ENHANCING MULTICLASS CLASSIFICATION OF MANGO LEAF DISEASES USING DIMENSIONALITY REDUCTION & DEEP-BASED FEATURE EXTRACTION

Jocelyn L. Garrido
University of Science and Technology of Southern Philippines, Cagayan de Oro City, Philippines
Tel. +639656510866, Email: *jocelyn.garrido@ustp.edu.ph*

ABSTRACT: *The classification of mango leaf diseases was improved by employing dimensionality reduction and deep learning-based feature extraction. The objective was to reduce computational complexity and overfitting caused by high-dimensional input features. Results showed that using PCA with SVM and CNN models achieved higher accuracy rates: SVM with PCA reached 91% accuracy compared to 78% without PCA, and CNN with PCA achieved 93% accuracy compared to 90% without PCA. Furthermore, the models with PCA significantly reduced build times: SVM with PCA took 0.52 seconds compared to 37.6 seconds without PCA, and CNN with PCA reduced build time from 3.23 minutes to 1.78 seconds. The findings suggest that PCA effectively improves accuracy and efficiency in disease classification models. Implementing PCA allows for faster training and better resource utilization without compromising accuracy. This study emphasizes the strength of PCA in addressing challenges associated with high-dimensional data, benefiting the early detection and classification of mango leaf diseases and facilitating targeted treatments to minimize disease spread..*

Keywords: Principal Component Analysis (PCA), Support Vector Machine (SVM), Convolutional Neutral Network (CNN), Mango Disease Classification, Image Processing, Plant Leaf Disease

1. INTRODUCTION

Mango (Mangifera Indica) is a perennial evergreen tree of the family Anarcadiacea, native to South Asia and now globally distributed. Renowned as one of the most cultivated fruit trees in tropical regions, the mango tree exhibits remarkable longevity, with some specimens remaining productive even after three centuries of growth, earning them the moniker "Century Mango Tree". The tree's bountiful yield consists of sweet aromatic kidney-shaped drupe fruits, which have gained immense value as a globally prized commodity. Recognizing the diverse possibilities of the mango, the mango industry has capitalized on its versatility, producing an array of product forms, including fresh fruit, confectioneries and more. With its exceptional nutritional value and multifaceted applications, the mango continues to enthrall consumers and drive the thriving mango industry.

In the Philippines, mango holds the distinction of being the country's third most important fruit crop and ranks as the third most exported fruit crop, following banana and pineapple, according to the Department of Agriculture (DA).

However, based on the current report of the Philippine Mango Industry Roadmap 2021-2025, the mango industry has been confronted with a persisting and formidable challenge of unstable fruit production since 2008. In fact, the 2020 PSA report reveals that the industry has been on a continuous decline in all indicators of industry performance which includes production volume, productive area, as well as yield per unit area, and yield per tree. This decline in productivity and quality can be primarily attributed to the prevalence of pests and diseases, high post-harvest losses, and other related factors. As identified by many stakeholders of the Mango Industry in the Philippines, the worsening pests and diseases such as Cecid fly, Anthracnose, and Sooty Mold have been the root cause of many problems. These diseases affect young and mature fruits, twigs, leaves, and blossom spikes. Severely affected fruits, fall off while new shoots may defoliate. The presence of Cecid fly, Anthracnose, Sooty Mold, and other Mango diseases are manifested in the leaves. Shown in Figure 1 are some samples of the manifestation of the type of disease based on the leaf appearance.

Figure 1: Manifestation of disease based on leaf appearance

Moreover, pests and diseases affecting mango trees are not limited to the Philippines but are a shared concern among other countries involved in the mango industry. For instance, nations such as China[1], Indonesia[2], Thailand[3], India[4], and Pakistan[5] have also encountered significant challenges related to pests and diseases in their mango cultivation. The widespread occurrence of these problems highlights the global nature of the issue.

On the other hand, pest and disease infestations in plants are not uncommon, and numerous studies have been conducted to investigate effective strategies for pest and disease management in various crops. In recent years, there has been a surge of studies employing machine learning (ML) and deep learning (DL) techniques for disease classification in various fields, including agriculture. Several studies have focused on developing ML and DL models capable of accurately detecting and classifying plant diseases based on image analysis [6-10]. These advanced algorithms have shown promising results in early disease detection, enabling timely interventions and facilitating more efficient and targeted control measures.

Building upon this foundation, this current work shifted its focus towards enhancing these models through dimensionality reduction techniques to improve accuracy and processing time. By reducing the number of input features while retaining relevant information, dimensionality reduction methods such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) can effectively optimize ML and DL models for disease classification. These techniques allow for more efficient data representation, enabling faster computation and

improved model performance [13]. The integration of dimensionality reduction with ML and DL algorithms presents a promising avenue for advancing disease classification accuracy and reducing processing time, ultimately enhancing disease management and supporting informed decision-making in agriculture and other relevant domains.

Moreover, the use of Principal Component Analysis (PCA) as dimensionality reduction techniques shows better potential in reducing image space while preserving the most relevant features of the image[11-12]. Thus this study use PCA combined with deep-based classification models.

The utilization of machine learning (ML) and deep learning (DL) techniques for disease classification has shown promising results. However, there is a pressing need to address the challenges of accuracy and processing time by focusing on effective dimensionality reduction approaches. Handling high-dimensional data is very difficult in practice, commonly known as the curse of dimensionality. If the dimensionality of the input dataset increases, any machine learning algorithm and model becomes more complex. As the number of features increases, the number of samples also gets increased proportionally, and the chance of overfitting also increases. If the machine learning model is trained on high-dimensional data, it becomes over-fitted and results in poor performance. Thus, the problem at hand lies in identifying and implementing dimensionality reduction methods that can effectively reduce the feature space while retaining crucial information necessary for precise disease classification. This study aims to enhance multiclass classification model using Dimensionality Reduction & Deep Learning-Based feature extraction for early detection and classification of mango leaf diseases.

B. Overview of Principal Component Analysis (PCA)

In 1991 Turk and Pentland [14], introduced the concept of "eigenface". They harnessed the power of eigen-pictures and their corresponding weights as distinctive features for facial recognition. The paper not only presented an efficient method to compute eigen-pictures but also proposed an algorithm for operating a face recognition system. The work of [14] contributed significant advancement in the field of facial recognition by providing a comprehensive framework based on eigenfaces.

The initialization steps of the face recognition Eigenface approach, as outlined in [14], are as follows:

Let's consider a set of m images (dataset) of dimension n x n (size of the image). Each Image in the dataset can be presented as a Matrix Γ i of Pixels, so the image dataset is represented as vectorized images Γ 1 Γ 2 Γ 3 … Γm.

1. Compute the mean face (average face)

$$\psi = \frac{1}{m}\sum_{i=1}^{m}\Gamma_i \qquad \text{Eq. 1}$$

where

ψ = mean face vector

m = number of images in the dataset

Γ_i = image vector

2. Subtract the mean face ψ from each image vector Γi such that each face differs from the average vector

$$\Phi = \Gamma_i - \psi \qquad \text{Eq. 2}$$

The image dataset now is an Eigen face vector represented as

$$A = \Phi_1, \Phi_2, \ldots \Phi_M$$

$$C = A^T A \qquad \text{Eq. 3}$$

3. This set of large vectors undergoes principal component analysis (PCA) to identify a set of M orthonormal vectors, un that effectively represents the data distribution. Covariance Matrix is used to calculate the Eigenvalues and Eigenvectors in order to determine the Principal Components of the data.

The covariance matrix is used to figure out the similarity of the features. In the context of multivariate data analysis, covariance provides insights into the relationships and patterns between different features or variables. Covariance measures the co-movement or association between two variables. A positive covariance suggests a direct relationship, meaning that as one variable increases, the other tends to increase as well. A negative covariance indicates an inverse relationship, indicating that as one variable increases, the other tends to decrease. A covariance of zero suggests no linear relationship between the variables. However, covariance do not provide a standardized measure of the strength or scale of the relationship.

3. Express the Eigenvectors u in terms of Eigenvalues λ in the covariance matrix.

$$Au_i = \lambda u_i \qquad \text{Eq. 4}$$

To get the Eigenvalue λ the following steps are done:

$$Au_i = \lambda u_i$$

Step 1: Perform a subtraction property of equality.

$$Au_i - \lambda u_i = \lambda u_i - \lambda u_i$$

$$Au_i - \lambda u_i = 0$$

Step 2: Perform vectorization using an identity matrix since Aui is a matrix-vector and λui is a scalar-vector.

$$Au_i - \lambda I u_i = 0$$

$$(A - \lambda I)u_i = 0$$

Step 3: Each Eigenvalue is associated with a specific Eigenvector.

$$\det(A - \lambda I) = 0 \qquad \text{Eq. 5}$$

The eigenvalues associated with the eigenvectors provide a means to determine the relative significance of each eigenvector in representing the variations observed among the images. The columns of ui are ordered by how large their corresponding Eigenvalues. So in the column arrangement of ui, the principal component associated with the largest eigenvalue, representing the most dominant features, will always be positioned in the first column. The subsequent largest eigenvalue will follow in the second column, and so forth. In the data approximation, the dimensions corresponding to the smallest Eigenvalues are omitted or discarded.

2. METHODOLOGY

A workflow shown in Figure 2 was used as the research method to achieve the objectives set in this current work. All processes involved in the workflow were implemented using Google Colab, a cloud-based platform that allows users to run and execute Python code using Jupyter Environment.

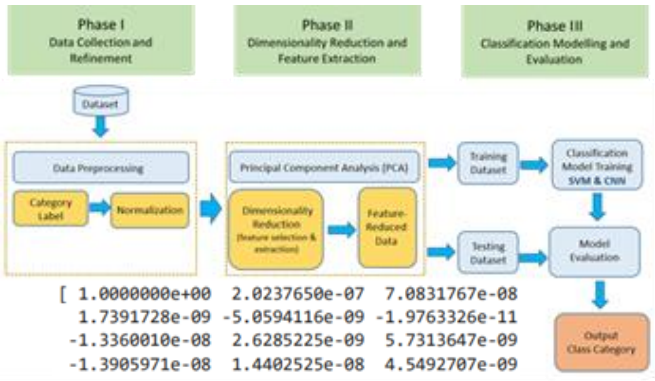


Figure 2: Workflow on Disease Classification using PCA and DL

Phase 1: Data Collection and Refinement

The available dataset for mango leaf diseases that is similar to the identified diseases mentioned in this study for classification was searched online and found a Mango Disease Dataset in Kaggle with 2267 leaf images for training and 611 leaf images for testing.

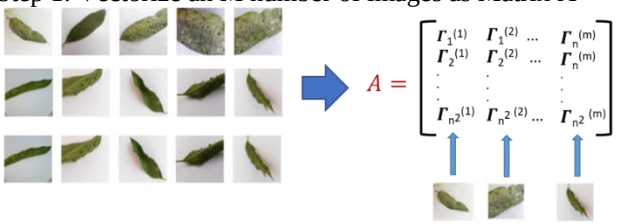
Category or data labeling is then performed in the dataset assigning appropriate class labels to each image. Data labeling establishes the ground truth or the correct reference for training machine learning models

Normalization was done for scale consistency using unit vector scaling. It ensures that the pixel values across different images are on a consistent scale. This allows the machine learning model to learn patterns and features from images without being biased by the differences in pixel intensities.

Phase 2: Dimensionality Reduction and Feature Extraction

The following steps were done to perform dimensionality reduction and feature extraction based on the Principal Component Analysis algorithm.

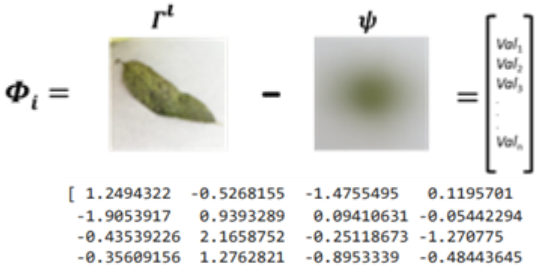
Step 1: Vectorize an M number of images as Matrix A



Step 2: Calculate the mean leaf image.

$$\psi = \frac{1}{m}\sum_{i=1}^{m} \Gamma^i$$

Step 3: Subtract the mean leaf image ψ from each image vector Γi using $\Phi = \Gamma^i - \psi$



This step produces a leaf vector matrix

$$A = \Phi_1, \Phi_2, \ldots \Phi_n$$

Step 4: Calculate the Covariance Matrix to figure out the similarity of the features using Eq.3.

Step 5: Express the Eigenvectors in terms of Eigenvalues in the covariance matrix using Eq.4.

Step 6: Calculate the determinant of the Eigenvector as associated with the Eigenvalues using Eq. 5

This step produces an Eigenvector matrix where the columns are ordered on how large their corresponding Eigenvalues. The principal component with the largest eigenvalue, which represents the most dominant features or variance in the data, is always assigned to the first column of the resulting matrix, followed by the second largest, and so on.

Step 7: For dimensionality reduction, each variable in the original dataset can be represented in terms of the picked K Principal Component or Eigenvectors.

Phase 3: Building the Model

A machine learning model Support Vector Machine (SVM) and a deep learning model Convolutional Neural Network were used as classification models in this study.

Support Vector Machines (SVM) is a machine learning algorithm that aims to find an optimal hyperplane to separate different classes of data points. By maximizing the margin between the hyperplane and the nearest data points, SVM achieves better generalization and improved classification performance. SVM can handle both linearly separable and non-linearly separable data by mapping the data to a higher-dimensional feature space using kernel functions. It is effective in high-dimensional spaces, robust to outliers, and can be used for binary and multi-class classification tasks. However, SVM's performance relies on the choice of hyperparameters and can be computationally demanding for large datasets. Overall, SVM is a versatile and widely used algorithm in various domains for accurate and reliable classification.

CNNs were employed in this context because they are well-suited for analyzing visual data, such as images. Due to their architecture inspired by the human visual cortex, CNNs can effectively extract local features from the input data using convolutional layers. This allows them to capture meaningful patterns and structures in the images. Additionally, the use of pooling layers enables down sampling and the extraction of the most important information. By incorporating fully connected layers, CNNs can further learn global features and perform classification tasks. With their ability to automatically learn and recognize complex visual patterns, CNNs have proven to be highly effective in various computer vision tasks, making them a go-to choice in this field.

Phase 4: Performance Evaluation

To evaluate the performance of the machine and deep learning models utilized in this study, various evaluation metrics were employed, including accuracy, precision, recall, F1-score, and confusion matrix. These metrics provided insights into the effectiveness of the models in accurately classifying the images and assessing their predictive capabilities. By analyzing these evaluation metrics, a comprehensive understanding of the performance and reliability of the machine and deep learning models could be obtained.

3. RESULTS AND DISCUSSIONS

Phase I: Data Collection.

The mango leaf disease dataset used in this study was sourced from Kaggle, consisting of 2,267 images for training. Each disease category contained approximately 400 images, while the testing set comprised 611 images, with around 100 images per category. The images had dimensions of 64 x 64 pixels and included three color channels (R, G, B).

There are 6 disease categories labeled Anthracnose, Bacterial Canker, Gall Midge, Healthy, Powdery Mildew, and Sooty Mold. Figure 3 shows sample dataset for each category.

Figure 3: Sample Dataset for each Disease Category

Phase II. Dimensionality and Feature Extraction using PCA.

Prior to applying PCA, the original dataset consisted of 2,267 images with a total of 12,288 principal components (PCs). To identify the optimal number of PCs that could effectively capture the essential features of the images while minimizing information loss, the proportion of variance explained by each PC was taken into account. Figure 4 reflects the cumulative proportion of variance explained by PCs.

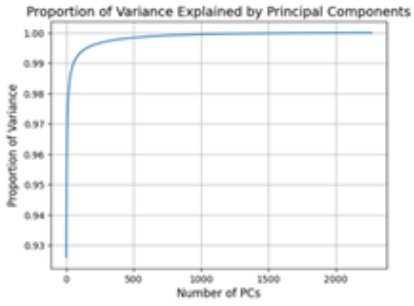


Figure 4: Proportion of Variance explained by PC

Based on Figure 4, it can be observed that by considering 300 principal components (PCs), the accumulated variance has already captured nearly 100% of the significant features that are capable of representing the image.

Based on Figure 4, a decision was made to utilize 300 principal components (PCs) instead of the original 12,288 PCs from the dataset. This resulted in a significant reduction of approximately 97% in the image space (Figure 5).

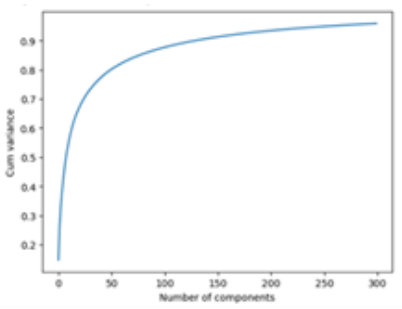


Figure 5: Reduced PC

Figure 6 displays the mean leaf of the leaf images, which represents the average appearance of the dataset. On the other hand, Figure 7 illustrates how the leaf images are reconstructed based on different numbers of principal components (PCs).

Figure 6: Mean Leaf

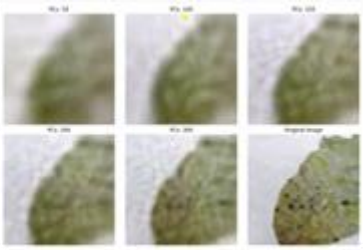


Figure 7: Leaf image on different number of PCs

Phase III. Building the Model and Evaluation

Table 1: SVM Model Classification Report without PCA

Category	Precision	Recall	F1-Score Support	
Anthracnose	0.91	0.86	0.89	100
Bacterial Canker	0.90	0.86	0.88	100
Gall Midge	0.74	0.70	0.72	102
Healthy	0.79	0.82	0.80	102
Powdery Mildew	0.84	0.61	0.71	105
Sooty Mold	0.61	0.85	0.71	102

SVM Accuracy: 78%

SVM Time of Execution: 37.61 seconds

Table 2: SVM Model Classification Report with PCA

Category	Precision	Recall	F1-Score Support	
Anthracnose	0.90	0.92	0.91	100
Bacterial Canker	0.96	0.92	0.94	100
Gall Midge	0.86	0.92	0.89	102
Healthy	0.96	0.92	0.95	102
Powdery Mildew	0.94	0.88	0.91	105
Sooty Mold	0.88	0.91	0.89	102

SVM-PCA Accuracy: 91%

SVM-PCA Time of Execution: 0.52 seconds

The classification result the SVM-PCA model achieved a higher accuracy of 91% in classifying the data, while completing its execution in a significantly shorter time of 0.52 seconds. On the other hand, the SVM model achieved an accuracy of 78% and took 37.61 seconds to complete its execution. This demonstrates that the SVM-PCA model outperformed the SVM model in terms of both accuracy and execution time.

The overall evaluation performance reveals that the SVM model with PCA (Table 2) shows improved performance

compared to the SVM model without PCA, as indicated by higher precision, recall, and F1-score values for most categories. In the SVM model with PCA, the highest precision is achieved for Healthy and Bacterial Canker with 0.96, while the highest recall is observed for Anthracnose with 0.92. The F1-scores range from 0.89 to 0.95, reflecting better overall classification accuracy across different categories. In contrast, the SVM model without PCA (Table 1) exhibits lower precision, recall, and F1-scores, indicating comparatively lesser performance in classification accuracy.

Table 3: CNN Model Classification Report without PCA

Category	Precision	Recall	F1-Score Support	
Anthracnose	0.94	0.95	0.95	100
Bacterial Canker	0.86	0.96	0.91	100
Gall Midge	0.94	0.80	0.87	102
Healthy	0.88	0.97	0.92	102
Powdery Mildew	0.86	0.96	0.91	105
Sooty Mold	0.94	0.75	0.83	102

CNN Accuracy: 90%

CNN Time of Execution: 3.23 minutes

Table 4: CNN Model Classification Report with PCA

Category	Precision	Recall	F1-Score Support	
Anthracnose	0.92	0.97	0.94	100
Bacterial Canker	0.99	0.94	0.96	100
Gall Midge	0.90	0.95	0.92	102
Healthy	0.94	0.91	0.93	102
Powdery Mildew	0.99	0.84	0.91	105
Sooty Mold	0.86	0.96	0.91	102

CNN-PCA Accuracy: 93%

CNN-PCA Time of Execution: 1.78 seconds

The CNN model achieved an accuracy of 90% with a time of execution of 3.23 minutes. However, by incorporating PCA (Principal Component Analysis) into the CNN model, the accuracy improved to 93% while significantly reducing the execution time to just 1.78 seconds. This indicates that the CNN-PCA model performed better in terms of both accuracy and efficiency compared to the regular CNN model.

The overall evaluation performance for the CNN model without PCA (Table 3) achieved the following results: For the categories Anthracnose, Bacterial Canker, and Gall Midge, the precision was high, indicating accurate classification. However, the recall and F1-score were lower for Gall Midge and Sooty Mold, suggesting some misclassifications.

On the other hand, the CNN model with PCA (Table 4) showed improved performance. The precision, recall, and F1-score were generally higher for all categories, indicating better classification accuracy.

4. CONCLUSIONS

The study found that employing PCA in SVM and CNN models resulted in higher accuracy rates and faster model build times. The SVM model with PCA achieved 91% accuracy compared to 78% without PCA, with a significantly reduced build time of 0.52 seconds. Similarly, the CNN model with PCA achieved 93% accuracy compared to 90% without PCA, with a reduced build time of 1.78 seconds. These results demonstrate the potential of PCA in improving the accuracy and efficiency of AI learning models for early disease detection and classification. By reducing the number of features, PCA enables faster model training and better resource utilization without compromising accuracy, making it a valuable technique for addressing high-dimensional data challenges and facilitating targeted treatments for mango leaf diseases.

REFERENCES.

[1] Liu B, Xin Q, Zhang M, Chen J, Lu Q, Zhou X, Li X, Zhang W, Feng W, Pei H, Sun J. (2023). Research Progress on Mango Post-Harvest Ripening Physiology and the Regulatory Technologies. Foods. 12(1):173. https://doi.org/10.3390/foods12010173.

[2] Zainuri Zainuri & Taslim Sjah. (2022). “Empowering Communities of Mango Agribusiness in North Lombok, Indonesia.” Proceedings of the 6th International Conference of Food, Agriculture, and Natural Resource (IC-FANRES 2021). DOI:10.2991/absr.k.220101.027

[3] Jiamin, Liu. (2019) . “Research on Competitiveness of Export Trade and Strategy of Thai Mango”.

[4] P. K. Shukla, Tahseen Fatima, and Nidhi Kumari. (2021). “First Report of Berkeleyomyces basicola Causing Mango Root Rot and Decline in India.” https://doi.org/10.1094/PDIS-10-20-2133-PDN

[5] Khaskheli, M.I. (2020). Mango Diseases: Impact of Fungicides. Horticultural Crops, 143.

[6] J. Ma, K. Du, F. Zheng, L. Zhang, Z. Gong and Z. Sun, (2018). "A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network", Comput. Electron. Agricult., vol. 154, pp. 18-24

[7] S. Kaur, S. Pandey and S. Goel. (2019). "Plants disease identification and classification through leaf images: A survey", Arch. Comput. Methods Eng., vol. 26, no. 2, pp. 507-530.

[8] X. Zhang, Y. Qiao, F. Meng, C. Fan and M. Zhang. (2018) "Identification of maize leaf diseases using improved deep convolutional neural networks", IEEE Access, vol. 6, pp. 30370-30377.

[9] R. Gandhi, S. Nimbalkar, N. Yelamanchili and S. Ponkshe (2018). "Plant disease detection using CNNs and GANs as an augmentative approach", Proc. IEEE Int. Conf. Innov. Res. Develop. (ICIRD)

[10] H. Durmuş, E. O. Güneş and M. Kırcı. (2017). "Disease detection on the leaves of the tomato plants by using deep learning", Proc. 6th Int. Conf. Agro-Geoinformatics, pp. 1-5.

[11] Gadekallu, T.R., Rajput, D.S., Reddy, M.P.K. et al. (2021). “A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. J Real-Time Image”. Proc 18, 1383–1396.

[12] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto and Y. E. Pratama Yudoutomo (2021). "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants,". International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Banyuwangi, Indonesia, pp. 46-50,

[13] Pandian, J. Arun, K. Kanchanadevi, V. Dhilip Kumar,

Elżbieta Jasińska, Radomír Goňo, Zbigniew Leonowicz, and Michał Jasiński. (2022). "A Five Convolutional Layer Deep Convolutional Neural Network for Plant Leaf Disease Detection" Electronics 11, no. 8: 1266. https://doi.org/10.3390/electronics11081266.

[14] Turk and Pentland. 1991. “Eigenfaces for Recognition”. Journal of Cognitive Neuroscience (1991) 3 (1): 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

[15] Ali, Safdar, Mehdi Hassan, Jin Young Kim, Muhammad Imran Farid, Muhammad Sanaullah, and Hareem Mufti. (2022). "FF-PCA-LDA: Intelligent Feature Fusion Based PCA-LDA Classification System for Plant Leaf Diseases" Applied Sciences 12, no. 7: 3514. https://doi.org/10.3390/app12073514.

[16] K. Roy et al., "Detection of Tomato Leaf Diseases for Agro-Based Industries Using Novel PCA DeepNet," in IEEE Access, vol. 11, pp. 14983-15001, 2023, doi: 10.1109/ACCESS.2023.3244499.

[17] Nigam, Aakrati, Avdhesh Kumar Tiwari, Akhilesh Pandey. 2020. “Paddy leaf diseases recognition and classification using PCA and BFO-DNN algorithm by image processing ”. Materials Today: Proceedings Volume 33, Part 7, 2020, Pages 4856-4862

.