International Journal of Computing and Digital Systems
2025, VOL. 17, NO. 1, 1–09

http://dx.doi.org/10.12785/ijcds/1571020612

Optimized Feature-Set In Classification of Plant Leaves Images Using Machine Learning Models

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

[1,2]*Department of Electronics and Communication Engineering KLS Gogte Institute of Technology Belagavi Karnataka India-590006 and affiliated to Visvesvaraya Technological University Belagavi Karnataka India-590008*

Received 18 April 2024, Revised 07 October 2024, Accepted 25 October 2024

Abstract: Saving the earth becomes the utmost priority and responsibility of any individual. Environmental and ecosystem health assessments 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 classifying plant leaves. The work uses fourteen different plant leaves, namely, apple, blueberry, cherry, corn, cotton, grape, groundnut, peach, pepper, potato, raspberry, soybean, strawberry, and tomato. Around 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, chi-squared and ANOVA. 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. 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.

Keywords: Ecosystem:Biodiversity:Classification:HSI: Wavelets:

1. Introduction

Smart Agriculture, a transformative approach to farming, integrates innovative technologies to revolutionize traditional agricultural practices. Utilizing 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. Plant leaf classification using image processing has taken a tremendous turn in the years, which hard press the reason to take up the study. 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 comprise of many attributes, namely, shape, color, texture, and margin characteristics, which serve as distinctive markers for species differentiation. These features enable 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 features 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, each having their unique strengths and limitations [5]. Ensemble methods, combining 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

To know the state-of-the-art methods in the related study, following literature survey is carried out and gist of the

E-mail address: njinamdar@git.edu, manjunathm@git.edu

papers are discussed. Many researchers have contributed towards leaf identification and detection of types, diseases and the like. The proposed method involves categorizing weeds by combining handcrafted shape and texture features at the feature level [7]. Support Vector Machine (SVM) is used for classification having given an accuracy of 93.67% using shape curvature features. Identifying and assessing the severity of PVY and TMV infections in tobacco leaves using hyperspectral imaging is given in [9]. Three preprocessing techniques—MSC, SNV, and SavGol—to spectral data spanning the full length of the leaves are adopted. The combination of SavGol with SVM proves highly effective, achieving a remarkable 98.1% average precision in distinguishing various PVY severity levels and 96.2 in classifying different TMV severity levels. While [9] have worked on classification of fig leaf diseases using 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. [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%.

Work on cotton verticillium wilt identification is done using SVM and BPNN classifiers. On the other hand, EfficientNet is used to obtain classification accuracy of 93%, while SG-MN-SPA-BPNN giving an accuracy of 93.78%. Notably, the SG-MN-SPA-FF-BPNN model achieves 98.99% [11]. Binary histograms are used for crop species classification. Combined Probabilistic Neural Network (PNN), k-Nearest Neighbors (KNN), and SVM reported an accuracy of 94.58%. A method is proposed on image classification through Flavia and Swedish datasets. GLCM, LBP and Hu invariant moments are used to train the algorithms [14]. Hardware project on identification of medicinal species is implemented using Raspberry Pi 3 Model B+ achieving the best result obtaining 99% recognition rate. A Convolutional Neural Network (CNN) named D-Leaf is introduced in [16] for leaf classification. The study compares three 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 performance comparable to the pre-trained AlexNet 93.26% and fine-tuned AlexNet 95.54% models. Another study [17] proposes an approach that uses morphological features, such as centroid, major axis length, minor axis length, solidity, perimeter, and orientation, 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%. The novelty of the work carried out lies in deploying features using the knowledge gained by farmers. Contributions of the research lies in Achieving 99.59% accuracy with SVM suggesting a highly effective method for plant leaf classification. Feature Set Optimization is carried out through techniques like XGBoost and Pearson correlation for feature reduction which contributed to a more efficient model by focusing on the most informative features. Automatic plant leaf identification can be a valuable tool for precision farming and disease detection, benefiting agricultural practices.

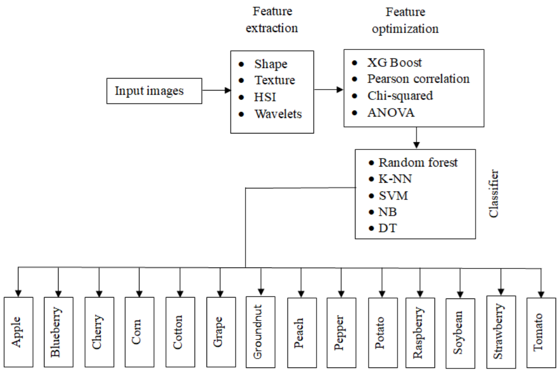


Figure 1. Block diagram of the proposed methodology

3. Methodology

The proposed methodology consists of four stages: preprocessing, feature extraction, feature optimization, and classification, as depicted in Fig. 1. Initially, various features, including shape, texture, HSI, and wavelets, are extracted from input images, totaling around 28 features. Due to the detrimental impact of a higher number of features on model performance, feature optimization techniques such as XGBoost, Pearson correlation, chi-squared, and ANOVA are employed to reduce the feature set. This process narrows down the features to about six for subsequent analysis. To identify the most suitable model for the input images, several classifiers are trained and tested, including random forest, k-nearest neighbor, support vector machine, naïve Bayes, and decision tree. Performance metrics are evaluated through varied hyperparameters, and classification metrics are analyzed to draw conclusions and determine the best performing model.

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.

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

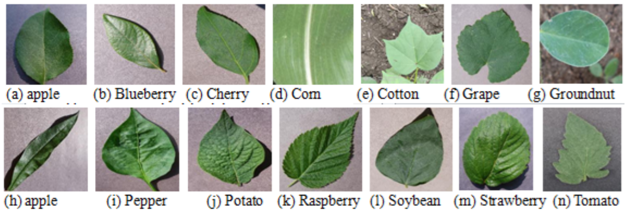


Figure 2. Sample leaf images used for classification of 14 plant types

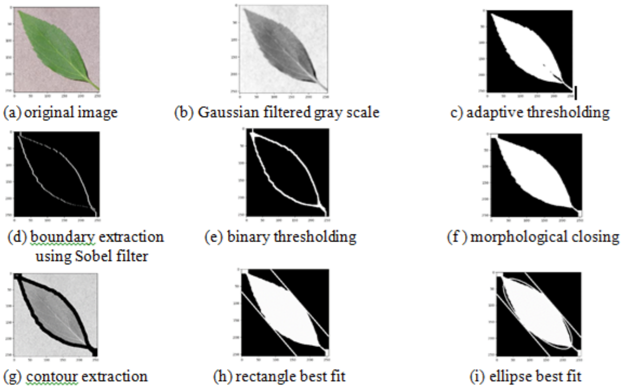


Figure 3. Stepwise output images for shape feature extraction

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] and are chosen due to facts provided by experts are oriented towards these features through image processing.

C. Shape Features:

These features describe the shape and structure of objects or regions within an image. Shape features gives vital information about the spatial arrangement. 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 to use the space optimally.

Five shape features are used to extract shape features 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.
P=$_{(}i = 1)^n l_i..Eq(1)$
$Wherelirepresentsthelengthofeach$
$boundarysegmentiandnisthetotalnumberofboundarysegments.$
$Theaspectratio, AR, iscalculatedas$
$theratioofthewidth(W)totheheight(H)oftheboundingboxofthe$
$object.$

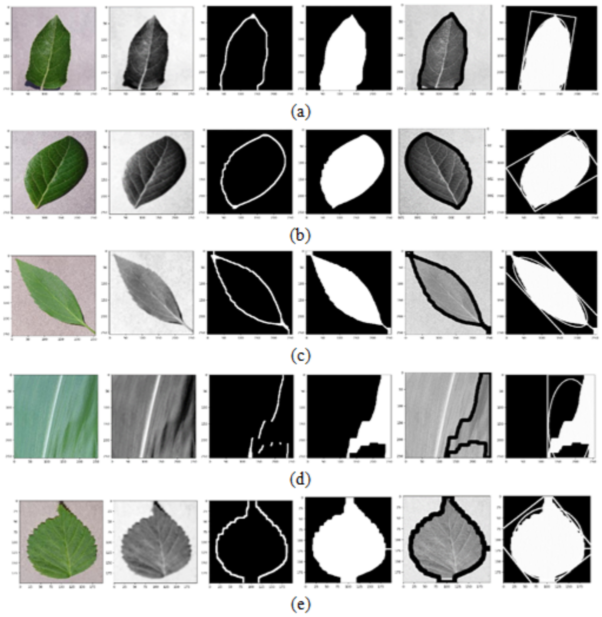


Figure 4. Stepwise output images for shape feature extraction

$AR = W/H.................Eq(2)$
$Rectangularity, R, iscalculatedastheratiooftheobject'sarea$
$(A)tothearea$
$ofitsboundingbox(BB).$
$R = A/BB........Eq(3)$
$Circularity, C, iscalculatedasafunctionoftheobject'sarea(A)$
$andperimeter(P).$
$C = 4A/P^2....................Eq(4)$
$Herearesomecommonshapefeaturesandareusedinthework,$
$perimeter, aspectratio, rectangularity, circularity, anddiameter.$
$Featurevaluesoftwoimagesof14varietiesaregiveninTable1.$

D. Texture Features:

Gray-Level Co-occurrence Matrix (GLCM) texture features are statistical metrics commonly employed 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, thus providing information about the texture and patterns present. Some common GLCM texture features include contrast, dissimilarity, homogeneity, energy, and correlation. In the work, these specific GLCM texture features are utilized, and the corresponding values are presented in Table 2.

E. HSI Color Feature:

The HSI (Hue, Saturation, and Intensity) color space offers an alternative representation of an image compared to the more commonly used RGB color space. The hue channel encodes color information, representing the dominant color of a pixel. The saturation channel represents the intensity or vividness of colors, with high saturation values indicating more vibrant and pure colors, and low values representing desaturated or grayscale regions. The

TABLE I. Various shape feature values

SL no	Plant type	Perimeter	Aspect ratio	Rectangularity	Circularity	Diameter
1	Apple	669.8478 — 663.5635	0.8750 — 1.0220	1.3145 — 1.2295	16.8524 — 16.1684	184.1195 — 186.2099
2	Blueberry	467.2447 — 465.6884	0.8101 — 0.5580	1.5747 — 1.6854	16.9989 — 19.9940	127.8758 — 117.5168
3	Cherry	596.1148 — 587.127	0.7550 — 0.8000	1.8043 — 1.7874	21.2309 — 19.2547	145.9824 — 150.9795
4	Corn	56.2426 — 0.1481	1.6000 — 46.8627	9.2705 — 24.4853	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	46.8701 — 1.0666	2.3188 — 21.2251	11.4795 — 41.3137	3.1666 — 1.7270	25.8609 — 9.1669
8	Peach	60.3848 — 1.7692	1.6032 — 19.5513	15.4097 — 91.6569	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

TABLE II. Various texture 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

intensity channel measures the brightness of the colors. In the work, various metrics are computed for each channel in the HSI color space, including energy, contrast, correlation, homogeneity, and entropy. These metrics, with their related values provided in Table 3, are useful for image recognition as follows: a) Energy: Measures the uniformity or texture of the image. High energy indicates less texture and more uniform regions. b) Contrast: Quantifies the difference in intensity between a pixel and its neighbor over the entire image. High contrast indicates a sharper image with more distinct features. c) Correlation: Assesses how correlated a pixel is to its neighbor over the entire image. High correlation implies a repetitive pattern. d) Homogeneity: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. High homogeneity indicates a smoother texture. e) Entropy: Represents the randomness in the image. High entropy indicates more complexity and diversity in the pixel values.

F. Wavelets:

Wavelets excel at localizing features, making them ideal for identifying specific regions of interest within the data. 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 decompose it into four sets of coefficients: approximation (cA), horizontal detail (cH), vertical detail (cV), and diagonal detail (cD). The values related to the approximation coefficients (cA), which are crucial for summarizing the main information content of the image, are provided in Table 4. These coefficients are essential for understanding the underlying structure and features of the image, and they play a significant role in tasks such as image compression, denoising, and feature extraction for image recognition. 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. Feature reduction helps focus on the key clues (features) that separate different plant types as too many features (like every vein and wrinkle) can confuse your detective (model). As there are many methods to optimize the number of features used, the work uses four optimization methods, namely, random forest, XG

TABLE III. Various texture feature values

Feature Energy	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

TABLE IV. Various texture feature values

SL no Entropy$_c A$	Plant type	Mean$_c A$	Std$_c A$	Corn
1	Apple	293.7464— 110.3517	13.2131— 207.7909	99.8450— 13.1590
2	Blueberry	364.4550— 91.7143	13.2931— 320.3815	114.5524— 13.2357
3	Cherry	330.1093— 60.4493	13.3190— 347.7514	67.3546— 13.3146
4	Corn	267.7135— 63.2476	13.3087— 296.4907	53.2749 —13.3232
5	Cotton	290.7258— 85.1987	13.2796— 295.5716	86.7186— 13.2790
6	Grape	262.2370— 73.0528	13.2847— 261.6206	72.2410— 13.2864
7	Groundnut	330.6849— 59.9341	13.3190— 303.4776	82.4168— 13.2840
8	Peach	212.0142— 112.2547	13.0816— 249.2547	107.5163— 13.1679
9	Pepper	218.4804— 64.9777	13.2793— 240.4474	76.0860 —13.2632
10	Potato	247.1427— 59.6585	13.2984— 234.9834	75.4747 —13.2628
11	Raspberry	252.5994— 114.7552	13.1810— 285.4574	82.2530— 13.2810
12	Soybean	247.6792— 89.3175	13.2446— 257.2906	112.6939— 13.1958
13	Strawberry	231.3158— 75.6729	13.2555— 284.4451	74.1916 —13.2906
14	Tomato	243.5522— 36.4301	13.3279— 233.6705	37.6624— 13.3254

Boost, Pearson correlation and Chi-squared. Random Forest and XGBoost are like magnifying glasses, highlighting the most important features for distinguishing leaves. Think lobed vs. unlobed edges, or net vs. parallel vein patterns. Leaf size and leaf perimeter might be practically the same information. Pearson correlation helps identify these redundant details, letting your detective focus on just one (e.g., perimeter). Leaf images do have texture (smooth, rough, hairy). Chi-squared helps see if this texture is truly a helpful clue for identifying the plant, or if it's just random noise that can be ignored [21]. Fig 5 and Fig 6 show the feature score using random forest and XG boost feature reduction techniques.

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

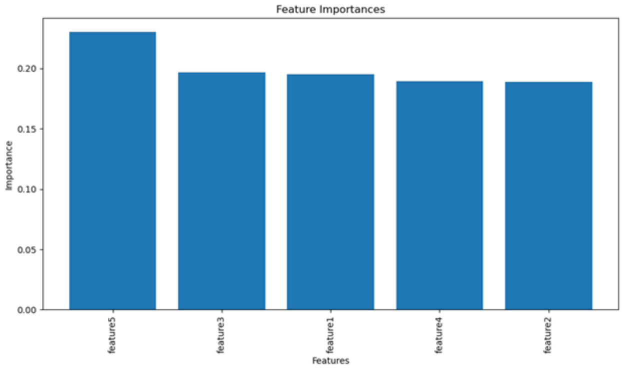


Figure 5. Feature optimization using random forest using feature importance

TABLE V. Values of feature optimizer using Pearson and spearman correlation

Pearson Rank Correlation	Pearson Rank Correlation	Spearman Rank Correlation	Spearman Rank Correlation
Feature	Values	Feature	Values
Feature	Values	Feature	Values
5	0.180034	4	0.208682
4	0.139825	5	0.169557
3	0.102321	3	0.075228
2	0.072335	2	0.062871
1	0.038105	1	0.054530

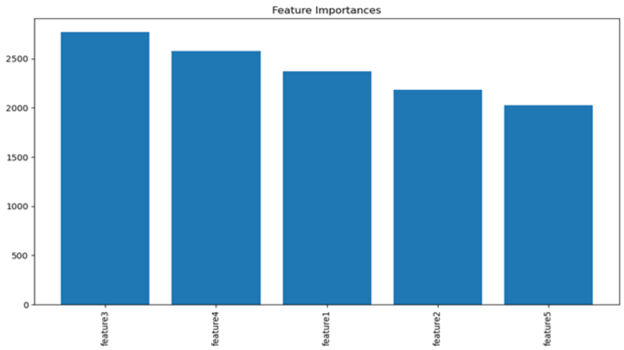


Figure 6. Feature optimization using XG Boost using feature importance.

H. 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.

I. Random Forest (RF):

Each tree in the ensemble contributes a unit vote for the most popular class when classifying an input vector [22]. In this study, the RF classifier utilizes randomly selected features or combinations of features at each node to grow a tree. Several key aspects characterize the RF classifier: 1. Full-Grown Trees: Grown trees will not reduce, which instead allow them to capture intricate relationships within the data. 2. Generalization and Overfitting: As the number of trees in the ensemble increases, the generalization error converges, even without pruning the trees. Overfitting is mitigated due to the strong law of large numbers, ensuring robust performance on unseen data. Hyperparameters are crucial in tuning the performance of the RF classifier. Table 6 provides details of the hyperparameters used in this study, which are set based on the model's performance and the characteristics of the dataset. Fine-tuning these hyperparameters ensures optimal performance and generalization of the RF classifier.

J. K-NN classifier:

The key concept involves measuring the distance between data points in a dataset and selecting the k-nearest neighbors to make predictions. The most critical hyperparameter is ”k,” which determines the number of neighbors to consider and influences the shape of decision boundaries. The algorithm is known for its simplicity, making no assumptions about data distribution, and handling multi-class classification effectively. However, it is computationally expensive when the dataset is large and is sensitive for the value of ”k.” k-NN finds applications in recommendation systems, image classification, and anomaly detection, particularly in cases where data distribution is not well-defined. It possesses less hyper parameter, with ‘n’ neighbors and weights is set to uniform [23].

K. Support Vector Machine (SVM) classifier:

The Support Vector Machine (SVM) classifier is a robust and versatile algorithm used for classification and regression tasks. Its primary goal is to find the optimal hyperplane that maximizes the margin between data points of different classes, making it suitable for linear and non-linear classification problems. SVM offers the kernel trick, enabling it to handle non-linear decision boundaries effectively by transforming data into higher-dimensional spaces using kernel functions like linear, polynomial, or Radial Basis Function (RBF) kernels. The regularization parameter (C) plays a crucial role in SVM, balancing the trade-off between maximizing the margin and minimizing classification errors. When appropriately set, SVM is effective for high-dimensional data and robust against overfitting. However, it can be computationally expensive for large datasets, requires careful hyperparameter tuning and kernel selection, and may pose challenges in interpretability, particularly with non-linear kernels. Despite these considerations, SVM remains widely used and effective in various machine learning applications [24].

L. Naïve bayes classifier:

The Naive Bayes classifier is a probabilistic machine learning algorithm that applies Bayes' theorem to predict the likelihood of data points belonging to specific classes. Its ”naive” assumption of feature independence simplifies calculations, making it computationally efficient. The classifier encompasses various variants such as Multinomial, Gaussian, and Bernoulli, each suited for different data types. It's particularly efficient, especially with high-dimensional data, and finds common use in text classification tasks like

TABLE VI. Values of feature optimizer using Pearson and spearman correlation

Random forest	Random forest	SVM	SVM
Hyper parameter	Value	Hyper parameter	Value
Number of estimators	200	C	1
Split criterion	gini	’kernel’	Poly
Maximum depth of trees	20	’degree’	4
Minimum samples for split	5	’gamma’	0.1
Minimum sample for leaf	4	’class$_{w}eight'$	Balanced
Maximum features	auto	’probability’	True

spam filtering and sentiment analysis. The Multinomial variant of Naive Bayes is often preferred due to its capability to handle imbalanced class distributions and provide interpretable probability scores. Its hyperparameters, including alpha, ’fit$_{p}rior', and'class_{p}rior', arerelativelyfew,$ $offeringeaseofmodeltuningandinterpretation$[25].

M. Decision tree Classifier:

Decision Trees are constructed recursively, with nodes representing decisions, branches representing possible outcomes, and leaves indicating class labels or numerical predictions. They are renowned for their simplicity and versatility in handling both categorical and numerical data. Decision Trees excel in capturing complex interactions between features, rendering them suitable for a diverse array of problems. However, they are susceptible to overfitting, especially when the tree depth increases excessively. To address these shortcomings, popular variants of Decision Trees have been developed, such as Random Forests and Gradient Boosted Trees. These variants enhance performance and robustness by aggregating multiple trees. Random Forests introduce randomness in the tree-building process, while Gradient Boosted Trees sequentially build trees, with each subsequent tree focusing on the errors of its predecessors. These techniques mitigate overfitting and improve the overall predictive accuracy of Decision Trees [26].

4. 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 lesser as shape features perform poor. The confusion matrix of the model (SVM classifier with set hyperparameters as given in Table 6) proposed is shown in Fig 9

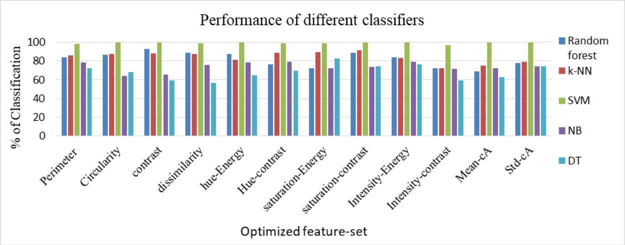


Figure 7. Feature optimization using XG Boost using feature importance.

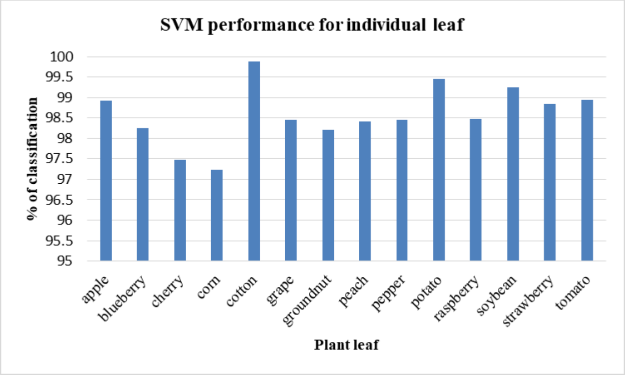


Figure 8. Performance of the model considering individual plant.

A. Comparison with Existing Methods:

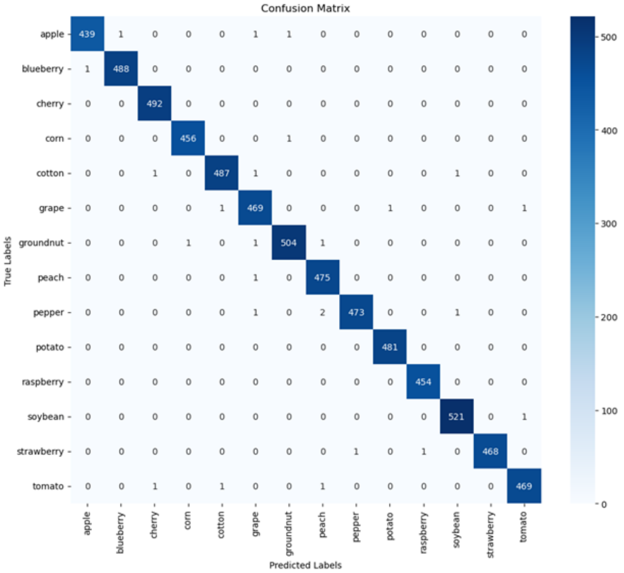
To prove the proposed method is best when compared to others, this section is used to showcase the dominance of the method adopted compared to other methods. In response, the proposed work adopts a machine learning framework, leveraging optimized features to achieve classification accuracy on par with deep learning models. This strategic approach seeks to reconcile the performance gap while circumventing the resource-intensive nature of deep learning methods.

5. Conclusions and Future Work

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

TABLE VII. Comparison with related work:

References	Type of class	No of classes	Accuracy	ML/DL
[27]	Vegetables	5	89	ML
[28]	Grass and weeds	4	88	ML
[29]	Plants	19	99.7	DL
[30]	Citrus	4	91.76	ML
Proposed	Plants leaf	14	99.59	ML



	apple	blueberry	cherry	corn	cotton	grape	groundnut	peach	pepper	potato	raspberry	soybean	strawberry	tomato
apple	439	1	0	0	0	1	1	0	0	0	0	0	0	0
blueberry	1	488	0	0	0	0	0	0	0	0	0	0	0	0
cherry	0	0	492	0	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	456	0	0	1	0	0	0	0	0	0	0
cotton	0	0	1	0	487	1	0	0	0	0	0	1	0	0
grape	0	0	0	0	1	469	0	0	0	1	0	0	0	1
groundnut	0	0	0	1	0	1	504	1	0	0	0	0	0	0
peach	0	0	0	0	0	1	0	475	0	0	0	0	0	0
pepper	0	0	0	0	0	1	0	2	473	0	0	1	0	0
potato	0	0	0	0	0	0	0	0	0	481	0	0	0	0
raspberry	0	0	0	0	0	0	0	0	0	0	454	0	0	0
soybean	0	0	0	0	0	0	0	0	0	0	0	521	0	1
strawberry	0	0	0	0	0	0	0	0	1	0	1	0	468	0
tomato	0	0	1	0	1	0	0	1	0	0	0	0	0	469

Figure 9. Confusion matrix of the proposed classifier.

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. The research, while achieving high accuracy, has room to grow. A wider range of plants and even more image data could make the model more adaptable. Future work could involve using even larger datasets with more plant varieties, precisely measuring the effectiveness of farmer-informed feature selection and exploring techniques to make the model's decision process more transparent. Additionally, pre-training the model on a vast leaf dataset and then specializing it for specific plant types, or even creating a user-friendly app for real-world use by farmers and researchers are all promising avenues for further development. In this paper we have presented an extensible CPU power measurement framework that supports our own research but is also generally applicable to the computer engineering community in general for accurate computational power consumption measurements.

References

[1] D. K. Kwaghtyo and C. I. Eke, "Smart farming prediction models for precision agriculture: a comprehensive survey," Artificial Intelligence Review, vol. 56, no. 6, pp. 5729–5772, 2023.

[2] S. K. Chakraborty, N. S. Chandel, D. Jat, M. K. Tiwari, Y. A. Rajwade, and A. Subeesh, "Deep learning approaches and interventions for futuristic engineering in agriculture," Neural Computing and Applications, vol. 34, no. 23, pp. 20 539–20 573, 2022.

[3] J. Yao, S. N. Tran, S. Sawyer, and S. Garg, "Machine learning for leaf disease classification: data, techniques and applications," Artificial Intelligence Review, vol. 56, no. Suppl 3, pp. 3571–3616, 2023.

[4] P. Bustios and J. L. Garcia Rosa, "Incorporating hand-crafted features into deep learning models for motor imagery eeg-based classification," Applied Intelligence, vol. 53, no. 24, pp. 30 133– 30 147, 2023.

[5] E. Odat and Q. M. Yaseen, "A novel machine learning approach for android malware detection based on the co-existence of features," IEEE Access, vol. 11, pp. 15 471–15 484, 2023.

[6] Y. Xu, Z. Yu, W. Cao, and C. P. Chen, "A novel classifier ensemble method based on subspace enhancement for high-dimensional data classification," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 16–30, 2021

[7] D. Agarwal, "A machine learning framework for the identification of crops and weeds based on shape curvature and texture properties," International Journal of Information Technology, vol. 16, no. 2, pp. 1261–1274, 2024.

[8] H. Chen, Y. Han, Y. Liu, D. Liu, L. Jiang, K. Huang, H. Wang, L. Guo, X. Wang, J. Wang et al., "Classification models for tobacco mosaic virus and potato virus y using hyperspectral and machine learning techniques," Frontiers in Plant Science, vol. 14, p. 1211617, 2023.

[9] S. Alzoubi, M. Jawarneh, Q. Bsoul, I. Keshta, M. Soni, and M. A. Khan, "An advanced approach for fig leaf disease detection and classification: Leveraging image processing and enhanced support vector machine methodology," Open Life Sciences, vol. 18, no. 1, p. 20220764, 2023.

[10] Q. Wu, X. Ma, H. Liu, C. Bi, H. Yu, M. Liang, J. Zhang, Q. Li, Y. Tang, and G. Ye, "A classification method for soybean leaf diseases based on an improved convnext model," Scientific Reports, vol. 13, no. 1, p. 19141, 2023.

[11] Z. Lu, S. Huang, X. Zhang, Y. Shi, W. Yang, L. Zhu, and C. Huang, "Intelligent identification on cotton verticillium wilt based on spectral and image feature fusion," Plant Methods, vol. 19, no. 1, p. 75, 2023.

[12] S. B. Jadhav and S. B. Patil, "Plant leaf species identification using lbhpg feature extraction and machine learning classifier technique," Soft Computing, vol. 28, no. 6, pp. 5609–5623, 2024.

[13] S. M. Hassan and A. K. Maji, "Deep feature-based plant disease identification using machine learning classifier," Innovations in Systems and Software Engineering, pp. 1–11, 2022.

[14] A. Ghosh and P. Roy, "An automated model for leaf imagebased plant recognition: an optimal feature-based machine learning approach," Innovations in Systems and Software Engineering, pp. 1–14, 2022.

[15] R. Shailendra, A. Jayapalan, S. Velayutham, A. Baladhandapani, A. Srivastava, S. Kumar Gupta, and M. Kumar, "An iot and machine learning based intelligent system for the classification of therapeutic plants," Neural Processing Letters, vol. 54, no. 5, pp. 4465–4493, 2022.

[16] J. Wei Tan, S.-W. Chang, S. Abdul-Kareem, H. J. Yap, and K.-T. Yong, "Deep learning for plant species classification using leaf vein morphometric," IEEE/ACM transactions on computational biology and bioinformatics, vol. 17, no. 1, pp. 82–90, 2018.

[17] M. Kumar, S. Gupta, X.-Z. Gao, and A. Singh, "Plant species recognition using morphological features and adaptive boosting methodology," IEEE Access, vol. 7, pp. 163 912–163 918, 2019.

[18] M. Aishwarya and A. P. Reddy, "Dataset of groundnut plant leaf images for classification and detection," Data in Brief, vol. 48, p. 109185, 2023.

[19] S. P. Mohanty, D. P. Hughes, and M. Salathe, "Using deep learning ´ for image-based plant disease detection," Frontiers in plant science, vol. 7, p. 1419, 2016.

[20] I. A. Talin, M. H. Abid, M. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques," Scientific Reports, vol. 12, no. 1, p. 20199, 2022.

[21] P. Dutta, S. Paul, K. Cengiz, R. Anand, and M. Majumder, "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. Elsevier, 2023, pp. 109–130.

[22] Z. Zhang, "Introduction to machine learning: k-nearest neighbors,"Annals of translational medicine, vol. 4, no. 11, 2016

[23] A. R. Ahmad, M. Khalid, and R. Yusof, "Machine learning using support vector machines," Centre for Artificial Intelligence and Robotics, 2002.

[24] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, na¨ıve bayes and knn machine learning algorithms for credit card fraud detection," International Journal of Information Technology, vol. 13, no. 4, pp. 1503–1511, 2021.

[25] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021.

[26] P. C. Ossani, D. C. de Souza, D. F. Rossoni, and L. V. Resende, "Machine learning in classification and identification of nonconventional vegetables," Journal of Food Science, vol. 85, no. 12, pp. 4194–4200, 2020.

[27] Y. Li, M. Al-Sarayreh, K. Irie, D. Hackell, G. Bourdot, M. M. Reis, and K. Ghamkhar, "Identification of weeds based on hyperspectral imaging and machine learning," Frontiers in Plant Science, vol. 11, p. 611622, 2021

[28] M. Nawaz, T. Nazir, A. Javed, M. Masood, J. Rashid, J. Kim, and A. Hussain, "A robust deep learning approach for tomato plant leaf disease localization and classification," Scientific reports, vol. 12, no. 1, p. 18568, 2022.

[29] H. Dang-Ngoc, T. N. Cao, and C. Dang-Nguyen, "Citrus leaf disease detection and classification using hierarchical support vector machine," in 2021 international symposium on electrical and electronics engineering (ISEE). IEEE, 2021, pp. 69–74.

[30] R. B. Koti, M. S. Kakkasageri, and R. S. Pujar, "Artificial neural network for safety information dissemination in vehicle-to-internet networks," ETRI Journal, vol. 45, no. 6, pp. 1065–1078, 2023.