Detection of Maize Disease Using Random Forest Classification Algorithm

[1]Ms. Deepika Chauhan, [2]Dr. Ranjan Walia, 3Dr. Chaitanya Singh, [4]Dr. M. Deivakani, [5]Mr. Makhan Kumbhkar

[1]Assistant Professor, Department of Computer Science & Engineering, Shivajirao Kadam Institute of Technology and Management, Indore, M.P, India

[2]Associate Professor, Electrical Engineering Department, Model Institute of Engineering and Technology, Jammu and Kashmir, India

[3]Associate Professor, Department of Computer Science & Engineering, Shivajirao Kadam Institute of Technology and Management, Indore, M.P, India

[4]Associate Professor, Electronics and Communication Engineering, Psna College Of Engineering and Technology, Dindigul, TamilNadu

[5]Assistant Professor, Department of Computer Science & Engineering, Christian Eminent college , Indore, M.P, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Plant sicknesses are the significant reason for low agrarian profitability. For the most part the farmers experience troubles in controlling and identifying the plant infections. Accordingly, early detection of these diseases will help to increase the productivity of crop. This paper projected early detection of disease in crop using AI methods like Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF). In this paper we compare all the method on the basis of accuracy and proposed the best model for the efficient detection of the disease. The Random forest model here achieve the accuracy of 80.68% when compare with other existing model.

Keywords: Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF).

I. Introduction

The demand for food is growing exponentially as agronomy production is very low. To address this, farmers, scientists, researchers, analysts, experts, and governments are trying to increase their efforts to increase agricultural productivity. There is great progress in the agricultural sector with the help of technology. There are various factors such as global climate change, as well as plant diseases that farmers face. Several reasons for the loss of crop production leading to suicidal conditions for farmers. The use of time, cost and accuracy of crop quality testing by visual inspection is a challenging task. To overcome this problem researchers came up with several solutions for the development of new technologies such as object detection, as well as image processing for quality tests. In this paper, image processing technology is used for the detection and classification of diseases [1-3] of plant diseases. This method of image processing requires high-resolution images to detect and classify diseases that were difficult to capture. For this reason, it is also a tedious task to predict disease effectively and efficiently. This paper aims to develop a model that accurately detects and differentiates leaf disease at the onset of disease by using algorithms to study [4] and takes the necessary steps to protect against such leaf diseases. This paper uses a variety of supervised machine learning methods such as NB [5], KNN [6, 7], DT [8], SVM [9], and RF [10, 11] to detect diseases and tree segregation from plant leaves and comparisons are made between of many classification techniques. It also offers a variety of strategies that will give a more accurate result compared to other strategies. These classification methods are successfully used in many applications [12] such as biomedical signal processing [13] and health care [14, 15]. The main causes of maize diseases can be caused by biotics (bacteria, fungi, todes, viruses) and abiotic (due to the effects of nutrient deficiencies, moisture and heat). Maize germs mainly affect the leaves, fruits and stems. Here, some of the major diseases are (a) Cercospora leaf spot or gray spot sent to Fig. 1. It is a small area with lumps of leaves that extend to rectangular wounds as the wound grows. They are sweetened and eventually turn gray. Management use of hybrid sterile resistance to the disease and the use of foliar fungicides. (B) General Rust presented in Fig. 2 the presence of brown pustules. It is found in both the upper and lower extremities of the leaves as it intensifies the eruption and the powdery red letters. When there is a severe infection of the pituitary glands from the wrists and ears and the sides begin to turn yellow. Management of the use of fertilizers resistant to the use of Fungicide sprays such as folicur, oxy-chloride, AMISTAR copper, and Bravo (c) Northern Leaf Blight shown in Fig. , can be seen. bacterial process, these lesions turn a light gray in tan color. Follow-up crop management and

crop management. The rest of the paper is organized as follows: Sect. 2 describes the parallel function of reducing maize separation processes, Sect. 3 presented a detailed description of many machine learning techniques, Sect. 4 introduces a detailed procedure for the proposed method of classifying corn leaf diseases, Sect. 5 presents a comparative analysis of all classification strategies and ultimately Sect. 6 concludes the paper with the future scope of the work.

Fig. 1 Cercospora leaf spot

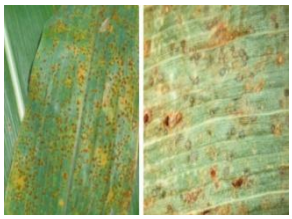


Fig. 2 Common rust

Fig. 3 Northern Leaf Blight

II. Literature Survey

Maize, also known as corn, is one of the most important crop yields under relatively few climates. Maize disease can be in a few areas such as stem, leaf or shock. In this section, we will only look at the infection identified by the leaves. Ishaket et al. [16] introduced the implant neural network (ANN) implantation of Phyllanthus Elegant Wall disease into two categories such as healthy or unhealthy. They have changed the color scheme of the vegetable photos using photo processing techniques. Images are classified according to color and location of the leaf. Dandawate et al. [17] detected soybean leaf infections using SVM. This algorithm has used a mutation-modifying process that has automatically detected plant diseases based on their composition. This helps the farmer with less effort online. Patil et al. [18] extracted features of tomato leaf. The leaf image is divided into red, green, and blue elements. These factors are used in the classification of diseases. Ghadge et al. [19] It focuses on assisting farmers in producing suitable crops based on soil quality. Machine learning algorithms are used to predict plants. The presence of nutrients in the soil is assessed and predicted for crop production in a particular area. Hong et al. [20] proposed a model for the development of accuracy in the agricultural sector. Soil moisture forecast is created to predict moisture based on natural conditions. These predictions result in greater accuracy over a longer period of time. Mohanty et al. [21] The convolution neural network (CNN) was used to diagnose leaf infections. They have used large database images of healthy and disease-resistant plant leaves. They have tried three types of stocks such as colored, grayscaled, and segmented. This CNN method is able to identify 26 diseases with 14 plant species easily. This algorithm has used a mutation-modifying process that has automatically detected plant diseases based on their composition. This helps the farmer with less effort online.

III. Machine Learning algorithm for classification

Naive Bayes Classifiers: Naive Bayes Classifiers are one of the easiest and most effective ways of organizing. This approach is based on the concept of Bayesian Networks which is a possible graphical model representing a set of random variables and their conditional autonomy. In Bayesian networks, there are many effective algorithms that make liking and learning. The only requirement is that the data features will work independently. The elements in the database are interdependent due to genetic origin but dependence does not appear to be strong. Therefore, considering that the characteristics of the data used are independent, the classification is based on the Naïve bayes process. Initially this approach creates a number of opportunities for each sequence in the sequence information, Sndb. When the inclusion data is provided, the number of probability sequences for each patient is calculated. This value is compared to the number of opportunities in the Sndb database. Sequence is therefore classified based on the number of probability sequences. The Naïve Bayes is a simple possible division based on applying the theory of the Bayes with a strong sense of independence. The possibilities for recording X data with a category Cj label are:

$$\boldsymbol{P(C_j \mid X) = \frac{P(X \mid C_j) * P(C_j)}{P(X)}}$$

the KNN algorithm is a method of classifying objects based on the closest examples of training in the feature space. KNN is a form of learning based on example, or lazy learning in which the work is limited only to the area and all calculations are taken up to section [22 - 25]. KNN is a basic and simple isolation method if there is little or no information about data distribution [25 - 28]. This rule simply preserves all the training set during the study and gives each question a category that represents the label of most of its closest neighbors. Neighborhood Neighborhood (NN) rule is the simpler way for KNN where K = 1. In this way each sample should be equally

separated from its surrounding samples. Therefore, if sample segregation is unknown, it can be estimated by considering the separation of the adjacent samples from neighbors. Given the anonymous sample and the training set, all distances between the unknown sample and all the samples in the training set can be calculated. The minimum value range is the same as the sample in the training set closest to the unknown sample. Therefore, an unknown sample can be classified based on the category of its immediate neighbor [28 - 33].

Decision tree: Decision tree method is the most widely used method of data mining to establish classification systems based on multiple covariates or to create predictive algorithms for target variability. This method divides the population into branch-like components that form a modified tree with a root node, internal nodes, and leaf nodes. The algorithm is not parametric and can handle large, complex databases without setting up a complex parametric structure. When the sample size is large enough, the study data can be broken down into training and validation data sets. A training database is used to construct a decision-making drug model and a confirmation database to determine the appropriate drug size required to achieve a complete final model.

Support vector machine: The vector support machine is based on the mathematical theory developed by Vapnik et al proposing a new method of learning, built on the basis of a limited number of samples from the information contained in the existing training text to obtain the best classification results. In recent years, many students have focused on research, which has become a practical problem for a solid theoretical foundation, but also for better resolution of small, non-linear sample, high magnitude and local minima points [34]. It is highly expected that the most distinctive features will be found in the large-size feature space. It can be complemented by modifying features using several common functions such as parallel, linear, polynomial and radial function. Modification of features can greatly increase the size of the feature space. Therefore, it increases the training time of the separation process. It can convert features into higher values by using computer products without changing the feature set.

Random forests: Random forests are a combination of predictable trees in such a way that each tree is based on random vector values sampled independently and with the same distribution of all the trees in the forest. The general deforestation error changes the a.s. the limit as the number of trees in the forest grows. The most common mistake of a tree divider tree depends on the strength of each tree in the forest and the interaction between them. Using random selection of features to separate each node produces error rates that compare well with Adaboost (Freund and Schapire [1996]), but are more robust in sound. Internal measurement monitors error, power, and correction and this is used to demonstrate the response to increasing the number of factors used in the division. Internal measurements are used to measure variable value. These ideas also apply to procrastination.

IV. Methodology

The proposed method consists of several elements such as image detection, pre-image processing, image classification, feature extraction, separation, and performance testing. The detailed descriptions of each process are as follows.

Image Acquisition: An image database, especially for photos of corn disease, is available on the plant's website. Maize plants are subsets with a total of 3823 photographs and labels for four categories of diseases such as common rust, gray hair, leaf damage to the north and health with 1192 images, 513 images, 956 images and 1162 images respectively. These labeled images are intended for disease classification and testing.

Image Pre-processing: This is required to achieve higher results in the resulting steps due to the presence of dew, dust, insect feces on plants. These effects are considered to be the sound of a corn image. To overcome these problems the RGB input image is converted to a gray image to provide better results. In this case, the size of the images is too large to reduce the size of the image. This image reduction also helps to reduce the memory size.

Image Segmentation: This plays an important role in the detection and classification of plant diseases. It simply divides the image into several objects or regions. Analyzes image data to extract useful information for continuous processing. This image class can be done in two ways depending on the similarity and the stop function. Similarly, images are classified according to certain pre-defined conditions. Therefore, the label-finding method is used in image classification and calculates the gradient of image sizes for each pixel within an image. But in contrast, the images are classified according to a sudden change in the intensity of the values as the edge detection.

Feature Extraction: This removes the elements of objects present in images. These extracted elements are used to illustrate the business. These elements are extracted and divided into three categories such as shape, color and texture. Diseases can vary in their appearance to various forms of the image due to disease. The model can easily identify diseases from the formation of traits. These feature structures vary in their axis, locations, and angles. The second parameter, that is, color is an important element of these three elements. It separates diseases from each other. The third parameter, that is, the texture describes how the color patterns are sprayed on the images. RGB feature extraction removes color details from images that are frequently used to process and identify patterns. RGB is highly recommended for object detection. It has a dramatic color change that easily identifies the images in the leaves. The RGB color value can determine all possible colors for three color lights such as red, green, and blue. The typical RGB value varies from 1 to 255 and tasks are performed in a range of 0-1. This test considers pixel grayscale values as analysis elements.

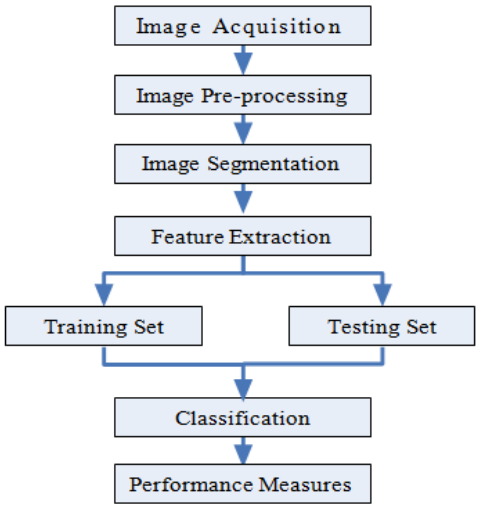


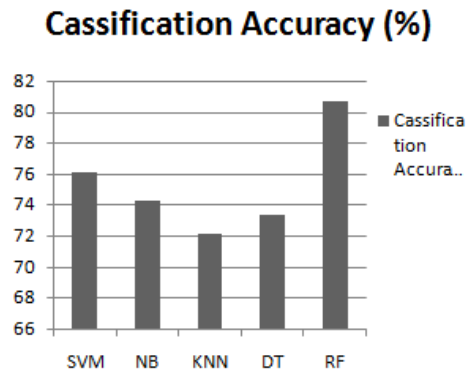
Fig. 4 Block diagram of the classification process

V. Experimental Result Analysis

This section investigates the effectiveness of the strategies of various categories such as NB, KNN, DT, SVM, and RF in the maize disease detection database and finds that the RF classification process is better than other classification methods. These maize data sets are divided into training data (90%) and test data (10%). The maize plant disease database contains a total of 3,823 photographs and four category labels. The details of the section label information about maize disease data are as follows: gray area, common rust, damage to the northern leaves, and healthy 513, 1192, 985, and 1162, respectively. The implementation of these partitions is done using Python 3.3 running on Windows 7 operating system. X64 based processor with a speed of 3.20 GHz, RAM size 8.00 GB and a 64-bit operating system type. Python software and python machine learning library and Pandas package are used. In this experiment, the image size was reduced by the original size to 100 100. The reason is that all the images have a unique size as the data set images were of different sizes. Images with gray labels are converted into cv2 image processing library in python. The formatted images are fed to the cucumber so that this change can be made more frequently. After this step, this cake file can be called to any split technique. After that certain code codes are applied and details are trained on the model and this model predicts image disease accordingly. The accuracy of the classification is presented in Table 1 and Fig. 5. Other operational methods such as precession, recall, and F-measure of all maize Database Database classification algorithms are shown in Fig. 6. We often differentiate between diseased or healthy leaves and if you are affected by a disease type.

Classifiers	**Accuracy (%)**
SVM	76.16
NB	74.35
KNN	72.16
DT	73.35
RF	80.68

Table 1 Comparison between classification techniques



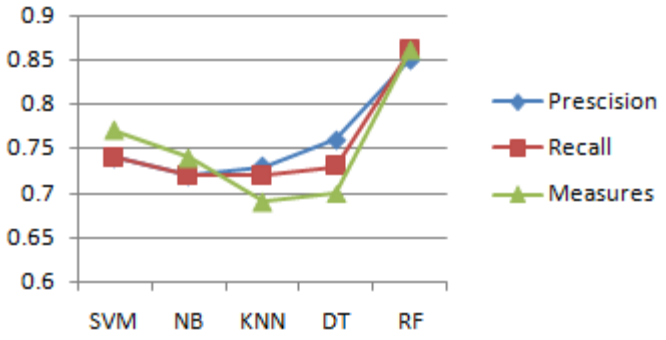


Fig. 5 Classification accuracy of classifiersFig. 6 Precision, recall, and F-measure of classifier

VI. Conclusion

In this work, surveyed machine learning methods NB, KNN, DT, SVM, and RF are used to diagnose various maize leaf diseases. The proposed method was applied using image data with a label to train the separation model. It is evident that the highest accuracy recorded in the RF classification among all other types of diagnostic classification in image data analysis. Farmers can take necessary steps depending on the early diagnosis to prevent maize diseases. However, there are certain pitfalls associated with each model in the classification process that may not apply to all data sets. In the future, these models can be implemented using high-dimensional data sets in many ways to differentiate.

Reference

1. Das, H., Naik, B., Behera, H.S.: Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In: Progress in Computing, Analytics and Networking, pp. 539–549. Springer, Singapore (2018)
2. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking, pp. 753– 764. Springer, Singapore (2018)
3. Das, H., Jena, A. K., Nayak, J., Naik, B., Behera, H.S.: A novel PSO based back propaga- tion learning-MLP (PSO-BP-MLP) for classification. In: Computational Intelligence in Data Mining, vol. 2, pp. 461–471. Springer, New Delhi (2015)
4. Murty, M.N., Devi, V.S.: Pattern Recognition: an Algorithmic Approach. Springer Science & Business Media (2011)
5. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, no. 22, pp. 41–46. IBM, New York (2001)
6. Fix, E., Hodges Jr, J.L.: Discriminatory Analysis-Nonparametric Discrimination: consistency Properties. California Univ Berkeley (1951)
7. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13(1), 21–27 (1967)
8. Quinlan, J.R.: Induction of decision trees. Mach. Learn. 1(1), 81–106 (1986)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)
10. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)
11. Barandiaran, I.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8) (1998)
12. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J.: Progress in computing, analytics and networking. In: Proceedings of ICCAN, p. 710 (2017)
13. Pradhan, C., Das, H., Naik, B., Dey, N.: Handbook of Research on Information Security in Biomedical Signal Processing, pp. 1–414. IGI Global, Hershey, PA (2018)
14. Sahoo, A.K., Mallik, S., Pradhan, C., Mishra, B.S.P., Barik, R.K., Das, H.: Intelligence-based health recommendation system using big data analytics. In: Big Data Analytics for Intelligent Healthcare Management, pp. 227–246. Academic Press (2019)
15. Dey, N., Das, H., Naik, B., Behera, H.S. (eds.).: Big Data Analytics for Intelligent Healthcare Management. Academic Press (2019)
16. Ishak, S., Rahiman, M.H.F., Kanafiah, S.N.A.M., Saad, H.: Leaf disease classification using artificial neural network. J. Teknologi, **77**(17) (2015)

17. Dandawate, Y., Kokare, R.: An automated approach for classification of plant diseases towards development of futuristic decision support system in Indian perspective. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 794– 799. IEEE (2015)
18. Patil, J.K., Kumar, R.: Color feature extraction of tomato leaf diseases. Int. J. Eng. Trends Technol. **2**(2), 72–74 (2011)
19. Ghadge, R., Kulkarni, J., More, P., Nene, S., Priya, R.L.: Prediction of crop yield using machine learning. Int. Res. J. Eng. Technol. (IRJET), **5** (2018)
20. Hong, Z., Kalbarczyk, Z., Iyer, R.K.: A data-driven approach to soil moisture collection and prediction. In: 2016 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 1–6. IEEE (2016)
21. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. **7**, 1419 (2016)
22. [Cover, T.M. (1968) "Rates of convergence for nearest neighbor procedures", In Proceedings of the Hawaii International Conference on System Sciences, Univ. Hawaii Press, Honolulu, 413–415.
23. Cover, T.M. & Hart, P.E. (1967) "Nearest neighbor pattern classification", IEEE Trans. Inf. Theory, 13: 21–27.
24. Devroye, L. (1981) "On the asymptotic probability of error in nonparametric discrimination", Ann. Statist, 9: 1320–1327.
25. Devroye, L. (1981) "On the equality of Cover and Hart in nearest neighbor discrimination", IEEE Trans. Pattern Anal. Mach. lntell. 3: 75- 78.
26. Devroye, L., Gyorfi, L., Krzyzak, A. & Lugosi, G. (1994) "On the strong universal consistency of nearest neighbor regression function estimates", Ann. Statist, 22: 1371– 1385.
27. Devroye, L. & Wagner, T.J. (1977) "The strong uniform consistency of nearest neighbor density estimates", Ann. Statist., 5: 536–540.
28. Devroye, L. & Wagner, T.J. (1982) "Nearest neighbor methods in discrimination, In Classification, Pattern Recognition and Reduction of Dimensionality", Handbook of Statistics, 2: 193–197. North-Holland, Amsterdam.
29. Domeniconi, C., Peng, J. & Gunopulos, D. (2002) "Locally adaptive metric nearestneighbor classification", IEEE Transactions on Pattern Analysis and Machine Intelligence. 24(9): 1281–1285.
30. Dudani, S.A. (1976) "The distance-weighted k-nearest neighbor rule", IEEE Transactions on System, Man, and Cybernetics, 6: 325-327.
31. Eldestein, H.A. (1999) "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation, USA, ISBN: 1-892095- 02-5.
32. Enas, G.G. & Choi, S.C. (1986) "Choice the smoothing parameter and efficiency of KNearest Neighbor classification", Comp & Maths with Apps, 12(2): 235-244.
33. Fix, E. & Hodges, J.L. (1951) "Nonparametric Discrimination: Consistency Properties", Randolph Field, Texas, Project 21-49-004, Report No. 4
34. Vapnik, V.N.: Statistical learning theory. Springer, New York (1995)