IJCDS 1571020612

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Optimized featureset in classification of plant leaves images using machine learning models

Nikhil J Inamdar[1] **and Manjunath Managuli**[2]

[1]*Electronics and communication department,Gogte Institute of Technology, Belagavi and affiliated to Visveswaraya Technology University, Belagavi,india*
[2]*Electronics and communication department,Gogte Institute of Technology, Belagavi and affiliated to Visveswaraya Technology University, Belagavi,india*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Saving the earth becomes the utmost priority and responsibility of any individual. Environmental and ecosystem health assessment studies require precision farming, enabling early identification of diseases and optimizing crop management. Automatic plant leaf detection will serve as one of the crucial contributions towards biodiversity research. The proposed work provides an optimized feature set in the classification of plant leaves using machine learning techniques. The work uses fourteen different plant leaves, namely, apple, blueberry, cherry, corn, cotton, grape, groundnut, peach, pepper, potato, raspberry, soybean, strawberry, and tomato. A total of 20, 357 images are taken for training and testing purposes. Features include shape, texture, HSI and wavelets. Features are reduced using feature optimization techniques such as XG Boost, Pearson correlation, and chi-squared. In search of the best classifier, five classifiers, namely, random forest, k-nearest neighbor, support vector machine, naïve Bayes and decision tree are varied with their hyperparameters. The SVM classifier gave the best results, achieving an accuracy of 99.59 with four-fold cross-validation. The novelty of the work lies in deploying features using the knowledge gained by farmers. Results reveal that the method outperforms the state-of-the-art works and are found encouraging. In this regard the techniques used here enable us to target the leaves and detect the diseases and further facilitate to opt for preventive measures.

Keywords: Ecosystem; Biodiversity; Classification; HSI; Wavelets; XG Boost; Pearson correlation; chi-squared; ANOVA; Classifiers

1. Introduction

Smart Agriculture, a transformative approach to farming, integrates innovative technologies to revolutionize traditional agricultural practices. Utilizing Internet of Things (IoT) sensors, data analytics, and automation, Smart Agriculture optimizes resource utilization, enhances efficiency, and promotes sustainability [1]. Precision farming, enabled by GPS and satellite technology, allows for accurate mapping and variable rate applications, optimizing the use of water, fertilizers, and pesticides. Despite challenges, Smart Agriculture holds promise for a resilient and sustainable food production system, addressing global demand while minimizing environmental impact.

The classification of plant leaves through image analysis has gained significant attention in recent years, driven by the intersection of advancements in computer vision and the pressing need for efficient plant species identification. Automated classification systems offer a promising solution to challenges in agriculture, environmental monitoring, and biodiversity conservation. This endeavor involves the fusion of plant biology, image processing, and machine learning, with the overarching goal of accurately identifying plant species based on leaf images [2].

Plant leaves exhibit a wide range of morphological features, including shape, color, texture, and margin characteristics, which serve as distinctive markers for species differentiation. Leveraging these features through advanced image processing techniques and machine learning models enables the creation of robust classification systems capable of handling diverse datasets [3]. The process of developing an optimized feature set for leaf image classification involves careful consideration of both handcrafted and deep learning features. Handcrafted features, derived from domain-specific knowledge, capture inherent botanical traits, while deep learning features, extracted from pre-trained convolutional neural networks, reveal complex hierarchical patterns within the images [4].

Selecting an appropriate feature set is only one aspect of the classification pipeline. Equally crucial is the choice of machine learning models and their configuration. Various algorithms, such as Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors (k-NN), and deep neural networks, can be employed, each with their unique strengths and limitations [5]. Ensemble methods, combining

E-mail address: author 1, author 2, author 3

multiple models, further enhance classification accuracy and robustness [6].

In this context, this exploration delves into the intricacies of creating an optimized feature set for the classification of plant leaves using machine learning models. Through a systematic approach encompassing data preprocessing, feature extraction, model selection, and iterative fine-tuning, the objective is to develop a highly accurate and generalizable classification system capable of addressing real-world challenges in plant species identification. As the synergy between plant science and computational methods continues to evolve, these efforts contribute to the advancement of precision agriculture, environmental monitoring, and biodiversity conservation.

2. Literature Review

To know the state-of-the-art methods in the related study, following literature survey is carried out and gist of the papers are discussed. [7] give machine learning-enabled weed classification system to categorize weeds based on a fusion of handcrafted shape and texture features at the feature level. The chosen classifier is a Support Vector Machine (SVM), and experimental results reveal a notable 93.67 % overall accuracy when utilizing shape curvature features. [8] for providing method for identifying and assessing the severity of PVY and TMV infections in tobacco leaves using hyperspectral imaging and machine learning.

The research involves applying three preprocessing techniques—MSC, SNV, and SavGol—to spectral data spanning the full length of the leaves. The combination of SavGol with SVM proves highly effective, achieving a remarkable 98.1 % average precision in distinguishing various PVY severity levels and a high recognition rate of 96.2 % in classifying different TMV severity levels. [9] have worked on classification of fig leaf diseases by combining support vector machine (SVM). The method uses Fuzzy C Means algorithm for segmentation, Principal Component Analysis for feature extraction, and a hybrid classification strategy involving Particle Swarm Optimization (PSO) with SVM.

The work in [10] provide deep learning network model designed for the more accurate recognition of soybean leaf diseases. The model incorporates a fully connected layer to integrate extracted features, resulting in an average recognition accuracy of 85.42 %. This outperforms six comparison deep learning models (ConvNeXt, ResNet50, Swin Transformer, MobileNetV3, ShufeNetV2, and SqueezeNet), which achieved lower accuracies ranging from 59.89 % to 77.00 %. [11] identifies cotton verticillium wilt by fusing spectral and image features and leveraging support vector machine (SVM) and backpropagation neural network (BPNN) models.

The results demonstrate high accuracy, with EfficientNet achieving an average accuracy of 93.00 %. Notably, the SG-MSC-BPNN model, using spectral full bands, and the SG-MN-SPA-BPNN model, using feature bands, both achieve a classification accuracy of 93.78 % while SG-MN-SPA-FF-BPNN model gives a remarkable classification accuracy of 98.99 %. [12] uses combined local binary histogram pattern of gradient (LBHPG) image feature extraction technique for classification Indian agricultural crop species. The LBHPG method identifies leaf objects within images, and for classification purposes, three shallow machine learning classifiers—PNN (Probabilistic Neural Network) and KNN (k-Nearest Neighbors) and SVM—are utilized. PNN Classifier achieves the highest accuracy at 94.58 % for the identification of crop species. [13] employs a Deep Convolutional Neural Network (CNN) for feature extraction, and the extracted features are subjected to classification using various classifiers, including Support Vector Machine (SVM), K Nearest Neighbor (KNN), Random Forest, Naive Bayes, and Logistic Regression (LR). [14] gives an optimal feature set for achieving higher classification accuracy, utilizing the Flavia and Swedish datasets. Various features, including Gray Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP), and Hu Invariant Moment, are combined in different ways to enhance accuracy. The current study aims to build on this success by exploring different feature combinations, seeking to further optimize and potentially exceed the accuracy levels achieved in the prior research.

In [15] proposes an intelligent system utilizing Raspberry Pi 3 Model B+ and the RPi camera to identify real-time images of Indian medicinal herbs and disclose their medicinal properties. Among the four developed machine learning models, one is specifically designed for identifying details of medicinal leaves, achieving an impressive top-1 accuracy of 98.98 % on a custom dataset containing 25 different medicinal species and 1500 leaf images. When implemented on the RPi, the model exhibits a real-time top-3 accuracy of 99 %. [16] introduces a new Convolutional Neural Network (CNN)-based method called D-Leaf for leaf classification. The study compares three different CNN models—pre-trained AlexNet, fine-tuned AlexNet, and D-Leaf—based on their feature extraction capabilities. The D-Leaf model achieves a testing accuracy of 94.88 %, demonstrating comparable performance to the pre-trained AlexNet (93.26 %) and fine-tuned AlexNet (95.54 %) models. [17] give framework for leaf categorization involving pre-processing, feature extraction, feature selection, and classification. Morphological features, such as centroid, major axis length, minor axis length, solidity, perimeter, and orientation, are extracted from digital images of leaves across various categories. The AdaBoost methodology is employed to enhance precision, resulting in an impressive precision rate of 95.42 %.

3. Methodology

It consists of four stages, namely, preprocessing, feature extraction, feature optimization and classification, and the block diagram of the proposed methodology is shown in Fig. 1. Different features are extracted from input images which include shape, texture, HSI and wavelets. Around 28 features together are extracted the images. As the higher number of features reduced the performance of the models, feature optimization techniques are used to reduce the features, such as XG Boost, Pearson correlation, chi-

squared and ANOVA. Around six features are finally used for further work based on the aforesaid methods. To find the best suitable model for the input images considered, the following classifiers are trained and tested, namely, random forest, k-nearest neighbor, support vector machine, naïve bayes and decision tree. Based on the performance through varied hyperparameters, classification metrics are drawn to write the inference.

A. Input Images

In total, 14 types of plant leaves are considered, namely, apple, blueberry, cherry, corn, cotton, grape, groundnut, peach, pepper, potato, raspberry, soybean, strawberry, and tomato. Out of 14 types, groundnut images are taken from [3] whereas cotton plant images are used form [2], with rest taken from [1]. Sample images of dataset considered are shown in Fig. 2.2.

B. Feature extraction:

As the trend is deep learning carried by every researcher in the related field, even though machine learning techniques have become absolute, there is a need for deployment of machine learning models which classifies different types of plant based on leaves images. The novelty of the work exists in developing machine learning mapping with the techniques used by subject experts involving farmers and agriculturists in identification of plants through leaves. Lots of features exist in literature used for extracting information from images. For the work, shape, texture, HSI and wavelets are used [1].

C. Shape features

These features describe the shape and structure of objects or regions within an image. Shape features provide valuable information about the spatial arrangement of pixels within objects or regions of interest. Stepwise output images making the images suitable for feature extraction is shown in Fig 3. Sample plant leaf variety taken for feature extraction is shown in Fig 4. Out of 14 different classes of plants, only few are shown so as to use the space optimally. Five shape features are extracted from the input images and the related mathematical representations are expressed as in Eq-1 through Eq-4. The perimeter, P, can be calculated by summing the lengths of all boundary segments in the object.
eq1
The aspect ratio, AR, is calculated as the ratio of the width (W) to the height (H) of the bounding box of the object.
eq2
Rectangularity, R, is calculated as the ratio of the object's area (A) to the area of its bounding box (BB).
eq3
Circularity, C, is calculated as a function of the object's area (A) and perimeter (P).
eq4
The diameter, D, is determined by finding the maximum Euclidean distance between any two points within the object. Here are some common shape features and are used in the work, perimeter, aspect ratio, rectangularity, circularity, and diameter. Feature values of two images of 14 varieties are given in Table 1.

An excellent style manual for science writers is [**?**].

D. Texture features

Gray-Level Co-occurrence Matrix (GLCM) texture features are a set of statistical features commonly used in image analysis to characterize the spatial relationships between pixel values within an image. GLCM is a matrix that quantifies how often pairs of pixel values at specific spatial relationships occur in an image. These features provide information about the texture and patterns present in the image. Here are some common GLCM texture features and the ones used in the work are contrast, dissimilarity, homogeneity, energy, and correlation and the related values are given in Table 2

E. HSI color feature:

The HSI (Hue, Saturation, and Intensity) color space provides a different representation of an image compared to the more common RGB color space. The hue channel encodes color information. It represents the dominant color of a pixel. The saturation channel represents the intensity or vividness of colors. High saturation values indicate more vibrant and pure colors, while low values represent desaturated or grayscale regions. The computed metrics include energy, contrast, correlation, homogeneity, and entropy for each channel and their related values are given in Table 3. Here's how each of these metrics can be useful for image recognition.

a) Energy:

Energy is a measure of the uniformity and smoothness of an image region. Higher energy values indicate regions with more distinct and pronounced patterns or textures.

b) Contrast:

Contrast quantifies the difference in intensity values within an image region. High-contrast regions have a wide range of intensity variations, which can be indicative of the presence of edges, boundaries, or sharp transitions.

c) Correlation:

Correlation measures the linear dependency between pixel values in an image region. High correlation values indicate that pixel values within the region are highly correlated and have a linear relationship.

d) Homogeneity:

Homogeneity represents the closeness of pixel value pairs in an image region to the diagonal. High homogeneity values indicate that pixel values within the region are similar and form a relatively homogeneous texture.

e) Entropy:

Entropy is a measure of the randomness or uncertainty of pixel values within an image region. High-entropy regions have a wide range of pixel values and exhibit more complex and unpredictable textures.

F. Wavelets

Wavelets excel at localizing features, making them ideal for identifying specific regions of interest within the data.

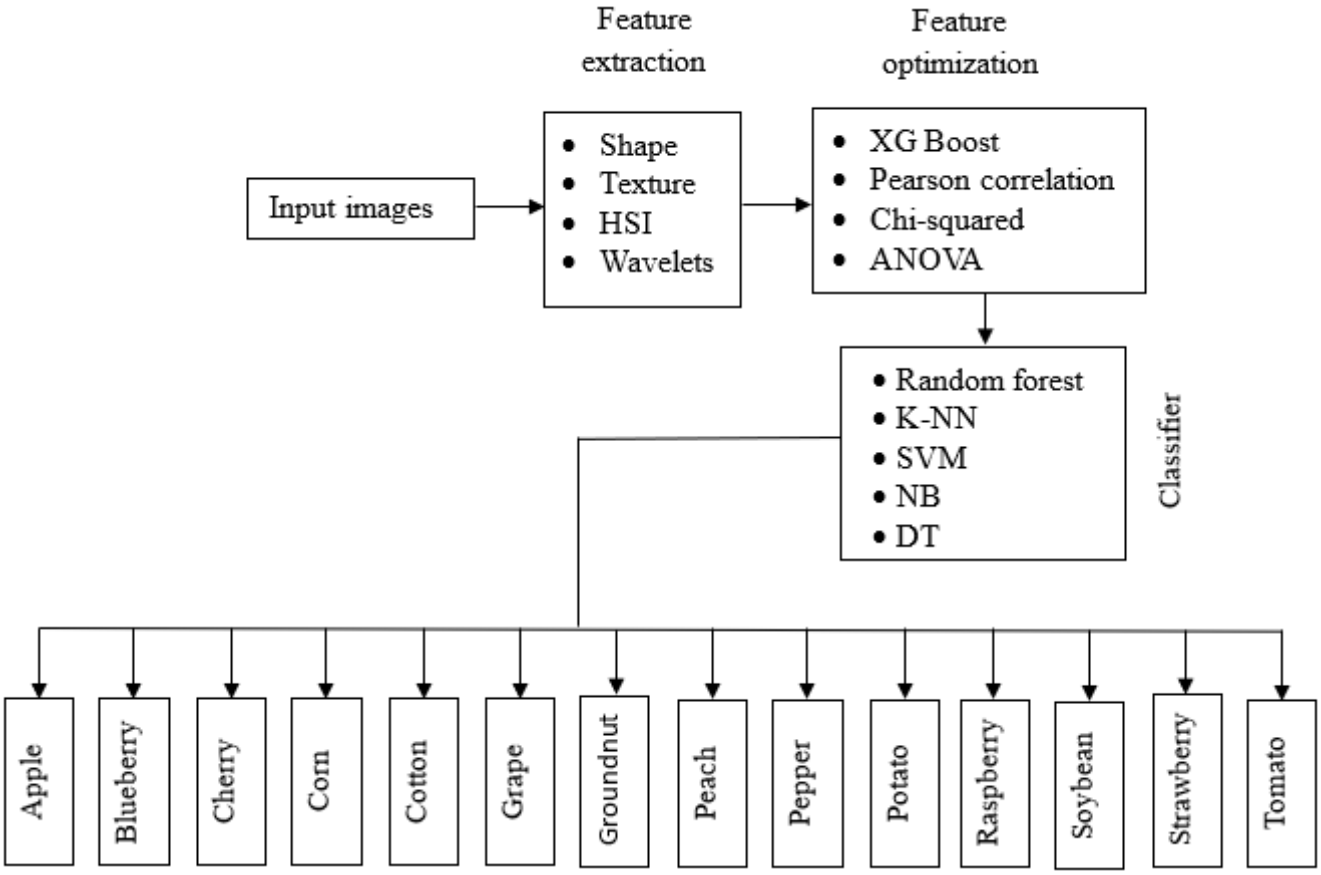

Figure 1. Block diagram of the proposed methodology

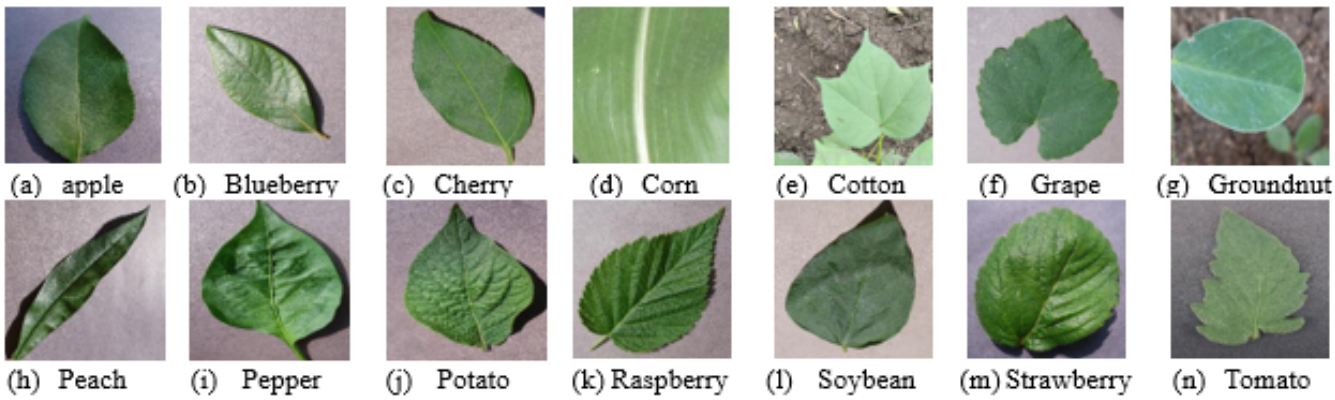

Figure 2. Sample leaf images used for classification of 14 plant types

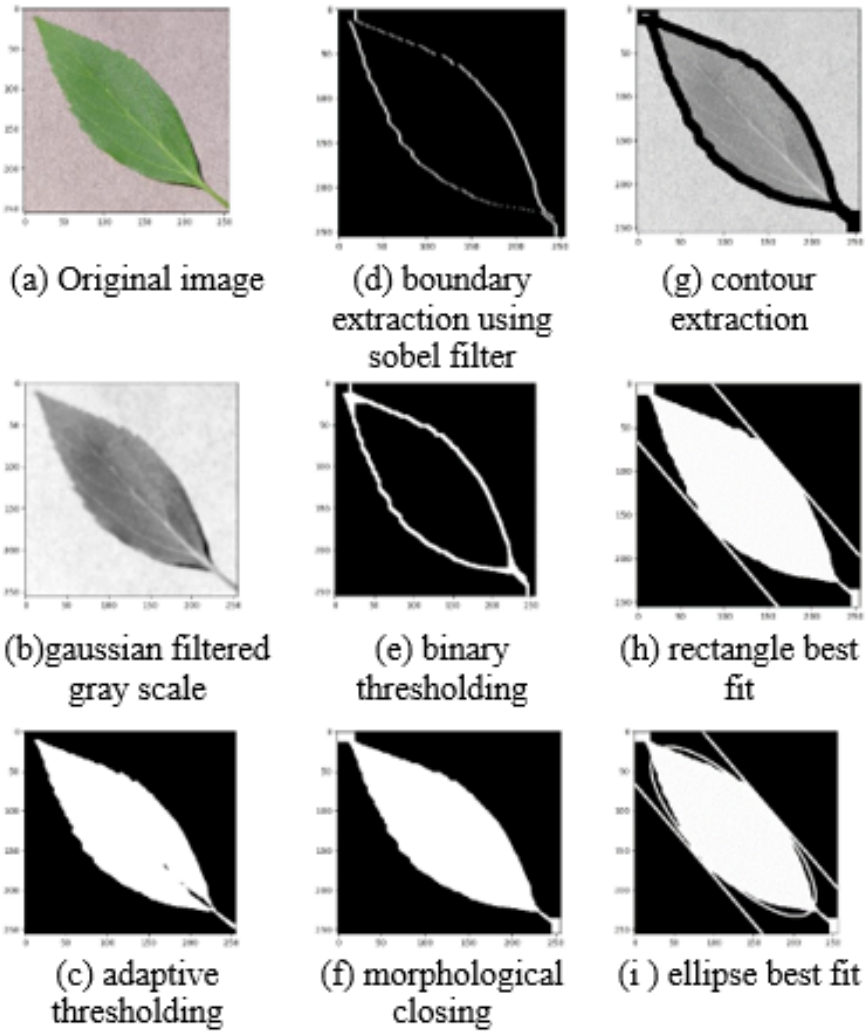

Figure 3. Stepwise output images for shape feature extraction

They offer time-frequency analysis for time-series data and can enhance the robustness of pattern recognition systems to variations in lighting conditions. Discrete Wavelet Transform (DWT) is applied to an image to split it into four sets of coefficients: approximation (cA), horizontal detail (cH), vertical detail (cV), and diagonal detail (cD). These coefficients represent different aspects of the image. The approximation coefficient (cA) captures the low-frequency information and provides an approximation of the original image at a coarser scale. All the related approximation coefficient values are given in Table 4.

DWT with the 'bior1.3' biorthogonal wavelet is used to decompose the grayscale image into its component coefficients, with a focus on the cA coefficient, which contains low-frequency information. Feature values, such as mean, standard deviation, and entropy, are calculated from the cA coefficient.

G. Feature optimization

With the 28 features extracted as discussed in the above section, there is a need to reduce the feature vector to improve the performance of the models. Optimizing features also aids in generalization to new data, avoiding the "curse of dimensionality," and reducing noise and redundancy. As there are many methods to optimize the number of features used, the work uses four optimization methods, namely, random forest, XG Boost, Pearson correlation and Chi-squared. These optimizers are involved as it covers feature importance, linear relationship between two continuous variables, measures the degree of association or independence between two categorical variables [21]. Fig 5 and Fig 6 shows the feature score using random forest and XG boost feature reduction techniques.

The output of Pearson and Spearman optimizers is given in Table 5.

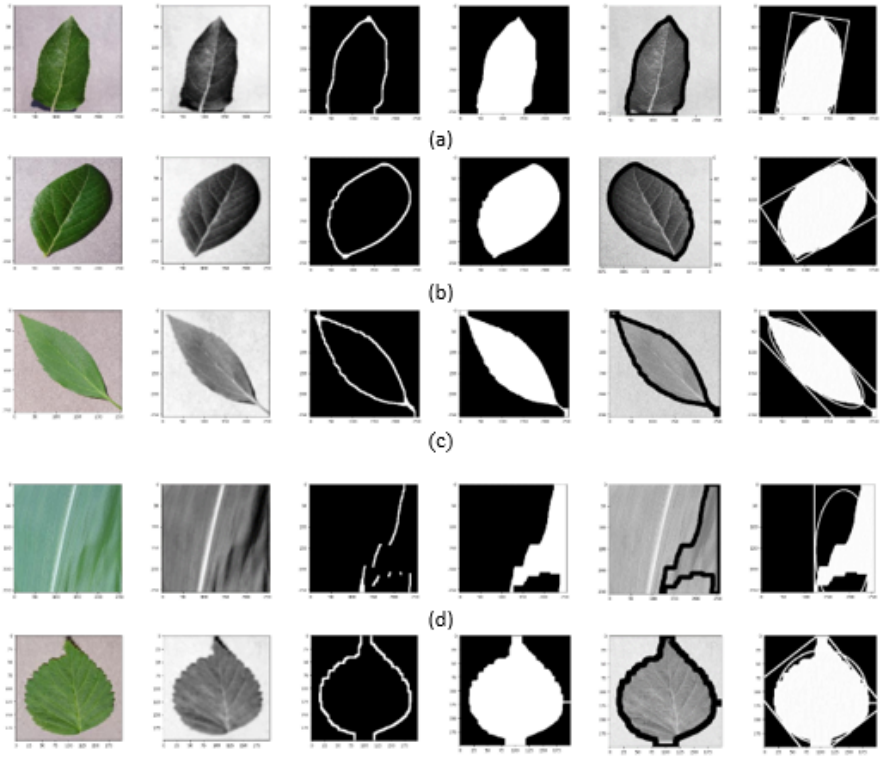

Figure 4. Stepwise output images for shape feature extraction

Sl no	Plant type	Perimeter	Aspect ratio	Rectangularity	Circularity	Diameter
1	Apple	669.8478	0.8750	1.3145	16.8524	184.1195
		663.5635	1.0220	1.2295	16.1684	186.2099
2	Blueberry	467.2447	0.8101	1.5747	16.9989	127.8758
		465.6884	0.5580	1.6854	19.9940	117.5168
3	Cherry	596.1148	0.7550	1.8043	21.2309	145.9824
		587.127	0.8000	1.7874	19.2547	150.9795
4	Corn	56.24264	0.1481	1.6000	46.8627	9.2705
		24.48528	0.3333	2.2857	28.5490	5.1708
5	Cotton	435.4214	0.7578	4.1804	63.8350	61.4940
		411.6396	2.0441	1.3137	23.5523	95.7095
6	Grape	720.5513	0.9100	1.2480	17.8016	192.7036
		740.2742	0.9891	1.4315	23.1724	173.5248
7	Groundnut	46.87006	1.0666	2.3188	21.2251	11.4795
		41.31371	3.1666	1.727	25.8609	9.1669
8	Peach	60.38478	1.7692	1.6032	19.5513	15.4097
		91.65685	8.8000	1.7187	65.6326	12.7661
9	Pepper	951.2447	1.0000	1.5943	36.0677	178.7261
		810.1737	0.9621	1.6949	33.7844	157.2804
10	Potato	685.9899	1.0752	1.3502	17.0804	187.2939
		693.3208	1.1111	1.2971	17.3201	187.9810
11	Raspberry	702.9016	1.0000	2.1256	26.2552	154.7896
		801.7645	1.0108	1.8349	34.0967	154.9335
12	Soybean	688.0904	1.1111	1.4043	18.4699	180.6623
		638.0732	1.0000	1.6018	18.8507	165.8295
13	Strawberry	101.6569	1.0000	1.0816	15.3325	29.2944
		127.5563	0.9166	1.0994	15.0584	37.0909
14	Tomato	149.3137	1.4687	1.0843	16.0739	42.0236
		299.7401	0.8961	1.3529	22.8785	70.7107

Figure 5. Various shape feature values

Sl no	Plant type	Contrast	Dissimilarity	Homogeneity	Energy	Correlation
1	Apple	70.2840	5.3863	0.2292	0.0236	0.9715
		72.9474	4.7926	0.2563	0.0221	0.9788
2	Blueberry	379.2394	12.0985	0.1634	0.0216	0.9570
		370.6507	12.4782	0.1384	0.0164	0.9303
3	Cherry	236.4059	11.0361	0.0967	0.0191	0.8764
		224.6254	10.0343	0.1229	0.0213	0.9018
4	Corn	18.4337	2.6298	0.3642	0.0400	0.9804
		34.0469	3.6284	0.3145	0.0372	0.9737
5	Cotton	184.9426	8.8845	0.1825	0.0212	0.9420
		130.4759	7.3272	0.1974	0.0245	0.9503
6	Grape	448.4420	14.4132	0.1163	0.0181	0.8853
		468.5253	14.7180	0.1213	0.0189	0.8671
7	Groundnut	27.7949	3.1552	0.3289	0.0381	0.9874
		29.2634	3.0941	0.3443	0.0435	0.9824
8	Peach	256.3978	9.9221	0.2014	0.0269	0.9621
		260.5193	10.2765	0.1507	0.0175	0.9139
9	Pepper	296.7717	10.5785	0.1480	0.0172	0.9254
		280.5926	10.1306	0.2068	0.0205	0.9428
10	Potato	1013.8350	23.6271	0.0513	0.0092	0.7696
		1009.6440	23.7471	0.0480	0.0092	0.7949
11	Raspberry	512.5751	14.8640	0.1412	0.0192	0.8472
		502.5683	16.0301	0.0883	0.0133	0.8709
12	Soybean	320.2494	11.7173	0.1333	0.0223	0.9161
		306.5886	12.3405	0.0975	0.0156	0.9405
13	Strawberry	522.6517	17.0725	0.0673	0.0115	0.7922
		509.6715	15.2711	0.1046	0.0153	0.8234
14	Tomato	992.8949	23.9689	0.0471	0.0102	0.6979
		877.5383	21.8491	0.0722	0.0141	0.7827

Figure 6. Various texture feature values

Feature	Apple	Blueberry	Cherry	Corn
Hue - Energy	97.4285	61.0480	191.7053	104.2124
Hue - contrast	181.8377	38.4365	0.8673	1.1970
Hue - correlation	0.9535	0.9861	0.8061	0.9914
Hue - Homogeneity	0.7624	0.7415	4.5945	0.8225
Hue - Entropy	5.6070	5.1675	96.2552	4.9124
Saturation - Energy	103.8204	107.6017	105.8958	68.5357
Saturation - contrast	70.6062	389.1749	0.9659	31.7439
Saturation - correlation	0.9741	0.9424	0.2634	0.9573
Saturation - Homogeneity	0.2303	0.1367	6.6064	0.2978
Saturation - Entropy	6.8486	7.6545	114.4194	6.0038
Intensity - Energy	84.5830	114.2419	163.1568	98.1944
Intensity - contrast	144.8161	502.1284	0.9192	25.8177
Intensity - correlation	0.9423	0.8768	0.3216	0.9021
Intensity - Homogeneity	0.4381	0.2064	6.4844	0.4386
Intensity - Entropy	6.1266	6.7933	191.7053	4.5537

Figure 7. Various HSI feature values

Feature	Apple	Blueberry	Cherry	Corn
Hue - Energy	97.4285	61.0480	191.7053	104.2124
Hue - contrast	181.8377	38.4365	0.8673	1.1970
Hue - correlation	0.9535	0.9861	0.8061	0.9914
Hue - Homogeneity	0.7624	0.7415	4.5945	0.8225
Hue - Entropy	5.6070	5.1675	96.2552	4.9124
Saturation - Energy	103.8204	107.6017	105.8958	68.5357
Saturation - contrast	70.6062	389.1749	0.9659	31.7439
Saturation - correlation	0.9741	0.9424	0.2634	0.9573
Saturation - Homogeneity	0.2303	0.1367	6.6064	0.2978
Saturation - Entropy	6.8486	7.6545	114.4194	6.0038
Intensity - Energy	84.5830	114.2419	163.1568	98.1944
Intensity - contrast	144.8161	502.1284	0.9192	25.8177
Intensity - correlation	0.9423	0.8768	0.3216	0.9021
Intensity - Homogeneity	0.4381	0.2064	6.4844	0.4386
Intensity - Entropy	6.1266	6.7933	191.7053	4.5537

Figure 8. Various wavelet approximation co-efficient values

Table 5. values of feature optimizer using Pearson and spearman correlation

4. Classifiers

These sections elaborate on the usage of different classifiers existing in literature. The work showcases the performance of classifiers, namely, random forest, k-NN, SVM, naïve bayes and decision tree.

A. *Random Forest (RF)*

The RF classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector [22]. The RF classifier used for this study consists of using randomly selected features or a combination of features at each node to grow a tree. There are many features in the RF classifier: (1) each time a tree is grown to the maximum depth on new training data using a combination of features. These full-grown trees are not pruned. (2) As the number of trees increases, the generalization error always converges even without pruning the tree, and overfitting is not a problem because of the strong law of large numbers. Hyper parameters used for RF is given in Table 6.

B. *K-NN classifier:*

The k-Nearest Neighbors (k-NN) algorithm operates on the principle that similar data points tend to belong to the same class or have similar numerical values. This algorithm involves measuring the distance between data points in a dataset and selecting the k-nearest neighbors to make predictions. The most critical hyperparameter in k-NN is ”k,” which determines the number of neighbors to consider and influences the shape of decision boundaries. One notable aspect of k-NN is its simplicity, as it makes no assumptions about data distribution and effectively handles multi-class classification tasks. However, k-NN can be computationally expensive for large datasets and is sensitive to the choice of distance metric and ”k” value. Despite these challenges, k-NN finds applications in various domains, including recommendation systems, image classification, and anomaly detection, especially in cases where data distribution is not well-defined. With fewer hyperparameters, typically involving the number of neighbors (”n”) and uniform weights, k-NN offers a straightforward yet versatile approach to pattern recognition and classification tasks. Additionally, k-NN can be adapted for regression tasks, where instead of classifying data points, it predicts numerical values based on the average or weighted average of the k-nearest neighbors. This flexibility further extends the applicability of k-NN in various domains, such as predicting housing prices or estimating stock prices. Despite its simplicity, k-NN remains a powerful and widely used algorithm in the field of machine learning, offering an intuitive and effective approach to data analysis and prediction tasks [23].

C. *Support Vector Machine (SVM) classifier:*

The Support Vector Machine (SVM) classifier stands as a cornerstone in the realm of machine learning, renowned for its versatility and robustness across a myriad of applications. It excels in tasks ranging from binary classification to multi-class classification and regression, owing to its ability to discern optimal hyperplanes that maximize the margin between distinct classes in the feature space. A key attribute of SVM lies in its adaptability to handle both linearly separable and non-linearly separable data through the utilization of kernel functions. By applying the kernel trick, SVM transforms the original feature space into a higher-dimensional space where linear separation is feasible. This flexibility enables SVM to tackle complex decision boundaries and nonlinear relationships between input features with remarkable efficacy. Among the various kernel functions available, such as linear, polynomial, and Radial Basis Function (RBF) kernels, the choice depends on the inherent characteristics of the dataset and the spe-

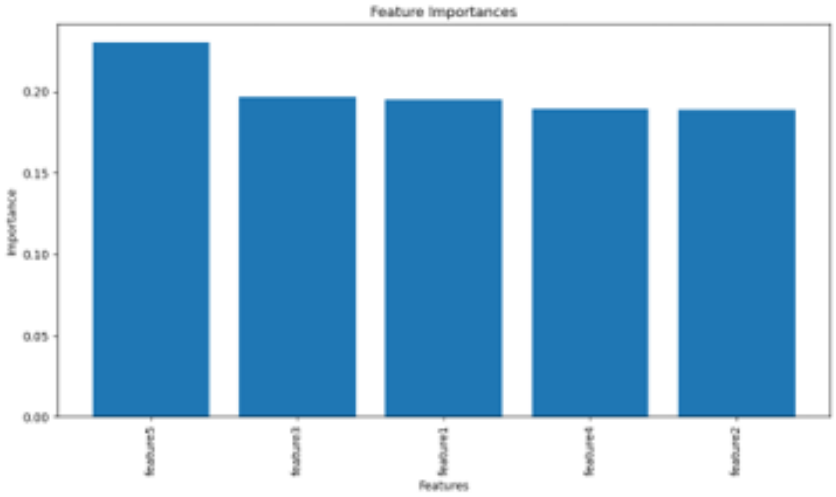

Figure 9. Feature optimization using random forest using feature importance

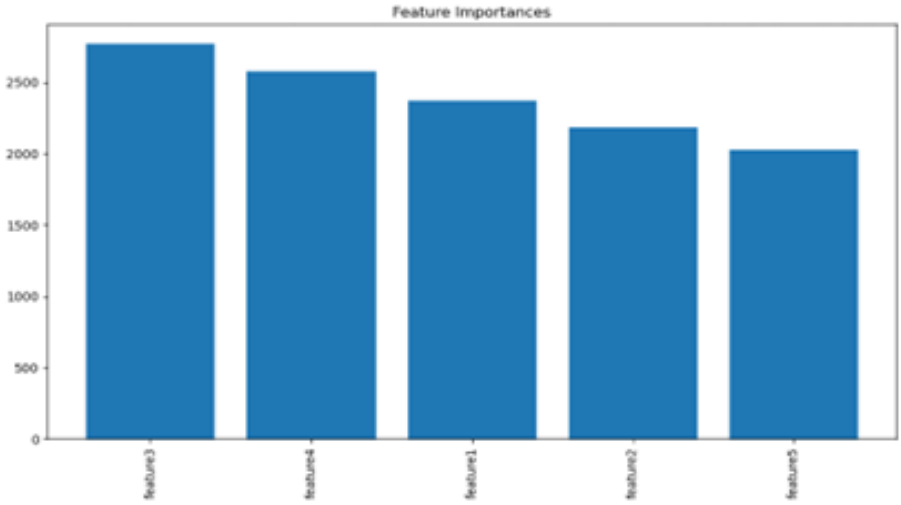

Figure 10. Feature optimization using random forest using feature importance

	Random forest		SVM	
Sl no	Hyper parameter	Value	Hyper parameter	Value
1	Number of estimators	200	C	1
2	Split criterion	gini	'kernel'	Poly
3	Maximum depth of trees	20	'degree'	4
4	Minimum samples for split	5	'gamma'	0.1
5	Minimum sample for leaf	4	'class_weight'	Balanced
6	Maximum features	auto	'probability'	True

Figure 11. hyper parameters of random forest classifier

Figure 12. Feature optimization using XG Boost using feature importance

cific problem at hand. While linear kernels suffice for linearly separable data, polynomial and RBF kernels offer greater flexibility for capturing intricate patterns in non-linear datasets. Furthermore, the regularization parameter (C) plays a pivotal role in SVM's optimization process, dictating the balance between maximizing the margin and minimizing classification errors. Proper tuning of the regularization parameter is crucial to achieving optimal model performance and preventing overfitting or underfitting. Despite its strengths, SVM does have its limitations. The computational complexity of SVM, particularly for large datasets, can pose challenges in terms of training time and resource utilization. Additionally, the selection of appropriate hyperparameters and kernel functions requires careful experimentation and tuning, which can be time-consuming and computationally intensive. Moreover, the interpretability of SVM models, especially when employing non-linear kernels, may be compromised, making it challenging to interpret the underlying decision-making process [24].

It also offers a kernel trick that enables it to handle non-linear decision boundaries effectively by transforming data into higher-dimensional spaces using kernel functions like the linear, polynomial, or Radial Basis Function (RBF) kernels. The choice of the regularization parameter (C) is crucial, as it balances the trade-off between maximizing the margin and minimizing classification errors. SVM is particularly effective for high-dimensional data, robust against overfitting when the C parameter is appropriately set. However, it can be computationally expensive for large datasets, requires careful tuning of hyperparameters and kernel selection, and may pose challenges in terms of interpretability, especially when non-linear kernels are used [24].

D. Naïve bayes classifier:

The Naive Bayes classifier is a probabilistic machine learning algorithm that leverages Bayes' theorem to make predictions based on the probability of data points belonging to specific classes. It simplifies calculations by assuming feature independence, making it "naive." This classifier comes in various variants, including Multinomial, Gaussian, and Bernoulli, each suitable for different types of data. Naive Bayes calculates the likelihood of features given a class and the prior probability of the class to estimate the probability of a data point belonging to that class.It's particularly efficient, works well with high-dimensional data, and is often used in text classification tasks like spam filtering and sentiment analysis.

Multinomial classifier of naïve bayes is chosen, as it can handle imbalanced class distributions and provide interpretable probability scores that enhances its utility. Hyper parameters of this classifier are few, namely, alpha, fit prior, class prior[25].

E. Decision tree classifier

The tree is constructed recursively, with nodes representing decisions, branches representing possible outcomes, and leaves indicating class labels or numerical predictions. Decision Trees are known for their simplicity and ability to handle both categorical and numerical data. They can handle complex interactions between features, making them suitable for a wide range of problems. However, they are prone to overfitting, particularly when the tree becomes too deep. Popular variants of Decision Trees include Random Forests and Gradient Boosted Trees, which improve performance and robustness by aggregating multiple trees [26].

5. Result and discussion

Even though the work seems to be simple as every researcher is behind deep learning, the novelty of the work lies in getting inputs from agriculturist and deploying machine learning models. Around 28 features are extracted from preprocessed images. Features are reduced using optimization techniques. In search of the best machine learning model, five different classifiers are used with exhaustive experimentation by varying hyper parameters. Instead of showing the performance of various classifiers individually, Fig. 7., shows the performance of all the classifiers using optimized feature-set. Since the SVM is found to classify better when compared to other classifiers as shown in Fig 7. Further, the behavior of SVM classifier to classify each plant is carried out and resulted performance is shown in Fig 8. Corn and cherry leaf classification accuracy are showing

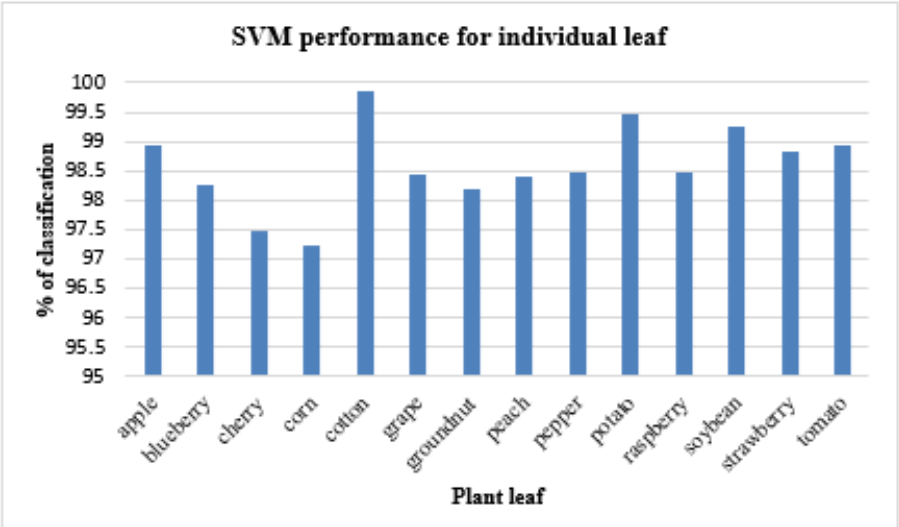

Figure 13. Performance of the model considering individual plant

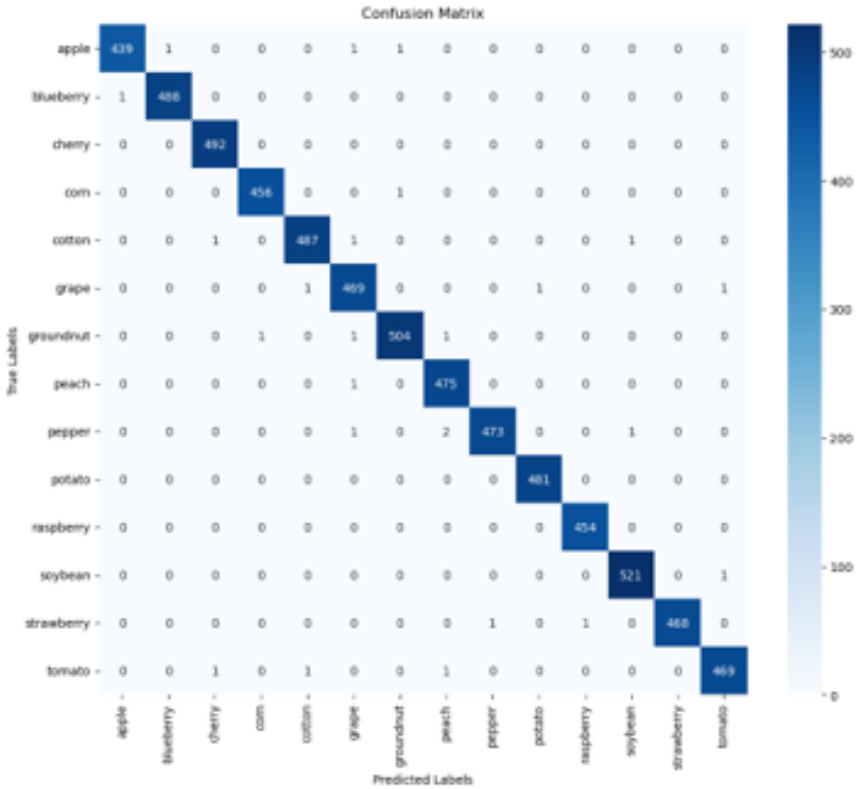

Figure 14. Confusion matrix of the proposed classifier

lesser as shape features perform poor. The confusion matrix of the model proposed is shown in Fig 9.

6. Conclusion

An optimized feature set is used to classify 14 different plants from 20,357 leaf images. Various features extraction techniques are used and are reduced using the most popular feature reduction methods. Although literature on this kind of work used deep learning for image classification, the novelty of the work identifies machine learning methodology with reduced number of features. Out of 5 different classifiers, SVM classifier is found to be the best performing achieving and accuracy of 99.59 %. The obtained results are compared with most recently cited related work which surpasses the cited works in terms of classification accuracy. The work finds applicable in smart agriculture and supports for maintain good ecosystem for mankind.

References

[1] Kwaghtyo, D.K., Eke, C.I. Smart farming prediction models for precision agriculture: a comprehensive survey. Artif Intell Rev 56, 5729–5772 (2023). https://doi.org/10.1007/s10462-022-10266-6

[2] Chakraborty, S.K., Chandel, N.S., Jat, D. et al. Deep learning approaches and interventions for futuristic engineering in agriculture. Neural Comput & Applic 34, 20539–20573 (2022). https://doi.org/10.1007/s00521-022-07744-x

[3] Yao, J., Tran, S.N., Sawyer, S. et al. Machine learning for leaf disease classification: data, techniques and applications. Artif Intell Rev 56 (Suppl 3), 3571–3616 (2023). https://doi.org/10.1007/s10462-023-10610-4

[4] Bustios, P., Garcia Rosa, J.L. Incorporating hand-crafted features into deep learning models for motor imagery EEG-based classification. Appl Intell (2023). https://doi.org/10.1007/s10489-023-05134-x

[5] E. Odat and Q. M. Yaseen, "A Novel Machine Learning Approach for Android Malware Detection Based on the Co-Existence of Features," in IEEE Access, vol. 11, pp. 15471-15484, 2023, doi: 10.1109/ACCESS.2023.3244656.

[6] Y. Xu, Z. Yu, W. Cao and C. L. P. Chen, "A Novel Classifier Ensemble Method Based on Subspace Enhancement for High-Dimensional Data Classification," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 16-30, 1 Jan. 2023, doi: 10.1109/TKDE.2021.3087517.

[7] Agarwal, D. A machine learning framework for the identification of crops and weeds based on shape curvature and texture properties. Int. j. inf. tecnol. (2023). https://doi.org/10.1007/s41870-023-01598-9

[8] Chen H, Han Y, Liu Y, Liu D, Jiang L, Huang K, Wang H, Guo L, Wang X, Wang J, Xue W. Classification models for Tobacco Mosaic Virus and Potato Virus Y using hyperspectral and machine learning techniques. Front Plant Sci. 2023 Oct 16;14:1211617. doi: 10.3389/fpls.2023.1211617. PMID: 37915507; PMCID: PMC10617679.

[9] Alzoubi S, Jawarneh M, Bsoul Q, Keshta I, Soni M, Khan MA. An advanced approach for fig leaf disease detection and classification: Leveraging image processing and enhanced support vector machine methodology. Open Life Sci. 2023 Nov 24;18(1):20220764. doi: 10.1515/biol-2022-0764. PMID: 38027230; PMCID: PMC10668111.

[10] Wu Q, Ma X, Liu H, Bi C, Yu H, Liang M, Zhang J, Li Q, Tang Y, Ye G. A classification method for soybean leaf diseases based on an improved ConvNeXt model. Sci Rep. 2023 Nov 6;13(1):19141. doi: 10.1038/s41598-023-46492-3. PMID: 37932395; PMCID: PMC10628197.

[11] Lu Z, Huang S, Zhang X, Shi Y, Yang W, Zhu L, Huang C. Intelligent identification on cotton verticillium wilt based on spectral and image feature fusion. Plant Methods. 2023 Jul 29;19(1):75. doi: 10.1186/s13007-023-01056-4. PMID: 37516875; PMCID: PMC10385904.

[12] Jadhav, S.B., Patil, S.B. Plant leaf species identification using LBHPG feature extraction and machine learning classifier technique. Soft Comput (2023). https://doi.org/10.1007/s00500-023-09358-4

[13] Hassan, S.M., Maji, A.K. Deep feature-based plant disease identification using machine learning classifier. Innovations Syst Softw Eng (2022). https://doi.org/10.1007/s11334-022-00513-y

[14] Ghosh, A., Roy, P. An automated model for leaf image-based plant recognition: an optimal feature-based machine learning approach. Innovations Syst Softw Eng (2022). https://doi.org/10.1007/s11334-022-00440-y

[15] Shailendra, R., Jayapalan, A., Velayutham, S. et al. An IoT and Machine Learning Based Intelligent System for the Classification of Therapeutic Plants. Neural Process Lett 54, 4465–4493 (2022). https://doi.org/10.1007/s11063-022-10818-5

[16] J. w. Tan, S. -W. Chang, S. Abdul-Kareem, H. J. Yap and K. -T. Yong, "Deep Learning for Plant Species Classification Using Leaf Vein Morphometric," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 1, pp. 82-90, 1 Jan.-Feb. 2020, doi: 10.1109/TCBB.2018.2848653.

[17] M. Kumar, S. Gupta, X. -Z. Gao and A. Singh, "Plant Species Recognition Using Morphological Features and Adaptive Boosting Methodology," in IEEE Access, vol. 7, pp. 163912-163918, 2019, doi: 10.1109/ACCESS.2019.2952176.

[18] Manvikar, Aishwarya; Reddy, Padmanabha (2023), "Dataset of groundnut plant leaf images for classification and detection", Mendeley Data, V3, doi:10.17632/22p2vcbxfk.3

[19] D3v (2020) Cotton Disease Dataset. https://www.kaggle.com/datasets/janmejaybhoi/cotton-disease-dataset.

[20] Mohanty S.P., Hughes D.P., Salathé M. Using Deep Learning for Image-Based Plant Disease Detection. Front. Plant Sci. 2016; 7:1419. doi: 10.3389/fpls.2016.01419.

[21] Talin, Iffat Ara, Mahmudul Hasan Abid, Md Al-Masrur Khan, Seong-Hoon Kee, and Abdullah-Al Nahid. "Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques." Scientific Reports 12, no. 1 (2022): 2019.

[22] Dutta, P., Paul, S., Cengiz, K., Anand, R., & Majumder, M. (2023). A predictive method for emotional sentiment analysis by machine learning from electroencephalography of brainwave data. In Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain.Academic Press.

[23] Zhang, Zhongheng. "Introduction to machine learning: k-nearest neighbors." Annals of translational medicine 4, no. 11 (2016).

[24] Ahmad, Abdul Rahim, Marzuki Khalid, and Rubiyah Yusof. "Machine learning using support vector machines." Centre for Artificial Intelligence and Robotics (2002).

[25] Itoo, Fayaz, Meenakshi, and Satwinder Singh. "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection." International Journal of Information Technology 13 (2021): 1503-1511.

[26] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2, no. 01 (2021): 20-28.

[27] Ossani PC, de Souza DC, Rossoni DF, Resende LV. Machine learning in classification and identification of nonconventional vegetables. J Food Sci. 2020 Dec;85(12):4194-4200. doi: 10.1111/1750-3841.15514. Epub 2020 Nov 10. PMID: 33174205. Li Y, Al-Sarayreh M, Irie K, Hackell D, Bourdot G, Reis MM, Ghamkhar K. Identification of Weeds Based on Hyperspectral Imaging and Machine Learning. Front Plant Sci. 2021 Jan 25;11:611622. doi: 10.3389/fpls.2020.611622. PMID: 33569069; PMCID: PMC7868399.

[28] Nawaz M, Nazir T, Javed A, Masood M, Rashid J, Kim J, Hussain A. A robust deep learning approach for tomato plant leaf disease localization and classification. Sci Rep. 2022 Nov 3;12(1):18568. doi: 10.1038/s41598-022-21498-5. PMID: 36329073; PMCID: PMC9633769.