Array 20 (2023) 100326

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Addressing agricultural challenges: An identification of best feature selection technique for dragon fruit disease recognition

Rashiduzzaman Shakil [a], Shawn Islam [a], Yeasir Arafat Shohan [a], Anonto Mia [a], Aditya Rajbongshi [b,*], Md Habibur Rahman [b,c], Bonna Akter [a]

[a] Daffodil International University, Bangladesh
[b] Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh
[c] Texas A&M University-Kingsville, Kingsville, Texas 78363, USA

ARTICLE INFO

Keywords:
Dragon
Feature selection
ANOVA
LASSO
GLCM
Random Forest
Performance metrics

ABSTRACT

Dragon fruit is a prominent substance in global agriculture. Despite this, it is gaining popularity and is a viable solution in resource-poor, environmentally degraded areas because of its many health benefits. Nevertheless, many dragon fruit plantations have been impacted by the disease, reducing their yield, and the detection system is still conventional. Farmers' lack of disease identification and management expertise diminished crop quality and products. As a result, little research was carried out to assist those specific farmers requiring adequate agricultural support. This research has proposed an autonomous agro-based system to recognize dragon diseases using in-depth analysis of feature selection techniques. After the collection of real-time images of the dragon, the images are preprocessed using various image-processing techniques. The two important features are retrieved after segmentation. The analysis of variance (ANOVA) and the least absolute shrinkage and selection operator (LASSO) are used as feature selection techniques to assess the feature rank based on the mutual score. To analyze the effectiveness of the machine learning algorithms that were used, six distinct machine learning classifiers were applied to the top-ranked feature sets, and their performance was measured using seven distinct performance evaluation metrics. AdaBoost and Random Forest classifiers for the LASSO feature ranking approach got the maximum accuracy, which is 96.29%, based on a comparison of classifiers based on the ANOVA and LASSO feature set. Despite this, we have optimized the computational resources of each classifier for the LASSO feature set.

1. Introduction

Dragon fruit (Pitaya), scientifically known as Selenicereus Undatus, contains resources of calories, protein, carbohydrates, fat, and fiber, which are highly gainful to human health. Dragon fruit originated in Mexico, a Central American country; currently, Vietnam and China are the world's biggest suppliers, with 70% [1]. The popularity of dragon fruit is increasing worldwide due to its unique appearance and potential as a profitable crop [2]. According to the study's findings [3], dragon fruit is more profitable than other crops with higher net returns. Based on observations in dragon fruit farming, the Compound Annual Growth Rate for the period(CAGR) 2023-2028 is estimated at 3.9% [4]. As a new growing place of dragon fruit farming, Bangladesh harvested almost 8660 tons of dragon fruit in 2020-21, more than double the previous year, 2019 [5]. The shape and exterior flaws determine the quality of dragon fruits. However, like other crops, Dragon fruit is infected with various diseases. Noninfectious damage, Fungi anthracnose, Botryosphaeria dothidea, and Bipolaris Cactivora diseases impacted dragon fruit the most.

One of the primary steps in feature engineering is feature selection. Remove redundant or unnecessary characteristics to decrease input variables via feature selection. The machine learning model's most significant characteristics are then selected as the best features for modeling the phenomena being examined. Feature selection improves machine learning model accuracy, selects the important variables, and elimination unnecessary that relate to the prediction capability.

Food, fiber, shelter, medicine, and fuel are just a few of the many things we rely on plants for. Despite their vital role in supporting life, plants face a number of obstacles. The quality and quantity of dragon products may suffer if infections are not properly diagnosed. The economic impact on a country is exacerbated. The use of fungicides and bactericides to protect crops from disease has a major influence on agricultural ecology. If farmers misdiagnose a disease, they may

* Corresponding author.
E-mail addresses: rashiduzzaman15-2655@diu.edu.bd (R. Shakil), shawonislamnm@gmail.com (S. Islam), shohanmd38@gmail.com (Y.A. Shohan), aanonto643@gmail.com (A. Mia), aditya0001@bdu.ac.bd (A. Rajbongshi), habibur0001@bdu.ac.bd (MH Rahman), bonna.diucse@gmail.com (B. Akter).

https://doi.org/10.1016/j.array.2023.100326
Received 25 September 2023; Accepted 21 October 2023
Available online 2 November 2023

treat it inefficiently, causing more damage to the plant. In addition, domain specialists' field excursions may be time-consuming and costly. As a result, an accurate and quick method of disease categorization is essential for the agroecosystem. System enhancements for early illness categorization on dragon leaves are possible with the help of advances in disease detection technologies including image processing and an ML-based model. When infections are diagnosed early on and the necessary steps are taken to halt their spread, the risk of output loss, processing expenses, and the negative environmental repercussions of chemical inputs (pollution with soil and water) are reduced.

In this study, several image processing methods are used to pre-process dragon real-time images. After segmentation, thirteen features are extracted. ANOVA and LASSO are also used to rank features based on mutual scores. Six machine learning classifiers namely, Decision Tree, Random Forest, KNN, AdaBoost, Logistic Regression, and SVM were applied to the top-ranked feature sets and evaluated using seven performance measures to assess their efficacy. A comparison of two feature selection techniques based on performance results is included to analyze the best feature selection technique adopted for dragon disease recognition. Finally, the computational cost is also estimated in order to find the significance of feature sets. The main contribution of our research is as follows:

1. **Agro-Based Feature Selection:** The research proposed an approach to select relevant features for identifying dragon diseases. It used two types of features, namely GLCM (Gray-Level Co-occurrence Matrix) and Statistical features.
2. **Feature Selection Techniques:** Two distinct feature selection techniques were employed: ANOVA (Analysis of Variance) and LASSO (Least Absolute Shrinkage and Selection Operator). These techniques were used to choose the top 10 features from an initial set of 13 features.
3. **Comparison of Classification Results:** The study compared the classification results achieved by ANOVA and LASSO for different subsets of features, including the top 5, 7, 9, and 10 features. The comparison was done using six different machine-learning algorithms.
4. **Research Objective:** The primary objective of this research is to determine the most effective feature selection approach for classifying dragon diseases. Additionally, the study aims to evaluate which specific features result in the highest accuracy in disease classification while keeping low computational resources.

In summary, the research aims to improve the accuracy and efficiency of dragon disease classification by selecting the most relevant features and assessing the performance of different feature selection techniques and machine learning algorithms.

This paper is organized into five sections. A brief introduction is mentioned in Section 1 and a literature review is discussed in Section 2. In Section 3 proposed method and material are described. The experimental results and evaluation of the proposed model and a comparative analysis and model validation have been analyzed in Section 4. The paper is concluded with the outcome of the study, limitation, and future work in Section 5.

2. Literature review

Dragon fruit is currently known as a healthy fruit all over the world. Many people are encouraged to plant dragon fruit, however, due to improper diagnosis, they are faced with losses. As a result, researchers are attempting to solve the problem using artificial intelligence and have also provided their proposed analysis.

L. Hakim et al. [6] presented a method for classifying dragon fruit based on their appearance (color and texture) using machine learning. They worked with three unique classes, totaling 81 images. The classification result was tested using two classic machine learning methods, KNN and SVM, with accuracy values of 73.33% and 87.5%, respectively.

A. Awate and S. Sonavane [7] utilized four distinct feature vectors, including color, texture, morphology, and structure, in order to address the challenge of detecting and classifying fruit diseases.

The categorization and prioritization of diseases affecting orange fruit were conducted by S.K. Behera et al. [8] through the utilization of fuzzy logic and machine learning techniques. The application of K-means clustering was utilized for the purpose of segmenting images. Subsequently, the GLCM feature extraction technique was employed to extract relevant features from the images. In the present investigation, the Support Vector Machine (SVM) demonstrated superior performance, with an accuracy rate of 90%.

A. Chakraborty et al. [9] proposed a plant leaf disease detection approach based on Fastai image classification. This study used a large data set called "PlantDoc", which includes 28 different classes.

B. Doh et al. [10] suggested a technique for disease identification in citrus fruits based on computer vision and machine learning. The dataset used in this study was not specified; nonetheless, the proposed study focused on categorizing six different types of citrus fruit. K-means clustering was employed for image segmentation. In addition, SVM and ANN were integrated into the image classification process. In comparison, the ANN model achieved an image classification accuracy of 93.12%, while the SVM model had an accuracy of 88.96%.

A. Bhargava and A. Bansal [11], worked with 12 features out of the 30 extracted features. The K-fold cross-validation technique was applied for image segmentation where $k = 10$. Several traditional machine learning models, including KNN, ANN, SVM, and a sparse representative classifier (SRC), were utilized to evaluate the classification result.

Local Binary patterns and Global color histograms are two color features used to classify papaya disease [12]. They obtained a dataset from the Mendeley database, which contains 214 images of five different types of papaya. In this work, they used k-means clustering for image segmentation and a support vector machine for categorizing the papaya classes with 96.3% accuracy.

S. Malathy et al. [13] introduced an image-processing approach for detecting fruit disease. They worked with an online dataset and mentioned an image restoration technique used to minimize image noise in this image preprocessing.

A hybrid machine-learning technique has been introduced by A.B. Mustafa et al. [14] to diagnose plant bacterial and fungal diseases. They conducted their research based on 500 samples with 3 classes namely fungal, bacterial, and healthy. The images were identified using KNN, SGD, and Random Forest. CNN-KNN and CNN-RF achieved 92.4% and 87.4% accuracy, respectively, while CNN-SGD achieved 93.6%.

W. N. Syazwani et al. [15] utilized contrast-limited adaptive enhancement to enhance the captured images. The researchers engaged in a study of three distinct feature extraction, namely color, texture, and shape. Additionally, the researchers employed the analysis of variance (ANOVA) feature selection approach to identify the most significant feature. And finally, six different machine learning algorithms are applied to classify the fruit crown and non-fruit crown. However, their combined ANN-GDX model outperformed with an accuracy rate of 94.4%.

3. Proposed methodology

There are several processes for classifying dragon fruit diseases, which are divided into numerous sections. Our first step was to collect the image from the field. Secondly, we need to preprocess the data before conducting this research. In addition, utilizing GLCM and statistical feature extraction, we also extracted 13 individual features from an image. ANOVA and LASO are two separate feature selection strategies used to determine the mutual score of each feature. Following the conclusion of the feature selection process, the top 10 features were chosen

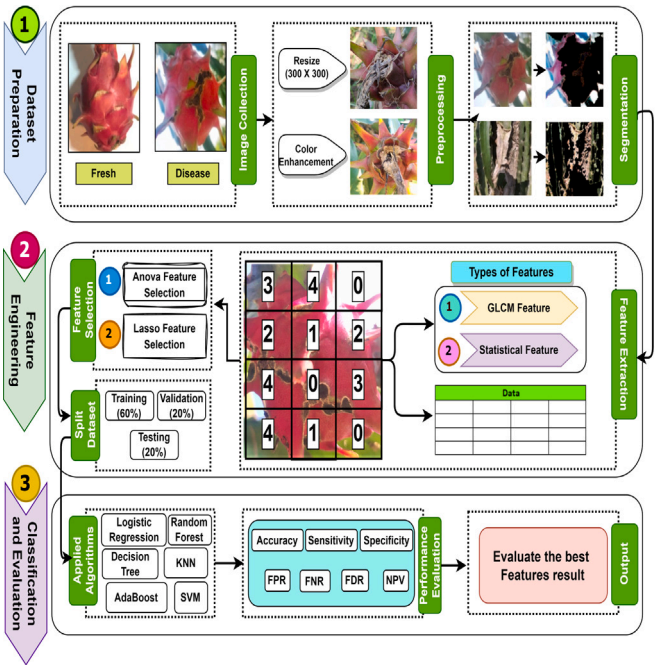
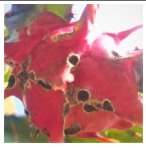


Fig. 1. Visual representation of the proposed system architecture in pictorial view.

Table 1
Dataset of Dragon fruit color images collected from a local filed in Dhaka, Bangladesh.

Class	Binary	Num. of images	Sample image
Defective	1	221	
Healthy	0	183	
Total		404	

for further analysis. These features include Contrast (CNT), Correlation (CRL), Skewness (SKEN), Kurtosis (KTS), Variance (VAR), Standard deviation (STD), Entropy (ENT), Energy (EG), Mean (MN), and Homogeneity (HGN). Then, we split the dataset, applied algorithms, and evaluated each algorithm using seven performance evaluation metrics. Fig. 1 depicts the architecture of our proposed work, including all of the processes mentioned above.

3.1. Image acquisition

The term "image acquisition" refers to the process of obtaining, compiling, and arranging images [16]. We acquired images taken from Poco X3 smartphone, which features a primary camera boasting a resolution of 65 megapixels and a storage capacity of 128 GB coupled with 6 GB of random-access memory (RAM). A total of 404 images of the dragon were collected from the dragon field in Dhaka, Bangladesh, including 183 healthy and 221 defective images. The frequency of the data is presented in Table 1.

3.2. Image prepossessing

Image preprocessing is essential for efficiently performing various computer vision-related tasks. To begin with, our original images were in 2160 × 2160 pixels, and resizing them to a standard size of 300 × 300 pixels helped with the Bilinear interpolation method [17]. This technique chooses four nearby pixels at random from the original image, and it uses the following formula to determine the value of each pixel in the compressed image: The variable x represents the distance between two places that are to be interpolated.

$$U(x) = \{0 \quad |x| > 1 \\ \{1 - |x| \quad |x| > 1 \tag{1}$$

Within our dataset, it has been observed that certain images exhibit suboptimal quality as a result of the presence of noise. In order to address the issue of noise reduction, a Gaussian filter was implemented, which is known for its ability to effectively remove noise and enhance the quality of the image by applying gamma correlation. After noise reduction, we employed histogram equalization to improve the contrast of each image, which is introduced as image enhancement in our study [18]. Assume that the range of the original image's intensities is [0, 1] and that the original image has been normalized. Let $p(x)$ represent the density function that describes the distribution of intensities in the normalized image, with x denoting the intensity value. The intended density function of the intensity distribution of the resulting image is equal to 1 following the process of equalization.

$$y = \int_{x}^{0} p(u)du \tag{2}$$

3.3. Image segmentation

Image segmentation segregates images into purposeful units, and the performance of feature extraction depends on segmentation quality. With a small dataset, K-means clustering executes well, assisting advanced analysis and perception of visual data [19]. For segmenting dragon fruit diseases, we used $k = 3$ to segment into 3 groups based on the correlation of the input feature, where k denotes the number of clusters [20]. In this research, we applied three clusters to improve the quality of the image segmentation.

3.4. Feature extraction

Typically, images are stored as massive matrices containing the values of individual pixels. It might be challenging for machine learning algorithms to learn from such large amounts of data. Feature extraction is the process of choosing the most informative features from a dataset in order to reduce the dimensionality of the data [21]. During the process of feature extraction, we identified a total of 13 distinct features, including GLCM and statistical features including Contrast (CNT), Correlation (CRL), Skewness (SKEN), kurtosis (KTS), Variance (VAR), Standard deviation (STD), Entropy (ENT), Energy (EG), Mean (MN), Homogeneity (HGN), Root means square (RMS), Smoothness (SM) and Inverse difference moment (IDM).

GLCM features have been proven effective for extracting textural features from images. The significant variance in the important pixel is obtained using the GLCM, which evaluates the relationship between both pixels. If $z(x; y)$ represents a two-dimensional digital image with $S \times T$ pixels and t_g gray levels, and (a_1, b_1) and (a_2, b_2) represent two pixels at $z(x, y)$ with an interval of d and angularity α of, then GLCM can be applied. GLCM equation can be expressed as follows for $Q(k, j, d, \alpha)$.

$$Q(j,k,d,\alpha) = [\{(a_1,b_1),(a_2,b_2)\, \epsilon\, S \times T : d, \alpha, z(a_1,b_1) \\ = j, z(a_2,b_2) = k\}] \tag{3}$$

GLCM features formulas derived from Eqs. (4) to (9). The relevant equations for these features are as follows: $Q(j, k)$ is the calculated GLCM's (j, k)th entry; t_g is the total number of grey levels in the image;

and μ_x, μ_y, and σ_x, σ_y are the means and standard deviations of the GLCM's row and column sums.

$$CNT = \sum_{k=0}^{t_g-1}\sum_{j=0}^{t_g-1}(k-j)^2 Q(k,j) \tag{4}$$

$$CRL = \frac{\sum_{j=0}^{t_g-1}\sum_{k=0}^{t_g-1} j.k.Q(j,k) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{5}$$

$$EG = \sum_{j=0}^{t_g-1}\sum_{k=0}^{t_g-1} Q(j,k)^2 \tag{6}$$

$$ENT = -\sum_{j=0}^{t_g-1}\sum_{k=0}^{t_g-1} Q(j,k)\log Q(j,k) \tag{7}$$

$$HGN = \sum_{j=0}^{t_g-1}\sum_{k=0}^{t_g-1} \frac{Q(j,k)}{1+(j-k)^2} \tag{8}$$

$$IDM = \sum_{j}\sum_{k} \frac{1}{1+(j-k)^2} Q(j,k) \tag{9}$$

Eqs. (10) to (14) demonstrate the formula for statistical features. From the below equation, in imperfect areas with P_x pixels and imperfect-free regions contain C pixels, where G is the number of pixels with exact color intensity and n is the summation index between 1 and P_x.

$$MN(\mu) = \frac{1}{P_x}\sum_{n=1}^{P_x} G_n \tag{10}$$

$$STD(\sigma) = \sqrt{\frac{\sum_{n=1}^{P_x}(G_n - C)^2}{P_x}} \tag{11}$$

$$VAR(\sigma^2) = \frac{1}{P_x}\sum_{n=1}^{P_x}(G_n - C)^2 \tag{12}$$

$$KTS(k) = \frac{\frac{1}{P_x}\sum_{n=1}^{P_x}(G_n - C)^4}{(\frac{1}{P_x}\sum_{n=1}^{P_x}(G_n - C)^2)^2} - 3 \tag{13}$$

$$SKEN(\gamma) = \frac{\sigma - G}{Q} \tag{14}$$

Here, In imperfect regions, the statistical measures used to describe grayscale color intensity are the mode, mean, and standard deviation, denoted as σ, Q, and G, respectively.

3.5. Feature selection

Feature selection is a data processing approach used to identify effective features from data and remove irrelevant features [22]. To determine the feature selection process, we used the Analysis of Variance (ANOVA) and Least Absolute Shrinkage Selection Operator (LASSO) to evaluate the score of feature ranking. However, after completing the feature ranking, we observed that three features, namely RMS, smoothness, and IDM, have nearly identical values. To mitigate the issue of redundancy, we eliminated these features and finally selected the top 10 features including CNT, CRL, SKEN, KTS, VAR, STD, ENT, EG, MN, and HGN to conduct this study.

3.5.1. ANOVA feature selection

ANOVA is a statistical approach examining whether there is any equal variance within groups of categorical features regarding numerical response [23]. The equation for ANOVA is as follows:

$$F = \frac{\frac{SSB}{K-1}}{\frac{SSW}{n-k}} \tag{15}$$

In this equation, ANOVA was represented by an F1-score and was determined by contrasting the variability within and across groups. SSW refers to the entire squares across groups, SSB refers to the entire squares among groups, k denotes the number of classes, and n represents the total sample.

Table 2
Hyperparameter configurations of six machine learning models employed in this evaluation.

Model	Hyperparameter	Value
AdaBoost	n_estimators	100
	min_samples_leaf	1
	criterion	gini
	min_samples_split	2
	max_features	sqrt
Decision tree	min_samples_leaf	1
	criterion	gini
	min_samples_split	2
KNN	n_neighbors	5
	weight	uniform
	algorithm	auto
	leaf_size	30
	p	2
	metric	minkowski
SVM	c	1
	kernel	rbf
	gamma	scale
Random forest	n_estimators	100
	min_samples_leaf	1
	criterion	gini
	min_samples_split	2
	max_features	auto
Logistic regression	c	1
	max_iteration	1000
	solver	lbfgs

3.5.2. LASSO feature selection

LASSO is mainly a regression method that also can be used for feature selection. The working procedure of the LASSO regularization strategy limits the sum of the actual values of each coefficient to impose picking characteristics and reduce overfitting [24]. LASSO is useful when the implemented dataset is small with many features and ranking the features more precisely. The equation for LASSO is as follows:

$$Minimize = (\frac{1}{2n_{samples}})((\|b - p_w\|)^2) + (a \times \|w\|_1) \tag{16}$$

In this equation, $n_samples$ denotes the total number of samples, target variable b, feature values matrix p, weight component factor w, $(\|y - x_w\|)^2$ squared error loss, α regularization parameter and $\|w\|_1$ sum of the absolute values of the coefficients.

3.6. Classifier training and testing

In this section, we mention the splitting ratio of train, validation, and test. The training, validation, and testing separation values are 60%, 20%, and 20% respectively. Training data is utilized to train a machine learning model and based on test data evaluate the efficiency of each model. As with validation data, it is used to compare the performance of various trained models.

3.7. Hyper parameter tuning

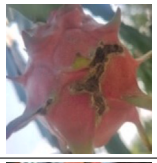
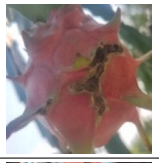
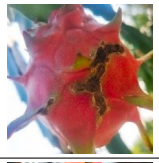
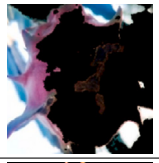
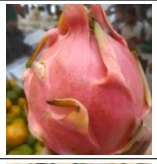
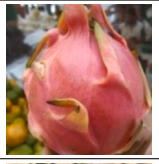
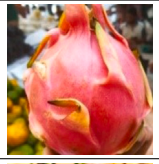
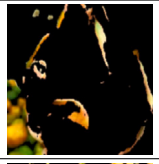
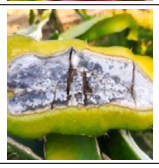
A hyperparameter is a type of machine learning parameter whose value is determined prior to algorithm training. Fine-tuning hyperparameters is a common technique for increasing the accuracy of a machine learning model [25]. In this research, the employed algorithm's hyperparameter values are all mentioned in Table 2.

Here, lbfgs means Limited-memory Broyden–Fletcher–Goldfarb-Shanno, gini refers to gini impurity and rbf means radial basis function.

3.8. Performance evaluation metrics

Performance evaluation metrics are needed to evaluate the model performance. In this research, we find out 7 (seven) performance evaluation metrics. These are gradual Accuracy (ACC), sensitivity (SEN),

Table 3
Representation of the feature extraction phases for four distinct Dragon images.

Image	Actual	Resized (300 × 300)	Enhanced	Segmented	Extracted feature values
Defective Fruit					0.24, 0.98, 0.37, 0.96, 68.42, 93.27, 4.13, 9.58, 4.67, 1.0, 2.28, 0.94, 255
Healthy Fruit					0.46, 0.93, 0.28, 0.94, 59.67, 74.37,4.79, 10.00, 3.87, 1.0, 2.32, 0.86, 255
Defective Leaf					0.29, 0.97, 0.42, 0.94, 56.66, 82.68, 3.84, 9.62, 6.77, 1.0, 2.44, 1.02, 255
Healthy Leaf					0.36, 0.93, 0.23, 0.93, 56.17, 56.06, 5.18, 11.53, 2.11, 1.0, 1.99, 0.46, 255

specificity (SPE), false positive rate (FPR), false negative rate (FNR), $F1\text{-}score$, and $precision$. All the metrics are calculated by Eqs. (17) to (23).

$$ACC = \left(\frac{TP+TN}{TP+FP+FN+TN}\right)\times 100\% \tag{17}$$

$$SEN = \left(\frac{TP}{TP+FN}\right)\times 100\% \tag{18}$$

$$SPE = \left(\frac{TN}{TN+FP}\right)\times 100\% \tag{19}$$

$$FPR = \left(\frac{FP}{FP+TN}\right)\times 100\% \tag{20}$$

$$FNR = \left(\frac{FN}{TP+FN}\right)\times 100\% \tag{21}$$

$$Precision = \left(\frac{TP}{TP+FP}\right)\times 100\% \tag{22}$$

$$F1-Score = \left(2\times\frac{SEN\times Precision}{SEN+Precision}\right)\times 100\% \tag{23}$$

4. Results and discussion

In order to identify dragon diseases, we completed our implementation in Jupyter Notebook with Python version 3.7.8 and MATLAB version 2017a. At first, all the images are resized into 300 × 300 pixels and then the contrast of the image is enhanced. After the completion of image segmentation, the two types of features are extracted [26]. Table 3, depicts the feature extraction procedure of an image.

Following the end of the feature extraction process, two feature-selection techniques, namely ANOVA and LASSO, are employed to determine the significance of features by considering their rank values. After completing the feature selection, we observed 3 features contain almost the same value. So based on the feature ranking score we eliminate these features to solve the redundancy issues and finally select the top 10 features to conduct this study. Tables 4 and 5 represent the mutual score of ANOVA and LASSO feature selection.

Six different machine learning (ML) algorithms were employed to conduct this study. The applying ML models are gradually Logistic regression (LR), Decision tree (DT), K-nearest neighbors (KNN), Random Forest (RF), Adaptive Boosting (AdaBoost), and Support vector machine (SVM). For calculating the classifier's performance, seven evaluation metrics are estimated. Tables 6, and Table 7 illustrate the performance of six ML models with different feature sets. Taking into consideration Table 6, it is possible to assert that the Random Forest classifier that used the top nine features of the ANOVA had the best level of accuracy, which was 95.06%. On the other hand, the accuracy for KNN when it adopted the top five characteristics was determined to be the lowest accuracy at 59.25%. Through the use of logistic regression, we were able to get a sensitivity of 71.79% using the five characteristics that were under investigation. We acquired the highest False Negative Rate (FNR) of 76.92% and False Positive Rate (FPR) of 4.76% attained by AdaBoost, out of the seven features that were picked. The ten-feature model had a f1-score that was as low as 81.57%, while the Support Vector Machine's (SVM) precision was as high as 83.78%.

Table 4
Top 10 feature ranking scores of the Dragon images found in the ANOVA feature selection technique.

Ranking	Name of feature	Ranking score	Ranking	Name of feature	Ranking score
1	HGN	0.1046	6	ENT	0.0923
2	KTS	0.1014	7	STD	0.0885
3	SKEN	0.0985	8	MN	0.0838
4	CRL	0.0973	9	CNT	0.0837
5	EG	0.0930	10	VAR	0.0814

Table 5
Top 10 feature ranking scores of the Dragon images found in the LASSO feature selection technique.

Ranking	Name of feature	Ranking score	Ranking	Name of feature	Ranking score
1	CNT	1.0	6	STD	0.2786
2	CRL	0.5305	7	ENT	0.2334
3	SKEN	0.3824	8	EG	0.1315
4	KTS	0.3222	9	MN	0.0069
5	VAR	0.2899	10	HGN	0.0021

Table 6
Comparative analysis of the performance of six machine learning models utilizing various feature sets selected through the ANOVA feature selection technique.

Feature number	Algorithm	Testing accuracy (%)	Sensitivity (%)	Specificity (%)	FPR (%)	FNR (%)	F1-Score (%)	Precision (%)
Five features	LR	76.54	71.79	80.95	19.04	28.20	74.66	77.77
	DT	80.24	82.05	78.57	21.42	17.94	79.48	79.48
	KNN	59.25	71.29	47.61	52.38	28.20	62.92	56
	RF	87.65	87.17	88.09	11.90	12.82	85.71	86.84
	AdaBoost	83.95	84.61	83.33	16.66	15.38	83.54	82.50
	SVM	77.77	74.35	80.95	19.04	25.64	76.31	78.37
Seven features	LR	88	100	78.57	21.42	0	89.65	81.25
	DT	92.59	97.43	88.09	11.90	2.56	92.68	88.37
	KNN	86.41	94.87	78.57	21.42	5.12	87.05	80.43
	RF	93.82	97.43	90.47	9.52	2.56	93.97	88.63
	AdaBoost	93.82	92.30	95.23	4.76	76.92	93.50	94.73
	SVM	88.88	100	78.57	21	0	89.65	81.25
Nine features	LR	90.12	100	80.95	19.04	0	90.69	82.97
	DT	88.88	92.30	85.71	14.28	7.69	90.24	86.04
	KNN	76.54	89.74	64.28	35.71	10.25	78.65	70
	RF	95.06	100	90.47	9.5	0	93.82	90.47
	AdaBoost	93.82	94.87	92.85	7.14	5.12	93.67	92.50
	SVM	90.12	100	80.95	19.04	0	90.69	82.97
Ten features	LR	83.95	82.05	85.71	14.28	17.94	83.11	84.21
	DT	83.95	79.48	88.09	11.90	20.51	83.78	88.57
	KNN	75.3	71.79	78.57	21.42	28.20	73.68	75.67
	RF	88.88	87.17	90.47	9.52	12.82	86.84	89.18
	AdaBoost	83.95	82.05	85.71	14.28	17.94	83.11	84.21
	SVM	82.71	79.48	85.71	14.28	20.51	81.57	83.78

Table 7
Comparative analysis of the performance of six machine learning models utilizing various feature sets selected through the LASSO feature selection technique.

Feature number	Algorithm	Testing accuracy (%)	Sensitivity (%)	Specificity (%)	FPR (%)	FNR (%)	F1-Score (%)	Precision (%)
Five features	LR	77.78	74.35	80.95	19.04	25.64	76.31	78.37
	DT	82.71	79.48	85.71	14.28	20.51	82.05	82.05
	KNN	90.12	84.61	95.23	4.76	15.38	89.18	94.28
	RF	92.59	92.3	92.85	7.14	7.69	87.17	87.17
	AdaBoost	83.95	87.17	80.95	19.04	12.82	83.95	80.95
	SVM	76.54	71.79	80.95	19.04	28.2	74.66	77.77
Seven features	LR	75.30	69.23	80.95	19.04	30.76	72.97	77.14
	DT	87.65	84.61	90.47	9.52	15.38	82.66	86.11
	KNN	61.72	58.97	64.28	35.71	41.02	59.74	60.52
	RF	90.12	89.74	90.47	9.52	10.25	89.74	89.74
	AdaBoost	86.41	84.61	88.09	11.9	15.38	85.71	86.84
	SVM	72.83	61.53	83.33	16.66	38.46	68.57	77.41
Nine features	LR	90.12	100	80.95	19.04	0	90.69	82.97
	DT	91.35	92.30	90.47	9.52	2.56	92.49	90.24
	KNN	76.54	89.74	64.28	35.71	10.25	78.65	70
	RF	93.82	97.43	90.47	9.52	2.56	96.29	92.85
	AdaBoost	**96.29**	**100**	**92.85**	**7.14**	**0**	**96.29**	**92.85**
	SVM	90.12	100	80.95	19.04	0	90.69	82.97
Ten features	LR	90.12	100	80.95	19.04	0	90.69	82.97
	DT	93.82	94.87	92.85	7.14	5.12	92.30	92.30
	KNN	76.54	89.74	64.28	35.71	10.25	78.65	70
	RF	**96.29**	**100**	**92.85**	**7.14**	**0**	**96.29**	**92.85**
	AdaBoost	95.06	97.43	92.85	7.14	2.56	95	92.68
	SVM	88.88	100	78.57	21.42	0	89.65	81.25

[a]The optimal value in each column is emphasized in bold.

On the other hand, it can be claimed from Table 7 that both AdaBoost and Random Forest classifier obtained the same highest which is 96.29% for nine and ten features adopting LASSO. The lowest accuracy is 61.72% achieved for the top seven features. Here the overall performance is found for the top ten features.

Accuracy comparison analysis between ANOVA and LASSO for the top 9 and 10 features are shown in Fig. 2. As a comparison of the results of ANOVA and LASSO above, ANOVA resulted in 95.06% accuracy for Random-Forest with nine features. In contrast, LASSO resulted in nine and ten features with the same higher accuracy for AdaBoost and Random Forest. AdaBoost has a higher accuracy of 96.29% and is the overall best performer in this proposed work. But the overall score is better in ten features over the nine features. Finally, it is observed that the number of features and the feature type significantly impact the performance of the classifiers, and LASSO features have a more significant impact than ANOVA features in recognizing the dragon disease.

The diagonal line is equal to the x-axis and the y-axis is followed by the ROC curve, which yields both true and false positive results simultaneously [27]. Here, The ROC Curve for ANOVA and LASSO for all applied algorithms considering nine features has been shown in Figs. 3 and 4 respectively.

For ANOVA, the highest AUC value is 0.9927 for Logistic Regression which was large enough. Also, frequently highest results were 0.9844, 0.9786, 0.8901, and 0.8245 achieved by Random Forest, AdaBoost, Decision Tree, and KNN classifiers. Lastly, the lowest AUC value achieved by the Support Vector Machine (SVM) is 0.6907 for all applied models.

In LASSO, the logistic regression model achieved the highlighted AUC value of 0.9927, indicating its intense discrimination. Other classifiers, such as Random Forest, AdaBoost, Decision Tree, and KNN,

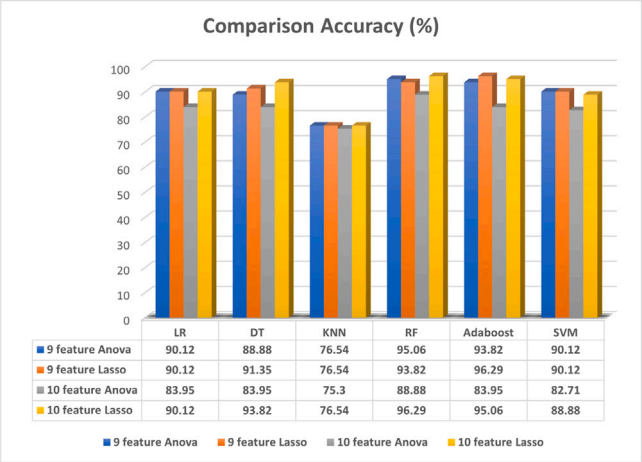


Fig. 2. Accuracy comparison of six machine learning models for top 9 and 10 features following ANOVA and LASSO feature selection techniques.

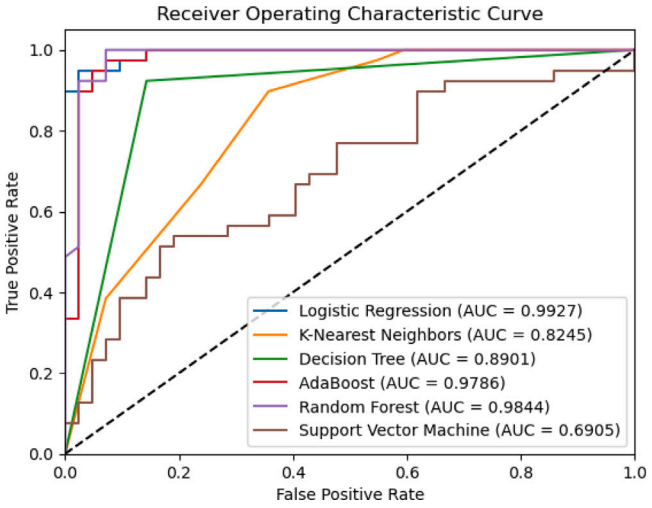


Fig. 3. Area under the curve of six machine learning models for top 9 feature sets selected through the ANOVA feature selection technique.

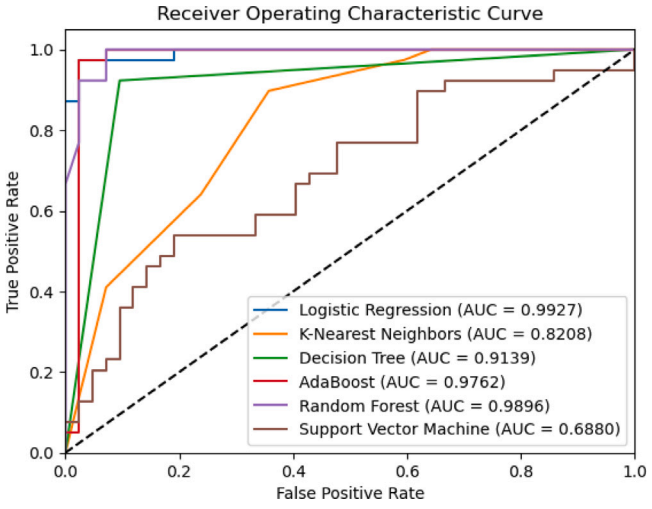


Fig. 4. Area under the curve of six machine learning models for top 9 feature sets selected through the LASSO feature selection technique.

Table 8
Execution time of six machine learning models utilizing various feature sets selected through the LASSO feature selection technique.

Number of features	Average elapsed time (ms)					
	Random forest	AdaBoost	Decision tree	Logistic regression	SVM	KNN
Five features	3270.26	3211.31	4.98	32.91	11.02	2.99
Seven features	3272.99	3216.98	6.98	61.83	10.01	2.99
Nine features	3287.14	3496.81	4.98	87.27	13.99	3.98
Ten Features	3247.22	3744.59	3.98	60.83	18.94	3.99

also performed well with AUC values of 0.9896, 0.9762, 0.9139, and 0.8208. The SVM had the lowest AUC value of 0.6907 among all models tested. The ROC curve shows the larger area of the model based on performance. After evaluating the ROC curve, the LASSO feature selection outperforms the ANOVA feature selection. To ensure and guarantee the reproducibility of the models presented in this work, all datasets utilized in the creation of predictive models are accessible on GitHub.

4.1. Average training time

In general, we worked with two feature selection methods and got two different tables of execution time for ANOVA and LASSO. After analyzing Tables 6 and 7 observed that LASSO scores well in all aspects. So here, we have computed the average training time for LASSO, presented in Table 8. Our attention is on AdaBoost because it outperforms other programs regarding accuracy and performance. Therefore, we may ignore the training term for the algorithms KNN, Logistic Regression, Random Forest, SVM, and Decision Tree. Lastly, Table 8, shows that the features selection strategy helps to reduce the classifier evaluating cost.

4.2. Comparative analysis

Numerous researchers are working on machine learning-based image classification systems in various areas. To evaluate the effectiveness of our expert system on dragon fruit disease classification, a comparison of recent work on similar machine learning systems has been necessary. We can directly compare our findings to any other work, as much research has yet to be done on dragon fruit disease classification. As mentioned, agriculture has the most significant impact on our global GDP. There is currently a lot of agricultural research, which needs more practical demands to be considered. We compared the result with other published works to evaluate our work, whereas Table 9, shows comparing results. Compared to previous research, our accuracy of 96.29% in identifying Dragon fruit disease from others is quite excellent.

5. Conclusion and future work

An agro-based feature selection approach to recognize dragon disease is presented in this study. A total of 404 images were used to conduct this research work, and two types of features were extracted from the segmented image data. ANOVA and LASSO are two distinct feature selection models applied to compute the feature rank and select the top $5 \leq N \leq 10$ features to demonstrate this work. We deployed six distinct machine learning algorithms depending on the feature rank to evaluate the efficiency of five, seven, nine, and ten feature sets respectively. Here, Random Forest obtained the best accuracy of 95.06% for the top nine features in the ANOVA feature selection. In comparison, Random Forest and AdaBoost both achieved the highest accuracy of 96.29% for the top nine and ten features in the Lasso feature selection. However, the research does have certain limitations, such as its dependence on the availability of representative

Table 9
Comparative analysis of the proposed method with state-of-the-art research work.

Research work	Dealing object	Class	Dataset frequency	Segmentation algorithm	Feature selection	Set of features	Employed classifier	Achieved accuracy
L. Hakim et al. [6]	Dragon root & Leaf	Three	81	k-means clustering	Color moment & GLCM	NM	SVM KNN	87.5%
A.B. Mustafa et al.[14]	NM	Binary	1000	KNN, SGD, Random Forest	NM	NM	CNN-SGD	93.6%
A. Rajbongshi et al. [28]	Cauliflower	Four	766	k-means clustering	NM	10	Random Forest	88.61%
Y. Lu et al.[29]	Rice	Ten	500	NA	NM	NA	CNN	95.48%
U. Mokhtar et al. [30]	Tomato	Binary	800	Manual Cropping	NM	36	Linear SVM	93.25%
M. T. Habib et al. [31]	Jackfruit	Four	480	k-means clustering	NM	10	Random Forest	89.59%
L. J. Rozario et al. [32]	Apple, Banana, Tomato	Three	63	k-means clustering, Modified k-means clustering, Otsu method	NM	NA	NA	NM
S. R. Dubey et al. [33]	Fruit and vegetable	Fifteen	2612	k-means clustering	NM	2	SVM	93.77
N. N. Kurniawati et al. [34]	Paddy	Three	NM	Local entropy thresholding		NA	NA	94.7%
R.Shakil et al. [35]	Cauliflower	Four	708	K-means Clustering	ANOVA	10	Logistic Regression	90.77%
Proposed Method	Dragon Fruit & Leaf	Binary	404	k-means clustering	**LASSO** ANOVA	**10**	**AdaBoost**	**96.29%**

NA: Not Application.
NM: Not Mentioned.

data, sensitivity to algorithmic choices, and the need to account for environmental variability. Additionally, model interpretability and the practical implementation of the proposed method in real-world agricultural settings may pose challenges. Acknowledging these limitations is crucial for effectively leveraging the method's practical advantages while addressing the complexities of agricultural disease management. The research not only holds promise for improving the detection of dragon diseases in agriculture but also has broader implications for the agricultural sector. Its findings, while initially focused on dragon diseases, may be applicable to the detection and classification of other agricultural diseases, thereby benefiting a wider range of crops and regions. Furthermore, the research serves as a valuable starting point for future investigations at the intersection of agriculture and machine learning, offering opportunities to refine and expand upon the proposed approach, ultimately leading to practical tools and solutions for addressing critical challenges in agricultural disease management. Our future goal is to develop an Android app based on research that can assist farmers in recognizing dragon diseases and providing essential data promptly.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Hossain FM, Numan SMN, Akhtar S. Cultivation, nutritional value, and health benefits of dragon fruit (Hylocereus spp.): A review. Int J Hortic Sci Technol 2021;8(3):259–69.

[2] Patil PU, Lande SB, Nagalkar VJ, Nikam SB, Wakchaure GC. Grading and sorting technique of dragon fruits using machine learning algorithms. J Agric Food Res 2021;4:100118.

[3] Saediman H, Mboe IS, Budiyanto B, Sarinah S, Hidrawati H. Smallholder adoption of horticultural crops: The case of dragon fruit in Southeast sulawesi. In: IOP conference series: Earth and environmental science. 819, (1):IOP Publishing; 2021, 012043.

[4] Dragon Fruit Market - Trends & Production by Country, Available link: https://www.mordorintelligence.com/industry-reports/dragon-fruit-market. [Accessed 16 July 2023].

[5] Dragon fruit cultivation gaining ground | The Daily Star, Available link: https://www.thedailystar.net/business/economy/news/dragon-fruit-cultivation-gaining-ground-3100246. [Accessed 18 July 2023].

[6] Hakim L, Kristanto SP, Yusuf D, Shodiq MN, Setiawan WA. Disease detection of dragon fruit stem based on the combined features of color and texture. INTENSIF J Ilm Penelit Penerapan Teknol Sist Inf 2021;5(2):161–75.

[7] Awate A, Deshmankar D, Amrutkar G, Bagul U, Sonavane S. Fruit disease detection using color, texture analysis and ANN. In: 2015 international conference on green computing and internet of things. IEEE; 2015, p. 970–5.

[8] Behera SK, Jena L, Rath AK, Sethy PK. Disease classification and grading of orange using machine learning and fuzzy logic. In: 2018 international conference on communication and signal processing. IEEE; 2018, p. 0678–82.

[9] Chakraborty A, Kumer D, Deeba K. Plant leaf disease recognition using fastai image classification. In: 2021 5th International conference on computing methodologies and communication. IEEE; 2021, p. 1624–30.

[10] Doh B, Zhang D, Shen Y, Hussain F, Doh RF, Ayepah K. Automatic citrus fruit disease detection by phenotyping using machine learning. In: 2019 25th International conference on automation and computing. IEEE; 2019, p. 1–5.

[11] Bhargava A, Bansal A. Automatic detection and grading of multiple fruits by machine learning. Food Anal Methods 2020;13:751–61.

[12] Islam MA, Islam MS, Hossen MS, Emon MU, Keya MS, Habib A. Machine learning based image classification of papaya disease recognition. In: 2020 4th International conference on electronics, communication and aerospace technology. IEEE; 2020, p. 1353–60.

[13] Malathy S, Karthiga RR, Swetha K, Preethi G. Disease detection in fruits using image processing. In: 2021 6th International conference on inventive computation technologies. IEEE; 2021, p. 747–52.

[14] BaniMustafa A, Qattous H, Ghabeish I, Karajeh M. A machine learning hybrid approach for diagnosing plants bacterial and fungal diseases. Int J Adv Comput Sci Appl 2023;14(1).

[15] Syazwani RWN, Asraf HM, Amin MAMS, Dalila KAN. Automated image identification, detection and fruit counting of top-view pineapple crown using machine learning. Alex Eng J 2022;61(2):1265–76.

[16] Mishra VK, Kumar S, Shukla N. Image acquisition and techniques to perform image acquisition. SAMRIDDHI J Phys Sci Eng Technol 2017;9(01):21–4.

[17] Fadnavis S. Image interpolation techniques in digital image processing: An overview. Int J Eng Res Appl 2014;4(10):70–3.

[18] Cheng H-D, Shi XJ. A simple and effective histogram equalization approach to image enhancement. Digit Signal Process 2004;14(2):158–70.

[19] Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. attern Recognit 2003;36(2):451–61.

[20] Rajbongshi A, Biswas AA, Biswas J, Shakil R, Akter B, Barman MR. Sunflower diseases recognition using computer vision-based approach. In: 2021 IEEE 9th region 10 humanitarian technology conference, R10-HTC. IEEE; 2021, p. 1–5.

[21] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 science and information conference. IEEE; 2014, p. 372–8.

[22] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing 2018;300:70–9.
[23] St L, Wold S, et al. Analysis of variance (ANOVA). Chemometr Intell Lab Syst 1989;6(4):259–72.
[24] Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. In: 2016 IEEE international conference on advances in computer applications. IEEE; 2016, p. 18–20.
[25] Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol Model 2019;406:109–20.
[26] Dubey SR, Jalal AS. Adapted approach for fruit disease identification using images. In: Image processing: Concepts, methodologies, tools, and applications. IGI Global; 2013, p. 1395–409.
[27] Hoo ZH, Candlish J, Teare D. What is an ROC curve? Emerg Med J 2017;34(6):357–9.
[28] Rajbongshi A, Islam ME, Mia MJ, Sakif TI, Majumder A. A comprehensive investigation to cauliflower diseases recognition: An automated machine learning approach. Int J Adv Sci Eng Inform Technol 2022;12(1):32–41.
[29] Lu Y, Yi S, Zeng N, Liu Y, Zhang Y. Identification of rice diseases using deep convolutional neural networks. Neurocomputing 2017;267:378–84.
[30] Mokhtar U, El Bendary N, Hassenian AE, Emary E, Mahmoud MA, Hefny H, et al. SVM-based detection of tomato leaves diseases. In: Intelligent systems' 2014: Proceedings of the 7th IEEE international conference intelligent systems IS'2014, September 24–26, 2014, Warsaw, Poland. Tools, architectures, systems, applications, vol. 2, Springer; 2015, p. 641–52.
[31] Habib MT, Mia MJ, Uddin MS, Ahmed F. An in-depth exploration of automated jackfruit disease recognition. J King Saud Univ, Comput Inf Sci 2022;34(4):1200–9.
[32] Rozario LJ, Rahman T, Uddin MS. Segmentation of the region of defects in fruits and vegetables. Int J Comput Sci Inf Secur 2016;14(5):399.
[33] Dubey SR, Jalal AS. Fruit and vegetable recognition by fusing colour and texture features of the image using machine learning. Int J Appl Pattern Recognit 2015;2(2):160–81.
[34] Kurniawati NN, Abdullah SNHS, Abdullah S, Abdullah S. Investigation on image processing techniques for diagnosing paddy diseases. In: 2009 international conference of soft computing and pattern recognition. IEEE; 2009, p. 272–7.
[35] Shakil R, Akter B, Shamrat FMJM, Noori SRH. A novel automated feature selection based approach to recognize cauliflower disease. Bull Electr Eng Inform 2023;12(6):3541–51.