Gongcheng Kexue Xueba

ISSN 2095-9389

Volume 11 || Issue 5 || 2026

Impact Factor: 6.7

https://kexuexuebao.org/

Editorial Board
GONGCHENG KEXUE XUEBAO

Table Of Contents

A Review On Digital Pharmaceutics: Integration Of Artificial Intelligence And Machine Learning In Drug Development
Page No : 01-13

A Lightweight Combined Feature Extraction Model for Tomato Leaf Disease Classification Using DWT–FCM and SVM
Page No : 14-29

A Lightweight Combined Feature Extraction Model for Tomato Leaf Disease Classification Using DWT–FCM and SVM

S. Ledbin Vini[#1], P. Rathika[*2], K. Manimala[#3]

[#1] *ECE Department, PSN College of Engineering and Technology, Tirunelveli, India, 627 152.*
[#2] *ECE Department, PSN College of Engineering and Technology, Tirunelveli, India, 627 152.*
[#3] *IST Department, College of Engineering, Guindy, Anna University, Chennai, India, 600025.*

***Abstract*—** Farmers are facing challenges in tomato leaf disease detection and recognition due to similar appearance of disease symptoms among the different diseases. False or late identification worsens the plant growth and hinders its control measures. Hence, researchers are embracing computerized systems inbuilt with an accurate feature extraction process to highlight the nature of diseases. But they are facing difficulties in categorizing the diseases due to the inseparable closeness among the extracted features. Hence, they are pushing towards the hybridization of various feature extraction models. In this study, a lightweight combined feature extraction model is proposed that reduces the state of inseparability among the features with 28 numbers of features and ensures effective classification of tomato leaf diseases. In the proposed framework, the disease signs are initially segmented by fuzzy c-means clustering techniques, from which 28 features are extracted using the hybrid feature extraction framework and fed to the SVM classifier for classification. The subsequent results are exposed in comparison with k-means clustering. Moreover, the proposed framework classifies the tomato leaf diseases without compromising minimum number of features and maximum number of diseases. Additionally, it demonstrates an accuracy of about 92.5% for the classification of real-time tomato leaf disease images. This represents a notable enhancement of 2.5%, 2.44%, and 1.25% in accuracy, specificity, and precision over k-means clustering. In addition, the proposed method overrides other previous studies expressing its effectiveness in classifying images from both plant village database and real-time images.

***Keywords*—** Image processing, Tomato leaf analysis, Fuzzy c-means clustering, Discrete wavelet transform, Machine learning, SVM (support vector machine).

1. Introduction

The tomato is a nutrient-dense, large-consumable vegetable plays a vital role in the global market. The UN Food and Agricultural Organization (FAO) states that China, India, the United States, Brazil, and Russia are the top competitors in the tomato production of several million metric tonnes annually (refer Fig. 1). On the other hand, such tomato production is influenced by the continuous and sudden changes in weather conditions, which may weaken the healthy growth of the plant, and influence the growth of several pathogens [1], leading to plant diseases. Mark Paul Selda Rivarez, et al., [2] stated that there are a large variety of plant diseases that are emerging and re-emerging in different parts of the world under various circumstances with rapid spreading.

The farmers in the field are recognizing the diseases based on the visual perception of developed symptoms on the leaf. But they are facing difficulties in recognizing the diseases because of its varying size and their similar colour. Also, it is impossible to predict very early signs of diseases and the prediction becomes useless at a higher degree of infection, as it almost destroys the crop. Therefore, the detection of diseases at the mid-early stage is extremely useful for preventing fast spreading disease by performing control measure. Hence, a computerized recognition system is necessary for examining the leaf for the occurrence of disease symptoms. The added advantages of using image processing mathematical tools are to visualize, detect and classify those symptoms linked to each plant disease automatically without the human intervention in the field of tomato leaf disease detection and classification.

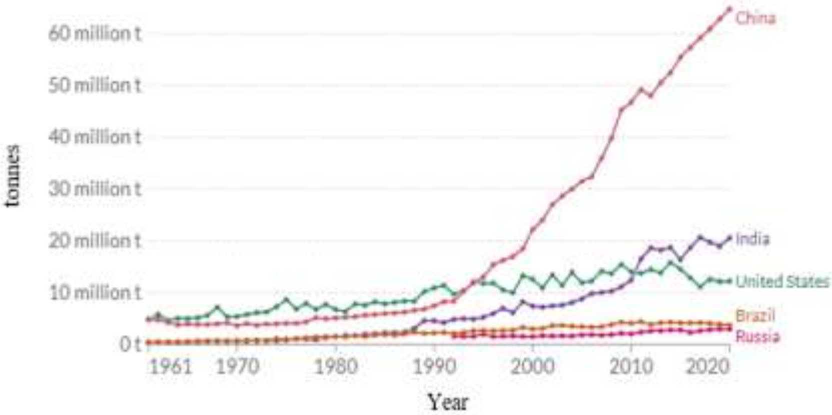

Fig. 1 Tomato Production, 1961 to 2020. (retrieved from https://ourworldindata.org/grapher/tomato-production?tab=chart)

Since 2000, the classification of these plant leaf diseases has been benefited from the development of machine learning based systems. It includes the hybridization of various methodological steps, including colour space conversion, region or sign-based segmentation, transform based operations, the feature extraction method, and the classification method. Usama Mokhtar et al. [3] tested an SVM-based machine learning approach that handled a total of 402 texture features through Gabor transform for classifying two categories of diseases with the help of RGB and HSV colour space conversion. Similarly, Raj Kumar et al. [4] utilized hybrid features—colour features, geometric features, and texture features—for the recognition of four categories of groundnut leaves through HSV colour space conversion. These research studies considered different colour space conversions and different forms of features for small-scale leaf categorization, leading to an increase in the number of features for every classification.

To reduce the number of features, researchers concentrated on the disease-infected region instead of extracting features directly from the entire leaf region. Saiqa Khan et al. [5] have chosen k-means clustering as a segmentation method for detecting diseased areas from the tomato images. This restricts the number of colour space conversions for feature extraction. However, the use of different types of features, such as texture features, shape features, and statistical features, are still considered by various researchers to improve classification. But it leads to an increase in the number of features. Studies proposed by different authors, Neelakantan et al. [6] and R. Karthickmanoj et al. [7], made the same observation, but their research differed in the choice of segmentation and feature extraction method.

Neelakantan et al. [6] employed texture and shape features with the addition of colour features from the segmented diseased region through a label edge detection algorithm. R. Karthickmanoj et al. [7] tried the double feature extraction process with texture features and ORB (oriented rotated and brief) features from the disease-infected area. Additionally, they utilized thresholding-based techniques called pixel replacement-based segmentation techniques for segmenting disease-infected regions. But these studies only concentrated on small-scale classification with a large number of features.

Recently, k-means clustering and wavelet transform were serially combined by the authors, Sunil S. Harakannanavar et al. [8], to extract the features through PCA and GLCM to produce the maximum classification result. But they employed their method to identify only six categories of tomato leaf diseases. In the same way, the authors Ledbin Vini S et al. [9] employed the Kapur thresholding technique, texture features, and various classifiers to acquire the maximum classification result, but they were concerned with four categories of tomato leaf diseases. Lately, Chimango Nyasulu et al. [10] extracted texture features from grayscale images with distances 1, 3, and 5 & angles at 0, 45, 90, 135, 180, 225, 270, and 315 degrees for the classification of real-time tomato leaf diseases with 5 categories. Venkata Lalitha Narla et al. [11] utilized multiple features, including texture, colour, and shape, directly from the diseased leaf for the classification of nine types of tomato leaf diseases. These research studies highlight the persistence of small-scale classification with a large number of features even under different feature extraction processes.

In order to enlighten readers about the previous work, Table I is formulated. According to it, researchers used various techniques for segmenting the disease-infected area. Out of which, k-means clustering is the most commonly used technique. Its result often segments the infected region with both lighter and darker signs, while other techniques concentrate on either lighter or darker signs. The major drawback of k-means clustering is that its ability to handle colour variation at the boundary is poor because of its crisp set behaviour [12]. As a result, the pattern of infected regions collapses, i.e., the disease-infected regions are not segmented properly at the edges. Hence, to utilize the advantages of k-means clustering and boost the performance of the segmentation process at the boundary, another form of k-means clustering called fuzzy c-means clustering is considered in this study, which helps to avoid crisp-set behaviour by adapting fuzziness at the boundaries [12 & 13]. As a result, the disease-infected regions are segmented properly without collapsing their pattern at the edges.

TABLE I

SUMMARIZATION OF PREVIOUS WORK IN MACHINE LEARNING BASED DISEASE CLASSIFICATION

Work Cited	Plant Species	Segmentation Algorithm	Name of the features extracted	No of features	Classifier Details	No of Disease Classified	Image Details
[3]	Tomato	-	**Texture features**	402	**SVM**	2	200 real time images
[4]	Groundnut	k-means clustering	Colour, geometric, **texture features**	23	BPN-FF	4	Real time images
[5]	Tomato	k-means clustering,	**Texture**, spatial shape, statistical features	180	LDA, KNN, CART, RF, **SVM**	4	633 real time, internet downloaded and plant village database
[6]	Tomato	Label edge-detection algorithm	Colour, **texture**, and shape features	Not defined	**SVM**, KNN, DT, RF, NB	5	1090 real time images
[7]	Tomato	Pixel replacement-based segmentation	**Texture** and ORB features	Not defined	**SVM**	4	4000 real time images
[8]	Tomato	k-means clustering	DWT, **Texture features**	Not defined	**SVM**, KNN, CNN	6	600 plant village database
[9]	Tomato	Kapur thresholding	**Texture features**	4	KNN, Ensemble, DA, DT, **SVM**, NB	4	400 plant village database
[10]	Tomato	-	GLCM **Texture features**	144	ANN, KNN, RF and **SVM**	5	3000 real time images
[11]	Tomato	k-means clustering	**Texture**, colour and shape	Not defined	**SVM**	6	300 Plant village database

Further, Table 1 explicitly states that mostly the colour, shape, and statistical features are extracted along with the texture features for classification. As the tomato leaf diseases are similar in shape, colour, and size, the exact categorization role is done through the texture features, which aid in the fundamental support for classification. Moreover, considering irrelevant features further increase inseparability among the features by increasing its count. Hence, in this study, only texture features are considered for classification. In addition, Table I explicitly states that the researchers employed different machine learning classifiers, including linear discriminant analysis (LDA), support vector machine (SVM), artificial neural network (ANN), back propagation feed forward network (BPN-FF), random forest (RF), ensemble, and naïve bayes (NB) for the classification. Out of these, the SVM classifier is commonly used due to its robust behaviour in handling high-dimensional feature sets. Due to its robustness, SVM is adopted as a classifier in this study for classification. Table II illustrates the relationship between previous works based on classification accuracy with respect to number of features and diseases classified by various classifiers.

TABLE II

SUMMARIZATION OF PREVIOUS WORKS BASED ON CLASSIFICATION ACCURACY WITH RESPECT TO NUMBER OF FEATURES AND NUMBER OF DISEASES CLASSIFIED BY VARIOUS CLASSIFIERS

Work Cited	Classifier	No of Features utilized	No of Disease Classified	Accuracy
[3]	SVM with cachy kernel	402	2	100%
[3]	SVM with Laplacian kernel	402	2	98%
[3]	SVM with Invmult kernel	402	2	78%
[5]	SVM	108	4	41.23%
[5]	LDA	108	4	89%
[5]	KNN	108	4	68%
[5]	CART	108	4	94%
[5]	RF	108	4	92.78%
[10]	SVM	144	5	86%
[10]	RF	144	5	86%
[10]	KNN	144	5	88%

Table II expresses that a large number of features are utilized by Usama Mokhtar et al., [3], Saiqa Khan et al., [5], and Chimango Nyasulu et al., [10] for small-scale (2, 4, and 5) disease classification. Additionally, it explores how the accuracy value decreases as the number of diseases increases. This is due to the existence of inseparable features among a large number of features. On a brief analysis of the existing methods, it is determined that the researchers tried to introduce separable features among the disease categories during feature extraction process for aiding accurate classification. But they are still facing the same issues while performing the classification on more disease categories. So, to overcome these drawbacks, a new feature extraction model is needed that must be able to increase the feature's separability among each disease category with a smaller number of features. Instead of going through complex learning techniques and to overcome the existing drawbacks, a lightweight combined feature extraction model is proposed in this study for the better classification of tomato leaf diseases. The primary contribution of this study encompasses

1. A lightweight combined feature extraction model is developed with the hybridization of fuzzy c-means clustering and discrete wavelet transform techniques. This novel approach eliminates the state of inseparability among the features with 28 features and improves accuracy for the classification of nine tomato leaf categories using an SVM classifier.
2. The superiority of the proposed feature extraction model is expressed in this study with various experimental analyses that are performed by answering the following research questionnaire.
   RQ1: Evaluating the combined feature extraction model (proposed framework) on existing database images.
   RQ2: Effectiveness of the combined feature extraction model in classifying real-time images.
   RQ3: Comparison with previous studies
   As fuzzy c-means clustering is a segmentation technique, each experimental analysis is performed and compared with the k-means clustering technique.

The rest of this study is organized as follows: In Section 2, the materials and methods employed by the proposed lightweight combined feature extraction model are discussed. Various experimental analyses, their results, and the practical applicability of the proposed framework are analysed in Section 3. Finally, the study is concluded in Section 4.

2. Materials and Methods

The layout involved in proposed SVM-based tomato leaf disease classification is shown in Fig. 2.

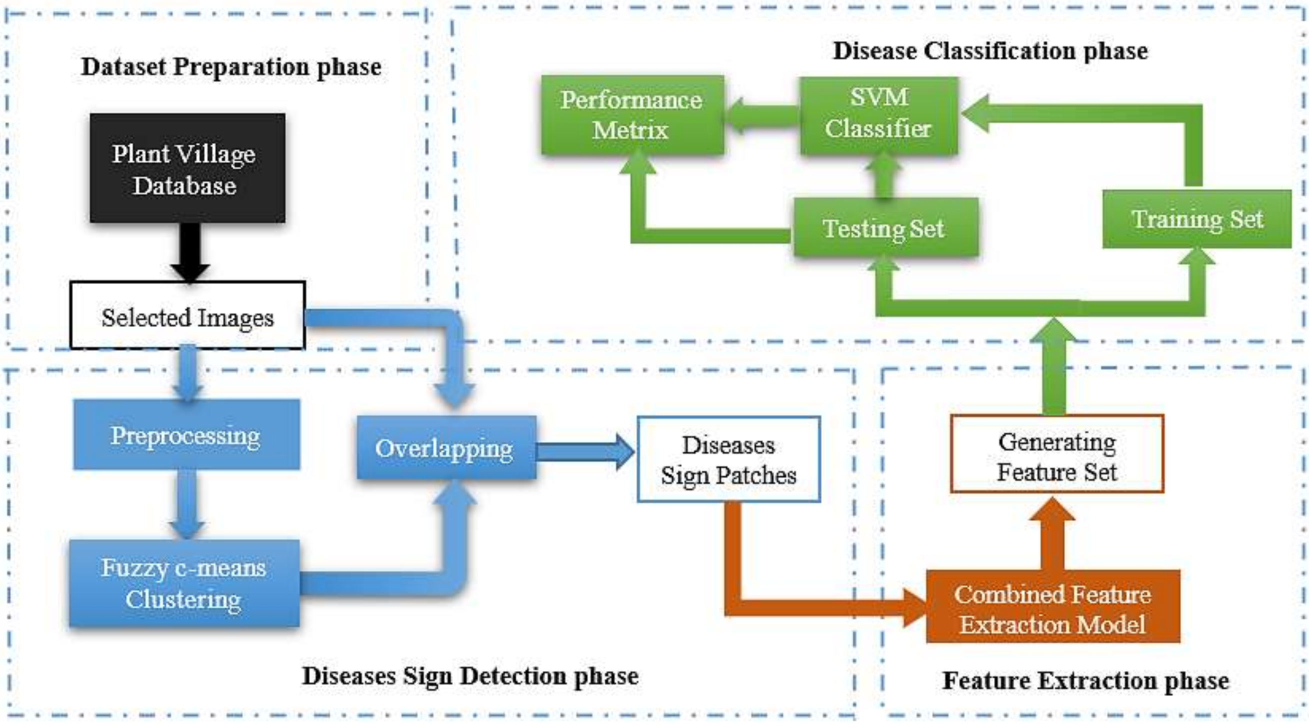

Fig. 2 Layout structure of proposed SVM based tomato leaf disease classification system.

Initially, the tomato leaf images are randomly selected under various disease categories from the plant village database. Then the preprocessing is performed on the chosen images, which are then subjected to fuzzy c-means clustering for segmenting the infected area of each disease. Once the disease-infected areas are segmented, they are overlaid on their original images to visualize only the segmented disease sign's area; they are simply referred to as spatial patches, or spatial domain patches, or simply patches. From the patches, the spatial and transform domain-based texture features are extracted through a combined feature extraction model. As a result, the feature sets are generated and fed to the SVM classifier for classification. Finally, the performance metrics are utilized to measure the performance of the resultant classification.

A. Dataset Preparation

The plant village database is an easily accessible dataset, available at "https://www.kaggle.com/emmarex/plantdisease" for plant leaf disease analysis that includes 256*256 size background-removed tomato leaf images, whose names and pattern indications are listed in Table III. In addition, some of the sample images are shown in Fig. 3.

TABLE III

Pattern indication of tomato leaf diseases [14]

S.no	Disease Name	Abbreviation	Pattern Indication
1	Bacterial Spot	BS	Dark circular spot [15 & 16]
2	Early Blight	EB	Dark circular pattern like bull eye [15]
3	Late Blight	LB	Large brown blotches [15]
4	Leaf Mold	LM	Pale green to yellow spot with indefinite margin
5	Septoria Leaf Spot	SS	Circular Dark brown spot with tan gray centre [15 & 16]
6	Spider Mites	SM	Pale yellow mark
7	Mosaic Virus	MV	Mosaic appearance [16]
8	Yellow Curl Virus	YC	Yellow between veins [16]
9	Healthy	HE	Greenish region spreads all over the leaf

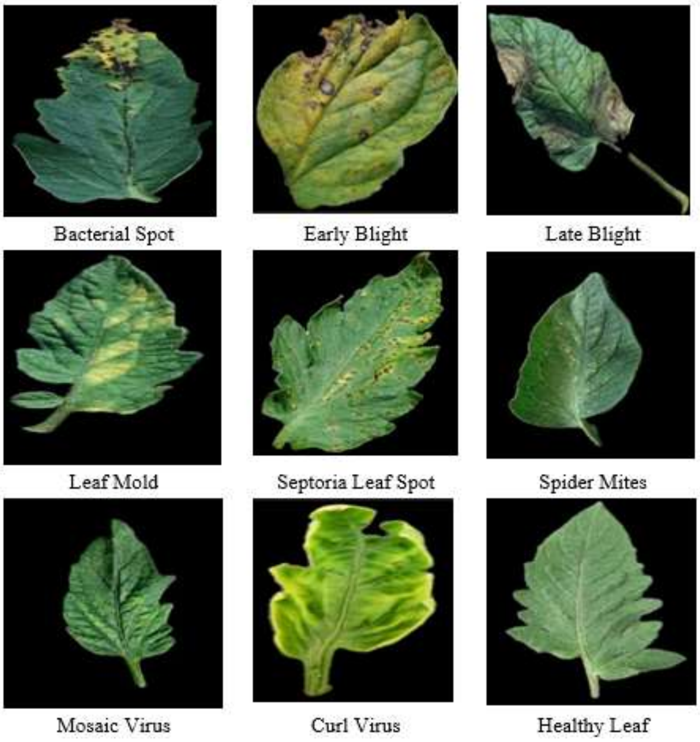

Fig. 3 Sample images for tomato leaf diseases under consideration

100 images from each category are selected based on the indication of the pattern listed in Table III for this analysis. Hence, a total of 900 images were selected for further processing.

B. *Preprocessing*

The colour conversion plays a major role in the detection of disease signs, which is performed during preprocessing. In this study, RGB to L*a*b* colour space conversion is utilized. Toran Verma et al., [17] stated that the a*b* colour information in L*a*b* colour space is primarily emphasize the affected area while performing clustering. Mathematically, they are expressed [18] as

$$L^{*} = 0.2126R + 0.7152G + 0.072 \quad (1)$$

$$a^{*} = 1.4749(0.2213R - 0.339G + 0.8006B) + 128 \quad (2)$$

$$b^{*} = 0.6245(0.1949R + 0.6057G - 0.8006B) + 128 \quad (3)$$

C. *Fuzzy c-means Clustering*

Fuzzy c-means clustering is another form of k-means clustering. It is capable of dealing with fuzziness and have similar parameters like k-means clustering [12]. The similar parameters include the choice of cluster number, centroid value, and distance measure between the pixels. In this study, the number of clusters is chosen as "3", as the colour of both healthy and disease-affected tomato leaf images are to be grouped under three similar categories. Further, the centroid values are initially chosen randomly based on the dissimilar groups of intensity values. Moreover, for grouping similar intensities values, the squared Euclidean distance is chosen whose value is given by

$$d(x,c) = (x-c)'(x-c) \quad (4)$$

where x is the current intensity value and c is the centroid value.

The fuzziness in the fuzzy c-means clustering is provoked by the membership function for minimizing the weighted sum of squared errors in the objective function [13] within the group is given in equation (5).

$$J_p = \sum_{m=1}^{D} \sum_{n=1}^{N} \delta_{mn}^{\;p} || x_m - c_n ||^2 \quad (5)$$

where D is the number of data points, N is the number of clusters, p is the fuzzy partition matrix exponent for controlling overlap, x_m is the m^{th} data point, c_n is the center of n^{th} cluster, and δ_{mn} is the degree of membership.

The cluster centers and the degree of membership function are updated during every iteration, which are calculated from equation (6) and (7)

$$c_n = \frac{\sum_{m=1}^{D} \delta_{mn}^{p} x_m}{\sum_{m=1}^{D} \delta_{mn}^{p}} \tag{6}$$

$$\delta_{mn} = \frac{1}{\sum_{k=1}^{N} \left(\frac{||x_m - c_n||}{||x_m - c_k||}\right)^{\frac{2}{p-1}}} \tag{7}$$

where c_k and c_n are the previous and current cluster centers, which are changeable during every iteration.

Since the membership function acts as a deciding factor, it avoids a crisp set of decisions, expanding the detection area and making it optimal for identifying disease symptoms in tomato leaf images. Moreover, fuzzy c-means clustering consumes more time than k-means grouping, but the resulting segmentation seems to be good for the non-crisp variation of intensity values. As a result of fuzzy c-means clustering, three clusters of images are created. Out of these, only the output with disease signs is selected manually for feature extraction.

D. *Discrete Wavelet Transform (DWT)*

The DWT itself, which holds spatial and temporal information [19], plays a greater role in the field of image processing than other transforms. The mother wavelet in DWT groups the various levels of frequencies into approximate and detail coefficient values, expressing the rate of change of information [20] in the tomato leaf images. Generally, the discrete wavelet transforms [21] of an image $I(x, y)$ of size M*N is given by

$$W_{\phi}(s_0, u, v) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y)\, \varphi_{s_{0,u,v(x,y)}} \tag{8}$$

$$W_{\psi}^{r}(s, u, v) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) \psi_{s,u,v(x,y)}^{r}, r = \{H, V, D\} \tag{9}$$

where equations (8) and (9) represent the approximate coefficient and detailed coefficient values. r identifies the directional wavelet, s_0 represents starting arbitrary scale, u, v represent the order of transform coefficient arranged, H, V, and D represent horizontal, vertical, and diagonal details for scales $s \geq s_0$.

$W_{\phi}(s_0, u, v)$ defines an approximation of the image $I(x, y)$ at scale s_0 and $W_{\psi}^{r}(s, u, v)$ defines horizontal, vertical, and diagonal details for scale $s \geq s_0$. The detail component expresses the high-frequency information of the images along vertical orientation, and the approximate component expresses its low-frequency information along vertical orientation [22]. The detailed and approximate coefficient of wavelet transformation is performed through the scaled and translated basis functions, which are given in equations (10) and (11).

$$\varphi_{s,u,v}(x, y) = 2^{\frac{s}{2}} \varphi(2^s x - m, 2^s y - n) \tag{10}$$

$$\Psi_{s,u,v}^{r}(x, y) = 2^s \psi^r (2^s x - m, 2^s y - n), r = \{H, V, D\} \tag{11}$$

The scaled and translated basis function acts as a filter bank that separates high-frequency and low-frequency information at each stage of filtering representing the family of wavelets in DWT. The Daubechies 'db4' transform is chosen as a wavelet for this study, which is a special mother wavelet whose iterations are overlapped to preserve the missing details. As it has the ability to avoid missing details during transformation, it secures more attention while choosing a mother wavelet.

E. Combined Feature Extraction Model

Once the disease sign area is visualized by overlaying the segmented output from the fuzzy c-means clustering on the original images, the resultant image is fed to the feature extraction process. In this study, a combined feature extraction model is proposed that combines the features from the spatial and transform domains of the disease sign area. Mathematically, it is represented by:

$$F_c = [\, F_s \mid F_t \,] \tag{12}$$

Here, | represents the concatenation operation, F_c represents resultant combined features, F_s represents spatial domain features and F_t represents transform domain features.

The features, which are extracted directly from the segmented disease sign area referred as spatial domain features (F_s) which hold spatial domain information. And the features, extracted through the transform from the segmented disease sign area referred as transform domain features (F_t) hold transform domain information. Here, DWT is chosen for the extraction of features in the transform domain. Especially, the DC value at the 8th level of DWT decomposition is chosen for extracting the transform domain features, as it shows vast improvements during experimentation. The schematic representation of the proposed feature extraction process is shown in Fig. 4.

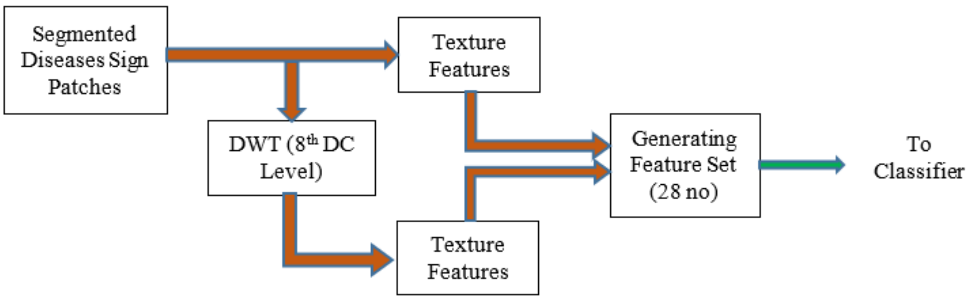

Fig. 4 Schematic representation of combined feature extraction process

Generally, texture features hold information related to the patterns associated with each tomato leaf disease. Haralick himself proposed [23] the texture features, which preserve and measure the information related to the pattern. Hence it is also named as Haralick features, which includes 14 features includes angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, difference variance, difference entropy, information measures of correlation-I, information measures of correlation-II (approximity), and maximal correlation coefficient.

Let $f(x, y)$ and $g(u, v)$ represent the spatial and transform domain features that are extracted from the images $I(x, y)$ through their grayscale co-occurrence matrices

$$f(x, y) = GLCM\,\{I(x, y)\} \tag{13}$$

$$g(u, v) = GLCM\,\{W(8_{dc})^r_\psi(s, u, v)\} \tag{14}$$

where, $I(x, y)$ is the segmented disease sign images, $W(8_{dc})^r_\psi(s, u, v)$ is the 8th level DC coefficient value of $I(x, y)$.

So, F_s holds 1x14 features for one image through its $f(x, y)$, similarly, F_t holds 1x14 features for that same image through its $g(u, v)$. As a result, F_c holds 1x28 features for one image i.e 28 number of features are extracted by the proposed combined feature extraction model for classification. This compact feature representation highlights the lightweight nature of the model, as it efficiently captures essential information with a minimal number of features, thereby reducing computational complexity and making it suitable for resource-constrained environments.

F. SVM Classifier

SVM is a binary classifier extended to classify the multiclass problem. For this research work, a one-vs-one classifier design is chosen. It classifies the dataset into one dataset of each category versus every other category for the classification [24] with the help of support vector and hyperplane. The support vectors are the training feature vectors that are lie nearer to the decision boundaries. The term "decision boundary" is refers to the hyperplane. The classifier finds a hyperplane using the support vectors by maximizing the margin of separation among each category. Depending on the support vectors, the position and orientation of the hyperplane gets vary. The number of support vectors n_{sv}, calculated from equation (15), highly depends upon the number of categories used for classification.

$$n_{sv} = \frac{numberofcategory * (numberofcategory - 1)}{2} \quad (15)$$

As nine types of tomato plant leaf categories were considered for the analysis, the classifier utilized 36 binary support vectors for classification. In multiclass problems, the machine tends to match the D dimension of the feature vector into the K dimensional vector by computing the score vector C.

$$C : X^D \rightarrow X^K \quad (16)$$

where D represents the number of feature vectors, whose value is $D = 28$ (as 28 features were considered for this study), and K represents the number of categories in the classification problem, whose value is $K = 9$. The vector representation of equation (16) can be expanded as

$$C(X_j, Q, b) = WX_j + b \quad (17)$$

where X_j is the input vector, Q is a matrix of dimension $K * D$ and the bias vector b is a vector of dimension K. Q and b are the parameters of the support vector machine. Q decides the orientation of the hyper plane, and b decides the position of the hype plane. After computation of these parameters, the score vector is calculated; hence, the equation (17) is written as

$$QX_j + b = s \quad (18)$$

In matrix form, the equation (18) can be expressed as

$$\begin{bmatrix} Q_{11} & Q_{12} \cdots & Q_{1D} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ Q_{K1} & Q_{K2} \cdots & Q_{KD} & b_K \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_D \\ 1 \end{bmatrix} = \begin{bmatrix} s_1 \\ \vdots \\ s_K \end{bmatrix} \quad (19)$$

Each row in the matrix Q represents the templates of each category involved in classification, and the X matrix represents the input vector. After computation, the class with the highest score is treated as the classification result.

G. Performance Measure

The performance of the classifier is measured using the statistical rates include accuracy, specificity, and precision for classifying tomato leaf diseases (see Table IV). These parameters are based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values calculated from the classification results.

TABLE IV

STATISTICAL RATES [25]

Parameter	Equation
Accuracy	$\frac{(TP + TN)}{(TP + FP + TN + FN)}$
Specificity	$\frac{TN}{(TN + FP)}$
Precision	$\frac{TP}{(TP + FP)}$

The overall accuracy is a good measure for measuring the performance of the tomato leaf disease image classification. Specificity explores the ability to predict the true negatives of each available category in a classification. Further, precision explores the ability to identify only the relevant category in a classification.

H. *Algorithm of Proposed Framework*

The overall algorithm involved in proposed method for the classification of tomato leaf diseases using SVM classifier as below:

Algorithm 1 Proposed framework

1. Start
2. Select images under each category for Tomato leaf disease classification.
3. for i = 1 to n (number of images selected)
   //Color Conversion
4. Extract a*b* channel information from RGB using (2) and (3)
   //Segmentation of disease sign
5. Assign number of clusters as 3, error as 0.001 and initialize δ_{mn}.
6. By random initialization of new cluster position. Calculate the cluster centers $c_n = \frac{\sum_{m=1}^{D} \delta_{mn}^{p} x_m}{\sum_{m=1}^{D} \delta_{mn}^{p}}$
7. Update degree of membership function, $\delta_{mn} = \frac{1}{\sum_{k=1}^{N}\left(\frac{\|x_m - c_n\|}{\|x_m - c_k\|}\right)^{\frac{2}{p-1}}}$
8. Calculate objective function, $J_p = \sum_{m=1}^{M}\sum_{n=1}^{N} \delta_{mn}^{\ p} \| x_m - c_n \|^2$
9. if $\|c_n - c_k\|$ > error, Repeat step 4.
10. else
11. Generate fuzzy membership matrix
    //Overlapping
12. Assign 0 to other value in disease sign segmented image (overlay)
13. Display segmented disease sign area $I(x,y)$ & calculate grayscale co-occurrence matrix $f(x,y)$
14. end
    //Generating spatial features
15. for j=1 to 14
16. Spatial feature $X(j)$ = Texture features ($f(x,y)$)
17. end
    //Generate transform features
    //Perform Wavelet decomposition on $f(x,y)$ to obtain Detailed coefficient $W(k)_{\psi}^{r}(s,u,v)$
18. for k = 1 to 8
19. $W(k)_{\psi}^{r}(s,u,v) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)\psi_{s,u,v(x,y)}^{r}, \quad r = \{H,V,D\}$
20. Calculate grayscale co-occurrence matrix $g(u,v)$ from $W(k)_{\psi}^{r}(s,u,v)$
21. k = k + 1, Repeat step 18
22. end
23. for j =15 to 28
24. Transform domain features $X(j)$ = Texture features ($g(u,v)$)
25. end
26. i = i + 1, repeat step 3
27. end
28. Create X^D vector representation for features $X(j)$, j ranges from 1 to D.
29. Compute Q, b and s from the training features using $QX_j + b = s$.
30. Calculate classification accuracy and performance measure (ref Table 4).
31. Display result
32. end

3. RESULTS

The simulation of SVM-based tomato leaf disease classification is performed with MATLAB software. Further, the research questionnaires are addressed in this section to explore the importance of the proposed combined feature extraction model.

A. *RQ1: Evaluating combined feature extraction model (proposed framework) on existing database images.*

In this section, the different combination of feature extraction model is analysed along with the proposed framework. Table V explicit individual analysis performed through the transform and the spatial domains are incapable of producing maximum accuracy values than their combined form of features. Also, it is evident from the Table V that the maximum accuracy is achieved in the analysis of spatial features with 8th level DC transform features (combined feature extraction model) rather than spatial features with 8th level AC transform features. This shows the ability of proposed combined feature extraction model in generating more separable features than other with minimum number of features. Besides, fuzzy c-means clustering with the combined feature extraction model outperformed the model with k-means clustering under different machine learning classifiers. Also, it is noted that SVM, KNN, DT and Ensemble model shows an incremental amount of 2.23%, 2.22%, 1.12%, and 3.33% in fuzzy c-means clustering over k-means clustering. Additionally, among the classifiers, SVM shows an incremental amount of 2.23%, 1.11%, and 2.23% over KNN, DT and Ensemble model respectively.

TABLE V

OVERALL CLASSIFICATION ACCURACY OF DIFFERENT FEATURE EXTRACTION MODEL WITH DIFFERENT MACHINE LEARNING CLASSIFIERS

Feature Extraction Model	Accuracy			
	SVM	KNN	DT	Ensemble
Normal spatial image	48.88%	42.89%	44.45%	46.72%
Spatial domain patches by k-means	65.55%	59.54%	60.02%	61.23%
Spatial domain patches by fuzzy c-means	64.44%	58.23%	61.19%	60.68%
8^{th} level DC transform domain analysis on normal images	17.77%	11.13%	13.57%	12.65%
8^{th} level AC transform domain analysis on normal images	15.55%	9.98%	9.09%	10.78%
8^{th} level DC transform domain analysis on k-means spatial patches	32.22%	29.89%	30.01%	28.32%
8^{th} level AC transform domain analysis on k-means patches	73.33%	72.45%	70.98%	71.23%
8^{th} level DC transform domain analysis on fuzzy c-means spatial patches	42.22%	40.78%	39.75%	38.97%
8^{th} level AC transform domain analysis on fuzzy c-means spatial patches	61.11%	58.23%	57.56%	56.89%
k-means spatial features with its 8^{th} level AC transform features	77.77%	64.44%	65.59%	67.78%
k-means spatial features with its 8^{th} level DC transform features	84.44%	82.22%	84.44%	81.11%
fuzzy c-means spatial features with its 8^{th} level AC transform features	76.67%	77.78%	68.89%	75.56%
fuzzy c-means spatial features with its 8^{th} level DC transform features (proposed model)	**86.67%**	**84.44%**	**85.56%**	**84.44%**

The novelty of the proposed framework lies in the incorporation of a combined feature extraction model that reduces the state of inseparability among the features for the classification. It allows the utilization of minimum number of features (28) for the classification of tomato leaf diseases, emphasizes its compact feature representation. Further, it highlights the lightweight nature of the model, as it efficiently captures essential information with a minimal number of features, thereby reducing computational complexity and making it suitable for resource-constrained environments. Additionally, the choice of fuzzy c-means clustering is an effective segmentation technique for tomato leaf diseases. It segments the disease infected area properly by facilitating fuzziness at the edges to take decisions without collapsing its pattern. Moreover, the utilization of a robust SVM classifier for efficient classification enhances the effectiveness of this study in classifying tomato leaf diseases.

In order to examine the disease-based classification of spatial patches through the combined feature extraction model, the specific parameters are listed in Table VI and its classification performance is illustrated in Fig 5.

TABLE VI

PERFORMANCE METRICS BY SVM CLASSIFIER FOR THE COMBINED FEATURE EXTRACTION MODEL WITH DIFFERENT CLUSTERING TECHNIQUES

S.No	Disease Name	k-means clustering			fuzzy c-means clustering		
		Accuracy	Specificity	Precision	Accuracy	Specificity	Precision
1	BS	80%	98.75%	88.8%	100%	100%	100%
2	EB	90%	98.75%	90%	80%	87.5%	88.88%
3	HE	100%	98.75%	90.90%	90%	100%	100%
4	LB	90%	98.75%	90%	100%	100%	100%
5	LM	60%	100%	100%	60%	96.25%	66.66%
6	SS	90%	98.75%	90%	100%	100%	100%
7	SM	80%	95%	66.66%	80%	97.5%	80%
8	MV	70%	95%	63.63%	70%	92.5%	53.84%
9	YC	100%	98.75%	90.90%	100%	100%	100%
Overall		84.44%	98.06%	85.67%	86.67%	98.33%	87.71%

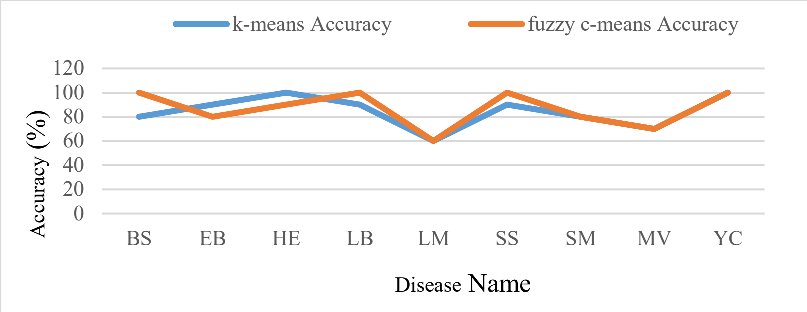

Fig. 5 Performance of combined feature extraction model on spatial image patches

Comparing to k-means clustering, all the disease categories show a significant increase in their detection value through fuzzy c means clustering (Fig. 5). It is observed that the LM disease category scores the same accuracy value for both clustering methods. On the other hand, in fuzzy c-means clustering, diseases categories including BS, LB, SS, and YC, are showing 100% of results. Further, SM, EB, and HE disease categories results vary between 70 and 90%. Also, LM and MV diseases categories secure the optimal value ranges between 60 and 70% for the classification.

B. *RQ2: Effectiveness of combined feature extraction model in classifying real time images.*

In this study, the real time tomato leaf disease images are taken from our study [14]. A total of 155 numbers of early blight images, 85 numbers of spider mite images, 85 numbers of curl virus images, and 85 numbers of healthy images were recorded. Out of which, 10 images from each category are randomly chosen for testing, and the remaining images are subjected to training. Thus, totally 370 images are used for training and remaining 40 images are used for testing in order to check the effectiveness of the proposed feature extraction model in classifying real-time images. The resultant performance metrics are tabulated in the Table VII.

TABLE VII

PERFORMANCE METRICS FOR THE COMBINED FEATURE EXTRACTION MODEL IN CLASSIFYING REAL TIME IMAGES

S.No	Disease Name	k-means clustering			fuzzy c-means clustering		
		Accuracy	Specificity	Precision	Accuracy	Specificity	Precision
1	EB	100%	100%	100%	100%	100%	100%
2	HE	90%	100%	97.5%	100%	90.9%	97.5%
3	SM	70%	87.5%	90%	70%	100%	92.5%
4	YC	100%	76.92%	92.5%	100%	83.33%	95%
Overall		90%	91.11%	95%	92.5%	93.55%	96.25%

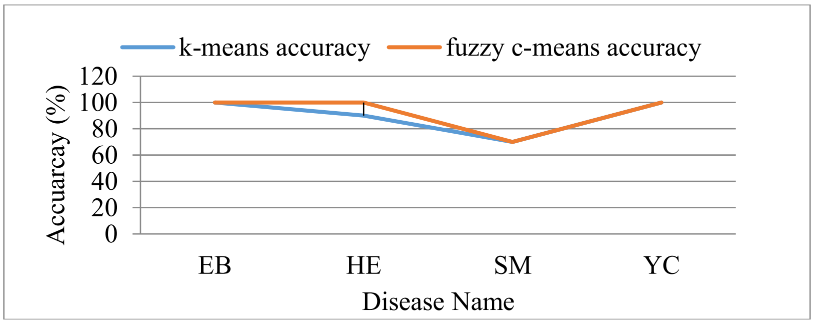

Fig. 6 Performance of combined feature extraction model in classifying real time images

From Fig. 6, it is understood that both methods show minimum result in classifying SM disease category. In addition, combined feature extraction model with k-means clustering shows less performance in classifying HE category. On analysing other category, the combined feature extraction model accurately classifies the real-time images with fuzzy c-means clustering rather than k-means clustering techniques.

Moreover, the result with fuzzy c-means clustering gives 100% accuracy in classifying images with disease categories EB, HE, and YC.

C. *RQ3: Comparison with previous studies*

On analysing the observations made under research questionnaires RQ1, and RQ2, it is proven that fuzzy c-means clustering is more effective than k-means clustering. However, the texture features (see Table VIII) are predominantly used for the classification of various diseases; they differ in the way of extraction. In previous studies, the researchers extracted the texture features either from the spatial domain or from the transform domain. Additionally, they combined the extracted texture features with other features like shape, colour, and statistical features to increase separability among the features. But this results in an increase in the number of features without improving separability among the features, resulting in a low accuracy value. These observations are made in the work marked with "*" in Table VIII. In a similar way, work marked with "&" explicit maximum accuracy for the classification of two diseases by compromising large number of features. On the other hand, work marked with "@" utilized a minimum number of features with a good accuracy value only for the disease with lighter signs. Even though the number of features is not defined in the other works cited in Table VIII, their classification is restricted to only six categories.

TABLE VIII

COMPARISON OF PREVIOUS SYSTEM ON TOMATO LEAF DISEASE CLASSIFICATION WITH PLANT VILLAGE DATABASE

Cited Work	Techniques	No of the features required	Disease Name	No of Disease Classified	Accuracy
[3] &	Texture features +SVM with Cauchy kernel	402	Powdery mildew and EB	2	100%
[3] &	Texture features +SVM with Invmult kernel	402	Powdery mildew and EB	2	98%
[3] &	Texture features +SVM with Laplacian kernel	402	Powdery mildew and EB	2	78%
[5] *	k-means clustering+ texture features + spatial shape features + statistical features + SVM	180	SS, LB, EB, HE	4	41.23%
[5] *	k-means clustering+ texture features + spatial shape features + statistical features + KNN	180	SS, LB, EB, HE	4	68%
[8] *	k-means clustering + DWT + PCA + GLCM +SVM	Not defined	HE, MV, LM, YC, SM, Target Spot	6	84.83%
[9] @	Kapur thresholding + texture features (energy, entropy, correlation and homogeneity) + SVM	4	HE, MV, YC, LM	4	80%
[9] @	Kapur thresholding + texture features (energy, entropy, correlation and homogeneity) + LDA	4	HE, MV, YC, LM	4	85%
[9] @	Kapur thresholding + texture features (energy, entropy, correlation and homogeneity) + DT	4	HE, MV, YC, LM	4	82.5%
[9] @	Kapur thresholding + texture features (energy, entropy, correlation and homogeneity) + NB	4	HE, MV, YC, LM	4	77.50%
[11]	Texture features + colour features + shape features + SVM	Not defined	BS, EB, LB, LM, SS, YC.	6	86.60%
Proposed framework	fuzzy c-means clustering + spatial texture features + temporal texture features (DWT) + SVM	28	BS, EB, HE, LB, LM, SS, SM, MV, YC.	9	86.67%

By considering these facts, the proposed method is designed without compromising the minimum number of features, the maximum number of disease classifications, and efficient classification accuracy. It is evident from Table VIII that the proposed framework considered only texture features with the help of a proposed lightweight combined feature extraction model that reduces inseparability among the features compared to previous studies. It utilized only 28 texture features by combining both spatial and transform domains with a larger number of disease classifications than previous studies, highlight its lightweight design. Moreover, the proposed framework effectively classifies a large number of diseases with an accuracy of 86.67% by overriding the performance of existing studies under comparison.

Table IX illustrates the effectiveness of the proposed framework in classifying real-time images than the previous studies. Even though different techniques are utilized by different researchers for the segmentation of disease-infected regions, the proposed framework with fuzzy c-means clustering is more attractive. Also, it is evident from the analysis recorded in Table IX, along with the attained accuracy and other performance metrics mentioned in research questionaries. Besides analysing existing studies on real-time images in Table IX, the method of feature extraction in the proposed framework seems to be unique in nature. Additionally, the number of features required by the proposed framework is 28, which is less than the previous studies.

At the same time, the proposed framework overrides the existing studies on real-time images with a maximum accuracy of about 92.5% in the classification of tomato leaf diseases.

TABLE IX

COMPARISON OF PREVIOUS SYSTEM WITH SVM ON TOMATO LEAF DISEASE CLASSIFICATION WITH REAL TIME IMAGES

Cited Work	Techniques	No of the features required	Disease Name	No of Disease Classified	Accuracy
[5]	k-means clustering + texture features + spatial shape features + statistical features	180	SS, LB, EB, HE	4	41.66%
[6]	Label edge detection algorithm + texture features + color + shape features	Not defined	SS, YC, BS, EB, and HE	5	83%
[7]	Pixel replacement-based segmentation + texture features + ORB features	Not defined	Alternaria Alternata, Anthracnose, Bacterial blight, Cerospora leaf spot	4	92.32%
[10]	Gray scale images with distance 1, 3 and 5 & angles at 0, 45, 90, 135, 180, 225, 270 and 315 degrees + texture features	144	Alternaria, Healthy, Curvularia, Helminthosporium and Lasiodiplodi	5	86%
Proposed framework	Fuzzy c-means clustering + spatial texture features + temporal texture features (DWT)	28	EB, HE, SM, and YC	4	92.50%

These observations of the proposed framework in classifying both plant village database images and real-time images show its robustness over previous studies. Also, the proposed method proved its effectiveness in classification without compromising the minimum number of features and performing the maximum number of disease classifications. These characteristics showcase its practical applicability in the field of tomato leaf disease classification by reducing computational complexity and making it suitable for resource-constrained environments.

4. CONCLUSIONS

The plants are to be monitored continuously, and their diseases are to be identified at an earlier stage to ensure their healthy growth. Ensuring the plant's healthy growth leads to a high production rate of healthy tomatoes, thereby maintaining its market value. Conversely, the tomato plant leaf and its image are non-uniform in terms of area, colour, appearance, shape, texture, etc. So, it is challenging to identify its status of growth and the diseases that attack on it. This paper shows the SVM-based tomato leaf disease classification using a lightweight combined feature extraction model with fuzzy c-means and wavelet transform for efficient classification. Due to the proposed combined feature extraction model, the inseparable closeness among the features has vanished with a minimum number of features. As a result, the proposed method provides maximum accuracy, high specificity, and good precision in each of the nine individual categories of diseases that are under test. Except for mosaic and leaf mold detection, all other categories scored maximum accuracy in disease detection. The mosaic virus and leaf mold disease detection score the least, as their sign appearances are more similar to each other, which reflects in the overall accuracy of 86.67%. Also, the proposed framework is more attractive as it effectively classifies real-time images with an accuracy of about 92.5%, indicating its practical applicability.

Competing Interests

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

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The dataset presented in this article are taken from readily available Plant village database that is accessible at "*https://www.kaggle.com/emmarex/plantdisease". The real-time images will be made available on request.*

REFERENCES

[1] Vnicius Bischoff, Kleinner Farias, Juliano Paulo Menzen, Gastavo Pessin, “Technological support for detection and prediction of plant diseases: A systematic mapping study”, Computers and Electronics in Agriculture 181, 2021. DOI: https://doi.org/10.1016/j.compag.2020.105922.

[2] Mark Paul Selda Rivarez, Ana Vucurovic, Natasa Mehle, Maja Ravnikar and Denis Kutnjak, “Global Advances in Tomato Virome Research: Current Status and the Impact of High-Throughput Sequencing”, Frontiers in Microbiology 12, 2021. DOI: 10.3389/fmicb.2021.671925.

[3] Usama Mokhtar, Mona A. S. Ali, Aboul Ella Hassenian, Hesham Hefny, “Tomato Leaves disease detection approach based on support vector machines”, IEEE Page No. 246-250, 2015. DOI: 10.1109/ICENCO.2015.7416356.

[4] Raj Kumar S, Sowrirajan S, “Automatic Leaf Disease Detection and Classification using Hybrid Features and Supervised Classifier”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 5(6): 4556-4563, 2016. DOI: https://www.ijareeie.com/upload/2016/june/8_Automatic.pdf.

[5] Saiqa Khan, Meera Narvekar, “Novel fusion of color balancing and superpixel based approach for detection of tomato plant diseases in natural complex environment”, Journal of King Saud University - Computer and Information Sciences 34(6): 3506-3516, 2020. DOI: https://doi.org/10.1016/j.jksuci.2020.09.006.

[6] Neelakandan. P, “Analyzing the best machine learning algorithm for plant disease classification”, Materials Today: Proceedings 80(3): 3668-3671, 2021. DOI: https://doi.org/10.1016/j.matpr.2021.07.358.

[7] Karthickmanoj R, Padmapriya J, Sasilatha T, “A novel pixel replacement-based segmentation and double feature extraction techniques for efficient classification of plant leaf diseases”, Materials Today: Proceedings 47(9): 2048-2052, 2021. DOI: https://doi.org/10.1016/j.matpr.2021.04.416.

[8] Sunil S. Harakannanavar, Jayashri M. Rudagi, Veena I Puranikmath, Ayesha Siddiqua, “Plant leaf disease detection using computer vision and machine learning algorithms”, Global Transitions Proceedings 3: 305-310, 2022. DOI: https://doi.org/10.1016/j.gltp.2022.03.016.

[9] Ledbin Vini S, Rathika P, “Thresholding based Tomato Leaf Disease Classification**”** 2023 4th International Conference on Signal Processing and Communication (ICSPC), Page No. 143-147, 2023. DOI**:** 10.1109/ICSPC57692.2023.10125865.

[10] Chimango Nyasulu, Awa Diattara, Assitan Traore, Cheikh Ba, Papa Madiallacke Diedhiou, Yakhya Sy, Hind Raki, Diego Hernan Peluffo-Ordonez, “A comparative study of Machine Learning-based classification of Tomato fungal diseases: Application of GLCM texture features”, Heliyon 9(11), 2023. DOI: https://doi.org/10.1016/j.heliyon.2023.e21697.

[11] Venkata Lalitha Narla, Gulivindala Suresh, (2023). Multiple Feature-Based Tomato Plant Leaf Disease Classification Using SVM Classifier. Machine Learning, Image Processing, Network Security and Data Science 946: 443-455. DOI: https://doi.org/10.1007/978-981-19-5868-7_33.

[12] Kai Tian, Jiuhao Li, Jiefeng zeng, Asenso Evans, Lina Zhang, “Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm”, Computers and Electronics in Agriculture 165, 2019. DOI: https://doi.org/10.1016/j.compag.2019.104962.

[13] Hind Rustum Mohammed, Husein Hadi Alnoamani, Ali AbdulZahraa Jalil, “Improved Fuzzy C-Mean Algorithm for Image Segmentation”, International Journal of Research in Artificial Intelligence 5(6): 7-10, 2019. URL: https://thesai.org/Downloads/IJARAI/Volume5No6/Paper_2-Improved_Fuzzy_C_Mean_Algorithm_for_Image_Segmentation.pdf.

[14] Ledbin Vini S, Rathika P, “TrioConvTomatoNet: A robust CNN architecture for fast and accurate tomato leaf disease classification for real time application”, Scientia Horticulturae, Vol. 330, 2024. DOI**:** https://doi.org/10.1016/j.scienta.2024.113079.

[15] Keerthi J, Suman Maloji, Gopi Krishna P, “An Approach of Tomato Leaf Disease Detection Based on SVM Classifier”, International Journal of Recent Technology and Engineering 7(6), 2016. URL: https://www.ijrte.org/wp-content/uploads/papers/v7i6/F2767037619.pdf.

[16] Jagadeesh Basavaiah, Audre Arlene Anthony, “Tomato Leaf Disease Classification using Multiple Feature Extraction Techniques”, Wireless Personal Communication 115: 633-651, 2020. DOI: https://doi.org/10.1007/s11277-020-07590-x.

[17] Toran Verma. Sipi Dubey “Impact of Color Spaces and Feature Sets in Automated Plant Diseases Classifier: A Comprehensive Review Based on Rice Plant Images”, Archives of Computational Methods in Engineering 27: 1611-1632, 2019. DOI: https://doi.org/10.1007/s11831-019-09364-6.

[18] Krishnaprasath V. T, Preethi, “Finite automata model for leaf disease classification”, Agricultural Economics 67(6): 220-226, 2021. URL: https://agricecon.agriculturejournals.cz/pdfs/age/2021/06/02.pdf

[19] Fang-Hsuan Cheng, Yu-Liang Chen, “Real time multiple objects tracking and identification based on discrete wavelet transform” Pattern Recognition 39(6): 1126-1136, 2005. DOI: https://doi.org/10.1016/j.patcog.2005.12.010.

[20] Kiran S M, Dr. Chandrappa D N, “Plant Disease Identification using Discrete Wavelet Transforms and SVM”, Journal of University of Shanghai for Science and Technology 23(6). 2021. DOI: 10.51201/JUSST/21/05226.

[21] Rafael C. Gonzalez, Richard E. Woods. “Digital Image Processing”, Third Edition, Pearson Publication, 2011.

[22] Mujeeb Ur Rehman, Arslan Shafique, Kashif Hesham Khan, Mohammad Mazyad Hazzazi, “Efficient and secure image encryption using key substitution process with discrete wavelet transform”, Journal of King Saud University – Computer and Information Sciences 35(7). 2023. DOI: https://doi.org/10.1016/j.jksuci.2023.101613.

[23] Robert M. Haralick, K. Shanmugam, and Its’hak Dinstein, “Textural Features for Image Classification”, IEEE transactions on Systems, Man and Cybernetic 3(6): 610-621.1973. DOI: 10.1109/TSMC.1973.4309314.

[24] Jason Brownlee (2020). One-vs-Rest and One-vs-One for Multi-Class Classification. [online], [accessed on 6 January 2023]. URL: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/.

[25] Dipak Kumar Patra, Tapas Si, Sukumar Mondal, Prakash Mukherjee, “Breast DCE-MRI segmentation for lesion detection by multi-level thresholding using student psychological based optimization”, Biomedical Signal Processing and Control 69. 2021. DOI: https://doi.org/10.1016/j.bspc.2021.102925.