

A Comprehensive Investigation to Cauliflower Diseases Recognition: An Automated Machine Learning Approach

Aditya Rajbongshi^{a,*}, Md. Ezharul Islam^a, Md. Jueal Mia^a, Tahsin Islam Sakif^b, Anup Majumder^a

^a Department of Computer Science and Engineering, Jahangirnagar University, 1342, Dhaka, Bangladesh

^b Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, United States

Corresponding author: *adityaraj.jucse@gmail.com

Abstract—Vegetables, a significant part of agriculture, are necessary for the general good health of human beings. The use of information technology can help vegetable farmers to reach high yields which can contribute to global food security and sustainable cultivation. Cauliflower (*Brassica oleracea var. botrytis*) is a popular vegetable that is easily affected by various diseases causing loss of production and quality. However, machine learning-based disease recognition has yet to be developed for cauliflowers which can help farmers to identify cauliflower diseases and enable them to take timely actions. In this paper, an online machine vision-based expert system for recognizing cauliflower diseases is proposed, where a captured image via a smartphone or handheld gadget is processed and then classified to identify disease to assist the cauliflower farmers. Based on the feature extraction, the system classifies four types of diseases namely ‘bacterial soft’, ‘white rust’, ‘black rot’, and ‘downy mildew’ in cauliflowers. A total of 776 images are utilized to implement this experiment. K-means clustering algorithm has been applied on captured images to segment the disease-affected regions before two-type features extraction namely statistical and co-occurrence feature. Six classification algorithms namely BayesNet, Kstar, Random Forest, LMT (Logistic Model Tree), BPN (Back propagation neural network), and J48 were used for disease classification, and we evaluated their performance using seven performance metrics. We found the Random Forest classifier outperforms all other classifiers for cauliflower disease recognition with accuracy approaching 89.00%.

Keywords— Agriculture technology; cauliflower; disease recognition; machine learning; random forest classifier.

Manuscript received 10 May 2021; revised 26 Aug. 2021; accepted 22 Sep. 2021. Date of publication 28 Feb. 2022.
IJASEIT is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



I. INTRODUCTION

The agriculture sector plays a vital role in providing food, revenue, and jobs to rural people, especially in developing countries. The contribution to world economic production from this sector is about 6.4% [1]. According to a survey in 2016, 65% percent of poor adults make a living through agriculture [2]. Moreover, agriculture is the source of many consumers demanded agricultural commodity markets, especially in rural areas. Therefore, it is important to have a worthwhile and feasible agricultural system to ensure stable food security for people. To sustain that condition, proper management of the agricultural production system is necessary.

Most farmers in rural areas are not experts in timely detecting and solving problems in their growing crops. They often fail to achieve the desired yield of their cultivated crops due to damage caused by various diseases. The production of any crop such as vegetables could be significantly increased

by timely and accurately detecting the farmers’ field’s diseases. Quick, accurate detection of the disease is a prerequisite for taking proper management options to control any diseases to ensure higher yield.

Cauliflower is a seasonal crop mainly cultivated in the farmland by the farmers, as shown in Fig. 1. It is a vegetable crop under the *Brassicaceae* family. It is rich in fiber and B vitamins [3]. It contains phytonutrients that can reduce the risk of cancer. As a cruciferous vegetable, it supplies fiber to lower the chance of cardiovascular problems and contains choline, a nutrient essential for helping with sleep, muscle movement, learning, and memory. Cauliflower is cultivated in many countries globally, such as China, India, the USA, Spain, Mexico, and Bangladesh. In India, cauliflowers’ acreage and annual production are about 2.5 lac hectares and 7,887,000 m tons, respectively [4].

On the other hand, acreage and annual cauliflower production in Bangladesh are about 9,400 ha and 73,000 m tons, respectively [5]. During cauliflower cultivation, it is infected by various diseases, namely bacterial soft rot, black

leg, black rot, downy mildew, powdery mildew, ring spot, white rust, etc. The growth and yield of cauliflower are significantly affected by these diseases. Early detection of a specific disease and application of the right control measure by the farmers are needed to increase the yield and profitability of cauliflower cultivation. Developing an automated cauliflower disease recognition system would greatly help farmers timely detect the disease and ensure a higher yield of cauliflower. Therefore, the development of an expert system is needed for the classification of cauliflower diseases.



(a)



(b)

Fig. 1 (a) Healthy cauliflower; (b) Cauliflower cultivation in farmland

The issue of machine vision-based disease or defect recognition can be disintegrated into two areas, specifically disease recognition and disease classification. Some research has been confined either in disease detection or disease classification whereas. Some others have been comprised both in disease detection and classification. Some exertions for automated recognition on various diseases in fruits and vegetables have been performed, namely apple, papaya, tomato, potato, and so on.

A framework for recognizing the application of machine learning to fruit and vegetable species was proposed by Dubey *et al.* [6]. For image segmentation, they used *k*-means

clustering and Multi-class Support Vector Machine for classification and training. In extraction, various features and colors are extracted. They only centered on species and variety recognition of fruits and vegetables. Krithika *et al.* [7] proposed a technique for detecting leaf diseases on cucumber applying Multi-class Support Vector Machine. They used *k*-means clustering for the segmentation and image processing techniques to extract the feature. The size of the dataset and obtained accuracy did not mention here. A CNN model is applied to the technique of cucumber infections diagnostic [8]. A total 48, 311 images were utilized, and obtained average accuracy was 95.5%. Similarly, Samajpati *et al.* [9] proposed an approach for detecting fruit disease, where a random forest classifier is used for classification. Only 70 images were utilized for implementation. Pulido *et al.* [10] presented a system for the classification of weeds and vegetables. The SVM classifier was applied for classification, and the gained accuracy was 90%. Wahab *et al.* [11] presented a method for detecting the diseases of chollis leaves using SVM. They used *k*-means clustering for segmentation and SVM for classification. The obtained accuracy was 57.1%. A technique for the recognition of Rose diseases is performed using the MobileNet model [12]. A total of 2000 images of four classes of diseases were utilized, and the obtained average accuracy was 95.63%.

On the other hand, Islam *et al.* [13] developed an automated method for potato disease detection. They applied an approach for segmentation and MSVM for the classification of potato disease. The dataset contained 300 images of 3 types of affected disease. They achieved 95% accuracy. Oppenheim *et al.* [14] demonstrated a sliding window classification in CNN to classify potato disease. They utilized 400 images for building a trained model and achieved accuracy from 80% to 90%.

A method was performed for the recognition of paddy diseases [15]. The KNN classifier was used to classify the diseases, and the obtained accuracy was 75.61%. Kurniawati *et al.* [16] have developed a prototype device with imaging techniques for the diagnosis of paddy diseases. They applied two methods such as OSTU, Threshold, and obtained 94.7% accuracy. Prajapati *et al.* [17] also introduced a prototype system for rice disease detection and classification. They worked with three diseases of rice plants. *K*-means clustering was applied for segmenting the disease and MSVM for classification. They obtained an accuracy rate of 93.33% and 73.33% on training and testing datasets, respectively. An approach for rice disease identification based on CNN was introduced by Lu *et al.* [18]. Five hundred images were adopted for building the trained model. They obtained 95.48% accuracy. Ferentinos [19] also developed a deep learning-based model for detecting the disease of the plant. They applied CNN models for training 87,848 images and accomplished an accuracy of 99.53%.

Jiang *et al.* [20] proposed a technique to classify the diseases of tomato leave using the ResNet-50 model. A total of 2700 images were utilized, and the obtained accuracy was 98.0%. On the other hand, Ashqar *et al.* [21] built a trained model using a deep CNN for detecting the tomato diseases. They used 9000 images for implementing the classification and achieved 99.84% accuracy. Durmus *et al.* [22] applied two network architectures of deep learning named AlexNet

and SqueezeNet for tomato leaves disease detection. They utilized ten disease categories of image data for training and testing the model. They obtained 95.65% accuracy for AlexNet and 94.3% for SqueezeNet. A system for recognition of papaya disease was developed by Habib *et al.* [23]. K-means algorithm was applied for segmentation and SVM for classification of disease. They practiced 500 images of papaya and achieved 90% accuracy.

In this study, we performed an exploratory analysis on cauliflower disease recognition applying the machine vision approach. Our proposed framework is an online machine learning-based system that can capture images by smartphones or handheld gadgets and utilize them as input. Then it recognizes four cauliflower diseases and provides instant feedback to the users. For cauliflower disease recognition, a *k*-means clustering algorithm was utilized on disease-affected images for segmenting the affected regions. Based on the results, a set of features were extracted applying image processing techniques. Then six prominent classification algorithms were used for training and testing to recognize the diseases. The performance of these algorithms was compared based on seven performance metrics. The main contribution of our research is as follows:

- The main attempt is to recognize cauliflower diseases in automated way.
- Systematically arrangement of the best solid feature for training and testing the classifiers in order to classify the diseases of cauliflower.
- An exhibit that our proposed approach conveys significant performance on the image data of cauliflower diseases.

II. MATERIALS AND METHOD

A. System Architecture

The system architecture of an online machine vision-based agro-medical expert system for recognizing cauliflower diseases is exhibited in Fig. 2. First, the system captured disease-affected or disease-free images through a smartphone application. Second, the captured images are transmitted to the backend core expert system over a cellular or Wi-Fi network. Third, the backend expert system analyzed the image through segmentation and feature extraction using its knowledge base, inference engine, and algorithms to recognize the disease (if any). Lastly, the result was dispatched back to the user through the network and shown on the user's device. The front-end software completed this.

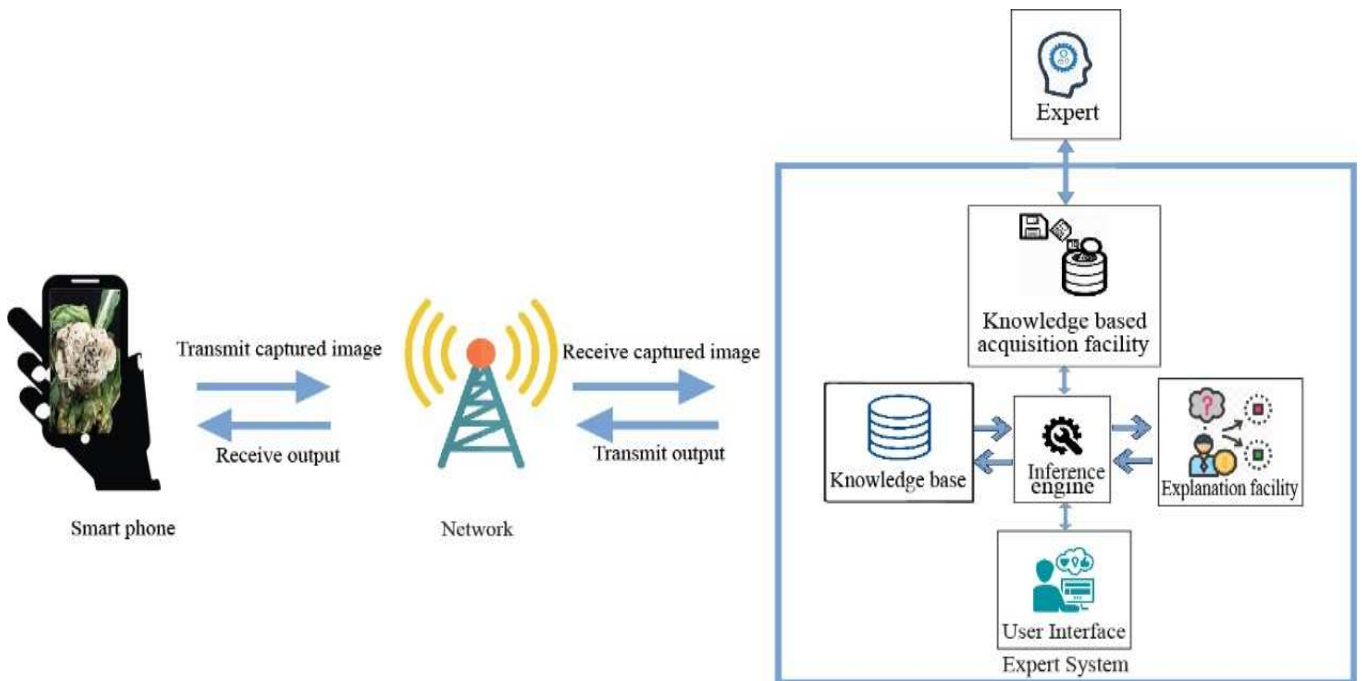


Fig. 2 Proposed architecture for cauliflower agro-medical expert system

B. Proposed Methodology

To clarify the methodology, the step-by-step working procedure of the proposed system has been discussed. After acquiring the image, this system resizes them into predestined size, then exacerbates their contrast using histogram equalization and converts their color space from RGB to

L^*a^*b . After preprocessing, this system segments the images using the *k*-means clustering algorithm, extracts their co-occurrence and statistical features, and then trains them with six prominent classifiers to finally generate the output, as shown in Fig. 3. The detailed descriptions of the above-mentioned steps are discussed in the following subsections.

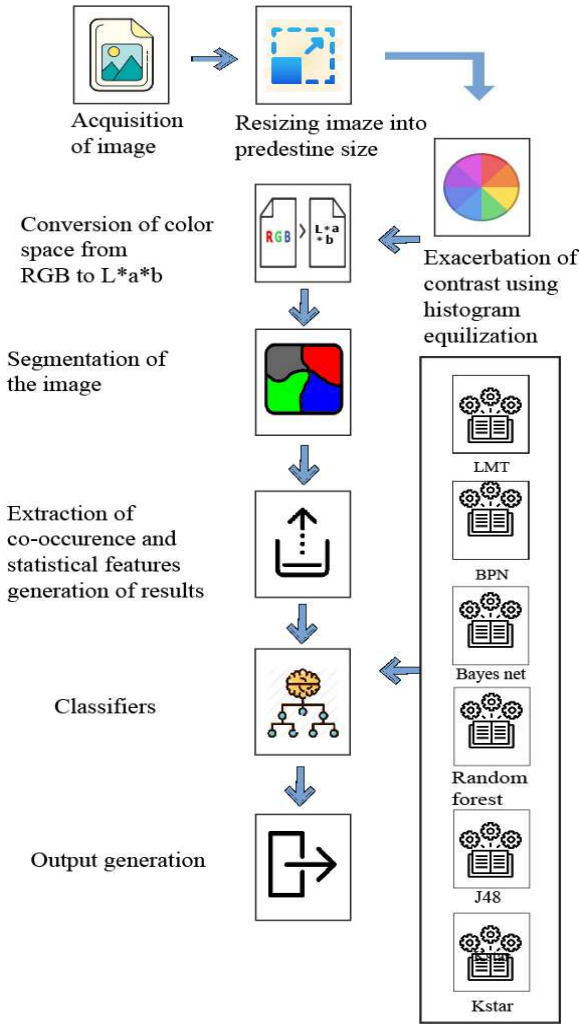


Fig. 3 Working procedure for cauliflower disease recognition.

1) *Acquisition of image*: In this system, the preprocessing phase outset with numerous images of cauliflowers. We collected a total of seven hundred and sixty-six (766) color images of cauliflower (disease-affected or disease-free). Among them, six hundred and sixty (660) images are collected locally, and one hundred and six (106) images are from the Internet.

2) *Resizing image into predestine size*: Bicubic interpolation is used to transfer the image into a predestine size image [24]. As the feature extraction varies on different image sizes, cauliflower images are resized into 300×300 pixels. Suppose the intensity value is I and derivatives are f_x , f_y , and f_{xy} . These values are inflicted for recognizing the four corners of a unit square which are (1, 1), (1, 0), (0, 1), and (0, 0). The intensity of the interpolation surface can be written as follows:

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

Where m_{ij} are coefficients.

3) *Exacerbation of contrast using histogram equalization*: The image contrast is intensified by utilizing the histogram equalization. Suppose P and Q are the row number (height) and column (width) in pixels, respectively. P_k is the color intensity of the pixel number n_k , and M is the intensity level of the image, which is measured by number. Then a mapping

method is performed between Color intensity P_k and each pixel intensity Q_k using the following equation.

$$Q_k = T(P_k) = \frac{M-1}{PQ} \sum_{j=0}^k n_j \quad (2)$$

Where $k=0, 1, 2, 3 \dots M-1$.

4) *Conversion of color space from RGB to L*a*b*: The conversion of RGB is exerted after contrast intensifies. This conversion is done mainly for achieving L*a*b color space. The reason to convert the image into L*a*b color space is that the converted image remains exactly the same. There is no chance to degeneration the image quality as L*a*b comprises all feasible colors. That is why it helps better for segmenting the images in k-means clustering [24]. At first, the RGB color space is converted to CIE, which is then converted to L*a*b color space.

For converting to CIE (XYZ) color space from RGB color space, the following equation is used

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 3.240479 & -1.537150 & -0.498535 \\ -0.969256 & 1.875992 & 0.041556 \\ 0.055648 & -0.204043 & 1.057311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3)$$

For transforming L*a*b color space, we can imagine the tri-stimulus values. Here P_w , Q_w , and Z_w are reference white of those values. We can assume more that

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

For the calculation of L*a*b, the following equation is used

$$L^* = \begin{cases} 116 \left(\frac{Q}{Q_w}\right)^{\frac{1}{3}} - 16 & \text{if } \frac{Q}{Q_w} > 0.00885 \\ 903.3 \left(\frac{Q}{Q_w}\right) & \text{if } \frac{Q}{Q_w} \leq 0.00885 \end{cases} \quad (5)$$

$$a^* = 500 \left(f\left(\frac{P}{P_w}\right) - f\left(\frac{Q}{Q_w}\right) \right) \quad (6)$$

$$b^* = 200 \left(f\left(\frac{Q}{Q_w}\right) - f\left(\frac{Z}{Z_w}\right) \right) \quad (7)$$

5) *Segmentation of the image*: The k-means clustering method is utilized for segmenting the images of cauliflower. This clustering algorithm is one of the most commonly used, where k represents the number of clusters. When the quality of segmentation is good, the features extraction performance is good. In this work, we have set $k=3$ for segmenting an image which means that the algorithm will identify 3 clusters in the image.

6) *Extraction of co-occurrence and statistical features*: After segmentation of disease-affected region, two types of features are extracted namely statistical and co-occurrence feature which explained in detail in the following sub-section C.

7) *Training with six classifiers*: The feature vectors that are generated from the segmentation were applied to classifier for training and testing. Six classifiers namely BayesNet, Back-Propagation Neural Network (BPN), J48, Radom Forest, Logistic Model Tree (LMT), and Kstar were selected among the off-the-shelf classifiers. These classifiers perused with testing dataset to generate numerous performance metrics to find the compatible classification model.

8) *Output generations*: To measure the performance of the classifiers, the testing dataset was used. An accurate matrix constructed from an imbalanced dataset was not suitable for finding or calculating the classification types i.e., it raised various observations for various classes that vary extensively. For evaluating the performance more accurately, some other metrics were used. For binary i.e., 2-class problems, the confusion matrix recounts the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). When the number of classes was more than 2 i.e., multi-class, the dimension of the confusion matrix was $n \times n$ ($n > 2$). In multi-class, there existed n rows, n columns and n^2 entries. The equation for confusion matrix (C) in multi-class is as follows:

$$C = [p_{ij}]_{n \times n} \quad (8)$$

For multi-class, the calculation of TPs, FPs, TNs and FNs cannot be calculated spontaneously. The following consequent equations are used to calculate those values for class i [23].

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

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

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

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

By applying this procedure, a 2×2 -dimension confusion matrix was formed. The confusion matrix was used to calculate the value of accuracy, precision, sensitivity, error rate, specificity, FNR (false negative rate) and FPR (false positive rate) of the expert system [25]. In this expert system, we have two types of data sets such as: training data set and testing data set. In the training dataset, maximum data were kept. The testing data were used to calculate the performance of this expert system focusing on evaluation metrics for calculation of accuracy, precision, sensitivity, specificity, FNR and FPR in percentage, the following equations are used.

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

$$\text{Precision} = \left(\frac{TP}{TP+FP} \right) \times 100\% \quad (14)$$

$$\text{Sensitivity} = \left(\frac{TP}{TP+FN} \right) \times 100\% \quad (15)$$

$$\text{Error rate} = \left(\frac{FP+FN}{TP+FP+FN+TN} \right) \times 100\% \quad (16)$$

$$\text{Specificity} = \left(\frac{TN}{FP+TN} \right) \times 100\% \quad (17)$$

$$\text{FNR} = \left(\frac{FN}{TP+FN} \right) \times 100\% \quad (18)$$

$$\text{FPR} = \left(\frac{FP}{FP+TN} \right) \times 100\% \quad (19)$$

The above equation from (13) to equation (19) is utilized to evaluate the overall performance of the classifiers. ROC (Receiver Operating Characteristic) curve was also employed to visualize the comparison among the relative performance of six classifiers.

C. Disease and Feature Description

1) *Description of Diseases*: Diseases are the major hindrances to growing bountiful cauliflower in recent times. Due to the lack of proper knowledge and guidelines, the farmers cannot make timely decisions. Our approach is to analyze cauliflower diseases, assist them in grasping the symptoms of disease, and apprehend proper clues. Many diseases are accountable for lowering the yield of cauliflower. In our work, visual symptoms of four major diseases of cauliflower were used for recognition, namely bacterial soft, white rust, black rot, and downy mildew (Fig. 4). A brief description of these diseases is given in the following subsections for a better understanding.

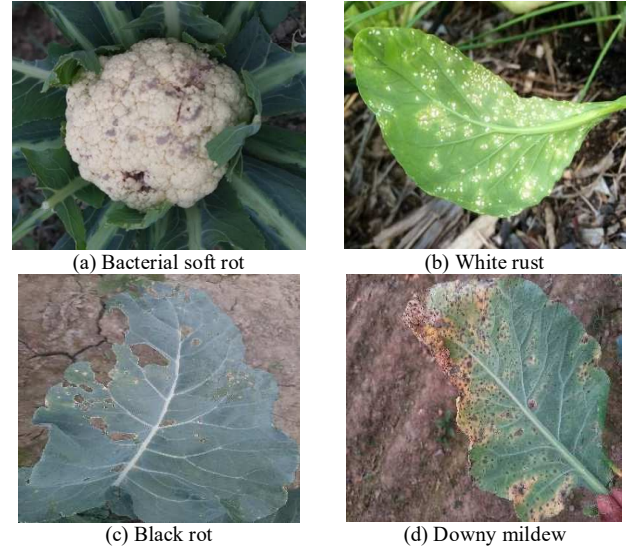


Fig. 4 Four most common diseases that affect Cauliflowers

Bacterial soft rot: Bacterial soft rot is caused by a bacterium, *Erwinia carotovora* subsp. *carotovora*, which results in various symptoms. The leaves and flower heads are injured by water-soaked spots that expand to form a large rotting mass. The injured portion of the leaves fractures and conveys slimy liquid, which turns into black or dark brown with the presence of air which is shown in Fig. 4(a). The emergence of bacterial soft rot depends on moist and warm conditions.

White rust: White rust is caused by the oomycete *Albugo candida*. It is a disease that causes white pustules on leaves and stems that cause various infections. The leaves become thick and curved, which is shown in Fig. 4(b). The emergence of white rust depends on dry conditions.

Black rot: Black rot is mainly caused by a bacterium, *Xanthomonas campestris* pv. *campestris*. V-shaped lesions are created in the middle area of leaves, which is shown in Fig.

4(c). The emergence of black rot depends on humid and warm conditions.

Downy mildew: Downy mildew is mainly caused by an oomycete *Peronospora parasitica*. It enters the plant through wounds and natural openings and first appears on older leaves, as white, yellow or brownish spots on the upper surfaces and downy grayish mold on the corresponding undersides (this eventually releases more spores). These spots eventually turn darker in color and the leaf dies (Fig. 4(d)). The emergence of downy mildew depends on moist and cool conditions.

2) Extraction of Features: Two types of feature vectors are extracted from this segmentation, namely GLCM and Statistical features. The quality of statistical features was utilized in textile defect-recognition [26]. To recognize the disease of cauliflower, some statistical features were utilized among numerous statistical features. The features that were utilized for this work are described below.

Mean (μ): If P is the number of the pixels in imperfect regions, Q is the number of the pixels in imperfect-free regions, and G is pixels number of color intensity with grayscale in the imperfect region, the equation of mean is written as follows:

$$\mu = \frac{1}{P} \sum_{i=1}^P G_i \quad (20)$$

Standard deviation (σ): If P is the number of the pixels in imperfect regions, G is pixels number of color intensity with

grayscale in the imperfect region, and Q is the color intensity in mean grayscale, the equation of standard deviation is written as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^P (G_i - Q)^2}{P}} \quad (21)$$

Variance (σ^2): If P is the pixels number in imperfect regions, G is pixels number of color intensity with grayscale in imperfect region, and Q is the color intensity in mean grayscale, the equation of variance is written as follows:

$$\sigma^2 = \frac{1}{P} \sum_{i=1}^P (G_i - Q)^2 \quad (22)$$

















Kurtosis (κ): If P is the number of the pixels in imperfect regions, G is pixels number of color intensity with grayscale in the imperfect region, and Q is the color intensity in mean grayscale, the equation of kurtosis is written as follows:

$$\kappa = \frac{\frac{1}{P} \sum_{i=1}^P (G_i - Q)^4}{\left(\frac{1}{P} \sum_{i=1}^P (G_i - Q)^2\right)^2} - 3 \quad (23)$$

Skewness (γ): If mode, mean, and standard deviation in imperfect regions for grayscale color intensity are σ , Q, and G respectively, the equation of skewness is written as follows:

$$\gamma = \frac{\sigma - G}{Q} \quad (24)$$

TABLE I
SEGMENTATION PROCEDURE OF THE ORIGINAL IMAGE

Disease name	Original image	Image resizing (300×300-pixel)	Images with contrast enhancement	Segmented image
Bacterial soft rot				
Black rot				
Downy mildew				
White rust				

Those statistical features are utilized for recognizing the disease of cauliflower. According to Habib *et al.* [26], GLCM features were also utilized for this purpose. For extracting textural features from pictures, these GLCM features were verified as efficient methods. Based on the relationship between two pixels, GLCM inflicts a measurement. This measurement is intensity variance in the interest's pixel. Suppose $f(m, n)$ is the digital image, which is also two-dimensional, the dimension of the image is $S \times T$ pixels, and the number of gray levels is L . Besides, consider two pixels $(x1, y1)$ and $(x2, y2)$ is in $f(m, n)$, the distance and angle between these pixels are r and ϕ respectively. Then according to the [26], the equation for GLCM, $P(i, j, r, \phi)$ is written as

$$P(i, j, r, \phi) = \{ \{(m_1, n_1), (m_2, n_2) \in S \times T : r, \phi, f(m_1, n_1) = i, f(m_2, n_2) = j\} \} \quad (25)$$

Five GLCM features have been utilized in this work, namely homogeneity (H), contrast (C), correlation (ρ), energy (E), and entropy (S). The corresponding equations of those features are given below

$$\text{Contrast: } C = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 P(i, j) \quad (26)$$

$$\text{Correlation: } \rho = \frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} i \cdot j \cdot P(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (27)$$

$$\text{Energy: } E = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i, j)^2 \quad (28)$$

$$\text{Entropy: } S = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i, j) \log P(i, j) \quad (29)$$

$$\text{Homogeneity: } H = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{P(i, j)}{1 + (i-j)^2} \quad (30)$$









Where μ_x, μ_y, σ_x and σ_y denotes summation values of row and column entries, respectively.

III. RESULTS AND DISCUSSION

A. Experimental Result Analysis

For the experimental analysis of our approach, the machine vision-based agro expert system was utilized, that exhibits in Fig. 2. To implement this system, at first disease-affected images of cauliflowers with different sizes and orientations were collected. These images are then sent to the expert system presented in Fig. 3. Then the perceived images were resized into 300×300 pixels. The mapping of color intensity was applied to enhance the contrast of images. Then the color images were segmented into three clusters. This segmentation was performed by using k-means clustering.

TABLE II
PROCEDURE OF CAULIFLOWER DISEASE PREDICTION WHILE TESTING THE DATASET

Actual disease	Acquisition image	k-means segmented image	Extraction of features	Predicted disease
Black rot			(1.00, 0.87, 0.26, 0.91, 64.90, 71.61, 4.61, 4229.55, 1.64, 0.47)	Black rot
Downy Mildew			(1.36, 0.82, 0.22, 0.86, 61.25, 69.09, 5.08, 4591.77, 2.79, 0.86)	Black rot
Bacterial soft rot			(0.66, 0.66, 0.66, 0.66, 0.66, 0.66, 0.66, 1779.06, 6.45, 2.04)	Bacterial soft rot
Disease-free			(0.48, 0.92, 0.24, 0.87, 62.07, 67.03, 4.71, 1708.43, 1.94, 0.55)	Downy Mildew

K-means clustering outperformed other algorithms which were utilized for segmentation [24]. Table 1 exhibits the procedure of segmentation of the original image. After completion of segmentation, we extracted feature vectors from each segmented image that were utilized for the training of the classifiers. Then we tested the proposed system using testing data. The system first took the image and performed the segmentation of that image. After segmentation, the feature vectors were extracted, which were compared to match the trained feature vectors, and then the system predicted the disease based on the analysis of similarities. Table 2 represents the procedure of prediction of four cauliflower diseases where two are predicted accurately, and two are not predicted accurately.

We collected a total of seven hundred and sixty-six (766) color images of cauliflower (disease-affected or disease-free). Six hundred and sixty (660) images were collected locally, and one hundred six (106) images from the Internet varied in resolutions for capturing by various handheld devices. The class-wise distribution of image data is exhibited in Table 3. This entire collected image data was then divided into two sections: training and testing. Five hundred and twenty-five (525) images were selected as training datasets, and two hundred and forty-one (241) images were for testing datasets. We used the cross-validation method for evaluation because this method provides less optimistic and biased estimation about the models. For the training purpose, each feature vector was utilized nine times, while each feature vector was utilized exactly one time for the testing purpose. According to the method, we applied ten-fold cross-validation where the error was reduced. The testing of the six classifiers was performed after the completion of the classifier's training. After completion of the testing procedure, 5×5 confusion matrices were generated by the six classifiers. Then for the computational purpose, we converted the obtained 5×5 confusion matrices to class-wise binary format.

TABLE III
DISTRIBUTION OF DATASET ACCORDING TO DISEASE CLASS

Class	Frequency
Bacterial Soft	185
Black Rot	128
White Rust	126
Downy Mildew	155
Disease-free	172
Total	766

TABLE IV
SPECIFICATIONS DETAILED FOR UTILIZED SIX CLASSIFIERS

Classifier name	Specifications
Random forest	Each bag Size= Data size training percentage=100% Tree height(Maximum)= unrestricted integer Randomly chosen attributes number=0. Network topology: Two-layer based on fully connected feedforward (10-15-5)
BPN	Sigmoid Function Rate of learning, $r=0.27$ Rate of momentum, $b=0.21$ Normal distribution: Variance, $\sigma^2=1$ Mean, $\mu=0$
BayesNet	Probability density function, $N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Probability category: Posterior class probability
LMT	Class Boundary: Hyperplane The probability function: $j^* = \operatorname{argmax}_j \Pr(G = j X = x)$
<i>K</i> -star	Transforming probability: Entropic measure Function for <i>K</i> -star: $K * \left(\frac{b}{a}\right) = -\log_2 p * \left(\frac{b}{a}\right)$
J48	Split tests arrangement: Subset Data Gain: Standardized Counting Gain: Entropy Decision node: utilization of expected class

TABLE V
PERFORMANCE EVALUATION VALUES FOR EACH CLASSIFIER

Classifiers	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	Error Rate (%)	FPR (%)	FNR (%)
BayesNet	80.06	86.96	44.70	45.68	19.94	13.04	55.30
BPN	85.66	90.56	55.54	57.21	14.31	9.44	44.46
J48	86.44	92.05	55.12	57.66	13.30	7.95	44.88
Radom Forest	88.61	92.56	63.71	66.70	11.39	7.44	36.26
LMT	87.67	92.08	62.27	63.20	12.33	7.92	37.73
Kstar	88.27	92.15	64.59	64.07	11.68	7.85	35.41

For experimental analysis, six prominent classifiers were applied: BayesNet, Back-Propagation Neural Network (BPN), J48, Radom Forest, Logistic Model Tree (LMT), and Kstar. The value of the classification models' related parameters was tuned during the training procedure, as stated in Table 4. To evaluate the performance of the six classifiers, we calculated

seven performance metrics such as accuracy, error rate, specificity, precision, sensitivity, FPR, and FNR. The obtained result of the seven-performance mentioned above metrics by employing the equations (Eq. 13 to Eq. 19) is presented in Table 5. Class-wise accuracy is presented in Table 6.

TABLE VI
ACCURACY AS CLASS-WISE FOR EACH CLASSIFIER

Classifier	Class (Disease and Disease-free)				
	Bacterial Soft (%)	Black Rot (%)	White Rust (%)	Downy Mildew (%)	Disease-Free (%)
BayesNet	83.04	71.53	88.94	74.34	82.45
BPN	86.27	85.40	89.82	81.27	85.55
J48	86.58	90.12	87.32	82.49	85.74
Radom Forest	90.56	89.09	90.56	83.48	89.38
LMT	88.20	87.91	89.38	84.96	87.90
Kstar	87.61	90.45	90.12	85.10	88.05

TABLE VII
COMPARISON OF RESULTS OF PROPOSED CAULIFLOWER DISEASE RECOGNITION APPROACH WITH OTHERS RELATED WORKS

Completed work	Object (s) dealt with	Problem Domain	Dataset size	Segmentation algorithm	Classify performed	Feature set	Applied Classifier	Obtained Accuracy
This work	Cauliflower	Recognition	766	<i>k</i> -means clustering	✓	10	Random Forest	88.61%
Dubey <i>et al.</i> [6]	fruit and vegetable	Recognition	NM	<i>k</i> -means clustering	✓	NA	SVM	NA
Samajpati <i>et al.</i> [9]	Apple	Detection	70	<i>k</i> -means clustering	✓	NA	Random Forest	NA
Pulido <i>et al.</i> [10]	weed	Recognition	320	<i>k</i> -means clustering	✓	10	SVM	90%
Islam <i>et al.</i> [13]	potato	Detection	300	Histogram based thresholding	×	10	SVM	95%
Oppenheim <i>et al.</i> [14]	potato	Detection	2465	NA	✓	NA	CNN	83~96%
Kurniawati <i>et al.</i> [16]	Paddy	Recognition	NM	Local entropy thresholding	×	NA	NA	94.7%
Prajapati <i>et al.</i> [17]	Rice plant	Detection	NM	<i>k</i> -means clustering	×	88	SVM	73~93%
Lu <i>et al.</i> [18]	rice	Detecion	500	NA	✓	NA	CNN	95.48%
Habib <i>et al.</i> [23]	Papaya	Recognition	243	<i>k</i> -means clustering	✓	10	SVM	90%

1_{NM}: Not Mentioned.

2_{NA}: Not Applicable.

Table 5 shows that the performance for Random Forest classifier outperforms other classifiers. We achieved the highest accuracy and lowest error rate for Random Forest classifiers, which were 88.61% and 11.39%, respectively. Among the classifiers, the promising accuracy and error rate after a Random Forest classifier was achieved by the Kstar classifier. The obtained results of LMT, J48, BayesNet, and BPN classifiers were also satisfactory in terms of defined metrics.

We obtained the highest accuracy (90.56%) for the individual disease class (Bacterial soft rot) by the Random Forest classifier (Table 6). On the contrary, the BayesNet classifier achieved the lowest accuracy of 71.53%. Finally, it can be stated that the overall performance of the Random Forest classifier was satisfied over other classifiers in terms of seven performance evaluation metrics in the field of automated cauliflower disease recognition. To understand the quality of the classifier's output, a ROC curve (receiver operating characteristics) curve is presented in Fig. 5. It revealed that the AUC of the Random Forest classifier was larger than other classifiers, which indicates the better output quality of the classifier (Fig. 5).

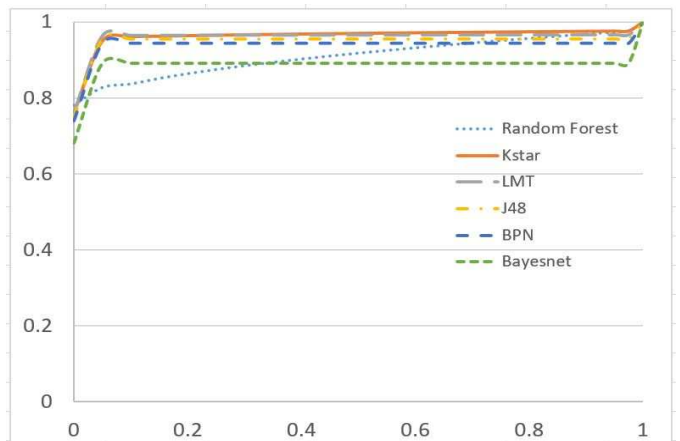


Fig. 5 Six experimental evaluated classifiers based ROC curve.

B. Comparative Analysis

Many researchers work with Machine learning techniques for recognition in various fields. To observe the performance of this expert system in cauliflower disease recognition, the comparison of various research relating to this field that has been published recently was required. As very little research has so far been conducted on cauliflower disease recognition, a direct comparison of our finding with any published

literature was possible. Nonetheless, there exists some restraint and failure in the theory of disease classification. Rare datasets were being utilized, and there existed different types of limitations in their works. In recent years, some exertions have been performed for disease recognitions in fruits and vegetables, but they were not bountiful considering the practical demand. To assess our work's excellence, we reviewed the numerical results with many other points of previously published findings with our results. Table 7 shows a comparative evaluation of our findings with previously performed works. Comparing with earlier findings, it can be concluded that our obtained accuracy of 88.61% is highly satisfactory in the recognition and discrimination of cauliflower diseases.

IV. CONCLUSION

In conclusion, we developed a machine vision-based expert system to recognize the diseases of a popular and economically important vegetable, cauliflower. A total of 766 images were utilized to implement this experimental study. Here, the k -means clustering algorithm was used to segment the disease-affected regions of the captured images. Two types of feature sets were utilized for recognizing the diseases of cauliflower. For extracting the features of the image, we performed image processing techniques. The extraction of features was then fed to six separated classifiers where the Random Forest classifier achieved the highest performance among them with 88.61% accuracy. Further research is needed to apply the deep learning models for improving the accuracy of cauliflower disease recognition with an extended image dataset. Our future research goal is to develop an integrated expert system for automated disease recognition on heterogeneous vegetables.

REFERENCES

- [1] GDP Sector Composition Countries List, Available online: "<http://statisticstimes.com/economy/countries-by-gdp-sector-composition.php>," [Last access: 21 April, 2021].
- [2] Food and Agriculture, Available online: "<https://www.worldbank.org/en/topic/agriculture/overview>," [Last access: 21 April, 2021].
- [3] S. K. Maria, S. S. Taki., M. J. Mia, A. A. Biswas., A. Majumder, F. Hasan, "Cauliflower Disease Recognition Using Machine Learning and Transfer Learning," In: Somani A.K., Mundra A., Doss R., Bhattacharya S. (eds) Smart Systems: Innovations in Computing. Smart Innovation, Systems and Technologies, vol 235, 2021.
- [4] Cauliflower production and export, Available Online: "<https://krishijagran.com/agripedia/know-about-the-production-marketing-and-export-of-cauliflower/>," [last access: 21 April, 2021].
- [5] Cauliflower, Available online: "<http://en.banglapedia.org/index.php?title=Cauliflower>," [Last access: 21 April, 2021].
- [6] S. R. Dubey, and A. S. Jalal, "Fruit and vegetable recognition by fusing colour and texture features of the image using machine learning," International Journal of Applied Pattern Recognition, vol. 2, no. 2, pp. 160-181, 2015.
- [7] P. Krithika and S. Veni, "Leaf disease detection on cucumber leaves using multi-class Support Vector Machine," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1276-1281, 2017.
- [8] H. Tani, R. Kotani, S. Kagiwada, H. Uga and H. Iyatomi, "Diagnosis of Multiple Cucumber Infections with Convolutional Neural Networks," 2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1-4, 2018.
- [9] B. J. Samajpati, and S. D. Degadwala, "Hybrid approach for apple fruit diseases detection and classification using random forest classifier," In 2016 International Conference on Communication and Signal Processing, pp. 1015-1019, IEEE, 2016.
- [10] C. Pulido, L. Solaque, and N. Velasco, "Weed recognition by SVM texture feature classification in outdoor vegetable crop images," Ingeniería e Investigación, vol. 37, no. 1, pp. 68-74, 2017.
- [11] A. H. B. A. Wahab, R. Zahari and T. H. Lim, "Detecting diseases in Chilli Plants Using K-Means Segmented Support Vector Machine," 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), pp. 57-61, 2019.
- [12] A. Rajbongshi, T. Sarker, M. M. Ahamad and M. M. Rahman, "Rose Diseases Recognition using MobileNet," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-7, 2020.
- [13] M. Islam, A. Dinh, K. Wahid, and P. Bhowmik, "Detection of potato diseases using image segmentation and multi-class support vector machine," In 2017 IEEE 30th Canadian conference on electrical and computer engineering, pp. 1-4, IEEE, 2017.
- [14] D. Oppenheim, and G. Shani, "Potato disease classification using convolution neural networks," Advances in Animal Biosciences, vol. 8, no. 2, pp. 244-249, 2017.
- [15] M. Suresha, K. N. Shreekanth and B. V. Thirumalesh, "Recognition of diseases in paddy leaves using knn classifier," 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 663-666, 2017.
- [16] N. N. Kurniawati, S. N. H. S. Abdullah, S. Abdullah, and S. Abdullah, "Investigation on image processing techniques for diagnosing paddy diseases," In 2009 international conference of soft computing and pattern recognition, pp. 272-277, IEEE, 2009.
- [17] H. B. Prajapati, J. P. Shah, and V. K. Dabhi, "Detection and classification of rice plant diseases," Intelligent Decision Technologies, vol. 11, no. 3, pp. 357-373, 2017.
- [18] Y. Lu, S. Yi, N. Zeng, Y. Liu, and Y. Zhang, "Identification of rice diseases using deep convolutional neural networks," Neurocomputing, vol. 267, pp. 378-384, 2017.
- [19] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311-318, 2018.
- [20] D. Jiang, F. Li, Y. Yang and S. Yu, "A Tomato Leaf Diseases Classification Method Based on Deep Learning," 2020 Chinese Control and Decision Conference (CCDC), pp. 1446-1450, 2020.
- [21] B. A. M. Ashqar, and S. S. Abu-Naser, "Image-based tomato leaves diseases detection using deep learning," International Journal of Academic Engineering Research, vol. 2, no. 12, pp. 10-16, 2018.
- [22] H. Durmuş, E. O. Güneş, and M. Kırıcı, "Disease detection on the leaves of the tomato plants by using deep learning," In 2017 6th International Conference on Agro-Geoinformatics, pp. 1-5, IEEE, 2017.
- [23] 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, 2018.
- [24] M. R. Mia, M. J. Mia, A. Majumder, S. Supriya, and M. T. Habib, "Computer vision based local fruit recognition," International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 2810-2820, 2019.
- [25] M. M. Rahman, A. A. Biswas, A. Rajbongshi, and A. Majumder, "Recognition of local birds of Bangladesh using MobileNet and Inception-v3," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, 2020.
- [26] M. T. Habib, and M. Rokonzaman, "Distinguishing feature selection for fabric defect classification using neural network" Journal of Multimedia, vol. 6, no. 5, pp. 416-424, 2011.