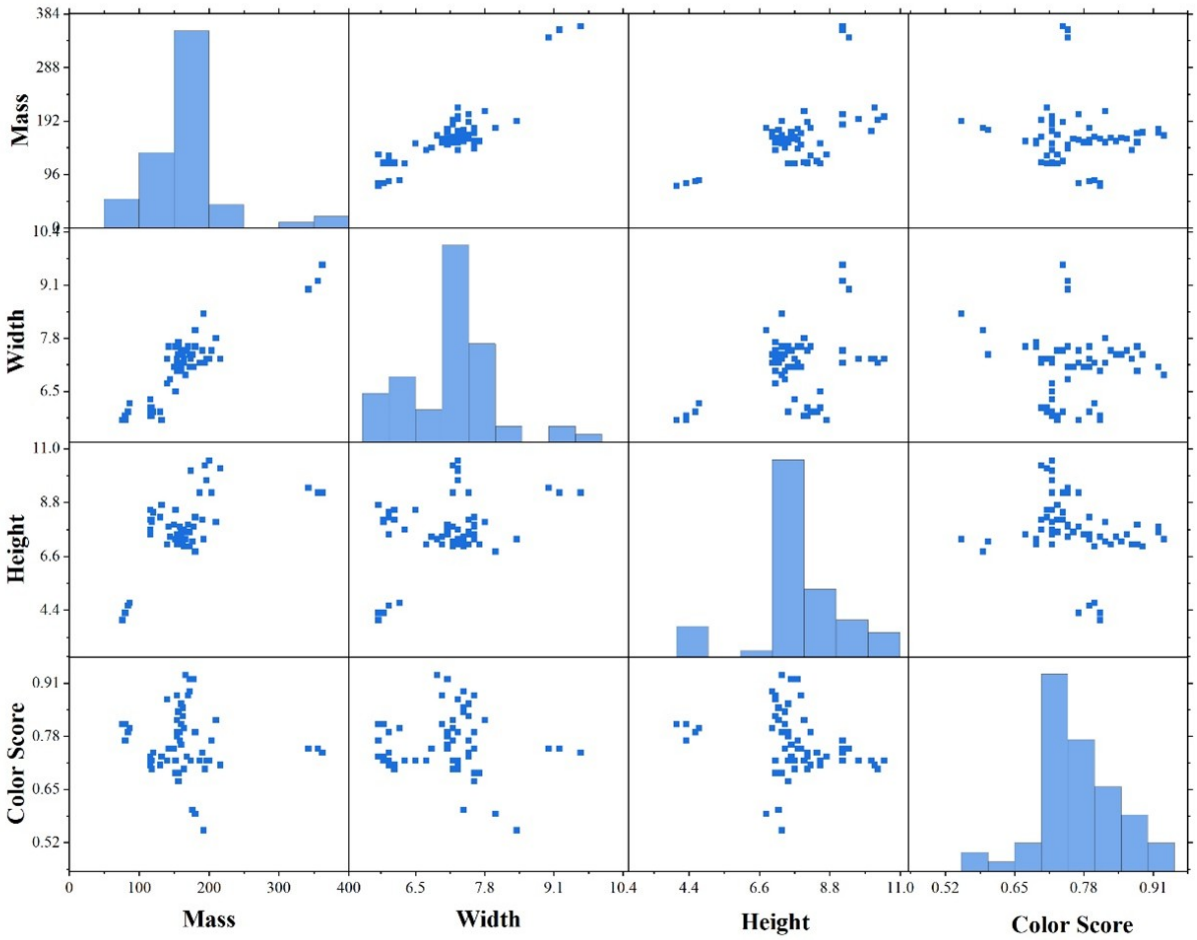

Figure 4. Scatter plot for input features.

Table 2. Accuracy of logistic regression classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.850	0.809
W	0.848	0.889
H	0.846	0.869
CS	0.842	0.829
M-W	0.852	0.829
M-H	0.855	0.859
M-CS	0.860	0.809
W-H	0.858	0.889
W-CS	0.861	0.819
H-CS	0.859	0.899
M-W-H	0.865	0.859
M-W-CS	0.868	0.889
W-H-CS	0.870	0.809
H-M-CS	0.864	0.849
M-W-H-CS	0.862	0.829

{4},{H-CS}-{DT}-{4},{M-W-H}-{DT}-{4},},{M-W-CS}-{DT}-{4},},{W-H-CS}-{DT}-{4},},{H-M-CS}-{DT}-{4},},{M-W-H-CS}-{DT}-{4}

Case:3

In this case {KNN}-{4} indicates the K-Nearest Neighbors with four labels such as apples, oranges, mandarins, and lemons. Using {KNN}-{4} with M, W, H and CS features there are total 15 possible combinations available which are listed below and are evaluated using K-Nearest Neighbors (KNN,K $=$ 5). The corresponding training and testing classifier accuracy is tabulated in the Table 4.

{M}-{KNN}-{4},{W}-{KNN}-{4},{H}-{KNN}-{4},{CS}-{KNN}-{4},{M-W}-{KNN}-{4},{M-H}-{KNN}-{4},{M-CS}-{KNN}-{4},{W-H}-{KNN}-{4},{W-CS}-{KNN}-{4},{H-CS}-{KNN}-{4},{M-W-H}-{KNN}-{4},},{M-W-CS}-{KNN}-{4},},{W-H-CS}-{KNN}-{4},},{H-M-CS}-{KNN}-{4},},{M-W-H-CS}-{KNN}-{4}

Case:4

In this case {LDA}-{4} indicates the linear discriminant analysis with four labels such as apples, oranges, mandarins, and lemons. Using {LDA}-{4} with M, W, H and CS features there are total 15 possible combinations

Table 3. Accuracy of decision tree classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.830	0.839
W	0.850	0.809
H	0.848	0.889
CS	0.844	0.849
M-W	0.854	0.849
M-H	0.857	0.879
M-CS	0.862	0.829
W-H	0.86	0.809
W-CS	0.863	0.839
H-CS	0.861	0.819
M-W-H	0.867	0.879
M-W-CS	0.807	0.809
W-H-CS	0.892	0.899
H-M-CS	0.866	0.869
M-W-H-CS	0.864	0.849

Table 4. Accuracy of KNN classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.955	0.959
W	0.952	0.929
H	0.905	0.909
CS	0.946	0.969
M-W	0.956	0.969
M-H	0.959	0.899
M-CS	0.964	0.949
W-H	0.962	0.929
W-CS	0.965	0.959
H-CS	0.963	0.939
M-W-H	0.969	0899
M-W-CS	0.972	0.829
W-H-CS	0.989	0.979
H-M-CS	0.968	0.889
M-W-H-CS	0.966	0.969

available which are listed below and are evaluated using linear discriminant analysis. The corresponding training and testing classifier accuracy is tabulated in the Table 5.

{M}-{LDA}-{4},{W}-{LDA}-{4},{H}-{LDA}-{4},{CS}-{LDA}-{4},{M-W}-{LDA}-{4},{M-H}-{LDA}-{4},{M-CS}-{LDA}-{4},{W-H}-{LDA}-{4},{W-CS}-{LDA}-{4},{H-CS}-{LDA}-{4},{M-W-H}-{LDA}-{4},},{M-W-CS}-{LDA}-{4},},{W-H-CS}-{LDA}-{4},},{H-M-CS}-{LDA}-{4},},{M-W-H-CS}-{LDA}-{4}

Case:5

In this case {GNB}-{4} indicates the Gaussian Naive Bayes (GNB)with four labels such as apples, oranges, mandarins, and lemons. Using {GNB}-{4} with M, W, H and CS features there are total 15 possible combinations available which are listed below and are evaluated using GNB. The corresponding training and testing classifier accuracy is tabulated in the Table 6.

{M}-{GNB}-{4},{W}-{GNB}-{4},{H}-{GNB}-{4},{CS}-{GNB}-{4},{M-W}-{GNB}-{4},{M-H}-{GNB}-{4},{M-CS}-{GNB}-{4},{W-H}-{GNB}-{4},{W-CS}-{GNB}-{4},{H-CS}-{GNB}-{4},{M-W-

Table 5. Accuracy of LDA classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.854	0.849
W	0.851	0.819
H	0.849	0.799
CS	0.845	0.859
M-W	0.855	0.859
M-H	0.858	0.889
M-CS	0.863	0.839
W-H	0.861	0.819
W-CS	0.864	0.849
H-CS	0.862	0.829
M-W-H	0.868	0.889
M-W-CS	0.871	0.819
W-H-CS	0.893	0.899
H-M-CS	0.867	0.879
M-W-H-CS	0.865	0.859

Table 6. Accuracy of GNB classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.849	0.799
W	0.847	0.879
H	0.845	0.859
CS	0.841	0.819
M-W	0.851	0.819
M-H	0.854	0.849
M-CS	0.859	0.849
W-H	0.857	0.879
W-CS	0.806	0.809
H-CS	0.858	0.789
M-W-H	0.864	0.849
M-W-CS	0.867	0.879
W-H-CS	0.899	0.896
H-M-CS	0.863	0.839
M-W-H-CS	0.861	0.819

H}-{GNB}-{4},},{M-W-CS}-{GNB}-{4},},{W-H-CS}-{GNB}-{4},},{H-M-CS}-{GNB}-{4},},{M-W-H-CS}-{GNB}-{4}

Case:6

In this case {SVM}-{4} indicates the Support Vector Machine (SVM) with four labels such as apples, oranges, mandarins, and lemons. Using {SVM}-{4} with M, W, H and CS features there are total 15 possible combinations available which are listed below and are evaluated using SVM. The corresponding training and testing classifier accuracy is tabulated in the Table 7.

{M}-{SVM}-{4},{W}-{SVM}-{4},{H}-{SVM}-{4},{CS}-{SVM}-{4},{M-W}-{SVM}-{4},{M-H}-{SVM}-{4},{M-CS}-{SVM}-{4},{W-H}-{SVM}-{4},{W-CS}-{SVM}-{4},{H-CS}-{SVM}-{4},{M-W-H}-{SVM}-{4},},{M-W-CS}-{SVM}-{4},},{W-H-CS}-{SVM}-{4},},{H-M-CS}-{SVM}-{4},},{M-W-H-CS}-{SVM}-{4}

The statistical analysis presented above concludes

Table 7. Accuracy of SVM classifier.

Fruit features	Accuracy on training set	Accuracy on testing set
M	0.850	0.809
W	0.848	0.789
H	0.846	0.869
CS	0.842	0.829
M-W	0.852	0.829
M-H	0.855	0.859
M-CS	0.806	0.809
W-H	0.858	0.789
W-CS	0.861	0.819
H-CS	0.859	0.799
M-W-H	0.865	0.859
M-W-CS	0.868	0.789
W-H-CS	0897	0.889
H-M-CS	0.864	0.849
M-W-H-CS	0.862	0.829

that the W-H-CS feature technique is the most effective model for classification. This proposed model is evaluated using various classification techniques, and the simulated training and testing accuracy are depicted in figure 5 and figure 6. The plots indicate that KNN with the W-H-CS feature technique outperforms basic Logistic Regression, decision tree, Linear Discriminant Analysis, Gaussian Naive Bayes, and Support Vector Machine, achieving accuracies of 0.989 and 0.979, respectively. Subsequent training and validation plots for KNN using the W-H-CS feature technique are shown in figure 7, while the decision boundary plot for the K-NN Classifier is displayed in figure 7. The suggested K-Nearest Neighbors (KNN) algorithm has been meticulously examined in relation to the previously published literature, with the intent of thoroughly validating the robustness and effectiveness of its performance, and the comparative results have been illustrated in the Table 8 presented below for better understanding and analysis. These results and discussions highlight the suitability of the proposed method for real-time fruit classification.

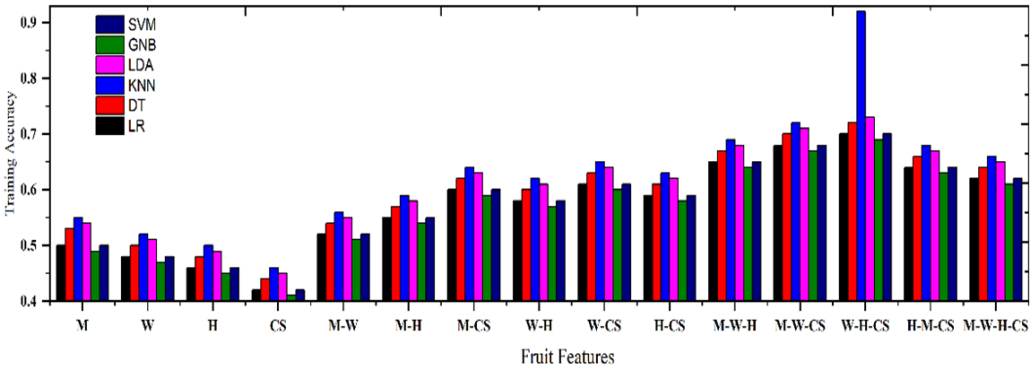

Figure 5. Training accuracy.

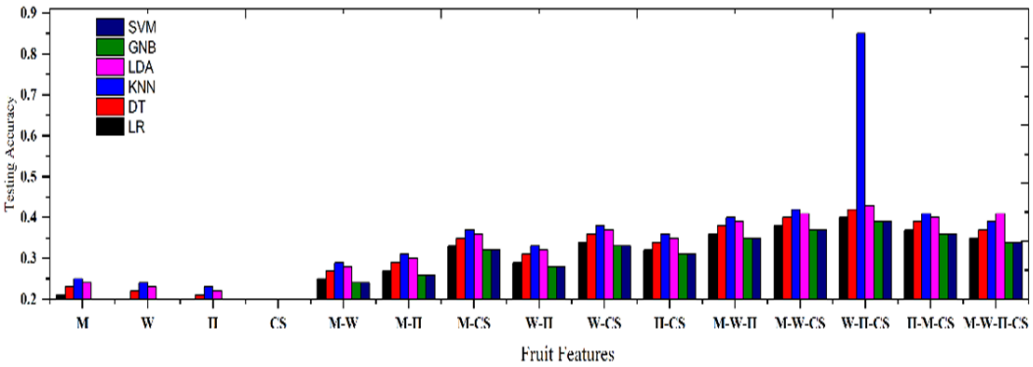

Figure 6. Testing accuracy.

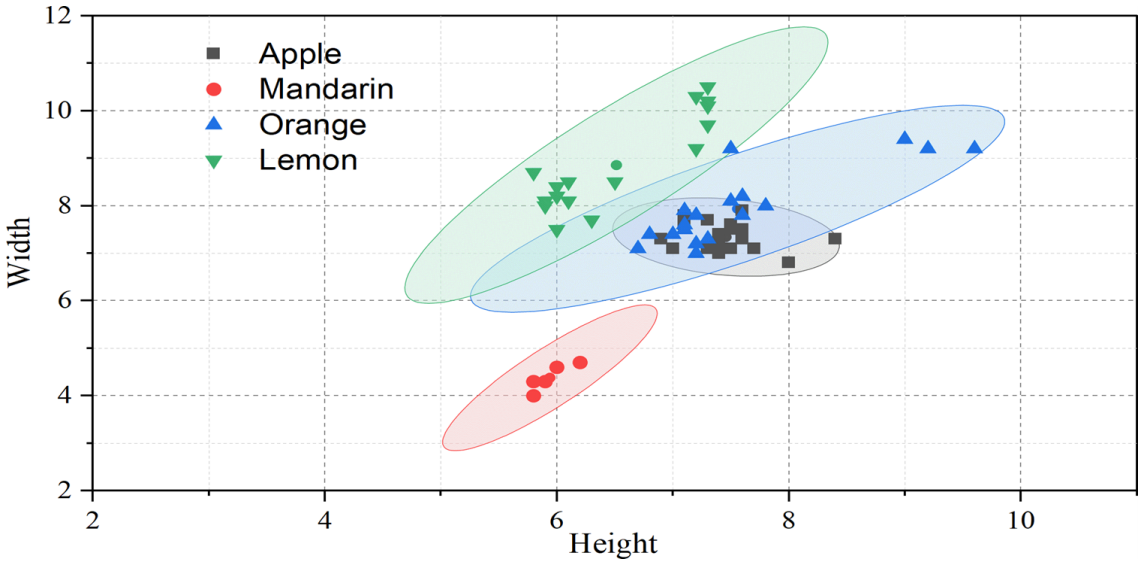

Figure 7. Decision boundary plot using KNN classifier.

Table 8. Comparison of case studies.

Method	Accuracy	Case studies
Edge detection	0.70	Blasco et al. (2009) – Citrus fruit sorting using color segmentation [27]
Transfer learning with VGG16	0.954	Mureşan & Oltean (2018) – CNN for fruit classificationTransfer learning with VGG16, ResNet [28]
W-H-CS-{KNN}-{4}	0.979	Proposed method

5. Conclusion

In this study, we conducted statistical analyses to determine the most effective combination of features and the most suitable model to achieve high classification accuracy. By applying the W-H-CS (Width-Height-Color Score) feature combination within the K-Nearest Neighbors (KNN) algorithm, we observed improvements in both training and testing accuracy on the fruit dataset. Despite the availability of numerous classification techniques in AI research, there remains a need for a robust approach for classifying dependent data, especially to support effective market analysis strategies. Our analysis of the fruit dataset highlights the importance of the W-H-CS feature combination for accurate label classification. Using KNN, we developed a feature model that achieved notably high accuracy, validated further by testing the method across additional classifiers: LR, DT, LDA, GNB, and SVM all tested with various feature combinations. Results demonstrate that pairing KNN with the W-H-CS feature combination enhances performance in classification tasks. However, the performance of the KNN algorithm combined with the W-H-CS feature combination technique is highly dependent on the choice of k and relevance of the selected features.

Authors contributions
Authors have contributed equally in preparing and writing the manuscript.

Availability of data and materials
The authors declare that the data supporting the findings of this study are available within the paper.

Conflict of interests
The authors assert that they do not have any identifiable conflicting financial interests or personal relationships that might be perceived to influence the work presented in this paper.

References

[1] T. B. Shahi, C. Sitaula, A. Neupane, and W. Guo. "Fruit classification using attention-based MobileNetV2 for industrial applications.". *PLoS ONE*, 17(2):e0264586, 2022.
DOI: https://doi.org/10.1371/journal.pone.0264586.

[2] X. X. Zhou, Y. Y. Li, Y. K. Luo, Y. W. Sun, Y. J. Su, C. W. Tan, and Y. J. Liu. "Research on remote sensing classification of fruit trees based on Sentinel-2 multitemporal imageries.". *Sci. Rep.*, 12 (1):11549, 2022.
DOI: https://doi.org/10.1038/s41598-022-15782-0.

[3] Y. D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad, and S. H. Wang. "Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation.". *Multimed. Tools Appl.*, 78:3613–3632, 2019.
DOI: https://doi.org/10.1007/s11042-018-6556-1.

[4] J. J. Bird, C. M. Barnes, L. J. Manso, A. Ekárt, and D. R. Faria. "Fruit quality and defect image classification with conditional GAN data augmentation.". *Sci. Hortic.*, 293:110684, 2022.
DOI: https://doi.org/10.1016/j.scienta.2022.110684.

[5] H. Wang, C. Xing, J. Yin, and J. Yang. "Land cover classification for polarimetric SAR Images based on vision transformer.". *Remote Sens.*, 14(18):4656, 2022.
DOI: https://doi.org/10.3390/rs14184656.

[6] S. Elango, D. Haldar, and A. Danodia. "Discrimination of maize crop in a mixed Kharif crop scenario with synergism of multiparametric SAR and optical data.". *Geocarto Int.*, 37(18):5307–5326, 2022.
DOI: https://doi.org/10.1080/10106049.2021.1927329.

[7] M. Imani. "Scattering and contextual features fusion using a complex multi-scale decomposition for polarimetric SAR image classification.". *Geocarto Int.*, pages 1–26, 2022.
DOI: https://doi.org/10.1080/10106049.2022.2061935.

[8] P. Lottes, J. Behley, A. Milioto, and C. Stachniss. "Fully convolutional networks with sequential information for robust crop and weed detection in precision farming.". *IEEE Robot. Autom. Lett.*, 3(4): 2870–2877, 2018.
DOI: https://doi.org/10.1109/LRA.2018.2846641.

[9] I. Sa, Z. Chen, M. Popović, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart. "WeedNet: Dense semantic weed classification using multispectral images and MAV for smart farming.". *IEEE Robot. Autom. Lett.*, 3(1):588–595, 2017.
DOI: https://doi.org/10.1109/LRA.2017.2778106.

[10] R. Berenstein and Y. Edan. "Automatic adjustable spraying device for site-specific agricultural application.". *IEEE Trans. Autom. Sci. Eng.*, 15(2):641–650, 2017.
DOI: https://doi.org/10.1109/TASE.2016.2636130.

[11] S. Surucu and A. Ecemis. "Classification of urban waste materials with deep learning architectures.". *SN Comput. Sci.*, 4(3):285, 2023.
DOI: https://doi.org/10.1007/s42979-023-01706-3.

[12] S. Bargoti and J. P. Underwood. "Image segmentation for fruit detection and yield estimation in apple orchards.". *J. Field Robot.*, 34(6):1039–1060, 2017.
DOI: https://doi.org/10.1002/rob.21709.

[13] R. K. Nath, H. Thapliyal, and T. S. Humble. "A review of machine learning classification using quantum annealing for real-world applications.". *SN Comput. Sci.*, 2:1–1, 2021.
DOI: https://doi.org/10.1007/s42979-020-00502-0.

[14] L. M. Abokaff. "Classification of Breast Cancer Diagnosis Systems Using Artificial Intelligence Techniques: Survey.". *SN Comput. Sci.*, 3(5):368, 2022.
DOI: https://doi.org/10.1007/s42979-022-01179-9.

[15] T. Wang, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao. "Learning rich features at high-speed for single-shot object detection.". *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1971–1980, 2019.
DOI: https://doi.org/10.1109/ICCV.2019.00206.

[16] H. Kang and C. Chen. "Fast implementation of real-time fruit detection in apple orchards using deep learning.". *Comput. Electron. Agric.*, 168:105108, 2020.
DOI: https://doi.org/10.1016/j.compag.2019.105108.

[17] A. Koirala, K. B. Walsh, Z. Wang, and C. McCarthy. "Deep learning for realtime fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'.". *Precis. Agric.*, 20:1107–1135, 2019.
DOI: https://doi.org/10.1007/s11119-019-09647-z.

[18] F. A. Kateb, M. M. Monowar, M. Hamid, A. Q. Ohi, and M. F. Mridha. "FruitDet: Attentive feature aggregation for real-time fruit detection in orchards.". *Agronomy*, 11(12):2440, 2021.
DOI: https://doi.org/10.3390/agronomy11122440.

[19] S. Wan and S. Goudos. "Faster R-CNN for multi-class fruit detection using a robotic vision system.". *Comput. Netw.*, 168:107036, 2020.
DOI: https://doi.org/10.1016/j.comnet.2019.107036.

[20] R. Kirk, G. Cielniak, and M. Mangan. "L*a*b* fruits: A rapid and robust outdoor fruit detection system combining bio-inspired features with one-stage deep learning networks.". *Sensors*, 20(1):275, 2020.
DOI: https://doi.org/10.3390/s20010275.

[21] Y. D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad, and S. H. Wang. "Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation.". *Multimed. Tools Appl.*, 78:3613–3632, 2019.
DOI: https://doi.org/10.1007/s11042-018-6556-1.

[22] A. Nasiri, A. Taheri-Garavand, and Y. D. Zhang. "Image-based deep learning automated sorting of date fruit.". *Postharvest Biol. Technol.*, 153:133–141, 2019.
DOI: https://doi.org/10.1016/j.postharvbio.2019.03.006.

[23] G. Chiesa, D. Di Vita, A. Ghadirzadeh, A. H. Herrera, and J. C. Rodriguez. "A fuzzy-logic IoT lighting and shading control system for smart buildings.". *Autom. Constr.*, 120:103397, 2020.
DOI: https://doi.org/10.1016/j.autcon.2020.103397.

[24] M. S. Munir, I. S. Bajwa, and S. M. Cheema. "An intelligent and secure smart watering system using fuzzy logic and blockchain.". *Comput. Electr. Eng.*, 77:109–119, 2019.
DOI: https://doi.org/10.1016/j.compeleceng.2019.04.008.

[25] A. A. F. Ogaili, M. N. Hamzah, and A. A. Jaber. "Integration of Machine Learning (ML) and Finite Element Analysis (FEA) for Predicting the Failure Modes of a Small Horizontal Composite Blade.". *Int. J. Renew. Energy Res.*, 12(4):2168–2179, 2022.
DOI: https://doi.org/10.20508/ijrer.v12i4.13354.g8589.

[26] L. A. Al-Haddad and A. A. Jaber. "Improved UAV blade unbalance prediction based on machine learning and ReliefF supreme feature ranking method.". *J. Braz. Soc. Mech. Sci. Eng.*, 45:463, 2023.
DOI: https://doi.org/10.1007/s40430-023-04386-5.

[27] J. Blasco, N. Aleixos, and E. Moltó. "Machine vision system for automatic quality grading of fruit.". *Biosyst. Eng.*, 103(1):29–39, 2009.
DOI: https://doi.org/10.1016/j.biosystemseng.2009.02.001.

[28] H. Mureșan and M. Oltean. "Fruit recognition from images using deep learning.". *Acta Univ. Sapientiae Inform.*, 10(1):26–42, 2018.
DOI: https://doi.org/10.2478/ausi-2018-0002.