DOI:https://doi.org/10.52756/ijerr.2023.v34spl.011 ***Int. J. Exp. Res. Rev.*, Vol. 34: 106-119(2023)**

Original Article | *Peer Reviewed* | *Open Access*

International Journal of Experimental Research and Review (IJERR)

ISSN: 2455-4855 (Online)
www.iaph.in
ISSN 2455-4855

MCIP: Mining Crop Image Data on PySpark Data Frame Using Feature Selection and Cluster-Based Techniques

Yashi Chaudhary* and Heman Pathak

Gurukul Kangri (Deemed to be University), Haridwar, Uttarakhand, India

E-mail/Orcid Id:

YC, mohita.chaudhary5@gmail.com, https://orcid.org/0000-0002-3959-4008; *HP,* hpathak@gkv.ac.in

Article History:

Received: 04th Jul., 2023
Accepted: 21st Oct., 2023
Published: 30th Oct., 2023

Keywords:

Agriculture, clustering, data mining, k-means, PCA, PySpark

How to cite this Article:

Yashi Chaudhary and Heman Pathak (2023). MCIP: Mining Crop Image Data on PySpark Data Frame Using Feature Selection and Cluster-Based Techniques. *International Journal of Experimental Research and Review*, *34*(Spl.), 106-119.
DOI:https://doi.org/10.52756/ijerr.2023.v34spl.011

Abstract: In India, the yearly economic losses incurred due to crop-related issues, including pests and diseases, surpass a staggering amount of $500 billion. Leaf blight constitutes a significant determinant in the remarkable economic ramifications, primarily affecting farmers engaged in cultivating forage and grain sorghum, who bear the brunt of its consequences. Numerous crops are affected by this disease, including maize, rice, tomato, potato, millet, and onion. However, crop variety, different disease kinds, and environmental conditions make early disease identification difficult. Due to the variety of crops and diseases, existing approaches for disease classification and prediction need more broad applicability. These techniques use image preprocessing and segmentation to handle datasets with specified inputs and outputs, frequently resulting in data loss and erroneous categorization. Additionally, they ignore specialized datasets. To address these challenges, this study proposes an innovative approach leveraging the PySpark-based mining crop image data (MCIP) framework. MCIP employs Principal Component Analysis (PCA) to extract relevant features, subsequently utilized by the K-means algorithm to identify distinct subgroups. This approach, initially demonstrated on potato leaves, proves valuable for disease identification. Notably, MCIP isn't restricted to potatoes; it's adept at detecting diseases across agricultural crops. To validate, an experiment was conducted on a rice disease dataset. Evaluation metrics, including Accuracy, Silhouette score, speed, and F1 score, affirm MCIP's robustness. Impressively, MCIP exhibits exceptional speed and accuracy, nearly achieving 100% accuracy. This innovative model signifies a significant advancement over existing techniques, offering a promising solution to the pressing issue of crop disease management.

Introduction

The economy of a nation is predominantly dependent on its agricultural sector. The attainment of food self-sufficiency is a critical factor in ensuring the survival and resilience of a nation. Researchers and growers devote their attention to various factors to enhance productivity. However, the issue of crop loss resulting from disease stands out as one of the most significant challenges they encounter. The monitoring of crop growth and the early identification of pest infestations continue to present a significant challenge. Manually monitoring and identifying insect and pest infestations is becoming progressively more difficult due to the expansion of farming into larger fields (Velusamy et al., (2021); Dhaliwal et al., (2021). Identifying crop loss at the field parcel scale through the utilization of satellite images presents several challenges. Firstly, crop loss can be attributed to many factors that occur throughout the entire growth cycle. Secondly, dependable reference data pertaining to crop loss is limited. Lastly, the definition of crop loss can vary across different contexts (Hiremath et al., 2021).

Helminthophobia, the pathogenic fungus known as turcicum Pass, is the causative agent of leaf blight. The disease on sorghum leaves manifests as reddish-purple or brown spots that coalesce into extensive lesions,

*Corresponding Author:mohita.chaudhary5@gmail.com

particularly in regions with high humidity. Seedlings are just as susceptible to damage as fully-grown plants. The disease commonly manifests in a moderate to severe manner on forage sorghum within the Indian states of Haryana, Rajasthan, Uttarakhand, and Uttar Pradesh (Muimba-Kankolongo, 2018). The disease can reach pandemic levels due to significant damage inflicted upon the leaf's photosynthetic machinery. This can result in adverse effects on forage productivity, quality, and grain yield. According to Das et al. (2016), it has been observed that during severe epidemics, there is a significant reduction in grain yields, potentially reaching a decline of 50% or even more.

The monitoring of plant health and timely identification of symptoms is essential for mitigating disease transmission, thereby facilitating farmers in implementing efficient management strategies and enhancing agricultural productivity. Therefore, the identification of crop diseases is of utmost importance in order to sustain agricultural productivity. Conventional methods of plant disease diagnosis predominantly depend on the subjective expertise of farmers, thereby possessing inherent limitations in terms of accuracy and precision. In a previous study conducted by Sasaki et al. (1998), by employing a spectrometer as a means to ascertain the health status of plant leaves. An alternative method involved the utilisation of the polymerase chain reaction (PCR) technique, as described by Henson et al. (1993), or the real-time polymerase chain reaction (RT-PCR) technique, as outlined by Koo et al. (2013), for the extraction of Deoxyribonucleic acid (DNA) from the leaves. In their research, Prasad et al. (2016) presented a methodology aimed at the detection and classification of plant leaf diseases through the utilisation of soft computing techniques. The authors employed a genetic algorithm for image segmentation to identify and classify plant leaf diseases. The proposed methodology was assessed using a diverse range of plant foliage and demonstrated efficacy in the early detection of diseases. In their study, Ali et al. (2017) conducted research on the detection and classification of plant diseases through the utilisation of a pattern recognition algorithm applied to images of crops. The Gabor Wavelet Transform (GWT) approach was employed in conjunction with pattern recognition techniques to identify plant diseases. The study conducted by Prasad et al. (2016) introduced a novel approach to disease detection known as automated mobile vision. In order to detect diseases in plants, the researchers employed a hybrid methodology known as GWT- (Gray-Level Co-Occurrence Matrix) GLCM. According to Ali et al. (2017), the DeltaE approach may be employed to identify diseases in citrus fruits accurately. In their study, the researchers employed the K-Nearest Neighbors (KNN) algorithm and the cubic Support Vector Machine (SVM) to classify diseases at both the image and disease levels.

Several approaches for leaf disease detection in various crops have been developed recently by various researchers (Singh et al., 2019, Zhang et al., 2019, and Lu et al., 2017). In most strategies, image processing techniques were used to take out features, which were then input into a technique based on classification. Deepa et al. (2021) suggested a method for detecting plant leaf disease. Authors use Kuan filter to remove noise before extracting colour, shape, and texture information using the Hough transformation. The plant leaf disease was classified using a reweighted linear programme boost classification. The suggested technique's performance was assessed using the Plant Village dataset. Hamuda et al. (2018) suggested a crop identification system based on image processing that used the Kalman filtering algorithm and the Hungarian algorithm. Over a mobile-acquired picture, Picon et al. (2019) employed the ResNet-50 architecture, a deep CNN architecture. They used stochastic gradient descent optimization to train the network. Ferentinos KP (2018) suggested a deep learning-based approach for detecting plant leaf disease using multiple CNN architecture models on an open dataset with 58 discrete classifications. Huang et al. (2018) proposed utilizing the RBF (radial basis function) kernel of a support vector machine to identify sugarcane borer illness choices made in a quick manner utilizing basic processors in this technique. It also uses less memory for data storage, i.e., data collected during the training process. There are various datasets on which the classification and prediction of disease is done. Advances in artificial intelligence have aided researchers in identifying and diagnosing plant disease utilising proper image processing and machine learning methodologies. Singh et al. (2019) classified mango leaves using CNN. Singh (2019) used image segmentation and Particle Swam Optimization (PSO). Convolutional Neural Network (CNN) and Deep CNN (DCNN) are widely used to classify and predict leaf diseases of wheat, tomato, corn, and seasonal crops (Sladojevic et al., 2016); Sharma et al., 2020); Agarwal et al., 2020; Mishra et al., 2020; Khamparia et al., 2020; Hussain et al., 2018). Researchers have also used Deep Neural Networks (DNN) to classify and detect plant leaf diseases Venkataramanan et al. (2019). Deep learning has also attracted researchers due to its high level of accuracy (Chandy, 2019; Karthik et al., 2020; Zhang et al. 2020).

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

Table 1. Summary of related work

Paper	Dataset	Technique	Advantage	Disadvantage
Khamparia et al. (2020)	Potato	Deep CNN	The system is capable of identifying and diagnosing various medical conditions. The achieved accuracy is 96.46%.	Requires large dataset and GPU. It is costly to train.
Nazki et al. (2020)	Tomato leaves	Activation Reconstruction loss Generative Adversarial Network (ARL-GAN) and CNN.	The provided information aptly demonstrates the process of generating synthetic images. Facilitates the process of categorization. Achieved an accuracy rate of 87.6%.	The procedure is complex and costly to get the desired results. There is an issue of mode collapse, too.
Ganatra et al. (2020)	Plant Village	ResNet 50 and 101	Achieve a significantly high accuracy in disease classification by utilising a reduced number of layers. A level of precision amounting to 99.7% has been attained.	The model works for certain epochs, but on increasing the epochs, it would suffer from an overfitting problem thus resulting in reduced accuracy.
Sambasivam et al. (2021)	Cassava	CNN	Detects multiple diseases. Achieves an accuracy of 93%	Suffers from overfitting problems. A large training dataset is needed. The position and orientation of an object are not encoded.
Geetharamani et al. (2019)	Maize, Potato, Tomato	CNN and Autoencoders	Can detect multiple diseases. Achieves an accuracy of 97.50%	Suffers from overfitting problem on a larger dataset. Requires big clean data to arrive at desired results.
Liang et el. (2019)	Potato	Resnet 50	Achieves an accuracy of 98%	Better techniques with 99.7% accuracy are available.
Khalifa et al. (2021)	Potato	CNN	Detects early and late phases of blight. Achieves a 98 % overall accuracy	Limited to small and specific datasets.
Rozaqi et al. (2020)	Potato	CNN	Early and late blight are identified using an accuracy of 92%.	Shows better results.

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

Sanjeev et al. (2020)	Potato	Feed Forward Neural Network (FFNN)	Accuracy of 96.5% for early and late blight disease detection	Loses neighborhood information as it cannot move back and learn.
Barman et al. (2020)	Potato	Simplified Bayesian CNN (SBCNN)	Blight condition in early and late stages is detected accurately with 96.75%.	The model requires more parameters to train.
Jhonson et al. (2021)	Potato	Mask R-CNN	Blight condition in early and late stages is detected accurately with 98%.	Limited to a single dataset only.
Lee et al. (2020)	Potato	CNN	Blight condition in early and late stages is detected accurately with 99%.	The model requires a lot of training data. It also does not encode the position and orientation of the leaf.
Islam et al. (2017)	Potato	Segmentation, and Multi SVM	Blight condition in early and late stages is detected accurately with 95%.	It is not suitable for large datasets. Segmentation could result in loss of features.
Rashid et al. (2021)	Potato	Yolov5 Segmentation, Deep learning using CNN	Blight condition in early and late stages is detected accurately with 99.75%.	The technique is limited to Early blight and late blight detection of a single dataset. The overfitting problem needs to be appropriately addressed, which is very common issue with deep learning.

Tarik et al. (2021) have used Image Processing picture division with machine learning for detecting potato leaf disease. Liu et al. (2022) have used machine learning techniques for tea plant leaves for classification, and Almoujahed et al. (2022) used machine learning techniques for identifying cereal blight conditions. de Oliveira Dias et al. (2023) used a random forest machine-learning technique for tomato late blight prediction. More findings of the related work are summarized in Table 1.

According to our survey, the techniques are complex, costly, and time-consuming, and they need pervasive operation, experimentation, and use of crop protection agents. The existing models have been trained, tested, and validated on benchmark datasets. The datasets have limited images. None of the techniques addresses the big problem of time taken in training and testing large image datasets. The detections are limited to a few diseases. If foreign leaf data without a label is introduced, the techniques will not return the desired results. MCIP tries to address these issues along with the issue of speed by using the Spark framework.

Spark is one of the most popular new technology trends. It is the framework with the best chance of realising the benefits of combining Big Data and Machine Learning (Jonathan et al., 2021). It is fast (up to 100x quicker than typical Hadoop MapReduce thanks to in-memory operations), delivers robust, distributed, fault-tolerant data objects (referred to as RDDs), and combines perfectly with machine learning and graph analytics. PySpark is an application program interface (API) that connects Spark with Python.

Multivariate data tables may be found on leaves. In varying quantities, the leaves might be exceedingly healthy or disease-ridden. The severity ranges from low to severe. PCA is a statistical process that summarizes the content of large data tables by having a lower number of "summary indices" that can be displayed and analyzed more easily. It aids in the identification of trends, leaps,

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

clusters, and outliers. PCA is used by MCIP to extract features from a dataset. To categorize the clusters, trends, and outlier features generated by PCA, MCIP employs the K-means clustering method.

MCIP is independent of a dataset and can take any image dataset and classify it. We have calculated the Silhouette score to understand how accurately images are classified. Clustering algorithms such as K-Means use the silhouette score to look at how well samples are grouped with other similar samples. The Silhouette score is calculated for each sample of unique clusters (Dutta et al., 2021).

Hardware setup

Since the proposed work requires parallel processing, the system requires a Nvidia graphic card and decent memory. We tested the proposed model using a gaming laptop from HP with 8 Gb RAM, Intel core i5, and 10th generation processor.

Dataset

Benchmark datasets of potato leaves and rice leaf diseases Rashid et al. (2021) and Sethy et al. (2020) are taken. The potato leave dataset contains 4962 images distributed in training, testing, and validation folders. The image folders are marked as late blight, early blight and healthy. The rice leaf data set contains 5932 images marked as bacterial blight, blast, tungro and brown spot. The dataset does not have a training, testing, and validation dataset. The images are distributed in the respective disease folders.

a.Early Blight

b.Late Blight

c.Healthy

Figure 1(a, b & c). Potato leaf diseases

a.Bacterial Blight

b.Blight

c.Brown Spot

d.Tungro

Figure 2 (a, b, c & d). Rice leaf diseases

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

Table 2. Potato leaf dataset

Class	Number of images
Early Blight	1928
Late Blight	1714
Healthy	1320
Total	4962

Table 3. Rice leaf dataset

Class	Number of images
Bacterial Blight	1584
Tungro	1308
Brown spot	1600
Blight	1440
Total	5932

Objective function

The objective of this research is to recognise the leaf disease accurately. The function is mathematically represented here:

$$D = \left\{ P\left\{ \sum_{i=1}^{n} L_i \xrightarrow{create\ dataset} \{L_d\} \rightarrow \right.\right.$$

$$PCA\{L_d\} \rightarrow K-means\{L_{PCA}\} \rightarrow classify(L_K)$$

Here,

P is predictive analysis

L_i *is leaf*

L_d *is leaf data*

L_{PCA} *is PCA features*

L_K *is K-means*

D is the predicted disease

The objective function is summarized as:

1. Read each folder name and data.
2. Normalize each image data by dividing them by 255.
3. Store the normalized data with labels in a list and convert it into a .csv file
4. Read the .csv file and drop all the Nan values.
5. Apply PCA on the data and get PCA features.
6. Create a K-means model.
7. Standardize PCA features using K-means model
8. Compile (fit) PCA features and standardized features with K-means model.
9. Transform output from 8 for predictive analysis
10. Take a new leaf
11. Extract features using PCA
12. Using transformed output, features of the new leaf are passed to get the disease.

Mathematical formulation for computing the results

MCIP uses clustering techniques for classification and prediction. To check the goodness, we calculated the Silhouette score. The value of Silhouette ranges between -1 and 1.

$$Silhouette = \frac{A_{ic} - A_{inc}}{max\left(A_{inc}, A_c\right)} \quad \ldots\ldots eq(2)$$

$Here,$

A_{ic} $is\ Average\ inter-cluster$

A_{inc} $is\ Average\ intra-cluster$

For checking the goodness of the classification of MICP, we calculated Precision, Recall, Accuracy, and harmonic mean or F1 score. The calculations are based on a percentage of correctly identified positives or True Positive (TP), correctly identified negatives or True Negative (TN), identified as positive but not positive or False Positive (FP), and identified as negative but not negative or False Negative (FN).

Precision

$$Precision = \frac{\#\ TP}{(\#\ TP + \#\ FP)} \ldots \text{eq(3)}$$

Recall

$$\text{Re}call = \frac{\#\ TP}{(\#\ TP + \#\ FN)} \ldots \text{eq(4)}$$

Accuracy

$$Accuracy = \frac{(\#\ TP + \#\ TN)}{(\#\ TP + \#\ TN + \#\ FP + \#\ FN)} \ldots \text{eq(5)}$$

F1 Score

$$F1 = 2 \times \frac{(\text{Pr}ecision \times \text{Re}call)}{(\text{Pr}ecision + \text{Re}call)} \ldots \text{eq(6)}$$

Research Methodology

The model is graphically represented to give an overview of the work. Detailed implementation is explained through an algorithm.

A. Model

The model is graphically represented in Figure3(a) and 3(b).

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

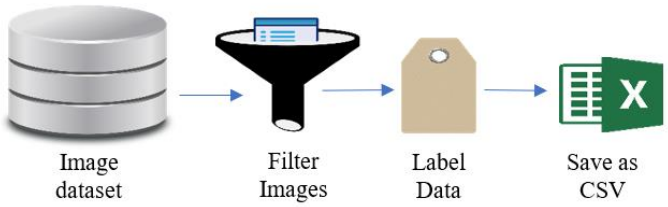

Figure 3(a). Dataset creation

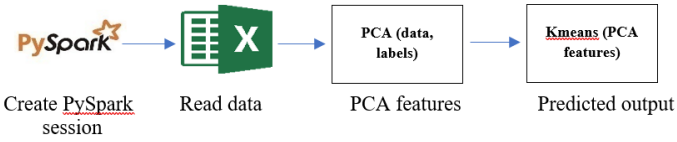

Figure 3(b). PCA and K-Means on data in the PySpark environment

The proposed model is represented using Figure 3 (a) and Figure 3 (b). MCIP reads the data from each folder and then filters it to scale the image evenly. The images are then labelled with their sub-folder names. The digitized images are then stored as a .cave file. To extract the features using PCA, we set up a PySpark environment. The data is read in PySpark framework. Once the data is read, PCA is applied with labels to extract the PCA features. Finally, K-means is applied to PCA features to classify and predict the data.

Implementation

Before applying Principal Component Analysis (PCA) and K-means clustering for disease classification and detection, we determined the optimal value of k by utilising the elbow technique. The elbow technique clusters the dataset using the k-means algorithm across k values (e.g., 1-32). Subsequently, at each value of k, the average score is calculated for all clusters. The distortion score is counted as the aggregate of the squared distances for each data point and its corresponding centroid. The MCIP algorithm transforms the input images into images with dimensions of 32x32 pixels. To determine the value of k, a dataset consisting of 1024*n data points was provided as input to the elbow visualizer. Where n represents the quantity of images and 1024 denotes the total number of pixels contained within each individual image. The potato dataset consists of three distinct classes, labelled 0, 1, and 2.

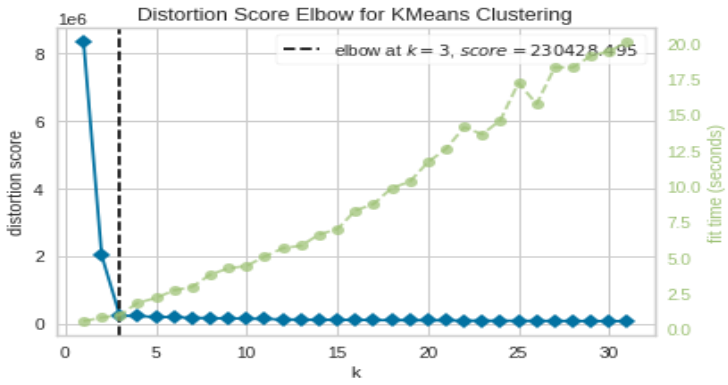

Figure 4. Plot of Elbow method

In Figure 4, the x-axis has different values of k and y-axis has a distortion score of s1^{e6}. Here, s is the score. The score is of all the pixels of the image dataset. Elbow method The Elbow technique identifies k clusters that best reflect the data points in their clusters. As a result, the data points and their respective cluster centroids' distance would be the assessment measure. In the plot, the bending starts at 3. This is where the reduction in Root Mean Square Error (RMSE) is no longer significant enough to justify raising the k value.

The equation can be represented mathematically:

$$k = min\left(RMSE\left(\sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{\|d_i - d_j\|}{2}\right)\right) \quad .....eq\cdot(7)$$

Here,

n is the number of clusters

d_i and d_j are the distances between two data points. The value of k we get here is 3, which is also the number of classes in the potato leaf dataset. This is used in PCA for a number of components and k-means clustering.

Algorithm: Classification

Setup:
Read dataset
Install required modules
Create spark session
Create context

//SparkContext is the entry gate of Apache Spark functionality.
Start:
Step 1. Prepare the dataset
validation_datagen ← Using Image Data Generator rescale validation image

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

test_datagen ← Using Image Data Generator rescale testing image

// Batches of tensor image data with real-time data are generated through Image data generator

//

Generate training set, testing set, and validation set

Resize image to 32X32 for faster processing

Step 2. Read the dataset

dataset ← read resized image dataset

label ←and label the data as: 0, 1, 2

Here,

0 is for Healthy

1 is for Late_blight and

2 is for Early_blight

ldata ← [dataset, label] #combines the data and list to have a labelled data.

Step 4. Convert data to an array

Step 5. Reshape the data to make it even.

Step 6. Divide by 255 to ensure that pixels are of same data type.

Step 7. Get the number of clusters

import k-means modules and other required modules.

Create a k-means object and fit the re-shaped data to predict.

Read re-shaped data and append labels to it.

Step 8. Clean data

Drop any Nan value from the dataset

Save the clean file as potato.csv

Step 9. Create a spark session.

Step 10. Read the CSV using spark

Step 11. Print the schema to understand the structure of the dataset

Step 12. Drop any Nan value if present

Step 13. Apply PCA

df ←Vectorize the labels

df ←Extract features

df ←Using vector assembler to generate features

assembled_data ← transform(df)

scale ← using Standard Scaler to generate standardized data

data_scale ← fit assembled_data

data_scale_output ← transform assembled_data

PCA ← apply PCA on data_scale_output

Here,

K ←3

Input ← features

Ouput ←PCA features

Step 14. Apply K-means on results obtained after PCA

scale ← generate standardized output usingfStandardScaler

Here,

input ←PCA features

output ←standardized

data_scale ← fit scale to PCA

data_sclae_output ← transform PCA using data_scale object

Step 15. Create an object of evaluator

Step 16. Create a kmeans object

fcol ←standardized

ncluster ←10

KMeans_Obj ←KMeans(featuresCol ←fcol, number of clusters ←ncluster)

KMeans_fit ←fit KMeans_Obj on data_scale_output

Step 17. Evaluate the silhouette score

silhouette coefficient, alternatively referred to as the silhouette score.It is a metric used to calculate the goodness of a clustering algorithm. Its value lies between -1 to 1.

1: Indicates separated and distinct customers from one another.

0: There is no substantial distance between clusters.

-1: incorrectly assigned clusters

The spark environment is set using Python's PySpark module. Reading image data in a PySpark environment is different. The output is a PySpark data frame, not the usual data frame, as in the case of Pandas. This implies that only commands from the PySpark module can be used on the output data frame. The challenge with image classification is the speed. Reading all the images and classifying them is time-consuming. We merged training and testing images in one folder. Parallel processing or PySpark helps us read the images into a data frame amazingly fast. To address the problem of slow classification, we first resized the images to 32X32. This helped in bringing uniformity to the data as well.

Further, we divided each pixel by 255 to get the pixel values between 0 and 1. Lower pixel values increase the speed of classification. The next task is to predict the disease.

Algorithm: Prediction

Step 1. Leaf ← upload a new leaf

Step 2. PCANew ← extract features leaf using PCA

Step 3. POutput ←KMeans_fit(PCANew) // Kmeans_fit is taken from algorithm 3.2.1

Step 4. If POutput==0:

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

Step 5. Disease ← Healthy
Step 6. else if POutput ==1:
Step 7. Disease ←Late_blight
Step 8. else:
Step 9. Disease ←Early_blight

For predicting the disease, the object of the compiled model is taken. The extracted features of the new image are passed through the classification output to get the predicted label.

III. Results

MCIP was evaluated on two datasets to ascertain the claim that the proposed work is independent of any dataset. The obtained results are compared with the results by existing techniques. We have designed three models to predict plant diseases.

PCA Features

Figure 5.PCA feature extraction

Figure 5 is the output of the PCA features extracted from the images. The first 20 results are reproduced here. The data is of the first 20 leaves in the image dataset.

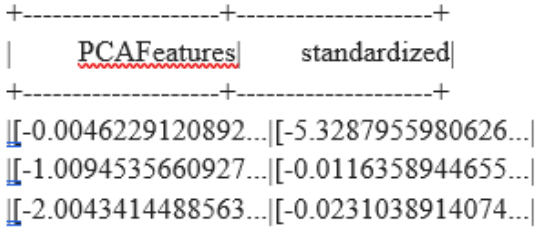

Figure 6.Standardized features

After PCA feature extraction, the data is standardized. The standardized features are shown in Figure 6. PCA and standardized features are used with k-means to classify the dataset. The outputs are of the potato dataset only.

Table 4. Silhouette score using only K-means (potato)

K values	Score
1	0.7966620104390114
2	0.7410152961759711
3	0.7764226529818571
4	0.7556580195635196
5	0.74193107935930953
6	0.7720033864837685
7	0.77958633539164464
8	0.7409168118624311
9	0.6966620104390114

Table 4 is the output received for different values of Silhouette for different values of K. The optimum value is K=7. The accuracy achieved is 98%.

Table 5. Silhouette score MCIP (potato)

K values	Score
1	0.8966620104390114
2	0.8410152961759711
3	0.9886226529818571
4	0.8556580195635196
5	0.84193107935930953
6	0.8720033864837685
7	0.87958633539164464
8	0.8409168118624311
9	0.7966620104390114

Table 5 is the score received when we used PCA and K-means. MCIP performs best at K=3. The accuracy achieved is close to 100%.

Table 6. Outcome of classifying potato dataset

Model	TP	FP	TN	FN
Rashid et al. (2021)	4875	10	40	35
Deep Learning	4886	9	36	29
K-means	4852	15	44	39
MCIP	4961	0	1	0

True positive is kept at 100, which is the number of images. All the input images are known to be correct; there are no images which are detected but are not correct; thus, False Positive is 0. The images which are Falsely detected negative are very low in numbering.

For comparison, we designed three models: Deep Learning, K-means, and K-means + PCA (MCIP). The other comparison is made with Rashid et al. (2021) for the potato dataset and Sethy et al. (2020) for the rice dataset.

The deep learning model has 6 Convolution 2D layers, 6 Maxpooling layers, 2 dropout layers, and two dense layers. Relu is being implemented as an activation

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

function. The activation function used on the output dense layer is softmax. All the three models run on PySpark. K-means (without PCA) re-scales the images to 32X32 as the feature input. The value of k is again 3. We wanted to see if PCA would make any difference in classification and prediction. The results show that although k-means can give good results, MCIP gives much better results.

Table 7. Accuracy using Table 6

Model	Precision	Recall	Accuracy	F1-score
Rashid et al. (2021)	0.9928717	0.8	0.9909274	0.8860616
Deep Learning	0.9940997	0.8	0.9923387	0.8865502
K-means	0.9920262	0.7457627	0.9890909	0.8514453
MCIP	1	1	1	1

The accuracy of MCIP is close to 100%. The accuracy would range between 99.5 and 100%. The algorithm was fixed on a random state to reproduce the same results every time. The results may vary if the state changes. The fall in the accuracy of Rashid et al. (2021) is due to the higher number of images in MCIP.

Table 8.Outcome of classifying rice dataset

Model	TP	FP	TN	FN
Sethy et al.	5726	43	92	71
Deep Learning	5854	11	34	33
K-means	5793	31	43	65
MCIP	5929	0	2	1

Table 9. Accuracy using Table 8

Model	Precision	Recall	Accuracy	F1-score
Deep learning	0.9943944	0.7555556	0.9925826	0.8586762
Sethy et al.	0.9877523	0.6814815	0.9807822	0.8065196
K-means	0.9889041	0.5810811	0.9838166	0.7320241
MCIP	0.9998314	1	0.9998314	0.9999157

The deep learning and MCIP can predict with 99.983% accuracy, again close to 100%. Without any changes to the algorithm with the rice dataset, we also get similar results. The model can take any image, classify it and predict the disease.

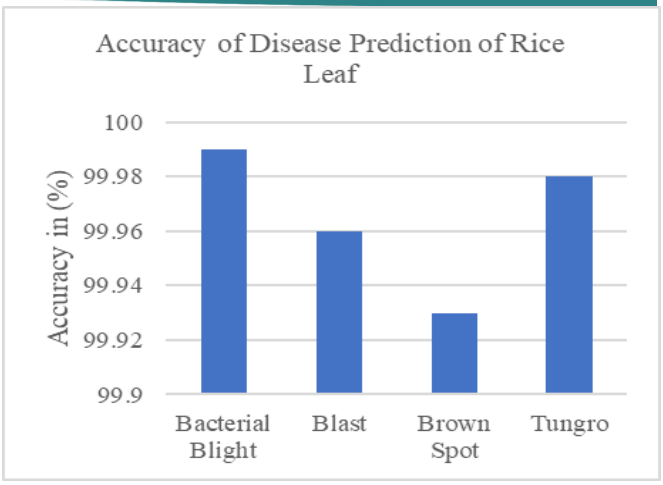

Figure 7. Disease classification accuracy potato dataset

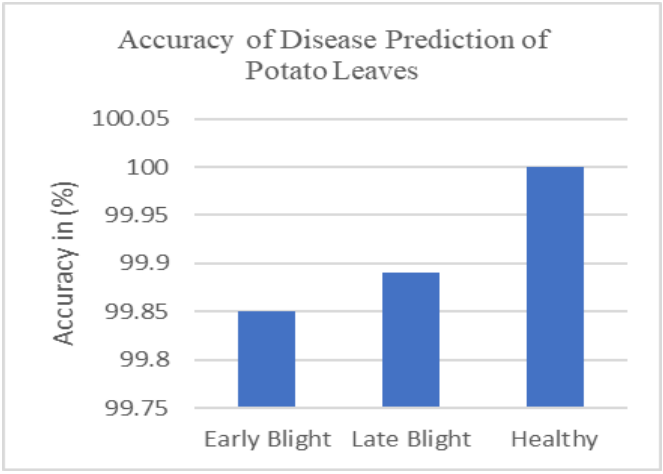

Figure 8. Disease classification accuracy rice dataset

Figure 7 and Figure 8 are plotted to show the accuracy of the predicted diseases. The accuracy varies for different diseases. The average classification accuracy is mentioned in Tables 7 and 9.

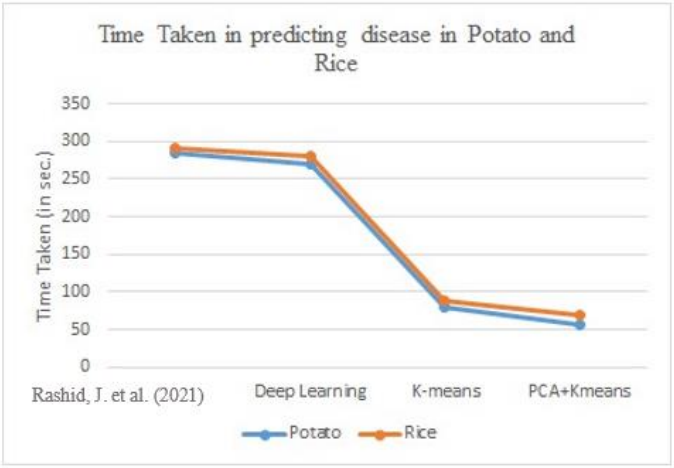

Figure 9. Time taken by potato and rice dataset

Time plays a vital role in the classification and prediction of an image dataset. We could reduce the time by 80% for potato dataset and 76% for rice leaf dataset.

DOI: https://doi.org/10.52756/ijerr.2023.v34spl.011

Paper Rashid et al. (2021) is on potato leaf; we implemented the model on rice leaf to arrive at the results.

Conclusion

This study presents an analysis of the PySpark-based mining crop image data (MCIP) framework and its efficacy in tackling the difficulties associated with crop disease diagnosis and classification. The focus is specifically on the significant economic losses caused by crop-related problems in India. Through the utilisation of deep learning methodologies and novel strategies such as Principal Component Analysis (PCA) and the K-means algorithm, MCIP demonstrates its proficiency in effectively and precisely identifying illnesses in diverse agricultural crops. The framework demonstrates a noteworthy capability of reducing processing time by 76-80% while achieving almost perfect accuracy, thereby distinguishing itself from conventional picture preprocessing and segmentation methods. Although there were early difficulties encountered in interpreting pictures inside the PySpark framework, this study effectively addresses and resolves these concerns. As a result, it paves the way for future investigations that explore the combination of Principal Component Analysis (PCA) with Long Short-Term Memory (LSTM) to improve plants' disease prediction capabilities. In general, the MCIP framework signifies a significant progression in the realm of crop disease management, demonstrating its capacity to transform the approach and mitigation of crop diseases within agricultural communities, leading to enhanced agricultural production and economic viability.

Conflict of interest

There is no known conflict of interest in this article.

References

Agarwal, M., Singh, A., Arjaria, S., Sinha, A., & Gupta, S. (2020). ToLeD: Tomato leaf disease detection using convolution neural network. *Procedia Computer Science*, *167*, 293-301. https://doi.org/10.1016/j.procs.2020.03.225

Ali, H., Lali, M. I., Nawaz, M. Z., Sharif, M., & Saleem, B. A. (2017). Symptom based automated detection of citrus diseases using color histogram and textural descriptors. *Computers and Electronics in Agriculture*, *138*, 92-104. https://doi.org/10.1016/j.compag.2017.04.008

Almoujahed, M. B., Rangarajan, A. K., Whetton, R. L., Vincke, D., Eylenbosch, D., Vermeulen, P., & Mouazen, A. M. (2022). Detection of fusarium head blight in wheat under field conditions using a hyperspectral camera and machine learning. *Computers and Electronics in Agriculture, 203*, 107456.https://doi.org/10.1016/j.compag.2022.107456

Barman, U., Sahu, D., Barman, G. G., & Das, J. (2020). Comparative Assessment of Deep Learning to Detect the Leaf Diseases of Potato based on Data Augmentation. IEEE.*In 2020 International Conference on Computational Performance Evaluation (ComPE)* (pp. 682-687). https://doi.org/10.1109/ComPE49325.2020.9200015

Chandy, A. (2019). Pest infestation identification in coconut trees using deep learning. *Journal of Artificial Intelligence*, *1*(01), 10-18. https://doi.org/10.36548/jaicn

Das, I. K., & Rajendrakumar, P. (2016). Disease resistance in sorghum. In *Biotic stress resistance in millets* (pp. 23-67). Academic Press. https://doi.org/10.1016/B978-0-12-804549-7.00002-0

de Oliveira Dias, F., Magalhães Valente, D. S., Oliveira, C. T., Dariva, F. D., Copati, M. G. F., & Nick, C. (2023). Remote sensing and machine learning techniques for high throughput phenotyping of late blight-resistant tomato plants in open field trials. *International Journal of Remote Sensing*, *44*(6), 1900-1921. https://doi.org/10.1080/01431161.2023.2192878

Deepa, N. R., & Nagarajan, N. (2021). Kuan noise filter with Hough transformation based reweighted linear program boost classification for plant leaf disease detection. *Journal of Ambient Intelligence and Humanized Computing*, *12*(6), 5979-5992. https://doi.org/10.1007/s12652-022-04156-6

Dhaliwal, G. S., Jindal, V., & Dhawan, A. K. (2010). Insect pest problems and crop losses: changing trends. *Indian Journal of Ecology*, *37*(1), 1-7.

Dutta, P., Shah, N., & Saha, S. (2021). A Multi-Objective Optimization-based Clustering Approach for CORD-19 Scholarly Articles. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1393-1398). IEEE. https://doi.org/10.1109/SMC52423.2021.9658719

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture,145*(1), 311-318.https://doi.org/10.1016/j.compag.2018.01.009

Ganatra, N., & Patel, A. (2020). Performance Analysis Of Fine-Tuned Convolutional Neural Network Models For Plant Disease

Classification. *International Journal of Control and Automation*, *13*(3), 293-305.

Geetharamani, G., & Pandian, A. (2019). Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Computers & Electrical Engineering*, *76*, 323-338. https://doi.org/10.1016/j.compeleceng.2019.04.011

Hiremath, S., Wittke, S., Palosuo, T., Kaivosoja, J., Tao, F., Proll, M., ...& Mamitsuka, H. (2021). Crop loss identification at field parcel scale using satellite remote sensing and machine learning. *PloS One*, *16* (12), e0251952. https://doi.org/10.1371/journal.pone.0251952

Henson, J. M., & French, R. (1993). The polymerase chain reaction and plant disease diagnosis. *Annual review of phytopathology*, *31*(1), 81-109. https://doi.org/10.1146/annurev.py.31.090193.000501

Huang, T., Yang, R., Huang, W., Huang, Y., & Qiao, X. (2018). Detecting sugarcane borer diseases using support vector machine. *Information Processing in Agriculture*, *5*(1), 74-82. https://doi.org/10.1016/j.inpa.2017.11.001

Hussain, A., Ahmad, M., Mughal, I. A., & Ali, H. (2018). Automatic disease detection in wheat crop using convolution neural network. In *The 4th International Conference on Next Generation Computing*. http://dx.doi.org/10.13140/RG.2.2.14191.46244

Islam, M., Dinh, A., Wahid, K., & Bhowmik, P. (2017, April). Detection of potato diseases using image segmentation and multiclass support vector machine. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1-4.

Jonathan, F., Yang, D., Gowing, G., & Wei, S. (2021, December). Machine Learning Framework for Detecting Offensive Swahili Messages in Social Networks with Apache Spark Implementation. In *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 293-297. https://doi.org/10.1109/PIC53636.2021.9687001

Johnson, J., Sharma, G., Srinivasan, S., Masakapalli, S. K., Sharma, S., Sharma, J., & Dua, V. K. (2021). Enhanced field-based detection of potato blight in complex backgrounds using deep learning. *Plant Phenomics*, *2021*. https://doi.org/10.34133/2021/9835724

Khamparia, A., Saini, G., Gupta, D., Khanna, A., Tiwari, S., & de Albuquerque, V. H. C. (2020). Seasonal crops disease prediction and classification using deep convolutional encoder network. *Circuits, Systems, and Signal Processing*, *39*(2), 818-836.https://doi.org/10.1007/s00034-019-01041-0

Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., &Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Computing*, *86*, 105933.https://doi.org/10.1016/j.asoc.2019.105933

Koo, C., Malapi-Wight, M., Kim, H. S., Cifci, O. S., Vaughn-Diaz, V. L., Ma, B., ...& Han, A. (2013). Development of a real-time microchip PCR system for portable plant disease diagnosis. *PloS one*, *8*(12), e82704. https://doi.org/10.1371/journal.pone.0082704

Khalifa, N. E. M., Taha, M. H. N., El-Maged, A., Lobna, M., & Hassanien, A. E. (2021). Artificial Intelligence in Potato Leaf Disease Classification: A Deep Learning Approach. Springer, Cham.In *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, pp. 63-79. https://doi.org/10.1007/978-3-030-59338-4_4

Lee, T. Y., Yu, J. Y., Chang, Y. C., & Yang, J. M. (2020, February). Health detection for potato leaf with convolutional neural network. IEEE,In *2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)* (pp. 289-293). https://doi.org/10.1109/Indo-TaiwanICAN48429.2020.9181312

Liang, Q., Xiang, S., Hu, Y., Coppola, G., Zhang, D., & Sun, W. (2019). PD2SE-Net: Computer-assisted plant disease diagnosis and severity estimation network. *Computers and Electronics in Agriculture*, *157*, 518-529. https://doi.org/10.1016/j.compag.2019.01.034

Liu, Z., Bashir, R. N., Iqbal, S., Shahid, M. M. A., Tausif, M., & Umer, Q. (2022). Internet of Things (IoT) and machine learning model of plant disease prediction–blister blight for tea plant. *IEEE Access*, *10*, 44934-44944. https://doi.org/10.1109/CCECE.2017.7946594

Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. *Computers and Electronics in Agriculture*, *142*, 369-379. https://doi.org/10.1016/j.compag.2017.09.012

Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2063-2079.

https://doi.org/10.1016/j.compag.2018.04.002

Mishra, S., Sachan, R., &Rajpal, D. (2020). Deep convolutional neural network based detection system for real-time corn plant disease recognition. *Procedia Computer Science*, *167*, 2003-2010. https://doi.org/10.1016/j.procs.2020.03.236

Muimba-Kankolongo, A. (2018). *Food Crop Production by Smallholder Farmers in Southern Africa: Challenges and Opportunities for Improvement*. Academic Press. https://doi.org/10.1016/B978-0-12-814383-4.00013-X

Nazki, H., Yoon, S., Fuentes, A., & Park, D. S. (2020). Unsupervised image translation using adversarial networks for improved plant disease recognition. *Computers and Electronics in Agriculture*, *168*, 105117. https://doi.org/10.1016/j.compag.2019.105117

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., & Johannes, A. (2019). Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Computers and Electronics in Agriculture*, *161*, 280-290.

Prasad, S., Peddoju, S. K., & Ghosh, D. (2016). Multi-resolution mobile vision system for plant leaf disease diagnosis. *Signal, Image and Video Processing*, *10*(2), 379-388. https://doi.org/10.1007/s11760-015-0751-y

Rashid, J., Khan, I., Ali, G., Almotiri, S. H., AlGhamdi, M. A., & Masood, K. (2021). Multi-Level Deep Learning Model for Potato Leaf Disease Recognition. *Electronics*, *10*(17), 2064. https://doi.org/10.3390/electronics10172064

Rozaqi, A. J., & Sunyoto, A. (2020, November). Identification of Disease in Potato Leaves Using Convolutional Neural Network (CNN) Algorithm. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 72-76. https://doi.org/10.1109/ICOIACT50329.2020.9332037

Sambasivam, G., & Opiyo, G. D. (2021). A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. *Egyptian Informatics Journal*, *22*(1), 27-34. https://doi.org/10.1016/j.eij.2020.02.007

Sanjeev, K., Gupta, N. K., Jeberson, W., & Paswan, S. (2021). Early Prediction of Potato Leaf Diseases Using ANN Classifier. *Oriental Journal of Computer Science and Technology*, *13*(2, 3), 129-134.http://dx.doi.org/10.13005/ojcst13.0203.11

Sasaki, Y., Okamoto, T., Imou, K., & Torii, T. (1998). Automatic diagnosis of plant disease-Spectral reflectance of healthy and diseased leaves. *IFAC Proceedings Volumes*, *31*(5), 145-150.

Sethy, P. K., Barpanda, N. K., Rath, A. K., & Behera, S. K. (2020). Deep feature based rice leaf disease identification using support vector machine. *Computers and Electronics in Agriculture*, *175*, 105527. https://doi.org/10.1016/j.compag.2020.105527

Sharma, P., Berwal, Y. P. S., & Ghai, W. (2020). Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Information Processing in Agriculture*, *7*(4), 566-574. https://doi.org/10.1016/j.inpa.2019.11.001

Singh, U. P., Chouhan, S. S., Jain, S., & Jain, S. (2019). Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. *IEEE Access*, *7*, 43721-43729. https://doi.org/10.1109/ACCESS.2019.2907383

Singh, V. (2019). Sunflower leaf diseases detection using image segmentation based on particle swarm optimization. *Artificial Intelligence in Agriculture*, *3*, 62-68. https://doi.org/10.1016/j.aiia.2019.09.002

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., &Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience, 2016*.https://doi.org/10.1155/2016/3289801

Velusamy, P., Rajendran, S., Mahendran, R. K., Naseer, S., Shafiq, M., & Choi, J. G. (2021). Unmanned Aerial Vehicles (UAV) in precision agriculture: applications and challenges. *Energies*, *15*(1), 217. https://doi.org/10.3390/en15010217

Venkataramanan, A, Honakeri, D. K., Agarwal, P. (2019). Plant disease detection and classification using deep neural networks. *Int. J. Comput. Sci. Eng.*,*11*(9), 40-6.

Zhang, K., Xu, Z., Dong, S., Cen, C., & Wu, Q. (2019). Identification of peach leaf disease infected by Xanthomonascampestris with deep learning. *Engineering in Agriculture, Environment and Food*, *12*(4), 388-396. https://doi.org/10.1109/ACCESS.2020.2982456

Zhang, Y., Song, C., & Zhang, D. (2020). Deep learning-based object detection improvement for tomato disease. *IEEE Access*, *8*, 56607-56614. https://doi.org/10.1109/ACCESS.2020.2982456

How to cite this Article:

Yashi Chaudhary and Heman Pathak(2023).MCIP: Mining Crop Image Data on PySpark Data Frame Using Feature Selection and Cluster-Based Techniques. *International Journal of Experimental Research and Review*, *34*(Spl.), 106-119.

DOI:https://doi.org/10.52756/ijerr.2023.v34spl.011