Forecasting diseases that affect plant leaves and moisture levels in the soil using a data mining approach

Ritu Shreea | Rupal Guptab | V. Srikanthc

aVivekananda Global University, Jaipur, India, Assistant Professor, Department of Computer Science & Application.
bTeerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Assistant Professor, College Of Computing Science And Information Technology
cJain (deemed to be) University, Bangalore, India, Associate Professor, Department of Computer Science and Information Technology.

Abstract The foundation of the global and Indian economies is agriculture. Since agriculture started millions of years ago, many environments, civilizations, and technical developments have fostered and defined the evolution of agricultural technology. In this study, we examine how we may analyze images of plants and soil to better keep tabs on their health, as well as how we can determine how much water each kind of plant needs. Images of the plants and soil are first taken using a digital camera with the necessary resolution. The form and geometric characteristics are extracted from the plant images using the inner distance shape context-based descriptor and geometrical descriptors. The soil images are also used to extract features and color properties. The botanical plant species dictionary is used to identify the plant type using the contour elements of the plant photos. Gradient structured random forest (GS-RF) classification is used to forecast leaf diseases. Principal Component Analysis (PCA) and Hierarchical Gradient Deep Neural Network (HG-DNN) classification techniques are used to determine the causes of a given plant disease based on the characteristics of soil images and plant disease images. The findings are communicated to the growers through text messages sent to their mobile phones on a daily and seasonal basis, along with any potential recommendations for preventative actions.

Keywords: crop production, geometrical descriptors, plant disease, PCA, GS-RF, HG-DNN

1. Introduction

India’s economy, as well as the whole world, is based on agriculture. The development of agricultural technology has been supported and characterized by many conditions, civilizations, and technological advancements since it began millions of years ago. Previously, the classification of agriculture was based on increased productivity, the substitution of synthetic fertilizers and insecticides for natural ones, and the allotment of more area. Due to the elimination of environmental reasons in contemporary society, macrobiotic and sustainable agricultural practices have emerged. Through modern prediction methods that give the ideal intensification environment under conditions of reproduction convenience protection, the maximum agricultural yield with the best quality is attained (Bradford 2020). The use of data mining methods and improved information and communication systems allow for the monitoring of the contemporary expansion of agricultural operations (Zhang 2021). The health and development of plants are strongly influenced by elements that are intimately related to plant leaves and soil moisture levels. While plant leaves serve as indications of a plant’s water status and general health, sufficient soil moisture is necessary for plants to absorb water and crucial nutrients. For irrigation and plant cultivation to be effective, both components must be monitored and managed (Joswig 2022). One may spot evidence of moisture stress by carefully examining the leaves. Leaves that are wilting, drooping, or turning yellow usually indicate inadequate moisture but leaves that are too wet may indicate over-irrigation or poor drainage. Irrigation decisions may be helped by the fast response provided by visual observation (Abioye et al. 2021). Different methods and tools may be used to assess soil moisture levels directly. Metal probes on portable plant moisture meters allow them to be put into the soil. They provide instantaneous moisture content measurements. Another option is to use tensiometers, which measure soil moisture based on the force needed to draw moisture from the ground. These tools may provide information on soil moisture at various depths (Wang 2019). To correctly detect moisture content, these sensors use a variety of technologies, including capacitance, time domain reflectometry, or gypsum blocks. Soil moisture sensors make it possible to gather data in real-time and may aid with irrigation scheduling (Kamath 2019). With the ability to monitor wide regions, remote sensing may provide important data on the condition of the vegetation and the water supply. One may locate places with insufficient soil moisture by looking at the vegetation indices produced from remote sensing data (Trugman 2019). The soil moisture levels may be inferred indirectly
from weather station data. Evapotranspiration rates, which affect soil moisture, are influenced by variables including rainfall, temperature, and relative humidity. It is possible to determine the moisture stress on plants and modify irrigation methods by taking weather station data into account and predicting water needs (Benos 2021). Effective plant management and irrigation depend on measuring soil moisture and plant leaves. Plant water stress may be immediately detected by looking at the leaves. Precision information on soil moisture content is provided by direct measuring methods such as plant moisture meters, tensiometers, and soil moisture sensors. Broader views on plant health and soil moisture patterns are made possible by remote sensing methods and weather station data. Making educated decisions and using effective water management techniques in plant agriculture are made possible by integrating information from these numerous sources (Millet 2019). Typically, the practice of extracting samples from huge databases is referred to as data mining. The fundamental goal of data mining methods is to extract the most important aspects from databases and arrange that information in a useful arrangement for further use. Depending on the agricultural uses, many data mining methods (Sapes 2019) are used in agriculture. Weather, pollution, and other environmental variables may be predicted with the help of the data mining tools supplied. Soil properties, weed identification, and water-core monitoring are only a few examples of the uses for data mining approaches.

However, the most pressing problem in agriculture is the control of total crop yields via the use of cutting-edge technologies. Therefore, an efficient data mining approach is suggested in this work to forecast the ideal humidity for plants and the onset of various illnesses. This paper’s primary objective is to determine the many plant diseases and the factors that contribute to their occurrence. The GS-RF classification technique is used to predict plant diseases. In addition, HG-DNN classification, which relies on probability values between characteristics of soil images and diseased plant images, is used to forecast the causes of plant illnesses.

The rest of the research is structured as follows: The various data mining methods used in agricultural domains are described in Section 2. The suggested agricultural production monitoring system is described in Section 3 employing unique data mining methods. The performance assessment of the suggested strategies is shown in Section 4. The research study is concluded in Section 5, which also includes recommendations for further work on improvement.

2. Related works

Study (Mishra 2021) examines how the Internet of Things (IoT) may be used to efficiently gather data on environmental factors such as temperature, pH, and precipitation using a variety of machine learning methods. Decision Tree to analyze agricultural data and provide recommendations about what may and cannot be planted. To improve agricultural yields on a large scale, a model is developed in this research that utilizes real-time data to make in-field monitoring decisions based on weather analysis and the identification of crop diseases. The software architecture of this platform is flexible enough to accommodate a variety of plant disease models and other precision agriculture applications, making it suitable for usage with a wide range of plant diseases (Khattab 2019). A study (Rigden 2020) showed that utilizing soil moisture data leads to more precise projections of maize output and suggests that explicitly accounting for fluctuations in water availability is necessary for precise estimations of how climate change will affect crop yields. The strengths and weaknesses of various methods and models offered in the current literature are highlighted in the article (Dhaka 2021). In addition to discussing how effective models are, this paper provides a summary of the datasets and performance measures utilized to do so. In the study (Archontoulis 2020), they accomplish the following: (1) explain the process behind the forecasts, (2) assess the accuracy of the model’s predictions using data from 10 sites over 4 years, and (3) pinpoint the variables most important for predicting future yields and soil N dynamics. The models, which were made using past, present, and future meteorological data, were made public four weeks after planting. Research (Albergel 2019) evaluated the land data assimilation system LDAS-Monde, created by Météo-France, for its ability to track how the 2018 summer heatwave in Western Europe affected the health of the region’s vegetation. Article (Sharma 2020) provided a comprehensive overview of the use of ML in farming. To keep an eye on the quality and production of crops, it is necessary to classify various photos of them. Improve livestock output by utilizing ML models trained on data gathered from collar sensors to make predictions about fertility, diagnose eating problems, and analyze the habits of cattle. Research (Pereira 2020) seeks to update their understanding of how to estimate crop coefficients using data on ground cover and vegetation height. Article (Abd El-Ghany 2020) offered a historical and futuristic perspective on remote sensing methods and their uses, particularly in the control of insect pests and plant diseases. The electromagnetic radiation reflected and emitted by the ground target is measured, recorded, and processed to perform remote sensing. The spectral characteristics of living things are crucial to remote sensing applications. Insect pests and plant diseases may now be detected, forecasted, and managed on a wide variety of fruit orchards and crops with the use of remote sensing. The primary goals of these apps were to gather information useful for making decisions about insect pest control and reducing chemical pesticide contamination. Each kernel’s development is influenced by the amount of water and nitrogen (N) available to the plant during the blooming period. The current research documents the changes in the dry weight of maize kernels in response to varying amounts of water and N. The effects of three irrigation regimes and five N treatment rates on weekly maize kernel growth were tested for two consecutive years. Findings add to the literature on
maximizing maize yields in semiarid environments by optimizing the use of nitrogen fertilizer and irrigation water (Hammad 2021).

3. Methodology

Our suggested method for enhancing the monitoring of agricultural yields is divided into three stages: image preprocessing, detection, and communication. Data mining methods are refined and adapted for various procedures at each stage. In the initial step, we prepare photos of the soil and plants. In the second stage, plant illness and its causes are predicted based on soil characteristics. The last step makes use of IT to warn farmers by disseminating data about plant varieties, plant illnesses, and their origins. This section provides a concise summary of these three stages of monitoring systems.

3.1. Image Pre-processing Phase

Initially, high-resolution digital camera input photos of several plant types and soil are gathered. To continue processing, the collected pictures are sent to an image processing unit, either over a wired or wireless network. The gathered pictures then undergo pre-processing, whereby noise and other disturbances are removed while the characteristics used in the prediction process are also improved. Images are cleaned up by removing noise and then transformed from RGB to their color space equivalents. After the enhancement process is complete, a Region of Interest (ROI) based segmentation technique is utilized for the resulting images. The goal of ROI segmentation is to divide a given image into many distinct sections or categories. It is denoted in equation (1):

$$Q_{ij} = \frac{\psi_j q_j}{\sum q_j} + \sum w_j$$

In equation (1), $\psi_j q_j$ refers to the weight of the two adjacent regions of the area, $\sum q_j$ is the length of the two regions, $\omega$ refers to the size of the area, $f$ denotes the boundary strength and $w_j, w_l$ represent the spectral values of two regions. The polygonal leaf model is used to recover the contour elements of plant pictures. Then, using a botanical plant species dictionary containing descriptions of various plant species, semantic representation is given to these contour data to identify the kinds of plants.

3.2. Plant Disease Prediction

The GS-RF makes its disease prediction based on the plant species and characteristics. When GS-RF is used in a training dataset, it may detect plant diseases that were previously undetected.

3.2.1. Gradient structured random forest

GS-RF combines the benefits of random forests with gradient boosting. To improve forecast accuracy and interpretability, it was added as an addition to the conventional random forest method. The goal of GS-RF is to identify the interactions and cumulative effects of features in a dataset.

Incorporating gradient boosting into the random forest framework is the main concept underlying GS-RF. A well-liked approach called gradient boosting creates a series of weak learners by fitting the residuals of the prior models repeatedly. By steadily improving forecasts, it focuses on reducing the loss function. Random forests, on the other hand, train several decision trees individually and combine their predictions using either majority voting or averaging.

A random forest model is trained on the training set after a subset of features is randomly chosen at each iteration. The residuals are then calculated using the predictions of these trees, and in the succeeding iteration, a fresh set of trees are used to fit the residuals. Until a certain number of iterations or a stopping condition is achieved, this process keeps going.

The boosting and random forest methods are combined in GS-RF to make use of each method’s advantages. The random forest component captures the non-linear correlations and interactions between the features, while the boosting component concentrates on improving the model by fitting the residuals.

Benefits for interpretability come from GS-RF. It enables feature significance analysis, allowing the discovery of factors that significantly influence the model’s predictions. Based on the average influence of each feature over the ensemble of trees, feature significance may be calculated. The underlying links between characteristics and the target variable may then be better understood, which makes GS-RF an important tool for data analysis and interpretation.

A random forest is a kind of ensemble learning machine that generates hypotheses by combining the major population of forecasts from several various base models. By reflecting important qualities, the RF is simple and efficiently specifies data from large datasets. The pixels of the leaf are constructed to generate a subdivision of $R^d$ at each level, and $R^d$ is the foundation of the corresponding tree. Each tree is created uniquely, with each node matching a rectangular subset of
$R^2$. One expanded tree leaf is picked at each step of construction. The dataset is randomly split into two pieces that serve different purposes in the construction of each tree. The structural elements that affect the shape of the tree are taken into consideration to estimate the split dimensions and split characteristics. The assessment nodes are used to fit the estimators in each tree leaf. By allocating each point to the structure or estimate component, each tree experiences a random division of the perceptions regarding data.

The testing examples for the $h_1(x), h_2(x), \ldots, h_k(x)$ classifiers were randomized from the distribution of the random vectors $Y, X$. The revenue feature is written in equation (2-5)

$$Mg(X, Y) = av_k(h_k(X)) = Y - \frac{\text{max}}{j \neq Y} av_k(h_k(X) = j) \quad (2)$$

where the measured value is $I(\cdot)$. The source of the mistake is

$$PE^* = P_{XY}(mg(X, Y) < 0) \quad (3)$$

The probability over the $XY$ dimension, indicated by where $X, Y$ space In RF, $h_k(X) = h(X, \Theta_k)$

The margin feature for an RF is

$$mr(X,Y) = P_a(h(X, \Theta) = Y) - \frac{\text{max}}{j \neq Y} P_a(h(X, \Theta) \text{)} \quad (4)$$

Moreover, the set of classifiers $\{h(X, \Theta)\}$ has a value of

$$S = E_{XY} mr(X,Y) / 4.5 \quad (5)$$

The GS-RF training procedure resolves this optimization issue. The training algorithm 1 is broken down into the following stages:

**Algorithm 1: Training Algorithm**

1. Set parameters $C$ and $C^*$
2. To create the first classifier, apply inductive GS-RF to the training set of data
3. Set the number of samples with a positive label depending on the rule
4. All unlabeled samples’ decision function values should be calculated
5. Indicate as affirmative samples those with the greatest decision function
6. Set temporary effect factor $C_{tmp}$
7. Overall samples retain the existing
8. One set of different-labeled unlabeled samples’ labels are switched using a specific rule to produce the value of the objective function by utilizing (5).
9. Repeat the procedure until there are no more pairs of samples that meet the switching requirement.
10. Increase the $C_{tmp}$ value before proceeding to step 7.
11. $f(C_{tmp} > C^*)$
12. Stop
13. Attain the output

In the testing phase, the data is compared to the training data, which successfully predicts the type of plant disease. Additionally, using the PCA approach to extract the characteristics from the soil images, the necessary water content level for the particular kind of plant is forecasted.

**3.3. Principal Component Analysis (PCA)**

The PCA is used to analyze the measurements of several variables in a group of individuals. The PCA operates in a space with fewer dimensions to facilitate comprehension and reduce the number of latent variables (principal components) from the original variables. By resolving an optimization issue specified in equation (6), the primary components are achieved.

$$\max (Y = x^2Ux), \quad (6)$$

Subject to $x^2Ux = 1$.

Where $V = (1/m) W^2W$ is the data matrix's sample covariance matrix, and $x$ is of unitary norm according to the requirement $X^2X = 1$, where $W$ is a matrix with $n$ elements and $o$ variables. The greatest eigenvalue $\lambda$ of $U$, or $UX = \lambda W$, is the answer to the optimization problem specified in equation (2) after performing a dimension reduction to $r$ components ($r < o$). The eigenvalues of $U$ are arranged according to $\lambda_1 > \cdots > \lambda_r$ in decreasing order. The first primary component $w_1$ is the unitary norm's eigenvector, or _1, which is connected to the greatest eigenvalue of $U$, that is, $\lambda_1$, when $x_1$ and $x_2$ are
orthogonal, as in $x_1^T x_2 = 0$, the second principal component, w2, $\lambda_2$ is the eigenvector associated with them, and so on for the other components.

It is possible to rotate the retrieved components to enhance the understanding. One of the most well-liked rotation techniques is Varimax, which is used for component loadings. see the varimax approach in detail. The varimax orthogonal rotation attempts to maximize the variation of the squared loadings in each component, resulting in big loadings for a select few variables and modest loadings for the other variables. Consequently, a subset of factors that influence each component may be found. The observed variables should therefore only have one loading that is significant in absolute value, ensuring that the variables are primarily connected to one component.

3.4. Hierarchical Gradient Deep Neural Network (HG-DNN)

DNN can extract latent learning from huge datasets, as seen in Figure. 1. With the help of the Python library Keras, users may quickly and flexibly build and test deep learning models. Wrapping Theano and Tensor Flow, two of the most potent numerical computing libraries presently accessible makes it simple to describe and create neural system models with only a few lines of code at most. An epoch is a whole loop over a particular dataset. By adjusting the number of epochs through backpropagation, errors may be significantly reduced. A confusion matrix is one tool for evaluating the effectiveness of a categorization system. The information about the existing classes and the predicted classes is presented in a confusion matrix in the form of rows and columns. Different components make up the HG-DNN’s many network levels. At each layer, information is received at the top, shown at the bottom, and kept hidden at all subsequent levels. The model is run, and the results are compiled. However, backpropagation is used to complete the HG-DNN model training. Or the loss function measures the discrepancy between outcomes that were anticipated and those that were achieved. Throughout the training phase, changes are made to the weight coefficient W and bias B.

$$I(V, a, y, x) = \frac{1}{2} \|x^K - x\|_2^2 = \frac{1}{2} \sigma(V^K x^{K-1} + a^K) - x \|_2^2 \quad (7)$$

In equation (7), The buried layer’s output is denoted by $x^K$.

3.5. Communication Phase

In this communication stage, farmers get letters or emails outlining the types of plant diseases that are likely to appear on their farms, as well as the soil conditions that are likely to cause those diseases. Farmers’ output may be boosted by the suggested monitoring system to better water management and reduced disease incidence. Algorithm 2 shows proposed steps.

Algorithm 2: Proposed Algorithm

Input: Plant or Soil images $I_1, ... , I_n$
Output: Type of plant & disease, and its causes.
1. Gather Images
2. Imagine noise reduction
3. Transform a picture into a representation in color space
   //Image segmentation
4. Select the image’s seed pixel
5. Set standards to help the area develop
6. If a pixel is 8-connected to at least one other pixel in the area, include it in the region
7. Test the percentage of each pixel
8. Label all the region
9. If two areas have the same label, combine them
10. Extrapolate the geometric details and form from the plant images
11. Apply semantic contour representation
12. Determine the plant’s species
//Prediction of plant disease
13. Define the optimization problem for the GS-RF classifier
14. Get training materials
15. Execute the GS-RF procedure
16. Attain the result
17. Compare the training and testing data.
18. Determine the disease’s form
//Prediction of the causes of plant diseases
19. Identify the soil image textures and colors.
20. Using PCA distribution, choose the topic mix for the description of the plant species
21. Select the subject to create each word in the description
22. Then create words based on a multinomial distribution for the subject
23. Calculate the likelihood that soil properties and words will occur
24. HG-DNN classification should be done using the calculated probabilities
25. Utilize equation (7) to determine the output function
26. Determine the root causes of the specific plant disease
27. Text messages may be used to inform farmers of the forecast findings
28. End

4. Results

In this section, we compare the accuracy and precision of current and proposed methods for predicting plant diseases and their underlying causes. Causes of plant illnesses are predicted and compared to those predicted using the principal component analysis and the similarity measure methods. The suggested methods of Convolutional Neural Network (CNN), K-Nearest Neighbor (KNN), and GS-RF are compared with one another and their respective plant disease prediction outcomes.

Precision, which is well-defined as the ratio of properly categorized cases to all occurrences of predictively positive data, is one of the most important criteria for accuracy. Equation (8) is used to compute the precision.

\[
\text{precision} = \frac{TP}{TP + FP} \quad (8)
\]

![Figure 2 Comparison of Precision in Plant Disease Causes Prediction.](image)

The comparison of the Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), Gabor filters (GF), and PCA techniques in terms of accuracy for predicting the causes of plant diseases are shown in Figure 2. The inspection of the aforementioned graph demonstrates that as the number of features rises, so does the accuracy of the PCA approach.
Methods are taken for the x-axis, and precision values are obtained for the y-axis. It demonstrates that the PCA prediction approach has improved accuracy.

![Figure 3](image3.png)

**Figure 3** A comparison of Precision for predicting plant leaf disease.

Figure 3 compares the accuracy of the CNN, KNN, and GS-RF approaches for predicting plant leaf disease. The examination of the graph above demonstrates that as the number of features rises, the accuracy of the GS-RF approach also increases. Methods are taken for the x-axis, and precision values are obtained for the y-axis. It demonstrates that the GS-RF prediction approach now has higher accuracy.

The system’s accuracy is measured by the proportion of samples for which the proposed strategy correctly anticipated outcomes. Equation (9) is used to determine accuracy.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)
\]

![Figure 4](image4.png)

**Figure 4** Comparison of Accuracy in Plant Disease Causes Prediction.

Figure 4 compares the accuracy of the GLCM, LBP, GF, and PCA approaches for predicting the causes of plant disease. The examination of the graph above demonstrates that as the number of features rises, the accuracy of the PCA approach also increases. Methods are listed on the x-axis, while accuracy (%) is shown on the y-axis. It demonstrates that the PCA prediction method’s accuracy has grown.

![Figure 5](image5.png)

**Figure 5** A comparison of Accuracy for predicting plant leaf disease.
Figure 5. Analyzes the accuracy of plant leaf diseases using CNN, KNN, and GS-RF methods in comparison. The examination of the graph above demonstrates that as the number of characteristics rises, the accuracy of the GS-RF approach also increases. Methods are listed on the x-axis, while accuracy (%) is shown on the y-axis. It demonstrates that the GS-RF prediction method's accuracy has grown.

5. Conclusion

This paper suggests a new data mining approach for use in agricultural field monitoring systems. The suggested data mining method is associated with the forecasting of plant diseases and their causes. The suggested approach uses contour analysis of extracted characteristics from leaf images in conjunction with a botanical plant species dictionary to determine the kind of plant. The GS-RF classification then makes a prediction of the plant disease based on the form and texture attributes of the plant images. The PCA method is used to create a model between the color and texture properties of soil photographs and images of infected plants. In addition, HG-DNN categorization is used to predict the origins of plant diseases. The farmers then get text messages on their mobile devices with the forecasted information. As a consequence, the agricultural output monitoring system is improved by the use of our proposed monitoring approach to better control irrigation and diseases. The experimental results validate the usefulness of the suggested monitoring system. Improving agricultural yields will need a further examination of the prediction method for tracking plant development.

Ethical considerations

Not applicable.

Declaration of interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References


