

A novel automated feature selection based approach to recognize cauliflower disease

Rashiduzzaman Shakil¹, Bonna Akter¹, F M Javed Mehedi Shamrat², Sheak Rashed Haider Noori¹

¹Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

²Department of Computer System and Technology, University of Malaya, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Nov 28, 2022

Revised Jan 2, 2023

Accepted Apr 17, 2023

Keywords:

Cauliflower disease

Decision tree

K-means cluster

K-nearest neighbor

Logistic regression

ABSTRACT

Cauliflower disease is a primary cause of reduced cauliflower yield. Preventing cauliflower disease requires early diagnosis. In the scope of this study, we suggested an agro-medical expert system that would make it easier to diagnose cauliflower disease. In this method, a digital image must be taken off the phone or handled device to diagnose cauliflower sickness. A data augmentation technique was initially used to construct a vast data set. The disease-affected parts of the cauliflower were then segmented using k-means clustering. Following that, ten statistical and gray-level co-occurrence matrix (GLCM) features were retrieved from the segmented pictures. After choosing the top n features (N ranged from 5 to 10), the synthetic minority oversampling technique (SMOTE) approach was used to handle training datasets with different amounts of each feature. After that, we utilized five machine learning (ML) algorithms and evaluated their performance using seven performance evaluation matrices for both augmented and non-augmented datasets. The same procedure was performed on both datasets. Then, we use both datasets to test how well the classifier works. Logistic regression (LR) is the most accurate method for the top nine features in the augmented dataset (90.77%).

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rashiduzzaman Shakil

Department of Computer Science and Engineering, Daffodil International University

Daffodil Road, Asulia, Dhaka 1341, Bangladesh

Email: rashiduzzaman.diucse@gmail.com

1. INTRODUCTION

Bangladesh's economy is dependent on agricultural production and this fact is widely recognized. Agriculture is a major driving force behind this progression in a developing country like Bangladesh. Bangladesh: employment in agriculture, a report by the World Bank, says that more than 39% of all jobs in Bangladesh are in agriculture [1] and agriculture also makes a big contribution to the country's gross domestic product (GDP), which is 12.92% [2]. Bangladesh: agricultural GDP share. Diseases affecting plants are a significant source of economic losses in the agricultural sector. Therefore, it is critical to detect plant diseases by observing their outward manifestations early before the infection can spread to the healthy plant.

Cauliflower, the scientific name *brassica oleracea*, has undergone many genetic changes and is now grown on every continent. China, India, the US, Spain, Mexico, and Bangladesh cultivate commercial cauliflower. Comparatively, Bangladesh produces 73,000 metric tons of cauliflower annually on 9,400 acres [3]. The aggregate nutrient density index (ANDI) score, which looks at how many vitamins, minerals, and phytonutrients are in a food, says that cauliflower is among the top 10 most nutrient-dense foods [4]. Cauliflower is a nutritional powerhouse with high levels of vitamins C and K [5] and can be eaten both cooked and raw in salads and relishes. Diseases can impede cauliflower's growth, lowering its quality and

yield. Traditional methods for diagnosing cauliflower infections are arduous, time-consuming, and costly making them unfeasible for large-scale farming operations. Farmers in less developed nations or rural Bangladesh may need to travel to meet with professionals.

This study addresses the application of machine learning (ML) to recognize and predict cauliflower diseases such as black rot, downy mildew, bacterial spot, and fresh leaves. Our framework is a cloud-based, ML-powered platform that uses mobile images as input. K-means clustering is used to categorize diseased samples. Then, five classifiers were used to train and assess disease recognition. Seven measures were used to evaluate the algorithms. The main goals behind developing these models are: i) recognize cauliflower disease early on in an automated way; ii) gray-level co-occurrence matrix (GLCM) feature extraction was used to pull features from the collected images. Analysis of variance (ANOVA) feature selection was used to rank features based on mutual information scores; iii) cauliflower diseases require a systematic organization of the most reliable features for training and testing classifiers; and iv) our model accurately predicted cauliflower disease from image data.

2. BACKGROUND STUDY

Disease in plants is a significant problem in the agriculture sector. Numerous investigations are undertaken to detect diseases in apples, rice, and cauliflower. Even though much work has been done to determine what's wrong with cauliflower, more must be done to make it work better. Sasirekha and Suganthy [6] suggested a k-means clustering algorithm for carrot disease. This study employed k-means clustering to segment images GLCM features help find the effect region to determine the standard deviation, IDM, entropy, root mean square (RMS), smoothness, variance, contrast, skewness, kurtosis, and correlation. Support vector machine (SVM) was used to classify carrot diseases in this article. But they can't mention the categorization accuracy of their model. Sari *et al.* [7] proposed an agro-medical method to identify papaya disease. For training and validation, they used 50 papaya trees. They used flexible Naïve Bayes classifier (FNBC) to determine how well this model was validated and compared it to the forward-changing technique. With FNBC, the validation accuracy was 88%, while it was 90% in the forward-changing phase. This study used specific data and forward change handles tiny amounts of data. So, its validity and accuracy may be questioned while managing massive amounts of information. Gu *et al.* [8] used hyperspectral imaging and ML to determine early on if a disease affected tomatoes. ML algorithms (boosted regression tree (BRT), classification and regression tree (CART), random forest (RF), and SVM) are used to find and confirm diseases. Switched parasitic array (SPA) and genetic algorithm (GA) are used to design environments. In this study, a BRT worked 85.2% of the time and had an area under curve (AUC) of 0.932.

Research by Chaudhary *et al.* [9] came up with the EnsPSO technique, which is a mix of voting, the particle swarm optimization (PSO) algorithm, the correlation-based feature selection (CFS) method, and random sampling. Its goal is to make it easier to find agricultural diseases. They used three datasets to train and test the proposed approach and the voting method. The EnsPSO method was more accurate than Vote (96%). Kianat *et al.* [10] developed a method for diagnosing cucumber leaf diseases that makes use of both a feature reduction method and a robust feature selection method. With the 900 samples from the six classes, quadratic SVM, cubic SVM, linear discriminant analysis (LDA), and linear SVM models were created. This method uses both feature reduction and robust feature selection methods. This method is PDbE-based. For this strategy, quadratic SVM was the best-fit model (93.5% accuracy). Islam *et al.* [11] constructed a novel ML-based papaya disease detection system. They used 214 samples from an online dataset to build their model and used RF, k-means, SVM, and CNN. The CNN algorithm's accuracy was 98.04%. Habib *et al.* [12] suggested a machine vision-based papaya disease diagnosis. They used an existing dataset with five diseases but excluded the areas. Two feature selection strategies and three ML classifiers are employed to identify papaya. SVM classifier accuracy was 95.2%. Panigrahi *et al.* [13] performed poor ML work on maize disease identification. An online dataset utilized Naïve Bayes (NB), decision tree (DT), k-nearest neighbor (KNN), SVM, and RF classifiers to classify the disease, but accuracy was poor. They got 79.23% accuracy, which was substantially lower than other work. Rajbongshi *et al.* [14] suggested a ML-based cauliflower disease detection method. To conduct this study, 766 disease images were used. K-means clustering segments of diseased areas. BayesNet, Kstar, RF, logistic model tree (LMT), back propagation neural network (BPN), and J48 classify diseases. The RF classifier scored 89% in this study. Methun *et al.* [15] presented a deep learning technique for carrot disease. This experimental CNN uses the VGG16, VGG19, MobileNet, and Inception v3 models. Inception V3 had the most accuracy 97.4%.

Based on the research of other authors, these studies were conducted to recognize cauliflower diseases and those authors' suggested methods were applied to the original data. It is remarkable that we applied our model to augmented and nonaugmented data. We also applied GLCM feature extraction method

to identify interested regions from image data. The accuracy of our model might be better if we apply different feature selection techniques to choose features.

3. METHOD

This section outlines the several methods utilized to implement ML-based cauliflower disease identification. Our approach comprises three main parts: the overall architecture, extracting features from the collected images, and using an ML-based strategy to find diseases in cauliflower. This method's justification and implementation are elaborated upon below.

In our proposed framework, images depicting cauliflower diseases are captured by smartphones and used in an online ML-based approach. Figure 1 illustrates the overall architectural design of our proposed machine vision-based expert system. Initially, consumers install our envisaged expert system app and capture images using their devices, which are then transferred as input through the application. The results are sent to the user through SMS when the analysis is finished using the proposed architecture. Finally, the user may see the outcome.

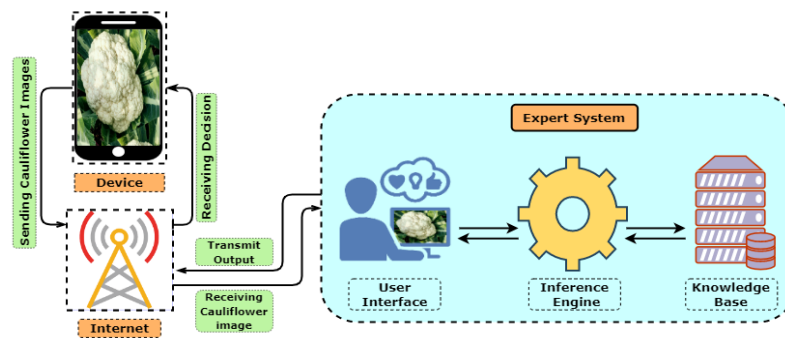


Figure 1. The system architecture for the recognition of cauliflower diseases using machine vision

3.1. Image collection

This dataset was collected by authors and it's already available in data in brief [16]. This dataset included a total of 1,920 pictures, divided into four categories: bacterial spots, downy mildew, black rot, and disease-free. To train the model using an extensive dataset. Additionally, the model's performance is compared to the model trained without the data augmentation technique. After adding new information, the total amount of data is presented in Table 1.

Table 1. Dataset overview

Class	Original data	Augmented data
Bacterial spot	173	300
Downy mildew	170	312
Black rot	160	280
Disease-free	205	320
Total	708	1,212

3.2. Preprocessing

In order to use the collected images effectively, it is essential to resize them to the correct dimensions, as they are all different. To begin with, the images were converted to a fixed length of 224×224 using bicubic interpolation. Assume i and f gradually, and the derivatives are f_x , f_y , and f_{xy} , which represent the four corners of a unit square (1, 1), (1, 0), (0, 1), and (0, 1), respectively, (0, 0), where m_{ij} denotes the coefficients. The stability of the interpolation surface [11] is defined using as (1):

$$f(x, y) = \sum_{i=0}^3 \sum_{j=0}^3 m_{ij} x^i y^j \tag{1}$$

3.2.1. Histogram equalization tends to increase contrast

Histogram equalization is used to intensify image contrast. Let's pretend that X and Y represent the number of rows (height) and columns (width) in pixels, that C_k is the color intensity of P_k pixels and that I is

the image's intensity level. The processed images [11] are explicitly defined as (2), where each pixel with C_k in is mapped to a pixel with color intensity S_k .

$$S_k = T(C_k) = \frac{l-1}{XY} \sum_{j=0}^k n_j \quad (2)$$

3.2.2. Convert the colour RGB to $L \times a \times b$

The k-means clustering algorithm is a form of unsupervised ML. An RGB color space image is converted into $L \times a \times b$ color space for better segmentation. This conversion is only used for $L \times a \times b$ color space. After the contrast is increased, the effort required to convert to RGB is calculated. Since the result of the $L \times a \times b$ color space conversion is identical to the original, there's a compelling reason to utilize it. Convert to CIA before transitioning to the $L \times a \times b$ color space in the RGB color space. In (3) used depending on [11]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 3.2405 & -1.5372 & -0.4985 \\ -0.9692 & 1.8759 & 0.0416 \\ 0.0556 & -0.2040 & 1.0573 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3)$$

The X_n , Y_n , and Z_n values of the reference white can be used to calculate the color space $L \times a \times b$. More information [11] is available in (4):

$$f(t) = \begin{cases} t^{\frac{1}{3}} & \text{if } t > 0.00885 \\ 7.787 + \frac{16}{116} & \text{if } t \leq 0.008856 \end{cases} \quad (4)$$

For $L \times a \times b$ can be calculated by using (5)-(7):

$$L^* = \begin{cases} 116 \left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16 & \text{if } \frac{Y}{Y_n} > 0.008856 \\ 903.3 \left(\frac{Y}{Y_n}\right) & \text{if } \frac{Y}{Y_n} \leq 0.008856 \end{cases} \quad (5)$$

$$a^* = 500 \left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right) \right) \quad (6)$$

$$b^* = 200 \left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right) \right) \quad (7)$$

Afterward, the k-means clustering method is applied to segment the images, which essentially chops out the diseased regions of the leaf while leaving the healthy ones to remain. During both the training and testing phases, the features were found using the above method were used to train and test a classifier. Then, five state-of-the-art, top-performing classifiers are chosen from among a broad range of classifiers. This group includes the methods of KNN, Adaboost, logistic regression (LR), and DT. Each candidate classifier's performance is compared using various evaluation matrices to narrow the field and concentrate on the optimal solution. During the performance analysis phase, accuracy is never a good way to determine how well the classifier performs. Because it may not be suited for examining categorization patterns on data sets that are otherwise imbalanced. A few other performance evaluation matrices for classifier performance analysis [17], [18]. Results from a two-class classification method can be described as true positives (TP), true negatives (TN), false positives (FP), or false negatives (FN). However, the matrix R can be expressed as (8) for multiclass classification:

$$R = [e_{ij}]_{N \times N} \quad (8)$$

The fact that R is a square matrix is immediately apparent in (8). Which included N (rows) times N (columns), where N was more than 2 and N^2 is the total amount considered. If we're talking about class i, the matrices can be computed as (9)-(12):

$$TP_i = e_{ii} \quad (9)$$

$$FP_i = \sum_{\substack{j=1 \\ j \neq i}}^n e_{ji} \quad (10)$$

$$FN_i = \sum_{\substack{j=1 \\ j \neq i}}^n e_{ij} \quad (11)$$

$$TN_i = \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{k=1 \\ k \neq i}}^n e_{jk} \quad (12)$$

After this procedure, the matrix R arrives at 2×2 dimensions. Consider this as a final result into action accuracy, sensitivity, error rate, specificity, precision, false positive rate (FPR), and false negative rate (FNR) are calculated as (13)-(19):

$$Accuracy = \left(\frac{TP+TN}{TP+FP+FN+TN} \times 100 \right) \% \quad (13)$$

$$TPR = \left(\frac{TP}{TP+FN} \times 100 \right) \% \quad (14)$$

$$TNR = \left(\frac{TN}{TN+FP} \times 100 \right) \% \quad (15)$$

$$FPR = \left(\frac{FP}{FP+TN} \times 100 \right) \% \quad (16)$$

$$FNR = \left(\frac{FN}{FN+TP} \times 100 \right) \% \quad (17)$$

$$Precision = \left(\frac{TP}{TP+FP} \times 100 \right) \% \quad (18)$$

$$Error Rate = \left(\frac{FP+FN}{TP+FP+FN+TN} \times 100 \right) \% \quad (19)$$

After using the cross-validation method, we used the receiver operating characteristic (ROC) to identify in (13)-(19). Finally, we select the most suitable classifier.

3.2.3. Feature extraction with gray-level co-occurrence matrix

We used image processing to pull out several statistical and GLCM features that help us spot diseases in cauliflower. We have selected the standard deviation (σ), mean (μ), variance (σ^2), the skewness (γ), and the kurtosis (k) [19]. If there are n pixels in the faulty region(s), where I is the gray-scale intensity of a pixel and I , I_m , and I_r are the mean, mode, and standard deviation of grey-scale intensity of all pixels correspondingly, then the related equations of these features are as (20)-(24). We used image processing to pull out several statistical and GLCM features that help us spot diseases in cauliflower.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (I_i - \bar{I})^2}{n}} \quad (20)$$

$$\mu = \frac{1}{n} \sum_{i=1}^n I_i \quad (21)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (I_i - \bar{I})^2 \quad (22)$$

$$\gamma = \frac{\bar{I} - I_M}{I_\sigma} \quad (23)$$

$$k = \frac{\frac{1}{n} \sum_{i=1}^n (I_i - \bar{I})^4}{\left(\frac{1}{n} \sum_{i=1}^n (I_i - \bar{I})^2 \right)^2} - 3 \quad (24)$$

You can also think of GLCM as a gray-level spatial dependence matrix. Each pair in (i, j) indicates how often the pixel was used. At the same time, i co-occurred horizontally with j 's pixel.

$$\text{Contrast: } \sum_{ij} |i - j|^2 p(i, j) \quad (25)$$

$$\text{Correlation: } \sum_{ij} \frac{(i - \mu_i)(j - \mu_j) p(i, j)}{\sigma_i \sigma_j} \quad (26)$$

$$\text{Energy: } \sum_{ij} p(i, j)^2 \tag{27}$$

$$\text{Homogeneity: } \sum_{ij} \frac{p(i, j)}{1+|i-j|} \tag{28}$$

$$\text{Entropy: } \sum_{i=0}^{L-1} (p(x_i) \log_2 p(x_i)) \tag{29}$$

This subsection explains how to apply the extracted feature for disease identification in cauliflower. First, the retrieved features were used as input, then a training set and a test set were made from them. Next, a ranking of features was carried out, utilizing a total of ten different features. The training set was then balanced using synthetic minority oversampling technique (SMOTE) and a ML model was applied to both the training and testing data. In the end, performance evaluation matrices are used to assess the efficacy of every classifier. In Figure 2, we depict all the procedures that are followed at this stage. The following is a comprehensive explanation of the technique that was discussed earlier.

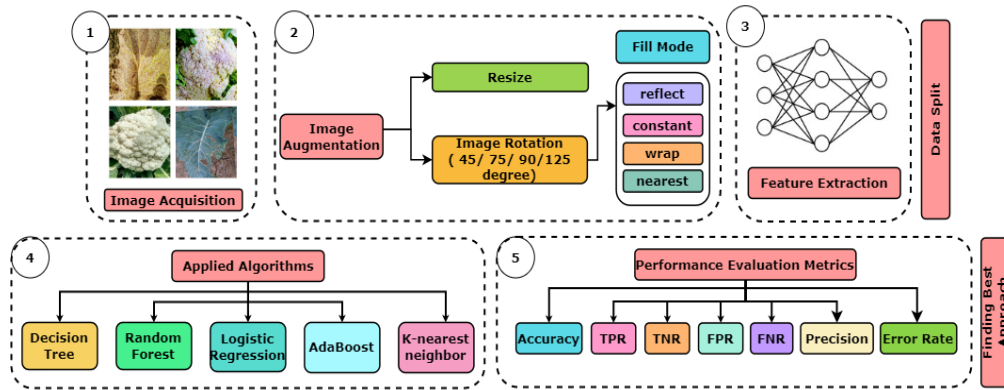


Figure 2. Working procedure of our proposed work

3.3. An evaluation of features using collective understanding

Different approaches are used for feature selection, such as the ANOVA correlation coefficient and the mutual information-based method [20]. This research was likewise conducted using a numerical input and a categorized output methodology. So, to rank the features used to diagnose cauliflower disease, we used mutual information and a target variable. Graphical representation of [21] mutual information between the two variables P1 and P2 (22). $r(l_1, l_2)$ is the joint probability distribution function.

$$I(P_1; P_2) = \sum_{l_2 \in L_2} \sum_{l_1 \in L_1} r(l_1, l_2) \log \left(\frac{r(l_1, l_2)}{r(l_1)p(l_2)} \right) \tag{30}$$

Mutual information details of the target variable represent in Table 2.

Table 2. The mutual information value of features

Rank	Name of features	Score of mutual information	Rank	Name of features	Score of mutual information
1	Entropy	0.13769223	6	Homogeneity	0.08052413
2	Mean	0.109962	7	Kurtosis	0.07979244
3	Standard deviation	0.10836427	8	Skewness	0.07792927
4	Contrast	0.1057053	9	Variance	0.07106182
5	Correlation	0.09430378	10	Energy	0.06997779

3.4. Choosing the most important N-features and carrying out synthetic minority oversampling technique

We have chosen the best N characteristics ($5 \leq N \leq 10$) based on the ranking. We divide the dataset again into a training set and a test set using the extracted N features. The model is trained using 80% of the data and then tested using the remaining 20%. The training set is very unbalanced based on these slices. Unbalanced datasets are not suitable for use in ML models. Hence, we used a method called the SMOTE. Using this method, the problem of the difference between classes can be fixed by making samples of the minority group.

3.5. Splitting up datasets

The extracted features of the dataset are divided into a train set comprising 80% of the data and a test set comprising 20% of the data. A significant proportion of the data is employed to train our model with this division. Afterward, the efficacy of the model is evaluated employing various classifiers on the test set.



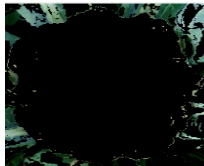





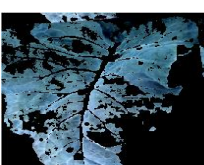



3.6. Selected classifier for cauliflower disease identification

Five machine learning classifiers, namely DT, RF, LR, AdaBoost, and KNN were utilised to identify maladies in cauliflower. These classifiers were utilised on augmented and unaugmented datasets. Their primary purpose is to accurately identify various cauliflower diseases.

4. RESULTS AND DISCUSSION

Since the farmer's smartphone and other hand-held devices will be used to take the sample images, the data we get to put into our model will be different in terms of size, viewing angle, and asymmetry. We changed the size of the original user-submitted image to work better for people from a wide range of situations and backgrounds. After that, the image is scaled down to a standard size of 300×300 pixels. We considered the wide variety of mobile device forms before settling on this standard size. Contrast mapping is employed to enhance the clarity of images. Then the resized images are segmented into 3 clusters using the k-means cluster algorithm. Feature extraction is a vital step for image-based classification. The quality of the segmented images determines the quality of the featured images. Cauliflower images are segmented, then the ten features are derived. The entire process of feature extraction is depicted in Table 3. We measure the performance of different applied ML algorithms using several performance evaluation matrices: accuracy, true positive rate (TPR), true negative rate (TNR), FPR, FNR, error rate, and precision. The model's effectiveness in identifying cauliflower diseases using various feature sets has been evaluated and compared in this study.

Table 3. Extraction procedure of captured cauliflower images

Selected class	Capture image	Contrast enhancement	Segmented image	Feature extracted value
Bacterial spot				0.132, 0.872, 0.801, 0.977, 24.433, 1.165, 4.423, 526.335, 18.743, 3.951
Downy mildew				2.709, 0.674, 0.348, 0.835, 70.138, 3.554, 9.090, 4.213, 2.519, 1.04
Black rot				1.048, 0.847, 0.273, 0.893, 59.859, 70.778, 4.450, 10.702, 1.990, 0.663
Disease free				2.659, 0.877, 0.353, 0.913, 111.394, 4.284, 10.460, 1.071, 1.199, 0.373

In order to analyze the effects of feature selection and augmentation strategy, we took the top ten features from the ranking and split them into four groups: 5, 7, 9, and 10. After that, we applied different ML classifiers on augmented and nonaugmented datasets. Tables 4 and 5 show how various ML models perform using a variety of feature sets with both selected datasets.

Table 4. Performance evaluation of nonaugmented data

Number of features	Model name	Accuracy (%)	TPR (%)	TNR (%)	FPR (%)	FNR (%)	Precision (%)	Error rate (%)
For top 5 features	DT	75.71	53.85	88.64	11.36	46.15	73.68	24.29
	RF	81.16	59.09	91.49	8.51	40.91	76.47	18.84
	LR	81.72	89.66	78.13	21.88	10.34	65	18.28
	AdaBoost	72.31	53.33	88.57	11.43	46.67	80.00	27.69
	KNN	88.89	73.91	95.92	4.08	26.09	89.47	11.11
For top 7 features	DT	82.19	69.57	88	12	30.43	72.73	17.81
	RF	85.71	73.91	91.50	8.51	26.09	80.95	14.29
	LR	87.30	68.75	93.62	6.38	31.25	78.57	12.69
	AdaBoost	75.76	56.67	91.67	8.33	43.33	85.00	24.24
	KNN	90.67	80	96	4	20	90.91	9.33
For top 9 features	DT	85.00	94.44	79.69	20.31	5.56	72.34	15.00
	RF	85.92	75.00	91.49	8.51	25.00	81.81	14.08
	LR	81.25	90.63	76.56	23.44	9.38	65.91	18.75
	AdaBoost	70.53	70.21	70.83	29.17	29.79	70.21	29.47
	KNN	84.93	69.23	93.62	6.38	30.77	85.71	15.07
For top all features	DT	80.60	59.09	91.11	8.89	40.91	76.47	19.40
	RF	87.5	79.17	91.67	8.33	20.83	82.61	12.5
	LR	81.25	90.63	76.56	23.44	9.38	65.91	18.75
	AdaBoost	67.02	75.86	63.08	36.92	24.14	47.83	32.98
	KNN	85.14	69.23	93.75	6.25	30.76	85.71	14.86

Table 5. Performance evaluation of augmented data

Number of features	Model name	Accuracy (%)	TPR (%)	TNR (%)	FPR (%)	FNR (%)	Precision (%)	Error rate (%)
For top 5 features	DT	85.29	85.71	84.91	15.09	14.29	84.00	14.71
	RF	87.88	97.30	82.26	17.74	2.70	76.59	12.12
	LR	80.85	93.55	74.60	25.40	6.45	64.44	19.15
	AdaBoost	75.53	75.00	75.93	24.07	25.00	69.77	24.47
	KNN	82.83	86.05	80.36	19.64	13.95	77.08	17.17
For top 7 features	DT	83.00	84.44	81.82	18.19	15.56	79.17	17.00
	RF	88.12	93.33	83.93	16.07	6.67	82.35	11.88
	LR	89.23	82.35	91.67	8.33	17.65	77.78	10.77
	AdaBoost	75.36	56.67	89.74	10.26	43.33	80.95	24.64
	KNN	87.13	95.00	81.97	18.03	5.00	77.55	12.87
For top 9 features	DT	88.57	80.95	91.84	8.16	19.05	80.95	11.43
	RF	87.13	86.54	87.75	12.24	13.46	88.24	12.87
	LR	90.77	78.95	95.65	4.35	21.05	88.24	9.23
	AdaBoost	76.47	61.76	91.18	8.82	38.24	87.50	23.53
	KNN	88.35	91.49	85.71	14.29	8.51	84.31	11.65
For top all features	DT	85.00	90.24	81.36	18.64	9.75	77.08	15.00
	RF	88.0	89.80	86.27	13.73	10.20	86.27	12.00
	LR	90.47	78.95	95.45	4.55	21.05	88.24	9.52
	AdaBoost	64.95	78.57	59.42	40.58	21.43	44.00	35.05
	KNN	87.50	91.30	84.48	15.52	8.70	82.35	12.50

When considering nine features (entropy, mean, standard deviation, contrast, correlation, homogeneity, kurtosis, skewness and variance) we noticed that the LR classifier resulted in the highest accuracy with augmented images, which is 90.77%, where 78.95%, 95.65%, 4.35%, 21.05%, 88.24%, and 9.23% are the TPR, TNR, FPR, FNR, error rate, and precision, respectively. In conclusion, statics are preferable to GLCM. After that, we visualize the performance of the applied classifier using a bar diagram based on top-ranked features in Figure 3. The top 5 features of accuracy are shown in Figure 3(a) and the top 7 features are presented in Figure 3(b). As the same as Figure 3(c), depict the top 9 and top 10 features of accuracy visualized in Figure 3(d).

Finally, we applied the ROC curve to compare the output quality from the augmented dataset to the nonaugmented dataset and to determine which classifier generated the better output. ROC curve comparison between the two sets of data is shown in Figure 4. Better classifier performance is often associated with a larger ROC curve area [22], [23]. Using the enhanced data and the best nine features, the LR classifier achieves a maximum area under the ROC of 93.34%, as shown in Figure 4(a). Alternatively, utilizing the top five features of nonaugmented data, the DT classifier achieved the lowest ROC value is 74.89% is visualized in Figure 4(b).

Comparative analysis is an essential part of the research. It helps a researcher find the research gap and make a new way to solve the problem efficiently [24]. A large number of research articles are available on agro-based systems. Table 6 represents the comparative analysis with other existing work.

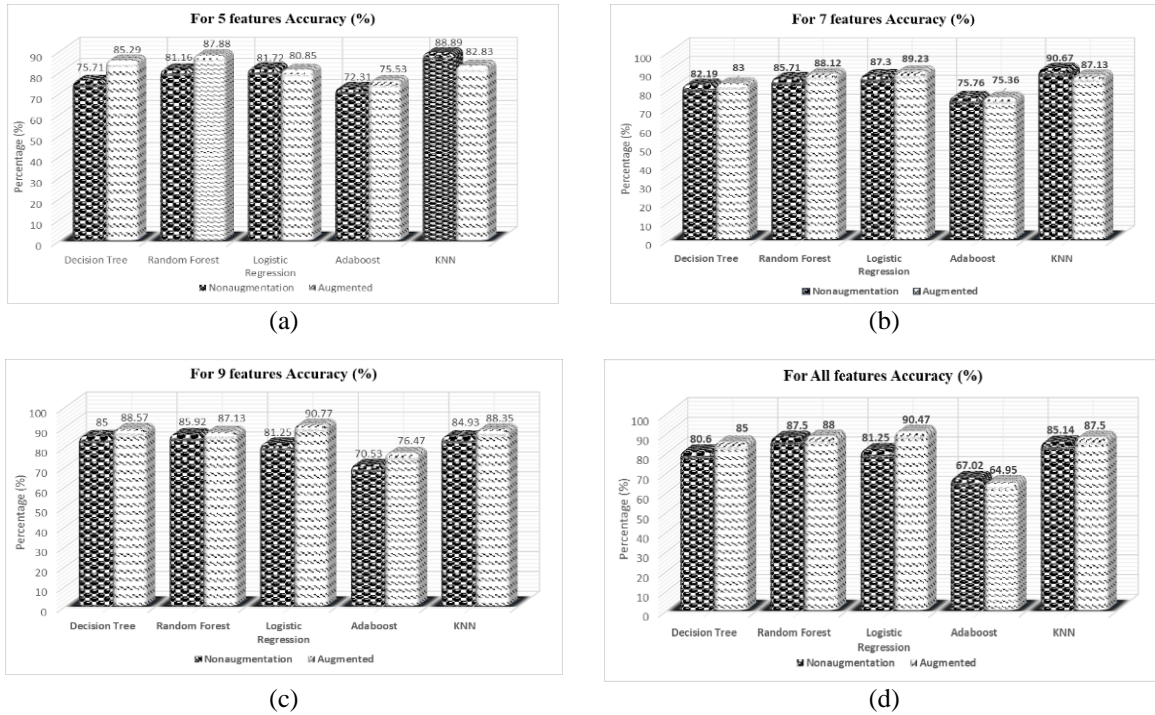


Figure 3. Comparison analysis between five to all features for; (a) 5 features, (b) 7 features, (c) 9 features, and (d) all features accuracy

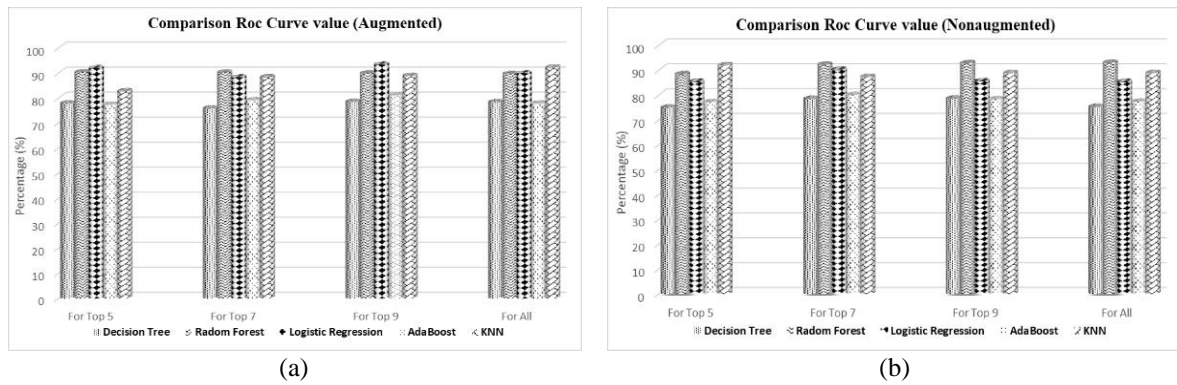


Figure 4. Comparison between ROC curve; (a) augmented and (b) nonaugmented

Table 6. Comparative analysis with other existing work

Related work	Object	Type	Dataset size	Segmented algorithm	Applied classifier	Accuracy (%)
Sari <i>et al.</i> [7]	Papaya	Detection	50	N/A	NB	88.00
Panigrahi <i>et al.</i> [13]	Maize	Classification	3,823	N/A	RF	79.23
					KNN	76.16
					SVM	77.56
					NB	77.46
					DT	74.35
Behera <i>et al.</i> [25]	Orange	Classification	N/A	N/A	SVM	90.00
Jaisakthi <i>et al.</i> [26]	Grape	Identification	5,675	N/A	SVM, AdaBoost, RF	Average: 93.03
Pulido <i>et al.</i> [27]	Weed	Recognition	320	K-means clustering	SVM	90
Mia <i>et al.</i> [28]	Mango	Recognition	8	K-means clustering	SVM	80
Proposed work	Cauliflower	Recognition	1,920	K-means clustering	DT	88.57
					RF	87.13
					LR	90.77
					AdaBoost	76.47
					KNN	88.35

5. CONCLUSION

Extensive research has been conducted on an agro-medical expert system based on machine vision, with a focus on cauliflower. We extract 10 features from cauliflower images by using the k-means clustering method. The mutual information-based selection method is then applied to rank the features. After selecting the top N features and applying five ML classifiers to train and test the dataset, we have used SMOTE to maintain balance in the data. We are confident that our system operates very well across the area. Using the top nine features, our model achieved the highest accuracy of 90.77% with a LR classifier, which is to say that it is outstanding and has amazing potential. In the future, we will be doing a lot more work on recognizing cauliflower disease, and a big part will be leveraging big data to detect different kinds of cauliflower disorders. In addition to tomatoes, cucumbers, carrots, and cabbage. This technology has broad applicability in agriculture.




REFERENCES

- [1] "Bangladesh: Employment in agriculture," *The Global Economy*, 2021. [Online]. Available: https://www.theglobaleconomy.com/Bangladesh/Employment_in_agriculture/. (accessed: Oct. 17, 2022).
- [2] M. Hossain and M. Islam, "Use of artificial intelligence for precision agriculture in Bangladesh," *Journal of Agricultural and Rural Research*, vol. 6, no. 2, pp. 81–96, 2022.
- [3] "Cauliflower," *Banglapedia: National Encyclopedia of Bangladesh*, 2021. [Online]. Available: <https://en.banglapedia.org/index.php?title=Cauliflower>. (accessed: Oct. 21, 2022).
- [4] J. M. -Castro, A. d. H. -Bailón, S. O. -Cano, I. M. G. Magdaleno, A. M. Ortega, and F. C.-Martos, "Bioaccessibility of glucosinolates, isothiocyanates and inorganic micronutrients in cruciferous vegetables through INFOGEST static in vitro digestion model," *Food Research International*, vol. 166, pp. 1–10, 2023, doi: 10.1016/j.foodres.2023.112598.
- [5] J. C. -González, M. C. Piñero, G. Otálora, J. L. -Marín, and F. M. D. Amor, "Merging heat stress tolerance and health-promoting properties: The effects of exogenous arginine in cauliflower (brassica oleracea var. botrytis l.)," *Foods*, vol. 10, no. 1, pp. 1–13, 2021, doi: 10.3390/foods10010030.
- [6] S. Sasirekha and K. B. Suganthy, "An Approach for Detection of Disease in Carrot using K-Means Clustering," *International Journal of Research in Engineering, Science and Management*, vol. 2, no. 2, pp. 527–530, 2019.
- [7] W. E. Sari, Y. E. Kurniawati, and P. I. Santosa, "Papaya Disease Detection Using Fuzzy Naïve Bayes Classifier," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2020, pp. 42–47, doi: 10.1109/ISRITI51436.2020.9315497.
- [8] Q. Gu *et al.*, "Early detection of tomato spotted wilt virus infection in tobacco using the hyperspectral imaging technique and machine learning algorithms," *Computers and Electronics in Agriculture*, vol. 167, 2019, doi: 10.1016/j.compag.2019.105066.
- [9] A. Chaudhary, R. Thakur, S. Kolhe, and R. Kamal, "A particle swarm optimization based ensemble for vegetable crop disease recognition," *Computers and Electronics in Agriculture*, vol. 178, 2020, doi: 10.1016/j.compag.2020.105747.
- [10] J. Kianat, M. A. Khan, M. Sharif, T. Akram, A. Rehman, and T. Saba, "A joint framework of feature reduction and robust feature selection for cucumber leaf diseases recognition," *Optik*, vol. 240, 2021, doi: 10.1016/j.ijleo.2021.166566.
- [11] M. A. Islam, M. S. Islam, M. S. Hossen, M. U. Emon, M. S. Keya, and A. Habib, "Machine Learning based Image Classification of Papaya Disease Recognition," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1353–1360, doi: 10.1109/ICECA49313.2020.9297570.
- [12] M. T. Habib, A. Majumder, A. Z. M. Jakaria, M. Akter, M. S. Uddin, and F. Ahmed, "Machine vision based papaya disease recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 3, pp. 300–309, 2020, doi: 10.1016/j.jksuci.2018.06.006.
- [13] K. P. Panigrahi, H. Das, A. K. Sahoo, and S. C. Moharana, "Maize Leaf Disease Detection and Classification Using Machine Learning Algorithms," *Advances in Intelligent Systems and Computing*, vol. 1119, pp. 659–669, 2020, doi: 10.1007/978-981-15-2414-1_66.
- [14] A. Rajbongshi, M. E. Islam, M. J. Mia, T. I. Sakif, and A. Majumder, "A Comprehensive Investigation to Cauliflower Diseases Recognition: An Automated Machine Learning Approach," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 12, no. 1, pp. 32–41, 2022, doi: 10.18517/ijaseit.12.1.15189.
- [15] N. R. Methun, R. Yasmin, N. Begum, A. Rajbongshi, and M. E. Islam, "Carrot Disease Recognition using Deep Learning Approach for Sustainable Agriculture," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, pp. 732–741, 2021, doi: 10.14569/IJACSA.2021.0120981.
- [16] U. Sara, A. Rajbongshi, R. Shakil, B. Akter, and M. S. Uddin, "VegNet: An organized dataset of cauliflower disease for a sustainable agro-based automation system," *Data in Brief*, vol. 43, pp. 1–8, 2022, doi: 10.1016/j.dib.2022.108422.
- [17] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. India: Pearson Addison-Wesley, 2006.
- [18] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. United States: Morgan Kaufmann, 2012, doi: 10.1016/C2009-0-61819-5.
- [19] P. Kulkarni, A. Karwande, T. Kolhe, S. Kamble, A. Joshi, and M. Wyawahare, "Plant Disease Detection Using Image Processing and Machine Learning," *Arxiv-Computer Science*, vol. 1, pp. 1–13, 2016, doi: 10.48550/arXiv.2106.10698.
- [20] J. Brownlee, *Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data*. Machine Learning Mastery, 2020.
- [21] Y. Wu, B. Liu, W. Wu, Y. Lin, C. Yang, and M. Wang, "Grading glioma by radiomics with feature selection based on mutual information," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 5, pp. 1671–1682, 2018, doi: 10.1007/s12652-018-0883-3.
- [22] R. Shakil, B. Akter, A. Rajbongshi, U. Sara, M. R. Barman, and A. Dhali, "A Transfer Learning Approach to the Development of an Automation System for Recognizing Guava Disease Using CNN Models for Feasible Fruit Production," in *Hybrid Intelligent Systems*, Cham: Springer, 2023, pp. 127–141, doi: 10.1007/978-3-031-27409-1_12.
- [23] P. Ghosh, F. M. J. M. Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2020, pp. 1–6, doi: 10.1109/iSAI-NLP51646.2020.9376787.




- [24] F. J. M. Shamrat, S. Azam, A. Karim, K. Ahmed, F. M. Bui, and F. D. Boer, "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images," *Computers in Biology and Medicine*, vol. 155, pp. 1–14, 2023, doi: 10.1016/j.compbiomed.2023.106646.
- [25] S. K. Behera, L. Jena, A. K. Rath, and P. K. Sethy, "Disease Classification and Grading of Orange Using Machine Learning and Fuzzy Logic," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, 2018, pp. 0678–0682, doi: 10.1109/ICCSP.2018.8524415.
- [26] S. M. Jaisakthi, P. Mirunalini, D. Thenmozhi, and Vatsala, "Grape Leaf Disease Identification using Machine Learning Techniques," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1–6, doi: 10.1109/ICCIDS.2019.8862084.
- [27] C. Pulido, L. Solaque, and N. Velasco, "Weed recognition by SVM texture feature classification in outdoor vegetable crop images," *Ingenieria e Investigacion*, vol. 37, no. 1, pp. 68–74, 2017, doi: 10.15446/ing.investig.v37n1.54703.
- [28] M. R. Mia, S. Roy, S. K. Das, and M. A. Rahman, "Mango leaf disease recognition using neural network and support vector machine," *Iran Journal of Computer Science*, vol. 3, no. 3, pp. 185–193, 2020, doi: 10.1007/s42044-020-00057-z.

BIOGRAPHIES OF AUTHORS






Rashiduzzaman Shakil    is studying his bachelor of science (B.Sc.) degree in the Department of Computer Science and Engineering at Daffodil International University, Bangladesh. He predominantly works on machine learning, deep learning, and image processing. He has been a collaborator in collaborative research projects with researchers from Bangladesh and Australia. His some of research paper were published in the prestigious journals (Scopus) and conferences (Scopus). He can be contacted at email: rashiduzzaman.diucse@gmail.com.






Bonna Akter    is a B.Sc. undergraduate student in the Department of Computer Science and Engineering at Daffodil International University, Bangladesh. She has been involved in cooperative research activities with researchers from Bangladesh and Canada. Specially, her research areas are machine learning, deep learning, and image processing. She is an expert at writing scientific research articles. Her several research articles were published in the renowned conference (Scopus) and Journals (Scopus). She can be contacted at email: bonna.diucse@gmail.com.



F M Javed Mehedi Shamrat    is currently pursuing the master of computer science in the Department of Computer System and Technology at the University of Malaya, Kuala Lumpur, Malaysia. He is a formal research associate at Daffodil International University. Also, he worked as a lecturer in the Department of Computer Science and Engineering at the European University of Bangladesh. Prior to joining Daffodil International University. He completed his B.Sc. in the Department of Software Engineering from Daffodil International University. He has achieved the best student research award for the Department of Software Engineering from Daffodil International University. Also, he has enlisted among the researchers in Bangladesh on AD science index. He has more than 55 publications in IEEE, Springer, Elsevier, and PubMed-indexed journals. His primary research interest includes the intersection of the internet of things, deep learning, data science, image processing, neural networks, artificial intelligence, bio-informatics, and machine learning. He is also an associate member of the Bangladesh Computer Society. He can be contacted at email: javedmehedicom@gmail.com.



Sheak Rashed Haider Noori    is the department associate head and a professor in the Department of Computer Science and Engineering at Daffodil International University in Dhaka, Bangladesh. His areas of expertise in study include AI, gamification, distributed systems, information and knowledge management, data mining, and pervasive and mobile computing. His several research articles have been published in prestigious journals and conferences. He can be contacted at email: drnoori@daffodilvarsity.edu.bd.