Robust crop recommender system using Random Forest and CNN

[1]Charukeerthi N Bhavikatti, [2]Adithya S Iyer, [3]Akshay Kumar T, [4] J B Abdul Azeez Ur Rahaman

[1]charukeerthi.nb@gmail.com, [2]adithyas20@gmail.com, [3]akshayklr1999@gmail.com, [4]jbazeez1999@gmail.com

[5]Chaitra H K, Asst Professor, SJB Institute of Technology, chaitrahk@sjbit.edu.in

Abstract — **India's economy is dominated by agriculture. The majority of its population relies on agriculture for their livelihood either explicitly or implicitly. Today the population is growing tremendously and with this, the demand for food is also increasing. Therefore, it is now time to introduce advanced farming methods. The farmers are not choosing the right crops for cultivation, which is a major and serious hindrance to crop productivity. To improve crop productivity, a crop recommendation system is to be developed that uses machine learning techniques for recommending the optimal solution. In this paper, we present the data preprocessing techniques for the recommendation system of crops, fertilizers, and the prediction of plant diseases. In addition to soil composition and environmental factors such as temperature and rainfall, the model intends to assist Indian farmers and the agriculture department in making informed decisions about which crops to grow.**

Keywords* — *Crop recommendation, Fertilizer recommendation, Plant disease prediction, Machine learning, Random Forest, CNN, Deep learning.

I. Introduction

Agriculture has an extensive history in India. Recently, India is ranked second in the farm output worldwide. Agriculture-related industries such as forestry and fisheries contributed 16.6% of 2009 GDP and around 50% of the total workforce [1]. In agriculture, crop yields are very much dependent on natural factors and are highly uncertain. We cannot expand the earth to create more land for agriculture i.e., we have to "Do More Within Less". Therefore, AI, Machine learning, and Deep learning techniques enable us to make effective cultivation [2].

The reason behind this is that the farmers often take the wrong decision about crop selection. For example, selecting a crop that won't give much yield for the particular soil, planting in the wrong season, and so on. The farmer might have purchased the land from others so without previous experience the decision might have been taken. Wrong crop selection will always result in less yield. If the family is fully dependent on this income, then it's very difficult to survive [3].

In the proposed system the environmental parameters such as rainfall, temperature, and geographical location in terms of the state along with soil characteristics such as soil type, pH value, and nutrients concentration are being considered to recommend a suitable crop and fertilizer to the user. In addition to this, the plant disease prediction provides the appropriate prevention and cure of the plant disease.

II. LITERATURE REVIEW

More and more researchers have begun to identify this problem in Indian agriculture and are increasingly dedicating their time and efforts to help alleviate the issue.

Shilpa Mangesh Pande and et al., highlighted the limitations of current systems and their practical usage on yield prediction and proposed a recommendation system. A proposed system provides connectivity to farmers via a mobile application to demonstrate a viable yield prediction system. The mobile application includes multiple features that users can leverage for the selection of a crop [1].

Narayani Patil and et al., proposed a system that provides a solution to the uncertainty in the biological nature of agriculture by assisting framers to reduce uncertainty in farming and improving productivity [2].

Priyadharshini.A and et al., proposed a system that helps the farmers to choose the right crop by providing insights that ordinary farmers don't keep track of thereby decreasing the chances of crop failure and increasing productivity. It also prevents them from incurring losses [3].

Vaishnavi.S and et al., highlighted the significance of management of crops was studied vastly. Recent technology can assist farmers in growing their crops.

DOI : 10.35291/2454-9150.2022.0181

Proper prediction of crops can be informed to agriculturists on a time basis. Agriculture parameters have been analyzed using various Machine Learning techniques [4].

A.M.Rajeswari and et al., proposed a method, the fuzzy-based rough set approach is implemented to help the farmers in deciding on crop selection in their agricultural land. The proposed method is tested for twenty-four different crops [5].

Aoqi Liu and et al., proposed a minority synthesis optimized algorithm that achieves excellent performance on soil data analysis and crop recommendation. Compared with the case without C-SMOTE processing, the algorithm accuracy grows by 27%. The accuracy, precision, and f1-score of the model are 98.7%, 97.4%, and 97.8% respectively. The disadvantage is that the RBF kernel is easy to over-fitting when the parameters are inappropriate [6].

Apporva Chaudhari and et al., proposed a single platform for crop recommendation and its optimal pricing. The crop is recommended using data mining techniques focusing on the essential parameters that affect the crop's cultivation. Dedicated web-scrapers are developed for finding an optimal website for the crop seeds to be purchased. This system would prove beneficial for farmers to purchase products for their farms [7].

Nidhi H Kulkarni and et al., proposed a crop recommendation system where a soil dataset has been included which takes into account four crops, Rice, Cotton, Sugarcane, and Wheat. An Ensembling technique is used to classify the four crops after a soil dataset is preprocessed. The individual base learners used in the ensemble model are Random Forest, Naive Bayes, and Linear SVM [8].

Zeel Doshi and et al., proposed and implemented an intelligent crop recommendation system, which can be easily used by farmers all over India. This system would assist the farmers in making an informed decision about which crop to grow depending on a variety of environmental and geographical factors. The high accuracies provided by both these models make them very efficient for all practical and real-time purposes [9].

Jenshkie Jerlin I. Haban and et al., implemented the fuzzy logic system which was successfully developed and simulated to give appropriate fertilizer recommendations. The input parameters of the fuzzy system include the level of nitrogen, phosphorus, and potassium, as well as the season. Different fertilizer combinations are produced depending on the input parameters selected [10].

Akshai K.P and et al., proposed a system that facilitates the early diagnosis of plant diseases to prevent crop loss and the spread of diseases. Several pre-trained CNN models are evaluated for their accuracy in predicting different plant diseases, including VGG, ResNet, and DenseNet, and then based on performance metrics, the DenseNet model is found to be more accurate. With 98.27% accuracy, the DenseNet model was the best model [11].

Rubini P.E and et al., proposed a machine learning model to determine whether a plant is healthy or diseased. The accuracy achieved with this model is satisfactory. To improve this model, we can adopt other machine learning algorithms and try to obtain a more efficient classifier. VGG-16's main flaw is that it has nodes that are fully connected and has a file size of over 533MB. This makes deploying VGG a tiresome task. The DenseNet architecture proves better accuracy than VGG16 because of its more diversified features [12].

Ms. Deepa and et al., proposed a method that used machine-learning techniques to detect leaf diseases in plants. Farmers do not have easy access to expert advice. It is beneficial to have an automated mechanism to identify plant diseases [13].

Jithy Lijo and has done a comparative study on the effectiveness of augmentation techniques on various transfer learning techniques. The study provided insight into the effectiveness of transfer learning methods and the results showed the technique to be highly effective when proper optimization and augmentation techniques are done. The result of the study proved that this ResNet50 transfer learning model can be used to classify healthy and unhealthy leaves in the data set with high accuracy [14].

III. METHODOLOGY

To eliminate the aforementioned drawbacks, we propose a system that takes into consideration all the appropriate parameters, including temperature, rainfall, location, and soil condition, to recommend crop, fertilizer, and predict plant disease. The various preprocessing methods are represented in the below figure. They are data cleaning, data normalization, data transformation, missing values imputation, data integration, noise identification.

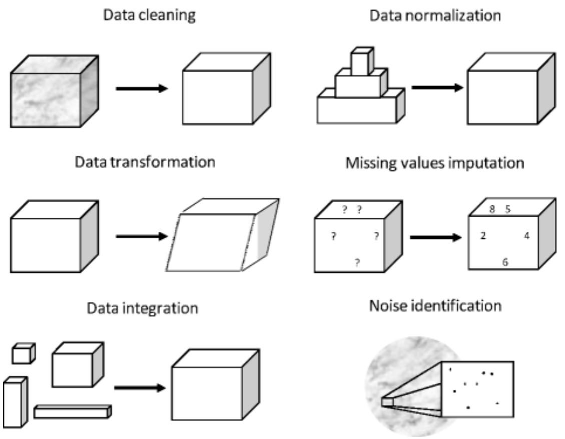


Fig 1: Methods of data preprocessing

DOI : 10.35291/2454-9150.2022.0181

3.1 Crop recommendation system

This sub-system is fundamentally concerned with performing the primary function of the system which is, providing crop recommendations to farmers. The steps involved in this sub-system are:

3.1.1. Acquisition of training dataset

The accuracy of any machine learning algorithm depends on the number of parameters and the correctness of the training dataset. For the first sub-system, we have made use of the 'Crop Recommendation Dataset' from Kaggle.

This dataset consists of records of soil and natural parameters, which were accumulated over the ten years (from 1997-98 to 2006-17). It consists of a total of 2200 entries. We provide the aforementioned parameters for 22 crops in total.

The schema of the training dataset is as follows:

- Soil composition: N values ranging from (0<N<140), P values ranging from (5<P<145), and K values ranging from (5<K<205).
- Temperature: Month-wise temperature (8.83<°C<43.7).
- Humidity: Month-wise Humidity (14.3<Humidity<100).
- Soil pH: Ranging from (3.5<pH<9.94).
- Rainfall: Month-wise rainfall (20.2<mm<299).

3.1.2. Data preprocessing

In Machine Learning, data preprocessing is an important step, since the quality of data and the amount and type of information it contains directly influence how well our model can learn. The input dataset is subject to various preprocessing techniques such as filling of missing values, encoding of categorical data, and scaling of values in the appropriate range.

The first step is to remove the missing values which were represented by a white space (' ') in the original dataset. By having these missing values, the data would be less valuable, which would subsequently hinder the performance of machine learning models. Hence, to deal with these missing values, we replace them with large negative values, which the trained model can easily treat as outliers.

To be able to apply machine learning algorithms to the data, we must first create class labels. Since we intend to use supervised learning, class labels are essential. The original dataset did not come with labels, and hence we had to create them during the second step data preprocessing phase.

The third step is that the values in the dataset are in string format. To pass the input to the machine learning model, this should be converted into integer values. Further to reduce the amount of data going into the random forest model the crops are being filtered based on the required nutrients present in the soil. If the nutrient content of the soil is below that required by the crops, then the respective crop will be discarded. This has resulted in a considerable reduction in training time.

The fourth step is that our dataset was originally quite large and imbalanced. If a dataset is not balanced this may lead to the problem of overfitting data for the categories whose number of images is large relative to others. Our solution has been to use data augmentation to preprocess the data.

The data augmentation process generates slightly different results for each data set that is processed by the model. The process prevents the model from learning the characteristics of each data set.

The last part of the process is dividing the dataset into training and testing. We will divide the dataset in a proportion of 4:1 for training and testing.

3.1.3. Random Forest Algorithm

The reason Random Forest algorithm is used in this process is due to the reason this algorithm had 95% accuracy in the reference paper [1]. This algorithm is very accurate since we are integrating multiple decision trees in a single tree. The dataset is subjected to train test split in 4:1 ratio. This ratio means 80% of the rows are subjected to training and the rest 20% rows are subjected to testing. The we are going to trading the model using the Random Forest algorithm which is in-built in sklearn python module. Using this model, we can predict the crops to recommend for the user.

3.2 Sub-system 2: Fertilizer recommendation system

A commercial fertilizer typically has three major nutrients, nitrogen, phosphorus, and potassium, each of which contributes to plant nutrition.

Nitrogen is the first key nutrient of commercial fertilizers since it is absorbed by plants more than any other element. Nitrogen is important in making sure plants are healthy while they grow and for making sure they are nutritious after they are harvested. This is due to the fact that nitrogen is necessary for the formation of protein, which is the basis for nearly all living tissue.

Phosphorus is the second key nutrient of commercial fertilizers and it plays an important role in a plant's ability to use and store energy, including photosynthesis; phosphorus is also required for normal growth and

DOI : 10.35291/2454-9150.2022.0181

development. In commercial fertilizers, phosphate rock is used as a source of phosphorus.

Potassium is the third key nutrient of commercial fertilizers. In addition to strengthening plants' resistance to disease, it plays a vital role in improving crop yields and quality. The potassium found in plants protects them from cold and dry weather conditions by strengthening their root systems and preventing wilt.

Thepredicted output of this subsystem can then be fed to sub-system 1 for the prediction of crop suitability. The steps involved are:

3.2.1. *Acquisition of training dataset*

For this sub-system, we used the soil composition dataset provided by Kaggle called 'Fertilizer recommendation dataset'.

This dataset consists of records of soil and natural parameters, which were accumulated over the five years (from 2012 to 2017).

3.2.2. *Data preprocessing*

In some cases, a particular element may not be present because of corrupt data, incomplete information, or inability to load the information. Making the correct decision on how to handle missing values is one of the most difficult challenges analysts face because robust data models are derived from the right decision.

Similar to the data pre-processing step done for sub-system 1, here the missing values are eliminated by being replaced with large negative values (-9999).

The overfitting of a model makes it relevant to just one dataset and irrelevant to all other datasets. Ensembling is a machine learning method that is used to combine predictions from two or more separate models.

Boosting increases aggregate complexity by using simple base models. This method trains many weak learners in a sequence so that each learner in the series learns from the previous learner's mistakes. Using bagging, many strong learner pairs are trained in parallel and are then combined to improve prediction accuracy.

Our categorical data contains a fertilizer column that is in a string format. Therefore, we are going to give incremental values to each fertilizer. So that we can predict and determine which fertilizers are commonly used for a particular crop in that location.

We will split the data in the ratio 4:1 as we did for the pre-processing in the previous subsystem. Finally, we will split the dataset into training and testing.

3.2.3. *Random Forest Algorithm*

We can use the same method which was used in the previous subsystem for the fertilizer dataset. So Random Forest algorithm is used in this process as well. This algorithm is very accurate since we are integrating multiple decision trees in a single tree. The dataset is subjected to train test split in 4:1 ratio. This ratio means 80% of the rows are subjected to training and the rest 20% rows are subjected to testing. The we are going to trading the model using the Random Forest algorithm which is in-built in sklearn python module. Using this model, we can predict the fertilizers to recommend for the users.

3.3 Sub-system 3: Plant disease detection

This sub-system is fundamentally concerned with performing the function of providing the plant disease prediction for the farmers.

3.3.1. *Acquisition of training dataset*

For this sub-system, we used the plant disease dataset provided by Kaggle. There are 87,000 RGB images of healthy and diseased crop leaves in this training dataset, categorized into 38 different classes. An additional directory containing 33 test images is included for prediction purposes.

3.3.2. *Data preprocessing*

The preprocessing techniques are applied to remove any noise enhancements to the images. The images are preprocessed using contrast enhancement. It enhances image features where the contrast of the image is increased by mapping input intensity to the new value.

The raw image taken from the database had gone through preprocessing before being fed into the CNN model. The images are reconstructed and normalized to establish a base size for all images and remove noise. An image is conceptualized as a three-dimensional vector of P, Q, and R, where the width and height of an image are represented by P and Q, respectively, and the RGB channels are represented by R. The images obtained after pre-processing is shown in Fig. 2

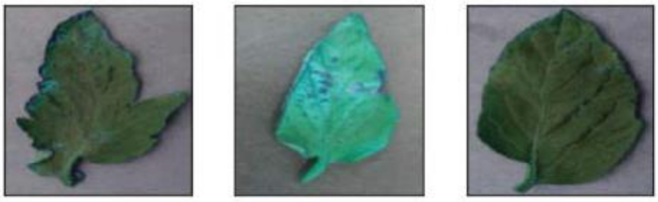


Fig 2: Preprocessing of the images

The convolutional neural network requires inputs of fixed size so that all the images in the data set are resized to 256x256 which is a suitable size that avoids loss of any

DOI : 10.35291/2454-9150.2022.0181

information from the image. If the fixed size is large enough then shrinking is not required so that the features can be preserved, and this will increase the accuracy of the classification. But if the size of the image is too large it will increase the time and space complexity also an appropriate size 256x 256 is selected for resizing in this paper. Another major preprocessing technique used in our paper is image augmentation which is explained in the coming section.

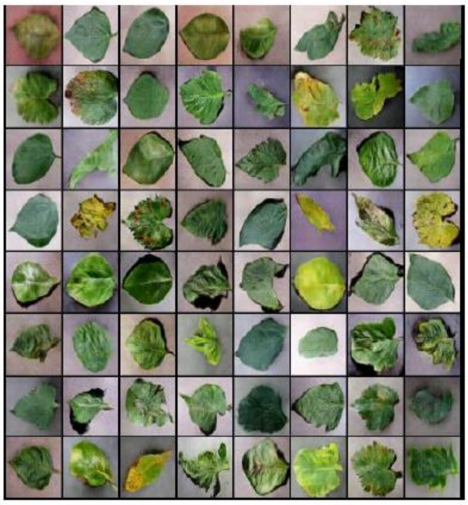


Fig 3: Sample dataset used

Overfitting is a major problem in deep learning techniques which indirectly reduce the accuracy of classification, to overcome this problem we had done data set augmentation. The 80 percent of data that is used for training is augmented with the help of various augmentation techniques after the splitting process. It is done with the help of the Keras library in Python which is used for implementing deep learning. Augmentation techniques that are commonly used include image rotation, image noise reduction, image contrast enhancement, and image brightness enhancement. The degree of rotation is between 0 to 45 and it is chosen randomly with help of augmentation the number of images increased between 1500 to 3500 in each class and it helped to achieve a balance in the distribution of samples. As a result, the model can also be avoided from overfitting. The augmentation technique applied is illustrated in Figure 4

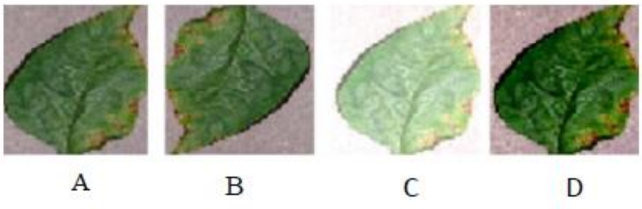


Fig 4: Augmentation

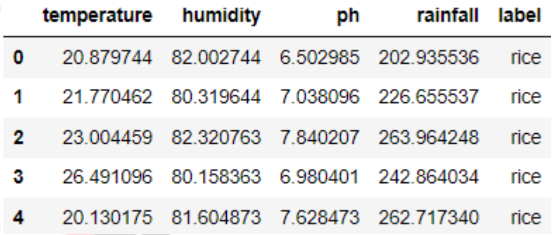
a) Original b) Rotation c) Noising d) Contrast

3.3.3. Convolution Neural Network

The reason Convolution Neural Network algorithm is used in this process is due to the reason this algorithm had 94.58% accuracy in the reference paper [11]. This algorithm is very accurate since the parameters is reduced and it is easy to perform computations. The dataset is subjected to train test split in 4:1 ratio. This ratio means 80% of the rows are subjected to training and the rest 20% rows are subjected to testing. The we are going to trading the model using the Convolution Neural Network (CNN) algorithm which is in-built in TensorFlow python module. Using this model, we can detect the plant disease in the image for the users.

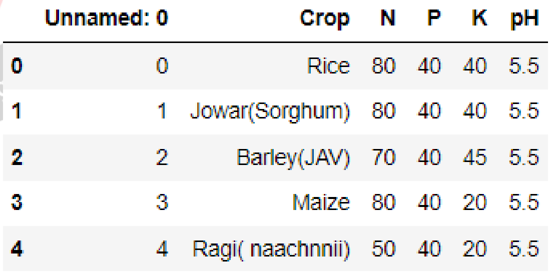
IV. RESULTS

The preprocessing had been applied to the datasets of the crops, fertilizers datasets, and the plant disease images. The below 2 figures shows the first 5 rows of the crop recommendation data set and fertilizer recommendation data set respectively. The first data set includes features such as temperature, humidity, pH, rainfall, and a label for the crops. The second data set includes features such as unnamed, N, P, K, pH, and a feature for the crops.



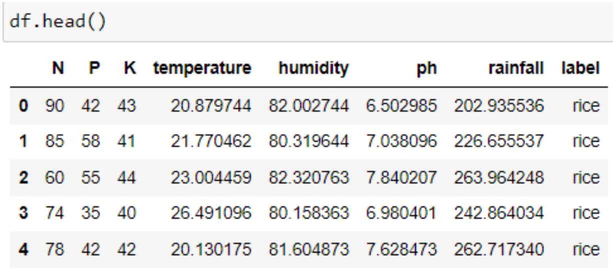
	temperature	humidity	ph	rainfall	label
0	20.879744	82.002744	6.502985	202.935536	rice
1	21.770462	80.319644	7.038096	226.655537	rice
2	23.004459	82.320763	7.840207	263.964248	rice
3	26.491096	80.158363	6.980401	242.864034	rice
4	20.130175	81.604873	7.628473	262.717340	rice

Fig 5: Dataset of crop recommendation before preprocessing



	Unnamed: 0	Crop	N	P	K	pH
0	0	Rice	80	40	40	5.5
1	1	Jowar(Sorghum)	80	40	40	5.5
2	2	Barley(JAV)	70	40	45	5.5
3	3	Maize	80	40	20	5.5
4	4	Ragi(naachnnii)	50	40	20	5.5

Fig 6: Dataset of fertilizer recommendation before preprocessing



```
df.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Fig 7: Dataset of crop and fertilizer recommendation after preprocessing

Figure 5 shows the crop recommendation system dataset. The dataset consists of many null values and is also

DOI : 10.35291/2454-9150.2022.0181

missing important features that are required for the recommendation of crops to the users which are N, P, and K values. Fig. 6 shows the fertilizer recommendation system dataset. The dataset consists of unwanted column names Unnamed: 0 which is to be removed and the names of the crops are not accurate since it contains the regional name of the crop mentioned within the parenthesis.

Figure 7 shows the final dataset after preprocessing for the crop and fertilizer recommendation system. In the final dataset, the data integration of the initial crop and fertilizer datasets has occurred. The unwanted column Unnamed: 0 is dropped using the Pandas library. The labels of crops have been preprocessed and a single label has been used from both datasets. All the null values have been replaced by the mean value of the feature which is known as the imputation of the missing values.

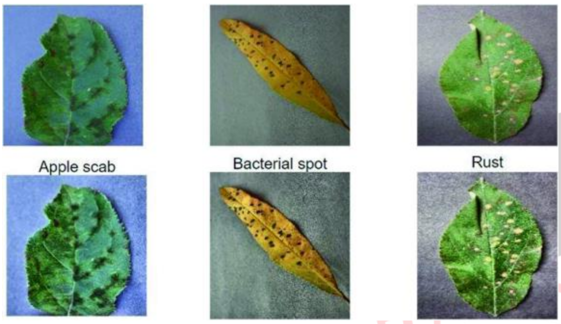


Fig 8: Dataset of plant diseases during preprocessing

Figure 8 shows us the images before and after preprocessing. The above 3 images show us the images which contain noise and other impurities which makes it hard for the model to classify the disease. Since the computer only understands the data in RGB format we need to transform it into numerical data. So, we are going to preprocess the images by image scaling, contrast enhancement, color space transformation, restoration from noise, and restoration from blur. Thus, the below 3 images represent the preprocessed images.

Thus, using the preprocessed datasets we have built a system to recommend the crops and fertilizers. In addition to this, the system can predict the diseases in plants.

V. CONCLUSION

In this paper, we have successfully proposed the preprocessing and machine learning methods for intelligent crop recommendation, fertilizer recommendation, and plant disease prediction which can be easily used by farmers and agriculture departments all over India. Using this system, farmers would be able to make informed decisions based on the conditions of the environment and the geography of their area.

The accuracies of Random Forest for crop recommendation and CNN for plant disease prediction are in the below table:

Random Forest	99.0909 %
Convolution Neural Network	99.2000 %

Table 1: Comparison of accuracies of Random Forest and CNN algorithm

The preprocessing of the model proposed in this paper can be further extended in the future by incorporating different datasets to increase the accuracy of the model. This would help in predicting the correct crop to grow, fertilizer to use and steps to take in order to cure the plant disease. Increasing the system's accuracy further will require the many types of machine learning algorithms to be used in the system.

ACKNOWLEDGEMENT

This research is supported by the SJB Institute of Technology, Bangalore, India through a final year project from the computer science department.

REFERENCES

[1] Shilpa Mangesh Pande, Dr. Prem Kumar Ramesh, Anmol, B.R Aishwaraya, Karuna Rohilla & Kumar Shaurya. (2021). *"Crop Recommender System Using Machine Learning Approach"*. Paper presented at 5th International Conference on Computing Methodologies and Communication (ICCMC).

[2] Narayani Patil, Shubham Kelkar, Mitali Ranawat & Dr. M. Vijayalakshmi. (2021). *"Krushi Sahyog: Plant disease identification and Crop recommendation using Artificial Intelligence"*. Paper presented at 2nd International Conference for Emerging Technology (INCET).

[3] Priyadharshini A, Swapneel Chakraborty, Aayush Kumar & Omen Rajendra Pooniwala. (2021). *"Intelligent Crop Recommendation System using Machine Learning"*. Paper presented at 5th International Conference on Computing Methodologies and Communication (ICCMC).

[4] Vaishnavi.S, Shobana.M, Sabitha.R & Karthik.S. (2021). *"Agricultural Crop Recommendations based on Productivity and Season"*. Paper presented at 7th International Conference on Advanced Computing & Communication Systems (ICACCS).

[5] A.M.Rajeswari, A.Selva Anushiya, K.Seyad Ali Fathima, S.Shanmuga Priya & N. Mathumithaa. (2020). *"Fuzzy Decision Support System for Recommendation of Crop Cultivation based on Soil Type"*. Paper presented at 4th International Conference on Trends in Electronics and Informatics (ICOEI).

DOI : 10.35291/2454-9150.2022.0181

[6] Aoqi Liu, Tao Lu, Bufan Wang & Chong Chen. (2020). *"Crop Recommendation via Clustering Center Optimized Algorithm for Imbalanced Soil Data"*. Paper presented at 5th International Conference on Control, Robotics and Cybernetics.

[7] Apoorva Chaudhari, Manasi Beldar, Riya Dichwalkar & Surekha Dholay. (2020). *"Crop Recommendation and its Optimal Pricing using ShopBot"*. Paper presented at International Conference on Smart Electronics and Communication (ICOSEC).

[8] Nidhi H Kulkarni, Dr. G N Srinivasan, Dr. B M Sagar & Dr.N K Cauvery. (2018). *"Improving Crop Productivity Through a Crop Recommendation System Using Ensembling Technique"*. Paper presented at 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions.

[9] Zeel Doshi, Subhash Nadkarni, Rashi Agrawal & Prof. Neepa Shah. (2018). *"AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms"*. Paper presented at 4th International Conference on Computing Communication Control and Automation (ICCUBEA)

[10] Jenskie Jerlin I. Haban, John Carlo V. Puno, Argel A. Bandala, Robert Kerwin Billones, Elmer P. Dadios & Edwin Sybingco. (2020). *"Soil Fertilizer Recommendation System using Fuzzy Logic"*. Paper presented at IEEE Region 10 Conference (TENCON)

[11] Akshai KP & J.Anitha. (2021). *"Plant disease classification using deep learning"*. Paper presented at 3rd International Conference on Signal Processing and Communication (ICPSC)

[12] Rubini PE & Dr.Kavitha P. (2021). *"Deep Learning model for early prediction of plant disease"*. Paper presented at 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)

[13] Ms. Deepa, Ms. Rashmi N & Ms. Chinmai Shetty. (2021). *"A Machine Learning Technique for Identification of Plant Diseases in Leaves"*. Paper presented at 6th International Conference on Inventive Computation Technologies (ICICT)

[14] Jithy Lijo. (2021). *"Analysis of Effectiveness of Augmentation in Plant Disease Prediction using Deep Learning"*. 5th International Conference on Computing Methodologies and Communication (ICCMC)

DOI : 10.35291/2454-9150.2022.0181