ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20230530000682
Preliminary communication

A PCA-SMO Based Hybrid Classification Model for Predictions in Precision Agriculture

Mo DONG, Haiye YU, Lei ZHANG, Yuanyuan SUI, Ruohan ZHAO*

Abstract: The human population is growing at an extremely rapid rate, the demand of food supplies for the survival and sustainability of life is a gleaming challenge. Each living being in the planet gets bestowed with the healthy food to remain active and healthy. Agriculture is a domain which is extremely important as it provides the fundamental resources for survival in terms of supplying food and thus the economy of the entire world is highly dependent on agricultural production. The agricultural production is often affected by various environmental and geographical factors which are difficult to avoid being part of nature. Thus, it requires proactive mitigation plans to reduce any detrimental effect caused by the imbalance of these factors. Precision agriculture is an approach that incorporates information technology in agriculture management, the needs of crops and farming fields are fulfilled to optimized crop health and resultant crop production. The proposed study involves an ambient intelligence-based implementation using machine learning to classify diseases in tomato plants based on the images of its leaf dataset. To analytically evaluate the performance of the framework, a publicly available plant-village dataset is used which is transformed to appropriate form using one-hot encoding technique to meet the needs of the machine learning algorithm. The transformed data is dimensionally reduced by Principal Component Analysis (PCA) technique and further the optimal parameters are selected using Spider Monkey Optimization (SMO) approach. The most relevant features as selected using the Hybrid PCA-SMO technique fed into a Deep Neural Networks (DNN) model to classify the tomato diseases. The optimal performance of the DNN model after implementing dimensionality reduction by Hybrid PCA-SMO technique reached at 99% accuracy was achieved in training and 94% accuracy was achieved after testing the model for 20 epochs. The proposed model is evaluated based on accuracy and loss rate metrics; it justifies the superiority of the approach.

Keywords: machine learning; one-hot encoding; plant disease; precision agriculture; principal components analysis (PCA); SMO

1 INTRODUCTION

Human beings are likely to become the most dominant species on this planet with a count of almost 10 billion in 2050 as informed by the United Nations in their report [1]. The size of the planet would remain same, but ever increasing growth in population will thrust immense pressure on the natural resources and landmass that is required for food. It is also difficult to ignore the predominant issues of global warming, deforestation, reduction in water resources, soil erosion and various others which have significant effect on our eco system. The geographical and environmental factors have immense effect on agriculture and its related outcome. The major contributing factors include terrain, climate, soil and soil water properties. The climatic factors that impact agriculture include light, temperature, water, rainfall, air, wind, and relative humidity. On the other hand, agriculture also contributes significantly to many environmental issues namely climate change, deforestation, irrigation issues, soil degradation and various others.

The use of intelligent systems empowered with sophisticated technological advancements and smart techniques have transformed the functioning of agricultural activities. But all of these approaches do not contribute towards automation in daily care and monitoring of seedlings and crop conditions during the post grow period. The existing solutions focus more on development of hardware infrastructure rather than specific needs of the farmers and enhancement of their user experience. Ambient intelligence based infrastructures play a significant role in creating such user-centric frameworks that are able to provide services in response to user needs. The present world is focused immensely on developing smart environments which involves higher degree of automation in all verticals of human life. This would help in adaptation to the ever-evolving environment and in communication with the humans in the most seamless and user friendly way. The use of ambient intelligence helps in fulfilling this objective involving the use of ubiquitous computing and personalized interface. This would enable communication at any place and at any time. This type of smart environment provides adjustable and customized control over remote devices thereby proving them the ability to communicate with one another and respond accordingly. In case of agriculture the role of ambient intelligence is huge wherein the sensor devices in the agricultural field can provide relevant details about the precision agriculture parameters and based on the data, optimized decisions could be taken for necessary action. The technologies like deep learning, artificial intelligence, machine learning and data analytics have all contributed towards providing efficient agricultural solutions to ensure better human life on this planet. As discussed, with the rapid growth in human population the demand for food supplies has also increased globally putting immense pressure on the agricultural systems. The use of machine learning and related approaches has helped to meet the challenges of such ever growing needs. Artificial Intelligence (AI) and Machine Learning (ML) approaches have transformed the farms into smart farms enabled with precision agriculture frameworks [2-4]. These approaches are integrated with big data, cloud, federated learning and various other approaches to develop applications that enable automatic detection of droughts, predict crop and plant diseases and detect ripening patterns of the harvest. These technologies also add micro and macro level nutrients based on physical, chemical properties and other attributes like moisture, pH, density, porosity, consistence, temperature and various others [4]. The integration of machine learning and aforementioned approaches in precision agriculture helps to identify the matches with the diseases stored in the imagery database and enables prediction of the diseases. This enables farmers to detect diseases that hinder growth of the plants at an early stage or even proactively so that they can take appropriate mitigation measures. In the same context, it is important to mention that tomato is one of the most important agricultural crops standing second in the list to potato consumed in almost all food cuisines, almost every day across the world. The world population requires almost 100 million tons of fresh tomatoes. This crop grows rather rapidly in about 90 to 150 days in an average temperature setting of 18-25 degrees during daylight and 10-20 degrees during the night time. The tomato plant is quite sensitive to reduced sunshine and higher humidity which significantly affect the quality of this crop production [5]. The crop is ideal to be cultivated in dry climatic conditions and additional attention is required for pest control. In case the humidity is high, the crop is susceptible to pests, tomato wilt and diseases caused by fungi, bacteria and viruses resulting in even rotting of the entire plant. Some of the common diseases occurring in the tomato plant are Early blight disease, Gray Leaf soft disease, Late Blight disease, Septoria Leaf Soft disease, Southern blight disease, Verticillium Wilt and Anthracnose and Bacterial Speck disease [6]. All these diseases pose critical threats on the cultivation output which acts as a resource for survival of most of the species living in this world. Machine learning is considered as a popular approach for the early detection of plant diseases. The traditional machine learning approaches analyse smaller image dataset for classification of crop diseases. On the contrary, application of convolutional neural network on the images of larger dataset eliminates the challenges associated with hand-crafted features of the small-scale datasets [7, 8]. The implementation of visualization techniques has enabled efficient and accurate prediction of the plant diseases based on the disease symptoms and attributes [9, 10].

The present study also proposes a hybrid Principal Component Analysis (PCA) and Spider Monkey Optimization (SMO) based deep neural network approach to classify tomato plant disease datasets. The steps involved in the present study are:

- Transformation of the dataset values using one-hot encoding technique.
- Application of PCA for dimensionality reduction.
- Application of SMO to select optimal features from the dimensionally reduced dataset.
- Evaluation of the model based on accuracy and loss percentage comparing it with the traditional approaches.

The unique contribution of the paper lies firstly in conducting a detailed review of studies wherein deep learning and machine learning techniques have been implemented successfully in precision agriculture. Secondly, the paper implements a hybrid framework involving PCA and spider monkey optimization that ensures dimensionality reduction such that when the significant parameters are fed into the DNN model, enhanced accuracy is achieved in the classification results.

Although the present study uses data available from public database but in practical implementation, the precision agriculture related data would be collected using IoT devices. This framework caters to the need of farmers to resolve dilemmas in complex decision making problems. The use of ambient intelligence incorporating the proposed framework would enable farmers to conduct on-demand dibbling and planting of seeds intelligently ensuring precise irrigation. The images would be captured on a regular basis in association with measurement of air temperature, humidity and other parameters using sensors guiding creation of ambient intelligence based seedbed.

2 LITERATURE REVIEW

Numerous studies have been conducted focusing on disease prediction in plants using machine learning approaches. The implementation of such techniques has significantly affected the prediction of plant diseases by classifying their images based on their severity levels. The study in [11] proposed a deep learning based model that helped to detect plant diseases based on the images of the plant leaves. The study included use of augmentation on the dataset which increased its sample size and then it was fed into a CNN model constituting of multiple convolutional and pooling layers. The result of the model was promising in terms of its accuracy but it was not tested on actual images collected in real-time using drones or other IoT devices. The study in [12] developed a deep convolutional neural network model based on a dataset consisting of 9000 images of infected and healthy tomato plants. The model helped to identify 5 types of diseases based on images of tomato plants which were collected under controlled conditions. The model seemed to be feasible as it generated high level of accuracy and reduced grayscale loss and full colour loss. However, the model did not include the use of any optimization algorithm which could improve the performance of the same. The study in [13] presented a machine learning based application for the detection of plant diseases based on images captured using smart phone. The system classified the plants into 5 classes out of which 4 classes represented diseased and one plant represented healthy plant. The severity levels of the diseases were also assigned to the classes wherein 1 represented a healthy plant and 5 represented a severely diseased plant. The study in [14] considered three types of tomato namely unripe, ripe and damaged (overripe or rotten). The paper implemented Convolutional Neural Network(CNN), artificial neural network (ANN), self-organizing map (SOM), learning vector quantization (LVQ) and support vector machine (SVM) to develop, assess, optimize and compare the model for the purpose of effective sorting of the tomatoes. The CNN-ANN based algorithm yielded superior accuracy in testing and training.

A computer-vision based algorithm for the grading of golden delicious apple was proposed in [15]. The calyx region was detected using K-means clustering technique and defect segmentation was performed using a multi-layer perceptron (MLP) neural network. The classification results of the proposed model were evaluated against SVM, MLP and K-Nearest Neighbour (KNN) technique based on the recognition rate and accuracy level. The study in [16] proposed a high performance deep neural network model for the detection of tomato plant diseases and pest recognition using a refinement filter bank. The model helps to eliminate the false positive results from the study generating a successfully high recognition rate. The study in [17] presented a PCA and Whale Optimization Algorithm (WOA) based deep neural network that enabled classification of plant diseases. The study used one hot encoding technique for data transformation and then dimensionality reduction was done using PCA and WOA algorithm. The study in [31] implemented ambient intelligence in precision agriculture to improve farmers interaction with the relevant intelligent environments supporting all related agricultural activities channelized towards improvement of quality and quantity of cultivation. The study proposed two frameworks namely Intelligent Greenhouse and AmI (Ambient Intelligence) seedbed that focused on improving wide range of agricultural activities starting from planting of seeds to monitoring growth of sprouted plants and finally their transplantation in the greenhouse. The study in [32] developed a hybrid model using IoT for yield production. The work was implemented in three stages wherein the dataset was initially pre-processed using correlation-based feature selection technique and variance inflation factor algorithm (VIF) and then the selected features were subjected to a two-tier ML model. This smart agriculture system yielded promising results and was evaluated based on accuracy, explained variance score and root mean squared error.

The summary of the papers reviewed in this section is presented in Tab. 1. The review of these various studies highlighted the need of high-quality dataset with only relevant features to generate optimal accuracy in classification and prediction. Thus, the present study emphasizes implementing the best possible approach in selecting the optimal features from the dataset.

Table 1 Summary of studies reviewed on precision agriculture.

Ref	Methodology	Contribution	Limitation
[11]	Augmentation and Convolutional Neural Network	Detection of plant disease with 98.3% accuracy	Did not consider use of IoT or any other system in the framework
[12]	Deep Convolutional Network	Smart phone assisted diagnosis and detection of 5 plant diseases with 99.84% accuracy	Did not consider any dimensionality reduction technique or sensor based technique
[13]	Machine Learning based on images captured using smart phone	Prediction of the state of health of farming gardens	Not much emphasis given on quality enhancement of the collected data
[14]	Hybrid CNN-ANN algorithm	Classification of ripe, unripe and rotten tomatoes with promising accuracy	Not much emphasis given on quality enhancement of the collected data
[15]	Multi-Layer Perceptron Neural Network	Classification of healthy and defective applies with recognition rate of 92.5% and 89.2% respectively	Not all significant features are included in testing
[16]	Refinement Filter Bank framework using CNN classifier	Detection of plant diseases with 96% recognition rate	Not all significant features are included in testing
[17]	PCA-Whale Optimization based DNN model	Detection of tomato plant disease with promising accuracy	More emphasis could be given on dimensionality reduction
[31]	Ambient Intelligence technique for the development of Intelligent Greenhouse and AmI seedbed	Smart agriculture based planting, caring, monitoring and harvesting of plants, Enhancement in plant growth	Participation of actual farmers and agronomists not considered
[32]	Hybrid ML model with IoT	Estimation of soil quality and classification of soil samples	Not much emphasis given on quality enhancement of the collected data
Proposed Approach	Principal Component Analysis and Spider Monkey Optimization based DNN model	Detection of tomato plant disease with training accuracy of 99% and testing accuracy of 94%	The actual participation of farmers and related stake holders not included which acts as a future scope of research

3 BACKGROUND

3.1 Precision Agriculture

The increasing demand for food supplies has created the need for optimised crop production across the globe [18]. The latest technologies incorporated in farming use machine learning techniques to meet the every evolving need of the agriculture industry. AI and machine learning transform cultivation lands used in smart farmlands integrating various technologies such as big data, IoT, machine learning, cloud technology and related applications. These frameworks enable automatic detection of draught patterns, prediction of agricultural output, plan disease detection and also track ripening pattern of the crops. These technologies are used predominantly for ensuring safety, performing research analysis, terrain monitoring and scanning, checking of soil hydration and identification of yield issues in agriculture. Smart drones embedded with IoT help in data collection and further use of ML and AI techniques to perform various types of predictive analytics. As an example, the smart drones which are equipped with ML and AI models identify diseased plant and help in precise spraying of the insecticides [18-20].

3.2 Role of IoT in Collecting Data in Precision Agriculture

The precision agriculture terminology itself implies the use of technologies especially Internet of Things (IoT). The technologies which have proved to be extremely effective in precision farming using IoT are [21-23]:

Sensors: The sensors have the capability to detect chemical, optical, thermal, biomolecular and biological metrics providing complete information pertaining to the health of the crops. Also the health monitoring sensors implanted on animals help farmers to perform real-time track on the livestock status

Precision Farming Software: These are controller tools which help in automatic equipment updates, software maintenance and provide advanced solutions for the management of farming.

Connectivity Protocols: The popularly used network protocols function properly in case of short distance ranges. The long ranged connectivity protocols that are used by intelligent farming include cellular connection, LoRaWAN, LPWAN and various others.

Monitoring Tools: The satellites are used for this purpose to collect data relevant to soil water, crop biomass and various other metrics. The data collected using GPS satellites are analysed and used by crop insurance companies, scientists, policy makers and governmental bodies. Unmanned aerial vehicles embedded with hyper-spectral and multi-spectral sensors enable farmers to monitor health of plants and also measure water levels. The use of IoT helps in climate monitoring, crop monitoring, cattle monitoring and green house automation [21-23].

3.3 One Hot encoding Technique

One Hot encoding technique helps in the conversion of categorical data variables to values acceptable by the machine learning and deep learning algorithms thereby improving the accuracy of the predictions. It is a technique for pre-processing categorical attributes making it suitable for machine learning models. As part of the encoding process, a new binary feature is created for each possible category assigning a value of 1 to the sample features that map to its original category [17]. It is an extremely important part of the feature engineering process wherein the categorical data is converted to numerical form and applied to the integer representation of the data. The variable encoded as integer is removed and new binary variable is assigned for each of the unique integer values in the dataset. It generates multiple additional features depending on the total number of unique values existing in the categorical feature. Each unique value in the category is added as a feature and thus one hot encoding technique is also popular as the process of creating dummy variables [17, 24, 25].

3.4 Principal Component Analysis (PCA)

Imbalance in the dataset is a major issue when applying machine learning algorithms to the dataset which is often tackled using upsampling and downsampling. The sampling process is implemented only on the training dataset and it does not impact the testing data. Upsampling is a process wherein data is injected synthetically into a pre-existing dataset which equalizes the label counts and prevents the model from inclining towards a particular majority class. Downsampling on the contrary reduces the training sample count falling under a particular majority class equalizing the count of target categories. But in the process a lot of valuable information gets lost which hinders achievement of accurate results at the end of data analysis [33, 34].

Principal Component Analysis (PCA) is a dimensionality reduction technique that is used to the dimensionality of larger dataset. The large set of variables is transformed into smaller ones that hold most of the information of the larger dataset. The reduction in the number of variables in a dataset often compromises the accuracy of the results generated but this is accepted in exchange of the simplification achieved with the help of dimensionality reduction. The reason is that the smaller datasets are easier to be explored and visualized. It is also easier to analyse the same using machine learning algorithms generating enhanced accuracy [26]. The PCA is performed in five stages as mentioned below:

1. Standardization: The range of the continuous initial variables is standardized so that each of them contributes equally to the analysis. Standardization is done mathematically using the following Eq. (1). Value means Real-time data. Mean presents the average data. Standardization transforms all the variables to the same scale.

$$z = \frac{\left(Value - Mean\right)}{StandardDeviation} \tag{1}$$

2. Development of Covariance Matrix: This is developed in order to understand the relationship between the input data variables and also the difference of the same from the mean and each other. Variables often hold redundant information when they are highly correlated. The covariance matrix is a $N \times N$ matrix wherein N is the number of dimensions. The matrix holds the covariance related to the pairs of initial variables as shown in Eq. (2)

$$\begin{matrix} Cov(a,a) & Cov(a,b) & Cov(a,c) \\ Cov(b,a) & Cov(b,b) & Cov(b,c) \\ Cov(c,a) & Cov(c,b) & Cov(c,c) \end{matrix} \tag{2}$$

3. Computation of Eigen Vector and Eigen Values from the covariance matrix to find the principal components: Principal components are new variables that are developed as linear combination of initial variables. It is computed in such a way that the new variables stay uncorrelated and most of the information within the initial variables are compressed inside the first component. As an example n - dimensional data have n principal components. PCA compresses the major information in the first component and the majority of the remaining information in the second one and continues the same. Eigen value and eigenvectors are always paired in the sense, each eigenvector has an eigenvalue which adds up to the dimension of the data. The eigenvectors of the covariance matrix depict the directions of the axes wherein there exist most variances which are basically the principal components. The eigenvalues are the coefficient of the eigenvectors which present the variance held in each of the principal component. The ranking of the eigenvectors on the basis of the eigenvalues in descending order generates the principal components as per the order of significance [26, 27].

The reason for choosing PCA over other state of the art techniques is due to its advantages. Firstly PCA works on the basis of linear algebra which is trivial to be solved computationally. Secondly, the machine learning algorithms work faster when trained using the principal components in comparison to the original dataset. It is often observed that high dimensional data face overfitting issues when subjected to regression-based algorithms. But the application of PCA reduces the dimensions of the training dataset thereby preventing the predictive algorithms to suffer from overfitting.

3.5 Spider Monkey Optimization Technique

Spider monkey optimization (SMO) algorithm is a metaheuristic method which simulates the social behaviour of spider monkeys and also adopts the fission and fusion swarm intelligence technique for the purpose of foraging. The spider monkeys usually live in a group of 40-50 members known as swarm. A female monkey usually acts as the group leader and creates smaller mutable groups to provide food in case of any insufficiencies. There are two requirements of swarm intelligence that are satisfied by the SMO algorithm. Firstly they perform labour division wherein the monkeys divide their foraging work among multiple smaller groups. Secondly, they perform self-organization by selecting the size of the groups in order to meet the required food necessity [28-30]. The foraging behaviour is constituted of the following steps:

- Firstly the swarm initiates searching of food.
- The distance between the food and individual monkeys is calculated.
- The distance of the individuals in case of location change is also considered.
- The distance between the individuals and source of food is also calculated.

The SMO is a population based algorithm which involves trial and error method oriented collaborative approach consisting of six phases namely the local leader phase, local leader learning phase, local leader decision phase, global leader phase, global leader learning phase and global leader decision phase [28-30]. The steps involved in SMO implementation are mentioned below:

Initialization:

The population of X spider monkeys is uniformly distributed. SM_x denotes the x^{th} monkey of the population and $x = 1, 2, 3, \ldots, X$.

The monkeys are considered in M dimensional vector space wherein M represents the total number of variables. The algorithm initializes each monkey - SM_x as per the following equation - Eq. (3):

$$SM_{xy} = SM_{\min y} + UR(0,1) \times \left(SM_{\max y} - SM_{\min y}\right) \tag{3}$$

where, SM_{xy} is the y^{th} dimension of the x^{th} SM; $SM_{\min y}$ and $SM_{\max y}$ represent the lower and upper bounds of SM_x in the y^{th} direction wherein $y = (1,2,3, \ldots, M)$; $UR(0, 1)$ represents a random number which is evenly distributed within the range of [0,1].

In the Local Leader Phase (LLP) the spider monkey – SM changes its present location based on the past locations of the local leader and the local group member. The new location of the SM is updated only when the new location holds a fitness value higher than the previous location. The equation for updating the location is presented below in Eq. (4):

$$\begin{aligned} SMnew_{xy} &= SM_{xy} + UR(0,1) \times \left(LL_{ly} - SM_{xy}\right) + \\ &+ UR(-1,1) \times \left(SM_{zy} - SM_{xy}\right) \end{aligned} \tag{4}$$

where, LL_{ly} - represents the y^{th} dimension of the location of the lth group leader; SM_{zy} - represents the y^{th} dimension of randomly chosen l^{th} SM in the l^{th} local group, ensuring $r \neq p$.

In the global leader phase, the local group members experience and the global leaders experiences are used to update location of all the SMs. The equation of updating the location is presented below in Eq. (5):

$$\begin{aligned} SMnew_{xy} &= SM_{xy} + UR(0,1) \times \left(GL_{ly} - SM_{xy}\right) + \\ &+ UR(-1,1) \times \left(SM_{zy} - SM_{xy}\right) \end{aligned} \tag{5}$$

where, GL_{ly} - represents the location of the global leader in y dimension wherein $y = 1, 2, 3, \ldots, M$ considering an arbitrarily selected index. The fitness of SM is used to calculate the prb_x which is calculated using the following Eq. (6):

$$prb_x = \frac{fn_x}{\sum_{p=1}^{N} fn_x} \tag{6}$$

where, fn_x is the fitness value of the X^{th} SM.

In the Global Leader Learning Phase, a greedy selection method is used to update the location of the global leader. The optimum location is assigned to the global leader and in the absence of updates the Global Limit Count is incremented by 1.

In the Local Leader Learning Phase, greedy selection method is applied to update the local leader location in the local group. The optimum location is assigned to the local leader and in the absence of updates the Local Limit Count is incremented by 1.

In case the local leader does not update its present location within the fixed Local Leader Limit, the candidates of the local group modify their positions as per Step 1 using the past experiences of the global and local leader. The following Eq. (7) is used to update the new locations.

$$\begin{aligned} SMnew_{xy} &= SM_{xy} + UR(0,1) \times \left(GL_{ly} - SM_{xy}\right) + \\ &+ UR(0,1) \times \left(SM_{zy} - LL_{xy}\right) \end{aligned} \tag{7}$$

In the Global Leader Decision (GLD) Phase, the population is split into smaller groups when the global leader does not update its location as per the Global Leader Limit. The splitting of the groups continues until the maximum number of groups is received. A local leader is selected from the newly split-up groups. If the global leader does not update its position until the pre-decided allowed limit, the global leader merges the entire group into a single one [28-30].

The main advantage of SMO lies in its capability to converge quickly with optimal parameters to achieve enhanced accuracy in classification results. Thus it is a predominant choice for researchers when performing dimensionality reduction. The main advantage of using SMO is that it improves the balance between exploitation and exploration while searching for the optima. The local leader phase helps in exploring the search region whereas the global leader focuses on exploitation which makes SMO the best suited candidate among all the search based optimization algorithms. SMO also has an inbuilt mechanism that enables stagnation checks wherein the local and global leader learning phase checks on the stagnation of the search process. Thus the process of exploration and exploitation gets balanced in SMO while being consistent in maintaining the convergence speed.

The following section presents the methodology of the proposed framework.

4 PROPOSED FRAMEWORK

Fig. 1 represents the proposed framework consisting of the PCA-SMO model integrated with DNN for predicting diseases in tomato leaf plant. In real time, the data could be collected from agricultural and farming fields using IoT devices. However in the present study, since the focus is on the analytical approach, publicly available dataset is used for evaluating the performance of the model. The data is thus collected from the plant-village image dataset collected from the publicly available repository. At the outset, the tomato plant disease image as part of the precision agriculture dataset is encoded using the one-hot encoding technique [35]. This helps to transform the categorical value in the dataset into binary values. The inability of machine learning algorithms in processing categorical values has already been discussed earlier and hence the one-hot encoding scheme serves the said purpose. The integer coding approach could also be used but it has associated challenges of its inability to be used in relationships constituting of ordinal datatypes. The one-hot encoding approach converts the categorical values to "0" and "1" wherein "1" represents existent and "0" represents non-existent data values. This approach enables achieving optimum level of prediction results.

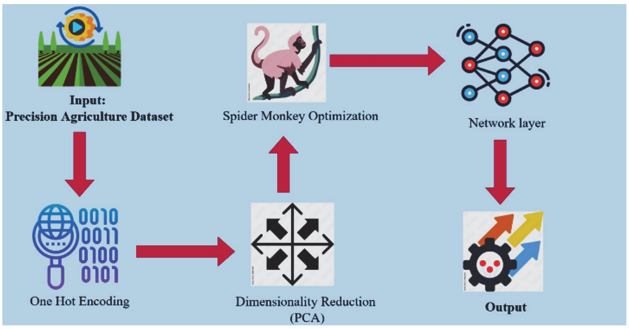


Figure 1 Proposed framework

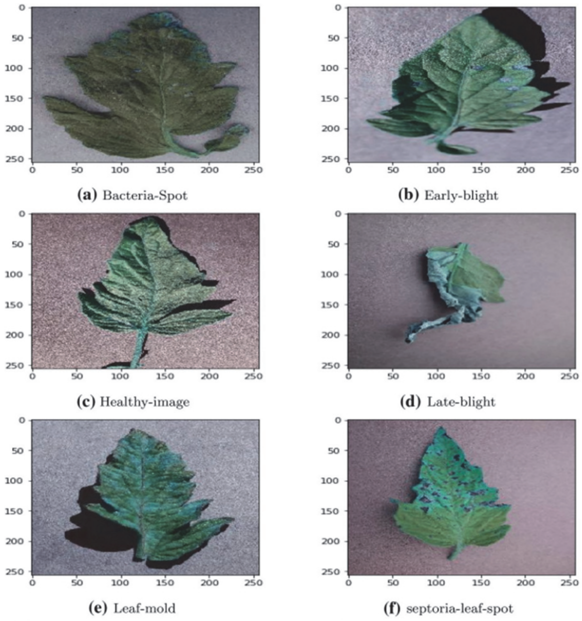
After transformation of the dataset values, PCA is used for dimensionality reduction. The preference of using PCA lies in its simplified usage, testing and comparative analysis in real-time. The PCA algorithm typically helps to eliminate correlated attributes in the dataset making the components independent of one another ensuring that the training time is reduced when fed into the machine learning algorithms. The selection of the most significant attributes using PCA reduces the possibilities of over-fitting thereby converting the high dimensional dataset into lower dimension. This enables enhanced visualization of the generated two dimensional plots and derivation of the inferences. To further select the optimal features, Spider Monkey Optimization (SMO) method was used which helped to select optimal features from the tomato plant disease dataset. To select the optimal parameters for the DNN model, grid search algorithm was used to perform hyper parameter tuning. The evaluation of the proposed SMO-DNN approach was performed using metrics such as accuracy and loss rate. The steps followed in the proposed approach could be summarized as:

- Loading of the tomato plant disease images into Google Colab which is GPU-enabled python notebook framework developed by Google.
- The one-hot encoding technique is applied to transform the categorical values of the dataset into binary digits.
- The principal components analysis (PCA) is implemented for dimensionality reduction and select the most significant attributes from the dataset.
- To further select the optimal features, SMO algorithm is used which helps in selecting the optimal pixels from the dataset. The grid search approach is used for hyper parameter tuning of the data to be fed into the DNN framework.
- The results of the DNN model are finally evaluated based on accuracy, loss and time complexity.

5 RESULTS AND DISCUSSION

The study was conducted using the tomato disease dataset collected from the publicly available plant – village data repository. The dataset included images of health and diseased plants and the study helped to segregate the healthy ones from the diseased [36]. The sample images of the healthy and diseased tomato plant leaves are shown in Fig. 2. The experiments of the study were conducted using Google Colab, which is a GPU framework having 50 GB hard disk and 25 GB GPU based RAM and accuracy and loss function metrics was used to evaluate the study.

The performance analysis of the model was conducted by evaluating the proposed DNN based model using a sequential approach. A random sample of 6000 images was used for performing cross validation wherein 5000 images in the dataset were used for training and the remaining 1000 images were used for testing purposes.



(a) Bacteria-Spot (b) Early-blight

(c) Healthy-image (d) Late-blight

(e) Leaf-mold (f) septoria-leaf-spot

Figure 2 Tomato leaf images

Fig. 3 represents the performance of the DNN model without implementing any dimensionality reduction technique considering accuracy as the metric. The best training and testing accuracy was achieved after running the model at 20 epochs. Also the model achieved training accuracy of 99% and testing accuracy of 88%. After running the model for 20 epochs the model started to get over-fitted. Fig. 4 presents the performance of the DNN model in terms of its loss percentage. Fig. 4 shows the loss percentage of the model without dimensionality reduction. The graph reflects a loss of 6.5% in testing when the model was run for 20 epochs.

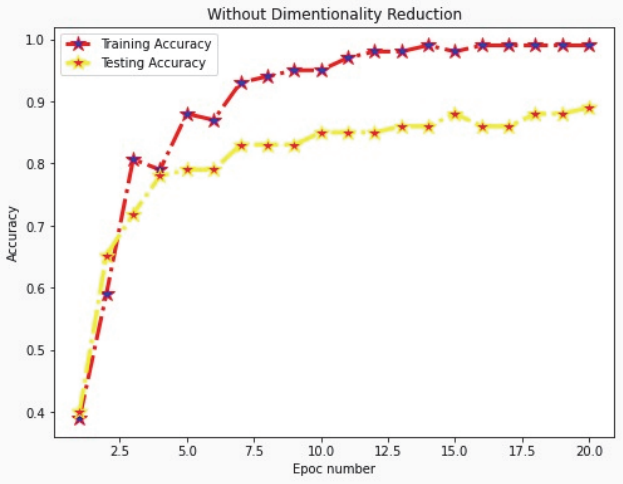


Figure 3 Accuracy without dimensionality reduction

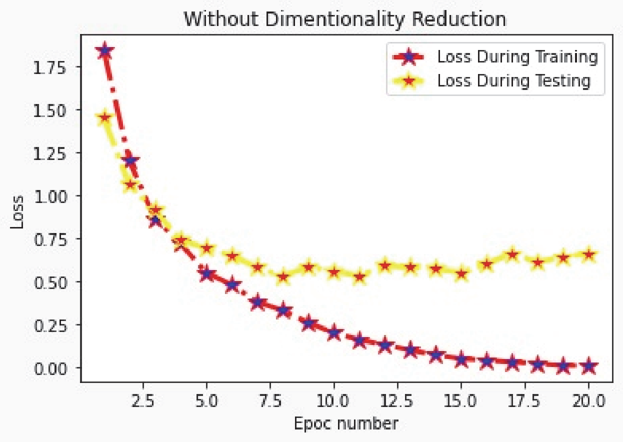


Figure 4 Loss without dimensionality reduction

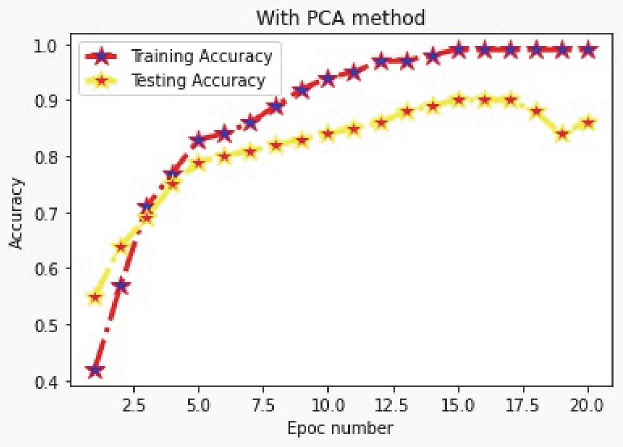


Figure 5 Accuracy using dimensionality reduction technique - PCA

Fig. 5 represents the performance of the DNN model after implementing dimensionality reduction technique - PCA. The model achieved an optimum accuracy of 99% at training and 85% at testing when run for 20 epochs. Also as shown in Fig. 6, the model generated a loss percentage of 3% at epoch 20 when PCA was implemented.

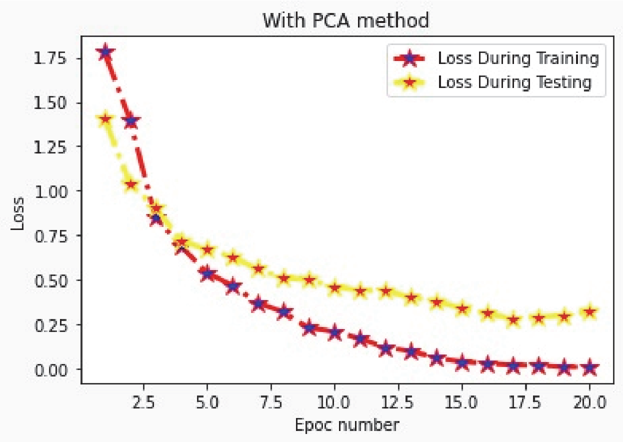


Figure 6 Loss using dimensionality reduction technique - PCA

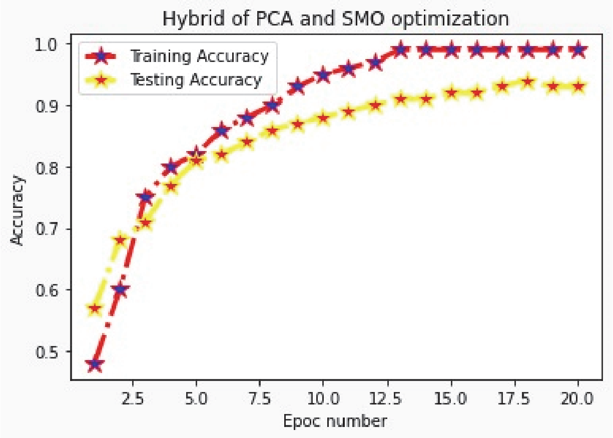


Figure 7 Accuracy using hybrid PCA and SMO

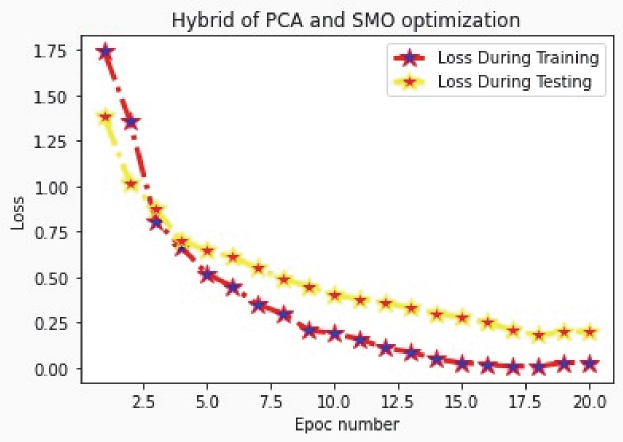


Figure 8 Loss using hybrid PCA and SMO

Fig. 7 represents the optimal performance of the DNN model after implementing dimensionality reduction using Hybrid PCA-SMO technique. The graph reveals that 99% accuracy was achieved in training and 94% accuracy was achieved after testing the model for 20 epochs. Considering the loss percentages, Fig. 8 reveals the testing loss of 0.24 after running the model for 20 epochs. Thus it is established that the Hybrid PCA-SMO technique yielded enhanced accuracy in training (99%) and testing (94%) in comparison to the other approaches. Similarly the testing loss was also found to be comparatively much lesser (0.24) in case of the proposed hybrid approach.

The time complexity of the model for three cases namely "Without Dimensionality Reduction", "With Dimensionality Reduction using PCA" and "With Dimensionality Reduction using Hybrid PCA-SMO" is shown in Fig. 9. As presented in the figure the Hybrid PCA-SMO based DNN model performed the analysis in lesser time in comparison to the other two cases.

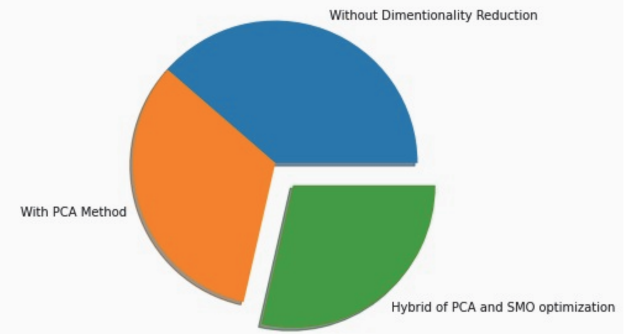


Figure 9 Time comparison of the models for three cases

The following is the summarized observation from the results generated.

- The training and testing accuracy improved after performing dimensionality reduction using PCA.
- The loss percentage was also reduced after performing dimensionality reduction using PCA.
- The training and testing accuracy were further enhanced when the model was run after performing dimensionality reduction using the Hybrid PCA-SMO approach.
- The loss percentage was also significantly reduced after implementing the Hybrid PCA-SMO.
- The time taken to perform the analysis also revealed that the Hybrid PCA-SMO method consumed lesser time in comparison to the other cases.
- The overall observation justified the superiority of the Hybrid PCA-SMO based DNN model considering accuracy and loss percentage.

6 CONCLUSION

The study proposes the use of a Hybrid PCA-SMO based deep neural network model for the classification of diseases in tomato plant. The data is collected from the plant - village dataset, which is a publicly available data repository. The study encompasses the use of a commercially available dataset, but the same could be acquired by IoT devices and related connectivity protocols. The study is a perfect example of the implementation of ambient intelligence wherein the farmers could respond based on prediction of the proposed framework which would include use of sensor technology, sensor networks and deep learning techniques. At the very initial stage of implementation, the one-hot encoding technique is applied on this dataset to transform it to be suitable for further analysis using machine learning approach. The data is converted to binary values and then PCA technique is applied on the dataset for dimensionality reduction. This helps to eliminate unwanted attributes from the dataset and ensures significant ones are only included in the study. To further improve the quality of the dataset, Spider Monkey Optimization (SMO) technique is applied on the dataset to select the optimal features. This dataset is fed into the deep learning model for further prediction on plant diseases. The major contribution of the proposed framework lies in the use of an efficient PCA-SMO based intelligent DNN framework that supported ambient intelligence yielding promising results in terms of accuracy in classification. The results of the hybrid technique classified the tomato diseases with optimal accuracy which enabled proactive risk mitigation in precision agriculture. The accuracy, loss percentage and time complexity of the model were evaluated, the results justified the superiority of the proposed framework.

Acknowledgments

This research was funded by National Key Research and Development Program Subproject of China (Grant No. 2022YFD2002305-2) and National Natural Science Foundation of China (NSFC) (Grant No. 32001418).

7 REFERENCES

[1] Bosona, T. & Gebresenbet, G. (2018). Life cycle analysis of organic tomato production and supply in Sweden. *Journal of cleaner production*, *196*, 635-643. https://doi.org/10.1016/j.jclepro.2018.06.087

[2] Cisternas, I., Velásquez, I., Caro, A., & Rodríguez, A. (2020). Systematic literature review of implementations of precision agriculture. *Computers and Electronics in Agriculture*, *176*, 105626. https://doi.org/10.1016/j.compag.2020.105626

[3] Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., & Iqbal, N. (2020). Precision agriculture techniques and practices: From considerations to applications. *Sensors*, *19*(17), 3796. https://doi.org/10.3390/s19173796

[4] Robert, P. C. (2020). Precision agriculture: a challenge for crop nutrition management. *Progress in Plant Nutrition: Plenary Lectures of the XIV International Plant Nutrition Colloquium*, 143-149. https://doi.org/10.1007/978-94-017-2789-1_11

[5] Lu, J., Ehsani, R., Shi, Y., de Castro, A. I., & Wang, S. (2018). Detection of multi-tomato leaf diseases (late blight, target and bacterial spots) in different stages by using a spectral-based sensor. *Scientific reports*, *8*(1), 1-11. https://doi.org/10.1038/s41598-018-21191-6

[6] Verma, S., Chug, A., Singh, A. P., Sharma, S., & Rajvanshi, P. (2019). Deep learning-based mobile application for plant disease diagnosis: A proof of concept with a case study on tomato plant. *Applications of image processing and soft computing systems in agriculture IGI Global*, 242-271. https://doi.org/10.4018/978-1-5225-8027-0.ch010

[7] Lu, J., Tan, L., & Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. *Agriculture*, *11*(8), 707. https://doi.org/10.3390/agriculture11080707

[8] 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

[9] Shrestha, G., Das, M., & Dey, N. (2020). Plant disease detection using CNN. *2020 IEEE Applied Signal Processing Conference (ASPCON)*, 109-113. https://doi.org/10.1109/ASPCON49795.2020.9276722

[10] Sembiring, A., Away, Y., Arnia, F., & Muharar, R. (2021). Development of concise convolutional neural network for tomato plant disease classification based on leaf images.

Journal of Physics: Conference Series IOP Publishing, *1845*(1), 012009. https://doi.org/10.1088/1742-6596/1845/1/012009

[11] Chohan, M., Khan, A., Chohan, R., Katpar, S. H., & Mahar, M. S. (2020). Plant disease detection using deep learning. *International Journal of Recent Technology and Engineering*, *9*(1), 909-914. https://doi.org/10.35940/ijrte.A2139.059120

[12] Ashqar, B. & Abu-Naser, S. (2019). Image-Based Tomato Leaves Diseases Detection Using Deep Learning. *International Journal of Engineering Research*, *2*, 10-16.

[13] Owomugisha, G. & Mwebaze, E. (2016). Machine learning for plant disease incidence and severity measurements from leaf images. *2016 15th IEEE international conference on machine learning and applications (ICMLA) IEEE*, 158-163.

[14] Haggag, M., Abdelhay, S., Mecheter, A., Gowid, S., Musharavati, F., & Ghani, S. (2019). An intelligent hybrid experimental-based deep learning algorithm for tomato-sorting controllers. *IEEE Access*, *7*, 106890-106898. https://doi.org/10.1109/ACCESS.2019.2932730

[15] Moallem, P., Serajoddin, A., & Pourghassem, H. (2017). Computer vision-based apple grading for golden delicious apples based on surface features. *Information processing in agriculture*, *4*(1), 33-40. https://doi.org/10.1016/j.inpa.2016.10.003

[16] Fuentes, A. F., Yoon, S., Lee, J., & Park, D. S. (2018). High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank. *Frontiers in plant science*, *9*, 1162. https://doi.org/10.3389/fpls.2018.01162

[17] Gadekallu, T. R., Rajput, D. S., Reddy, M., Lakshmanna, K., Bhattacharya, S., Singh, S., Alazab, M. et al. (2021). A novel PCA -whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing*, *18(*4), 1383-1396. https://doi.org/10.1007/s11554-020-00987-8

[18] Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture - a worldwide overview. *Computers and electronics in agriculture*, *36*(2-3), 113-132. https://doi.org/10.1016/S0168-1699(02)00096-0

[19] Gebbers, R. & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*, *327*(5967), 828-831. https://doi.org/10.1126/science.1183899

[20] Zhang, Q. (2016). *Precision agriculture technology for crop farming*. Taylor & Francis. https://doi.org/10.1201/b19336

[21] Khanna, A. & Kaur, S. (2019). Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture. *Computers and electronics in agriculture*, *157*, 218-231. https://doi.org/10.1016/j.compag.2018.12.039

[22] Dholu, M. & Ghodinde, K. A. (2018). Internet of things (IoT) for precision agriculture application. *2018 2nd International conference on trends in electronics and informatics (ICOEI)*, 339-342. https://doi.org/10.1109/ICOEI.2018.8553720

[23] Ahmed, N., De, D., & Hussain, I. (2018). Internet of Things (IoT) for smart precision agriculture and farming in rural areas. *IEEE Internet of Things Journal*, *5*(6), 4890-4899. https://doi.org/10.1109/JIOT.2018.2879579

[24] Nguyen, L. H. & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS computational biology*, *15*(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

[25] Okada, S., Ohzeki, M., & Taguchi, S. (2019). Efficient partition of integer optimization problems with one-hot encoding. *Scientific reports*, *9*(1), 1-12. https://doi.org/10.1038/s41598-019-49539-6

[26] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

[27] Vidal, R., Ma, Y., & Sastry, S. S. (2016). *Principal component analysis. Generalized principal component analysis.* Springer, New York, NY. https://doi.org/10.1007/978-0-387-87811-9_2

[28] Sultan, S., Javed, A., Irtaza, A., Dawood, H., Dawood, H., & Bashir, A. K. (2019). A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients. *Journal of Ambient Intelligence and Humanized Computing*, *10*(10), 4197-4206. https://doi.org/10.1007/s12652-019-01444-6

[29] Sharma, A., Sharma, A., Panigrahi, B. K., Kiran, D., & Kumar, R. (2016). Ageist spider monkey optimization algorithm. *Swarm and Evolutionary Computation*, *28*, 58-77. https://doi.org/10.1016/j.swevo.2016.01.002

[30] Sharma, H., Hazrati, G., & Bansal, J. C. (2019). *Spider monkey optimization algorithm. Evolutionary and swarm intelligence algorithms.* Springer, Cham. https://doi.org/10.1007/978-3-319-91341-4_4

[31] Stratakis, C., Menelaos Stivaktakis, N., Bouloukakis, M., Leonidis, A., Doxastaki, M., Kapnas, G., Evdaimon, T., Korozi, M., Kalligiannakis, E., & Stephanidis, C. (2022). Integrating Ambient Intelligence Technologies for Empowering Agriculture. *Engineering Proceedings*, *9*(1), 41. https://doi.org/10.3390/engproc2021009041

[32] Mhango, Joseph K. et al. (2021). Mapping Potato Plant Density Variation Using Aerial Imagery and Deep Learning Techniques for Precision Agriculture. *Remote Sensing*, *13*(14), 2705. https://doi.org/10.3390/rs13142705

[33] Akanksha, G. & Nahar, P. (2022). Classification and yield prediction in smart agriculture system using IoT. *Journal of Ambient Intelligence and Humanized Computing*, 1-10.

[34] Gaikwad, S. V. et al. (2021). An innovative IoT based system for precision farming. *Computers and Electronics in Agriculture*, 187.

[35] Gadekallu, T. R., Dharmendra Singh Rajput, M. Reddy, K. L., Bhattacharya, S., Singh, S., Jolfaei, A., & Alazab, M. (2021). A novel PCA - whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing*, *18*(4), 1383-1396. https://doi.org/10.1007/s11554-020-00987-8

[36] Reddy, G. T., Praveen Kumar Reddy, M., Lakshmanna, K., Kaluri, R., Singh Rajput, D., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

Contact information:

Mo DONG
College of Biological and Agricultural Engineering, Jilin University,
Changchun 130022, Jilin, China
Mudanjiang Medical University, Mudanjiang 157000, Heilongjiang, China

Haiye YU
College of Biological and Agricultural Engineering, Jilin University,
Changchun 130022, Jilin, China

Lei ZHANG
College of Biological and Agricultural Engineering, Jilin University,
Changchun 130022, Jilin, China

Yuanyuan SUI
College of Biological and Agricultural Engineering, Jilin University,
Changchun 130022, Jilin, China

Ruohan ZHAO
(Corresponding author)
Mudanjiang Medical University,
Mudanjiang 157000, Heilongjiang, China
E-mail: zhaoruohan@mdjmu.edu.cn