Prediction of Soil pH using Smartphone based Digital Image Processing and Prediction Algorithm

Article · April 2019

DOI: 10.26782/jmcms.2019.04.00019

1 author:

Utpal Barman
GIMT Guwahati

ISSN (Online) : 2454 -7190 **ISSN (Print) 0973-8975**

Prediction of Soil pH using Smartphone based Digital Image Processing and Prediction Algorithm

[1]Utpal Barman, [2]Ridip Dev Choudhury

[1]Department of Information Technology, Gauhati University, Assam, India

[2]Departent of Computer Science and Information Technology, IDOL, Gauhati University, Assam, India

[1]utpalbelsor@gmail.com, [2]rdc@gauhati.ac.in

Corresponding Author: Utpal Barman

Email: utpalbelsor@gmail.com

https://doi.org/10.26782/jmcms.2019.04.00019

Abstract

Soil pH is one of the major factors to be considered before doing any cultivation. Farmers always tested their soil pH either in soil pH laboratory, soil pH color chart or sometimes with the help of an expert. But these methods need time, labor and expertness. In this paper, a digital Smartphone image-based method is presented which predicts the soil pH in a simple and accurate way. Soil images are captured with the help of Redmi 3S prime Smartphone and store all the images as soil dataset. Soil images are processed through the different steps of digital image processing including soil image enhancement, soil image segmentation, and soil image feature extraction. During the feature extraction, Hue, Saturation and Value of the soil image are calculated and store Saturation and Hue plus Saturation as an index for the feature vector of the soil images. Prediction of soil pH is done with the help of Linear Regression, Artificial Neural Network, and KNN Regression. The coefficient of the linear regression is 0.859 for the Saturation feature of the soil image. Again, the coefficient of linear regression is 0.823 for Hue plus Saturation. The regression coefficient for ANN is 0.94064 with Levenberg-Marquardt algorithm and 0.92932 with Scaled Conjugate Gradient Backpropagation Algorithm. The regression coefficient of KNN is 0.89326 for K=5 with an RMSE value 0.1311. It is found that ANN always gives a better result as compare to another one.

Keywords : Soil pH, K Mean, HSV, Linear Regression, KNN, ANN

I. Introduction

India is very much rich in cultivation. In recent times, Image processing techniques are used in different areas of research including health and agriculture. But the application of image processing techniques in the agriculture filed is limited and especially in Assam, India. The main factor of any farming is soil. The Soil has different properties. Among the different physical and chemical properties of soil, soil's chemical properties always play an important role in farming. Soil pH is one of the major factors to be considered before doing any cultivation. A soil with 5.5 to 7.0 pH level is always good for cultivation. A soil with 5.5 to 7.0 pH level is always good for cultivation [VI] Farmers always tested their soil pH either in soil pH laboratory or using soil pH color charts. Sometimes, an expert helps farmers to determine the soil pH. But getting expert views in all times is not possible for all the times. Again all these methods need time, labor and expertness. Soil pH determination using soil pH chart is not an adequate process as it takes human perception and needs expertness. This paper presented a digital image based system which will accurately predict the pH of the soil with the help of different machine learning approaches including linear regression, ANN and KNN. In Assam, most of the farmers are from rural areas. So keeping in mind, soil images area captured using Smartphone and then images are processed through the different image processing steps for soil pH Prediction.

In recent years, many traditional and nontraditional methods are developed for the Prediction of soil pH using image processing but those methods are time-consuming and more expensive. But the computer process helps the human being to make a decision in an error-free manner [VII] Although, many researches are done for the prediction of different soil properties using machine learning approaches including digital image processing, linear regression, ANN, and digital pH meter but the application of machine learning in limited for soil pH. [I, III, V, VI, XI, XIII, XV] predicated the soil pH from the RGB color space of the soil images. They collected the soil images using a digital camera and predicted the soil pH using the following feature.

$$\text{Feature of Soil (pH Index)} = \text{Red/Green/Blue} \qquad (1)$$

Abu et al. [II] forwarded an expert system for controlling soil pH using fuzzy logic. During the process, they altered the level of soil pH so that farmer can replace the fertilizer and ensure a good quality of the plant. In the same year, Prediction of the paper microfluidic device pH is done by Ruiz et al. [XIX] with the help of a Smartphone. They captured the images by using a Smartphone and Open CV. Features of the images are extracted using the Hue (H) and Saturation (S) and related them to the pH of the devices. Babu et al. [V] presented a method which displays the physical properties of soil. They used fractal dimension technology using LabView and the equation of fractal dimension is presented as follows

$$\text{Fractal Dimension} = \log (X(s))/\log (1/s) \qquad (2)$$

They used 24-bit input color image in their model and converted those images to 8-bit images. For feature extraction, they used the equation as suggested by Kumar et al. [XIII]. Aziz et al. [I] used the same Red, Green and Blue values of images as given by Kumar et al. [XIII] and feed those values in the neural network for training and testing purpose. They used 10 hidden neurons in their hidden layer and got 80 % accuracy for their model. Gurubasava and Mahantesh S.D [XI] predicated soil pH from the mean values of the RGB values of the soil images. They compared the mean values of RGB with the actual soil pH and predicate it. Barman et al. [VI] presented a soil pH

Prediction methods using HSV color image processing. They computed the Hue, saturation, and value of the soil images and predicated the soil pH using Liner Regression, Logarithmic regression, E exponential regression and Quadratic regression. Sagar et al. [XX] presented a soil pH Prediction approach where images are captured with the help of a camera and raspberry pi and soil pH is determined using after processing the captured image.

Apart from soil pH, machine learning approaches are used to determine the other soil properties including soil moisture [XIV, XVI, XX, XXIII], prediction of Azotobacteria population in soil [X], soil mapping [VIII] and soil organic matter [IV,XV]. In these papers, authors presented the relation of soil properties using ANN and regression.

So from literature, it is come to know that machine learning approaches can be used to predict the soil different properties including soil pH. In this paper, the presented model predicts the soil pH. In the study, valuable works are found for soil pH prediction. But, most of the works are not in the aspect of rural farmers as images are captured using digital camera [XIII]. In this study, samples are captured with the help of a Smartphone which can reduce the cost of the system and determine the chemical characteristics of soil easily. But capturing soil sample using the Smartphone is the main challenging task for soil pH Prediction because the natural light can directly affect the quality of the soil sample. Taking all these factors, we develop our system in such a way that a rural farmer can easily capture a soil sample and use our system for actual soil classification Analyze and Prediction of soil pH is done using different machine learning Prediction algorithm including regression, ANN and KNN.

The rest of the sections are organized as follows: Section II covers the materials and methods; Section III gives result discussion and analysis; Section IV contains conclusion.

II. Materials and Methods

Soil Sample Collection Site

Soil sample collection site is present in Assam India. Total of 120 soil samples is collected from the West Guwahati, Assam. Samples are collected with the help of the ddepartment of Civil Engineering, GIMT, Guwahati. The sample collection site is in the latitude of 26.1445° N and in the longitude of 91.7362. For all 120 samples, 12 paddy fields are selected and 10 images per paddy fields are collected. The distance separation of two samples is 200 meter in each paddy field. Collecting samples are dragged out in depth of 0.5 feet in a vertical direction from the top position of the soil. All total of 120 soil samples is placed in white paper for image acquisition. Images are captured with the help of Smartphone and then soil samples are processed for pH calculation in Environmental laboratory of GIMT Guwahati. Fig. 1 shows the process the sample collection.

Fig. 1. Area for Sample collection [XIII].

Image Acquisition

A Smartphone is a good device to capture the agriculture image [VII, XVIII]. Soil image acquisition is a procedure for capturing different images with the help of digital image capturing devices [VII]. In most of the agriculture study, a digital camera is used to capture the sample images [XIII] but Smartphone is also used captured the agriculture images [XVIII]. Since the intention of doing this method is to predict the soil pH in a low-cost manner, a Smartphone is used to capture the soil images. A Redmi 3S prime Smartphone is used to capture all 120 soil samples and stored it as soil dataset. The image capturing is done under visible light in a closed room where only one window is open to enter the natural light. This is done to reduce the light effect during the image capturing. During the capturing time, all the properties of the camera setting of Redmi 3S is set to default such as F-stop=f/2 and exposure time = 1/ 60 sec. Soil images are captured in the vertical position from the 0.5 feet from the soil images. Fig. 2 Shows the process the of soil capturing.

Fig. 2. a) Soil Sample Collection b) Soil Sample

Deterination of soil pH using soil pH meter

The soil pH can be calculated using digital soil pH meter or soil pH color chart [XXI]. But, soil pH chart is not a adequate process as it require human perception during the time of color analysis. The actual soil pH value of 120 soil samples are done in the Environmental Laboratory of Civil Engineering Department of GIMT, Guwahati. In the laboratory test, a digital soil pH meter is used to determine the soil pH. The calibration of soil pH meter is done using a buffer solution with 7 pH value. A 20grms of soil sample is weighted and transferred it to a 100 ml glass beaker. 40 ml distilled water is mixed with the soil solution and stir with a glass rod. The solution stands for half an hour. Finally, the electrode of soil pH meter is immersed in the solution and determined the soil pH. The pH of the soil sample is displayed in the digital display of the digital soil pH meter and then, all the pH values are tabulated. Fig. 3 Shows the process the of soil pH calculation.

Fig. 3. (a-c) Soil pH calculation at laboratory

Copyright reserved © J.Mech.Cont.& Math. Sci., Vol.-14, No.2, March-April (2019) pp 226-249

Table 1: Different pH values of soil Samples as pr the Paddy Fields.

Sample No	Paddy Field 1	Paddy Field 2	Paddy Field 3	Paddy Field 4	Paddy Field 5	Paddy Field 6	Paddy Field 7	Paddy Field 8	Paddy Field 9	Paddy Field 10	Paddy Field 11	Paddy Field 12
1	8.25	8.24	8.01	7.11	7.12	7.91	7.27	7.25	7.62	8.21	8.19	7.22
2	8.23	8.21	7.90	7.69	7.15	7.82	8.02	7.45	7.63	7.63	8.19	7.22
3	8.31	8.26	7.85	7.08	7.25	7.84	8.03	7.50	8.10	7.65	8.2	8.25
4	8.25	8.15	7.95	7.09	7.09	7.96	8.2	8.00	8.16	7.60	8.17	8.25
5	8.29	8.16	7.98	7.07	7.08	7.99	8.10	7.10	8.13	7.64	8.12	8.27
6	8.24	8.00	7.99	7.11	7.91	7.92	7.26	7.61	8.1	7.55	7.50	8.05
7	8.19	8.10	7.95	7.09	7.92	7.90	7.21	7.62	8.15	8.25	8.13	8.00
8	8.17	8.15	7.82	7.09	7.85	8.02	7.25	7.64	8.10	7.65	7.40	8.20
9	8.16	8.17	7.89	7.15	7.98	8.08	7.28	7.66	8.12	7.7	7.20	8.10
10	8.21	8.19	7.84	7.16	7.97	7.86	7.20	7.65	8.10	7.72	7.24	8.25
Mean	**8.23**	**8.163**	**7.918**	**7.164**	**7.532**	**7.93**	**7.582**	**7.548**	**8.021**	**7.76**	**7.834**	**7.981**

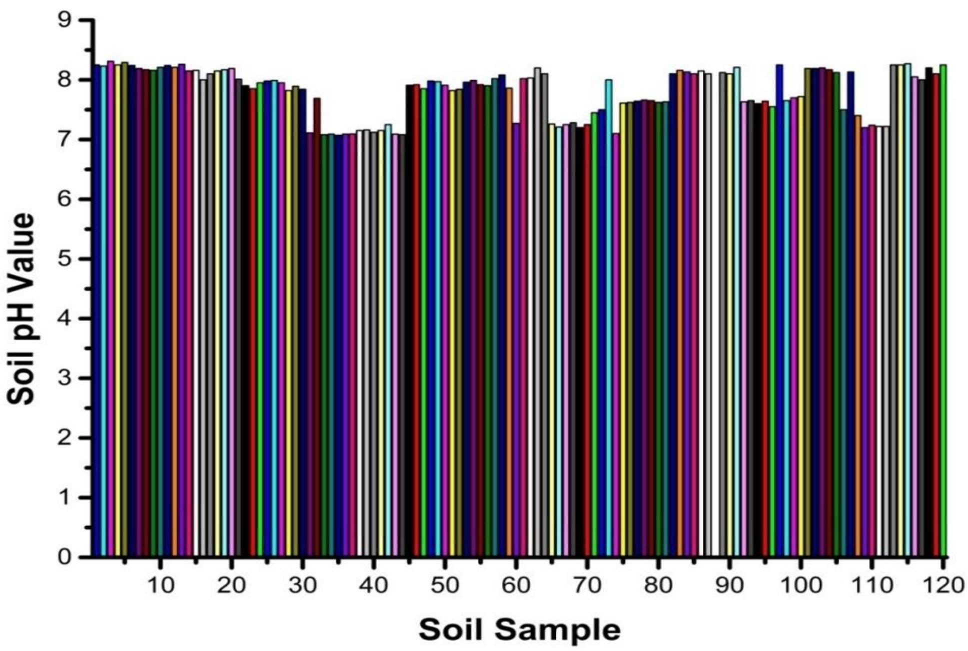


Fig. 4. Graphical value of soil pH values

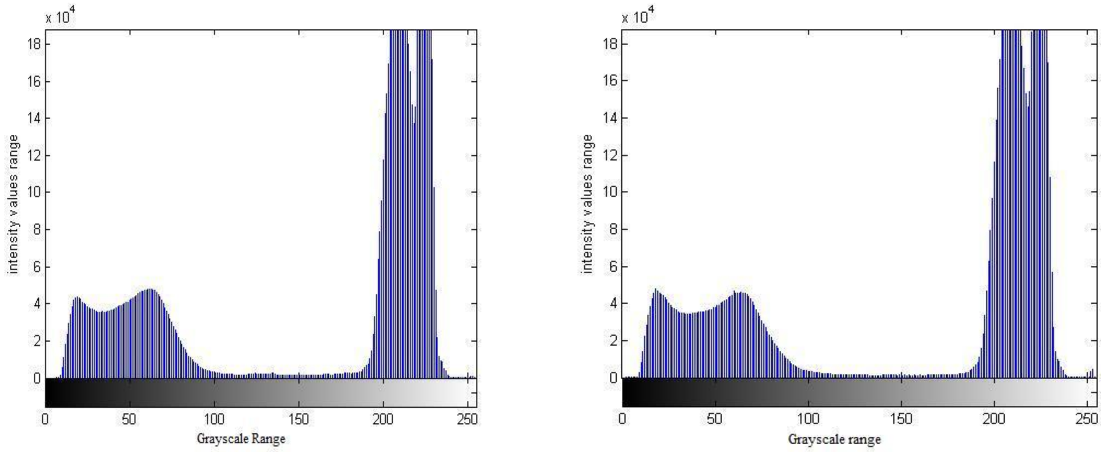
Copyright reserved © J.Mech.Cont.& Math. Sci., Vol.-14, No.2, March-April (2019) pp 226-249

Soil image Preprocessing, segmentation and feature extraction

Soil captured images are in 8-bit RGB mode and saved in jpeg format. All the images are in 4160 x 2340 dimensions with 72 dpi. After this, images are processed through the different image processing steps including image filtering, image segmentation, and feature extraction.

Image Filtering: As all the captured images are in high in dimension, images are resized to 300 x 400 dimensions. After that, Image filtering is used to filter the images. Image filtering is the process of enhancing the quality of soil images using filtering and other image enhancement techniques. Fig 5. Shows the original image and its filter image along with its equivalent histogram.

(a) (b)



(c) (d)

Fig. 5. a) Original Soil Image b) Enhanced Soil Image c) Histogram of Original Soil Image d) Histogram of Enhanced Soil Image.

Segmentation using K-Mean and masking: Image segmentation is the process of segmenting the original image or filter image into different small segment so that the features of the images can be easily calculated. In this paper, K-Mean clustering is used to segment the filter image into 2 clusters. K mean is unsupervised clustering technique which is used for image segmentation. It divides the images into different cluster or segment [IX]. Algorithm 1 explains the process of image segmentation.

Algorithm 1: Image segmentation.

Input: Enhanced Image

Output: Segmented image

Step 1: Steps for K mean

1.1 Choose the cluster (K value) and in our case K=2

1.2 Calculate the mean of the cluster.

1.3 Use Euclidian distance to calculate the distance in between mean of cluster and image pixel.

1.4 Based on the distance, pixel is move to that cluster.

1.5 Again recalculate the mean.

1.6 Repeat the process until the mean is not changed.

Step 2: Steps for masking

2.1 Computed 2 Class Index and determine the pixel belongs to each class index and crop the need.

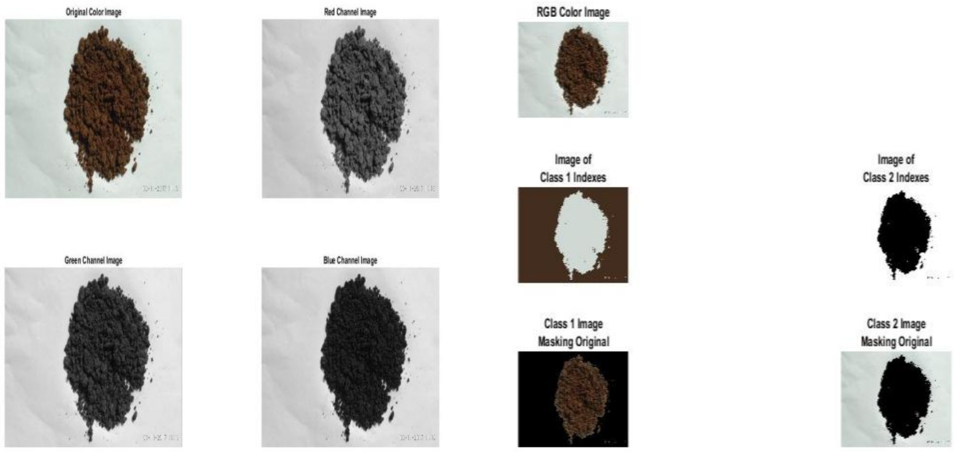
2.2 Compute an indexed image for comparison which contains at atleast 2 classes color

2.3 Display the classes, both binary and masking the original

2.4 Mask the image using MatLab function

2.5 Display masked image

In K-means clustering, images are partitions into different mutually exclusive clusters say K and based on each observation of the pixel data the accurate cluster is assigned. In our segmentation process, the number of classes is 2 and the clustering stages are shown in Fig 6. According to binary and masking the original.



(a) RGB components of original image (b) Image Masking and segmentation

Fig. 6. (a-b) Clustering Stages using K mean and masking.

Feature extraction of soil images: Feature extraction is the process of extracting different image features like color, texture, and shape. In this paper, the color image of soil is considered for feature extraction. Color is an important feature of an image and human can recognize any images using color. A human can determine the soil pH through naked eyes using soil pH chart but the process is not reliable as it is based on the color analysis of the soil images by a human. Color features of the soil images are extracted from the different soil images by using the following steps.

I. The mean Red, Green, and Blue (RGB) values for all the images are extracted and stored in m*n*3 matrix i.e. a 3D matrix by the following equation.

Mean_Red = double (original_image (:,:,1)); (3)

Mean_Green = double (original_image (:,:,2)); (4)

Mean_Blue = double (original_image (:,:,3)); (5)

II. The Red, Green and Blue (RGB) color space of all soil images are converted into Hue, Saturation and Value (HSV) color space by using Algorithm 2 and then, calculate the mean of Hue, Saturation and Value (HSV) using following equation.

Mean_Hue = double (hsv_image (:,:,1)); (6)

Mean_Saturation = double (hsv_image (:,:,2)); (7)

Mean_Blue = double (hsv_image (:,:,3)); (8)

III. After the calculation of RGB and HSV from each sample, we calculate two feature vectors as suggested by Kumar et al. [XIII] and Swapna et al. [XXII] and consider this index as a SI 1 and SI 2.

SI 1 = Mean_Red/ Mean_Green/ Mean_Blue (9)

$$SI\ 2 = Mean\_Hue / Mean\_Saturation / Mean\_Value \quad (10)$$

Algorithm 2: RGB to HSV

Input: RGB image

Output: HSV image

Step 1: Maximum = max([Red ,Green ,Blue]);

Step 2: Minimum = min([Red,Green,Blue]);

Step 3: Value = Maximum; // Calculation for Value

Step 4: Cost = Maximum-Minimum;

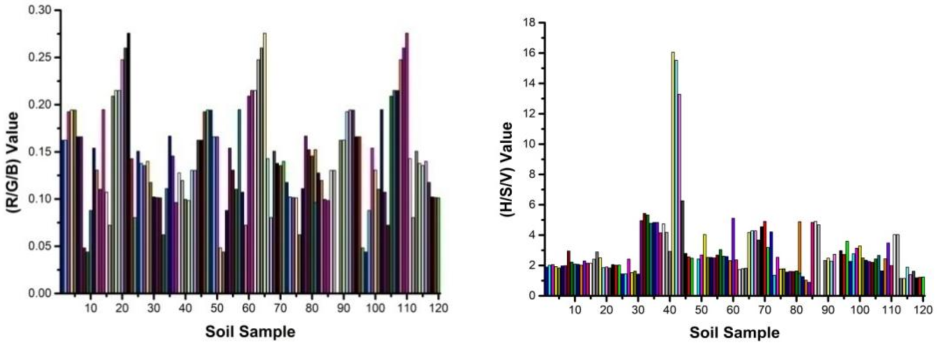
Step 5: if (Value == 0)   // Calculation for Saturation

```
    Saturation = 0;
else
    Saturation = Cost / Value;
end
```

Step 6: if (Cost==0)

```
    Hue=0;
elseif (Maximum==Red)
    Hue = 60 * mod(((Green-Blue)/Cost),6);
elseif (Maximum==Green)
    Hue = 60 * (((Blue-Red)/Cost) + 2);
elseif (Maximum==Blue)
    Hue = 60 * (((Red-Green)/Cost) +4);
End
```

The mean values of RGB and HSV values of the different soil images are presented in table 2 along with the mean values of SI 1 and SI 2. Fig.7 shows the graphical representation of SI 1 and SI 2.



(a) (b)

Fig. 7. Graphical representation of SI 1 b) Graphical Representation of SI 2

Table 2: Feature values of the Soil images.

Feature	Mean
Mean R	21.46636
Mean G	15.87252
Mean B	10.83996
Mean H	0.023307
Mean S	0.117823333
Mean V	0.086195
SI 1	0.144497381
SI 2	2.947872
SI 3	0.236758
SI 4	0.141131
SI 5	0.227326
SI 6	1.366939
SI 7	0.204018
SI 8	1.352423
SI 9	37.33888
SI 10	48.17884
SI 11	1.46426
SI 12	26.71247

III. Result Discussion and Analysis

Different values of Red, Green, Blue, Hue, Saturation and Values are calculated for the entire soil samples using the equation (iii), (iv), (v), (vi), (vii), and (viii). All the extracted values of RGB and HSV are stored in a .mat file for the Prediction. Again SI 1 and SI 2 are calculated for all 120 soil images as per the equation (ix) and (x). The prediction of soil pH is done with the help of different Prediction algorithm including Regression, ANN and KNN.

Soil pH prediction using Linear Regression

Linear regression is a good tool of machine learning to predict the unknown data in agriculture field and presented with the help of fitted line [VI, XVII]. Regression coefficient (R) gives the performance of the regression. The regression results showing an accuracy of 0.8-0.99 is accepted or highly correlated, otherwise, it is not accepted. In this paper, the following steps are performed.

- At first, the linear regression is performed in between the actual soil pH values and each value of Red, Green, and Blue separately and finds no correlation in all the cases.
- Secondly, the linear regression is performed in between the actual soil pH and each value of Hue, Saturation, and Value separately and finds a good correlation only in case of Saturation and Actual soil pH.
- Thirdly, the linear regression is performed in between the actual soil pH values and SI 1 and finds no correlation for SI 1.
- Fourthly, the linear regression is performed in between the actual soil pH values and SI 2 and finds no correlation for SI 2.

The result of all correlation is presented in Table 3 along the coefficient of linear regression.

Table 3: Coefficient of Linear Regression for Different features.

Feature	Regression	Coefficient of Linear Regression
R	No Correlation	0.077
G	No Correlation	0.046
B	No Correlation	0.008
H	No Correlation	0.005
S	**High Correlation**	0.858
V	No Correlation	0.077
SI 1	No Correlation	0.004
SI 2	No Correlation	0.301
SI 3	Moderate Correlation	0.613

SI 4	**High Correlation**	0.821
SI 5	Weak Correlation	0.389
SI 6	Moderate Correlation	0.430
SI 7	Moderate Correlation	0.443
SI 8	Weak Correlation	0.04
SI 9	Weak Correlation	0.064
SI 10	Weak Correlation	0.049
SI 11	Weak Correlation	0.072
SI 12	Weak Correlation	0.027

From table 3, it is analyzed that the extracted features of the images are not correlated with the actual pH values of the soil samples except the saturation value of the image. Barman et al. [VI] also found a strong correlation for the saturation values of the soil images. They achieved 86% accuracy for their model. SI 1 is suggested by Kumar et al. [XIII] and manually examined the soil pH from the different range of Red, Green and Blue values of the soil images. Again SI 2 is used by Swapna et al. [XXII] to find the pH of the sensor based soil image using K-mean clustering. In the paper [I], authors used the same dataset as used by Kumar et al. [XIII] and trained the model using ANN and found the 80% accuracy for their model. In this paper, some other soil indexes are defined using the following equation and their correlation with the soil pH values are presented in Table 3. Table 2 represents the mean value newly defined soil index.

SI 3 = Mean_Hue / Mean_Saturation (11)

SI 4 = Mean_Hue + Mean_Saturation (12)

SI 5 = Mean_Hue + Mean_Saturation + Mean_Value (13)

SI 6 = Mean_Saturation / Mean_Value (14)

SI 7 = Mean_Saturation + Mean_Value (15)

SI 8 = Mean_Red / Mean_Green (16)

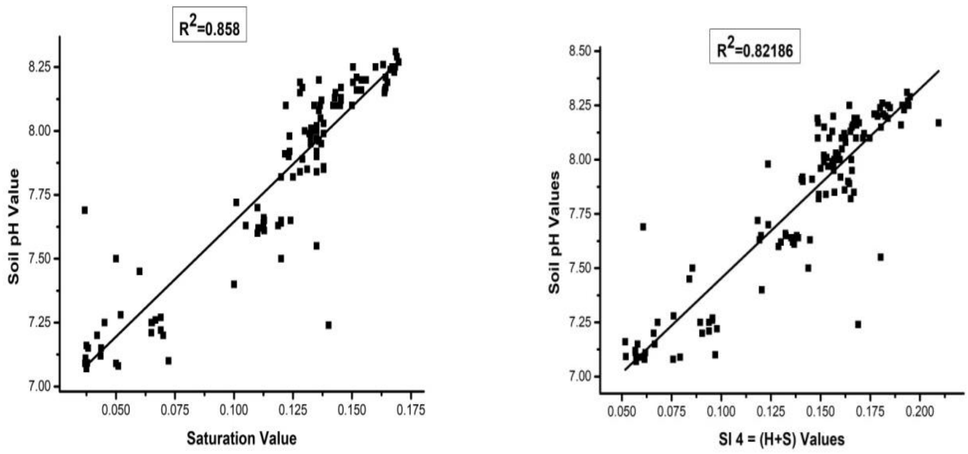
SI 9 = Mean_Red + Mean_green (17)

SI 10 = Mean_Red + Mean_Green + Mean_Blue (18)

SI 11 = Mean_Green / Mean_Blue (19)

SI 12 = Mean_Green + Mean_Blue (20)

From the Table 3, it is analyzed that SI 4 gives a high correlation with the soil pH values. The graphical representation of the correlation is presented in Fig 8



(a) (b)

Fig. 8. Regression in between different soil index and pH

Soil pH prediction using ANN

Different properties of soil can be estimated using Artificial Neural network. In paper [I], authors already used ANN for soil pH prediction with an accuracy of 80%. Paper [XXIV] used ANN for the Prediction of compaction parameter of soil. In this paper, an ANN model is developed to predict the soil pH using saturation and SI 4 as the input data. Saturation and SI 4 is used in input as these two features of soil already give a high correlation with the actual soil pH values. The used ANN model for this paper is a feed-forward neural network. The model consists of an input layer, one hidden layer, and an output layer. There are different Backpropagation algorithm is present in Neural Network Toolbox including Levenberg-Marquardt and Scaled Conjugate Gradient Backpropagation Algorithm. In this paper, both Levenberg-Marquardt and Scaled Conjugate Gradient Backpropagation Algorithms are used to train the model. Levenberg–Marquardt Backpropagation (LMBP) algorithm is the fastest neural network training and it uses less memory for the training set [XV]. Again, Scaled Conjugate Gradient Backpropagation algorithm (SCGBP) is a general-purpose training algorithm which requires no line search. The model is programmed in MATLAB (Release 2014b, MathWorks Inc., Natick, MA, Neural Network Toolbox™). The model is evaluated by considering 120 samples of soil. For the evaluation, 120 samples are divided into three categories of Training Set, Validation Set, and Testing Set.

The proportions of different sets are 70%-15%-15%. It means that the training set contains the 70% of the data and it is 84 samples. Validation set contains 15% of the data and it is 18 samples. Again, testing set contains 15% of the data and it is 18 samples. The performances of both the models are assessed using mean square error and coefficient of regression (R). Mean Square error (MSE) is the mean square error in between the actual values of soil pH and the predicted values.

For the Levenberg–Marquardt Backpropagation algorithm, 4 hidden neurons are applied in the hidden layer of the model. Initially, the performance of the model is not good, so the model is retrained again with weighting adjustments. The performance of the model after the retrain is presented in Table 4.

Table 4: MSE and R values of the LMBP algorithm

Set	Sample	MSE	RMSE	R
Training	84	0.0240224	0.1549916	0.927
Validation	18	0.0083895	0.091594	0.971
Testing	18	0.0086338	0.092918	0.970

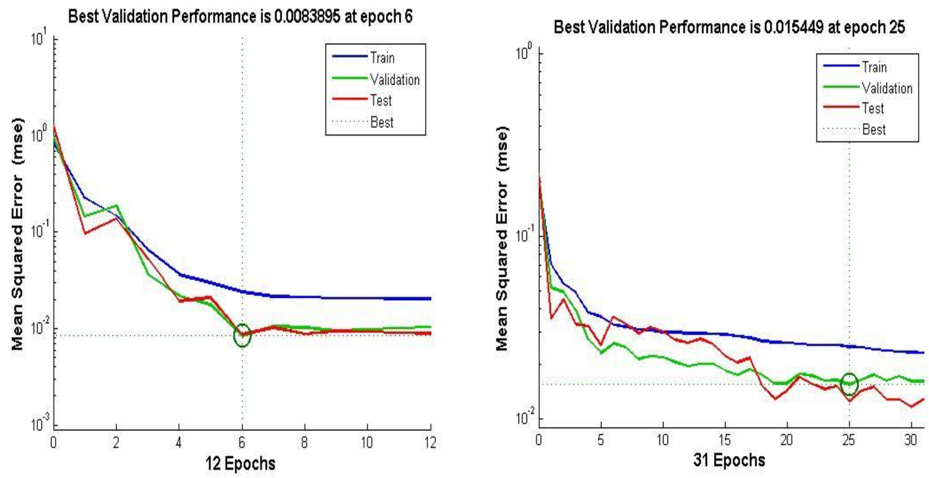
The best validation performance and the change of MSE with epoch are presented in Fig. 9. As per the Fig. 9a), it is observed that the best validation performance is obtained at epoch 6. After that, the validation graph with respect to MSE is not increasing and it is in the line of the testing graph. So, No overfitting occurred during the training of the Network. The post regression analysis of the model is presented in Fig.10. The coefficient of regression (R) of the training set, validation set, and testing are respectively 0.927, 0.971, 0.970. The accuracy of the validation set and testing set are almost similar and MSE error for both the set is minimum as compared to the training set.

For the Scaled Conjugate Gradient Backpropagation algorithm, 5 hidden neurons are applied in the hidden layer of the model and the performance of the model is achieved in the first training of the model. The performance of the model is presented in Table 5.

Table 5: MSE and R values of the LMBP algorithm

Set	Sample	MSE	RMSE	R
Training	84	0.024889	0.15776248	0.922
Validation	18	0.015449	0.1242940	0.948
Testing	18	0.012407	0.111386	0.967

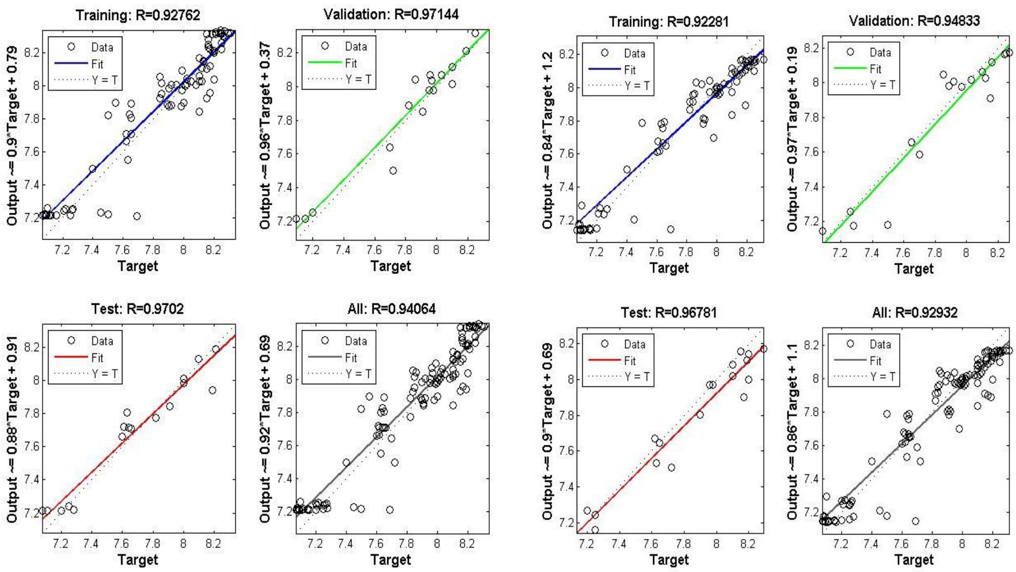
The best validation performance of the Scaled Conjugate Gradient Backpropagation algorithm and the change of MSE with epoch are presented in Fig. 9. As per the Fig. 9b), it is observed that the best validation performance is obtained at epoch 25 and it is 0.015449. After that, the validation graph with respect to MSE is not increasing and it is in a small difference with the line of the testing graph. So, No overfitting occurred during the training of the Network. The post regression analysis of the model is presented in Fig.10. The coefficient of regression (R) of the training set, validation set, and testing are respectively 0.922, 0.948, and 0.967. The difference between the accuracy of the validation set and the testing set is minimum and it is more than the accuracy of the training set.



(b)

Fig. 9. Best validation of ANN with LMBP and SCGBP

Fig. 10.



(a) (b)

Fig. 11. Regression of ANN with LMBP and SCGBP

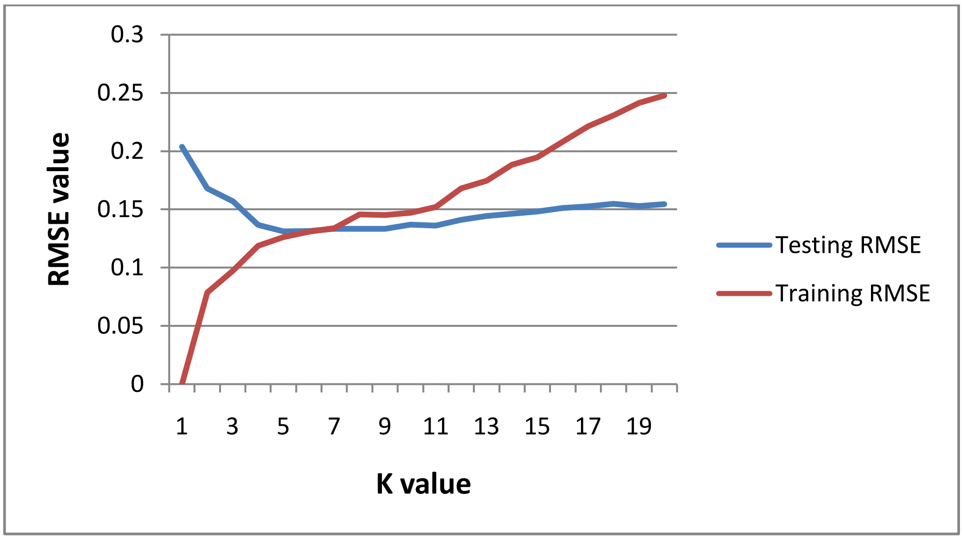
Soil pH Prediction using KNN

K nearest neighbor is a very popular algorithm which stores all the values of the training set and predicts the testing value based on the similarity measure. Various methods are present to calculate the distance between the training and testing set including Euclidian distance, Manhattan distance, Minkowski distance etc. The prediction of soil pH using KNN depends on the following steps. Division of entire dataset in between training and testing set.

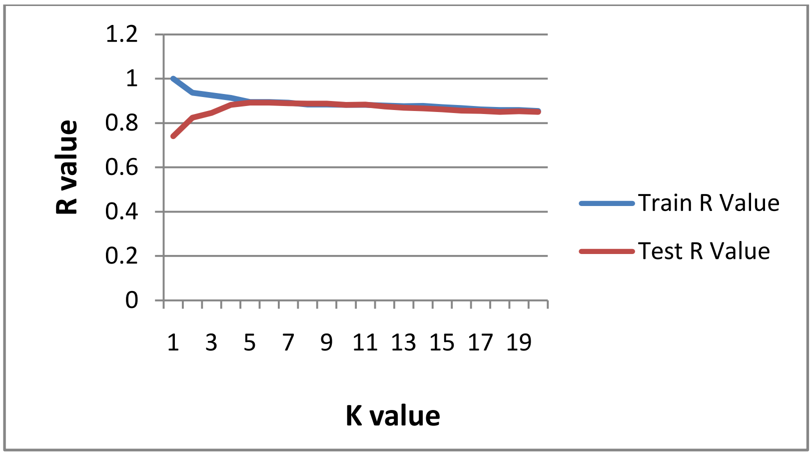
I. Determine the value of K.
II. Compute the distance between training and testing set using any distance algorithm
III. Sort all the training values as per the computed distance.
IV. Based on the RMSE value, find the appropriate K.

For KNN, the entire 120 values of saturation color space of the soil image and SI 4 of the soil image are considered as soil data. Soil data are divided into training and testing set. The ration of the division is 70:30. The training the similarity measurement of the model is performed using the Minkowski distance.

The performance of the model is assessed with the help of RMSE (Root Mean Square Error) and the coefficient of the regression (R). The RMSE and R values of the testing and training set for the different K values are presented in table 5. The model is executed for 20 different K values. The RMSE value of the training set is 0.00 for K=1 and it is 0.203640478 for testing set. So, for K=1, the training shows the overfitting. To reduce the overfitting, K values are changes and the accuracy of the model is determined in terms of the coefficient of regression. For K=5, the RMSE value of the testing set is minimum among all the K values and it is 0.131168255. Training data is not overfitting at K=5 as the training accuracy and testing almost same for K=5. The performance of training and testing sets of the model is presented in the Fig.11 and in Table 6.



(a) RMSE of KNN



(b) R Value of KNN

Fig. 12. Performance of KNN regression

Table 6: Performance of KNN with the changes of K.

K value	Train R Value	Test R Value	Testing RMSE	Training RMSE	Training MSE	Testing MSE
1	0.999846258	0.739673529	0.203640478	0	0	0.04146944
2	0.936805277	0.823186728	0.167827209	0.078691557	0.00619236	0.02816597
3	0.924263015	0.845446355	0.156907813	0.097231746	0.00945401	0.02462006
4	0.912899963	0.8825597	0.13677714	0.118649812	0.01407778	0.01870799
5	0.894530627	0.891994071	0.131168255	0.126200898	0.01592667	0.01720511
6	0.895087217	0.891676435	0.131360991	0.130998363	0.01716057	0.01725571
7	0.891843657	0.888433292	0.133312924	0.133791729	0.01790023	0.01777234
8	0.8833553	0.8883639	0.1333543	0.1458287	0.021266	0.017783

	04	15	67	2	02	39
9	0.882819719	0.888193958	0.133455839	0.145094218	0.02105233	0.01781046
10	0.881680422	0.882359519	0.136893661	0.147123248	0.02164525	0.01873987
11	0.881441863	0.883889427	0.136000601	0.152161726	0.02315319	0.01849616
12	0.879050479	0.875237845	0.140976387	0.167907361	0.02819288	0.01987434
13	0.876483815	0.869306848	0.144288381	0.174570243	0.03047477	0.02081914
14	0.876846514	0.865723948	0.146252812	0.188239735	0.0354342	0.02138988
15	0.87234857	0.862142444	0.148190455	0.194713877	0.03791349	0.02196041
16	0.867708391	0.856250988	0.151323848	0.208099426	0.04330537	0.02289891
17	0.861585882	0.853638782	0.152692586	0.221409322	0.04902209	0.02331503
18	0.858976209	0.84961696	0.154776269	0.230730263	0.05323645	0.02395569
19	0.858022127	0.853241571	0.152899643	0.241506793	0.05832553	0.0233783
20	0.854397646	0.850128366	0.154512872	0.24770554	0.06135803	0.02387423

Analysis of Result

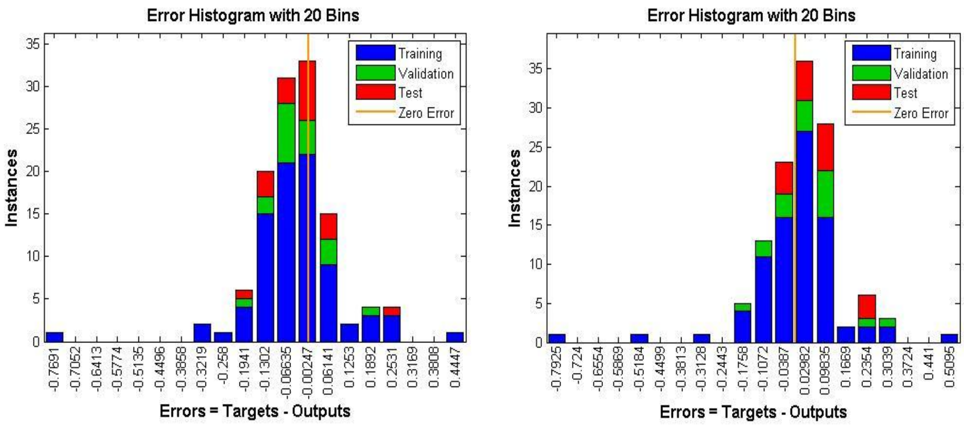
In this paper, soil pH prediction is done using linear regression, ANN and KNN regression. In the process, it is finding that soil saturation color is highly correlated with the actual value of the soil pH. Again from Table 3, it is found that SI 4 is also highly correlated with the actual value of the soil pH. Both the correlations are positive correlation and their coefficient of best fit is respectively 0.859 and 0.823. Apart from these two values, SI (3, 6 &7) gives a moderate correlation with the pH values and their

R-value is respectively 0.613, 0.430, and 0.443. The equations of highly correlated regression are presented by Table 7. As saturation and SI 4 are highly correlated with the values of the soil pH, so these two values are considered as input for ANN and KNN.

Table 7: Linear Regression equation of different Soil Index

Index	Equation
S	Y = 0.095X - 0.627
SI 1	Y = 0.009X + 0.073
SI 2	Y = - 3.161X + 27.62
SI 3	Y = - 0.303X + 2.605
SI 4	Y = 0.094X - 0.595
SI 5	Y = 0.072X – 0.336
SI 6	Y = 1.642X – 11.16
SI 7	Y = 0.073X – 0.367
SI 8	Y = 0.080X + 1.984
SI 9	Y = - 9.359 + 110.3
SI 10	Y = - 10.39X + 129.2
SI 11	Y = - 0.134X + 2.528
SI 12	Y = - 4.445X + 61.41

For KNN, Levenberg-Marquardt and Scaled Conjugate Gradient Backpropagation Algorithms are used as training algorithm. Levenberg-Marquardt model is used with 4 hidden neurons at the hidden layer and the R-value of the model is 0.94064. But the R-value of the tanning set, validation and testing set is presented in table 4. The accuracy of the training model is less compare to the testing and validation and gets the best validation result at epoch 6. The best validation is 0.0083895. Like Levenberg-Marquardt, Scaled Conjugate Gradient Backpropagation Algorithm also gives very good accuracy in the testing and validation set as compared to the training set and gets the best validation result at epoch 25. The best validation is 0.015449. The error histogram of both the model is presented in Fig. 12 with respect to the training, validation and testing set.



(a) Error Histogram of ANN with LMBP (b) Error Histogram of ANN with SCGBP

Fig. 13. (a-b) Error Histogram of ANN

KNN algorithm is widely used in the classification problem rather than regression. The performance of the KNN is depended on the optimum value of the K and optimum of K can be found by using the grid search. But in this paper, the KNN regression is implemented for the different value of K and finds the RMSE and MSE value of the model to find the optimum value of the K. The optimum value of the K is 5. The RMSE and MSE of the testing set of the KNN model are respectively 0.131168255 and 0.01592665. It is the less RMSE and MSE value among all the K values of RMSE and MSE of the testing set. At K =5, the accuracy of the testing and training is almost the same and it is presented in table 6. In the earlier study of soil pH prediction, authors find the different best fit of correlation such as R=0.849 [VI], R=0.998 [XIX], R=0.8 [I] for the different soil index. The comparative analysis of the different best values is presented in Table 8.

Table 8: Comparative Analysis of the Different Best values

Study	Index	Overall Best Fit Value	Tool
Barman et al. [VI]	Saturation (S)	0.849	Linear Regression
Aziz et al. [I]	R/G/B	0.80	ANN
Swpna et al [XXII]	H/S/V	Not Calculated	K-Mean
Ruiz et al [XIX]	Not Defined	0.988	Linear Regression
Kmar et al. [XIII]	R/G/B	Not Calculated	Not Performed
Proposed Model	Saturation	0.859	Linear Regression
Proposed Model	H+S	0.823	Linear Regression
Proposed Model	Saturation and (H+S)	0.94064	ANN with LMBP
Proposed Model	Saturation and (H+S)	0.92932	ANN with SCGBP
Proposed Model	Saturation and (H+S)	0.89326	KNN

From the table 8, it is analyzed that the soil pH prediction can be possible with the different prediction algorithm but ANN always gives a better result as compare to other one.

IV. Conclusion

Soil image color is a good indicator of soil pH. Previously farmers used soil color chart to predict the soil pH and this method is fully based on human perception. An Expert can also help farmers during this process but getting expert views in all time is not possible. Sometimes farmers used soil testing laboratory or soil pH digital meter to calculate the soil pH but these methods are very time and cost consuming. This paper gives a Smartphone based low-cost soil prediction model which will detect the soil pH easily. In this paper, different prediction algorithm is implemented but ANN with LMBP gives better accuracy for the prediction as compare to linear regression, ANN with SCGBP and KNN regression. Hence it is seen that Smartphone based technique for estimating the soil pH content is very helpful for rural farmers rather than other methods.

References

I. Aziz, M.M, Ahmed, D.R., Abraham, B.F, 2016. “Determine the pH of Soil by using Neural Network Based on Soil's Colour”. International Journal of Advanced Research in Computer science and Software Engineering, Vol.: 6, Issue: 11, pp: 51-54, 2018.

II. Abu, M.A., Nasir, E.M.M. and Bala, C.R, “Simulation of Soil PH Control system using Fuzzy Logic Method”, International Journal of Emerging Trends in Computer Image & Processing. Vol.: 3, Issue: 1, pp: 15-19, 2014.

III. Aditya, A., Chatterjee, N., Pradhan, C., “Computation and Storage of Possible Pouvoir Hydrogen Level of Soil using Digital Image processing”, International Conference on Communication and Signal Processing, India. pp: 205-209, 2017.

IV. Ayoubi, S., ShahriA, P., Karchegani, P.M., Sahrawat, K.L., “Application of Artificial Neural Network (ANN) to Predict Soil Organic Matter Using Remote Sensing Data in Two Ecosystems”. In: I. Atazadeh (Ed), Biomass and Remote Sensing Biomass. ISBN: 978-953-307. In Tech Publication. 2011.

V. Babu, C.S.M. and Pandian, M.A, “Determination of Chemical and Physical Characteristics of Soil using Digital Image processing”, International Journal of Emerging Technology in Computer Science & Electronics, Vol.: 20, Issue: 2, pp: 331-335, 2016.

VI. Barman, U., Choudhury, R., Talukdar, N., Deka, P., Kalita, I., & Rahman, N, "Prediction of soil pH using HSI colour image processing and regression over Guwahati, Assam", India. Journal of Applied and Natural Science, Vo.: 10, Issue: 2, pp: 805-809, 2018.

VII. Barman, U, Choudhury, R. D., Saud, A., Dey, S., Dey, B. K., Medhi, B.P., Barman, G.G., "Estimation of Chlorophyll Using Image Processing", Int J Recent Sci Res, Vol.: 9, Issue: 3, pp: 24850-24853, 2018

VIII. Bodaghabadi, M.B., Martínez-Casasnovas, J.A., Salehi, M.H., Mohammadi, J., Borujeni, I.E., Toomanian, N., Gandomkar, A., "Digital Soil Mapping Using Artificial Neural Networks and Terrain-Related Attributes", Pedosphere, Vol.: 25, Issue: 4, pp: 580-591, 2015.

IX. Dhanachandra, N., Manglem, k., Chanu y.J., "Image Segmentation Using K - means Clustering Algorithm and Subtractive Clustering Algorithm", Procedia Computer Science. Vol.: 54, pp: 764-771, 2015.

X. Ebrahimi, M., Sinegani, A.K.S., Sarikhani, M.R., Mohammadi, S.A., "Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses". Computers and Electronics in Agriculture. Vol.: 140, pp: 409-421, 2017.

XI. Gurubasava, Mahantesh S.D., "Analysis of Agricultural soil pH using Digital Image Processing" , International Journal of Research in Advent Technology, Vol.: 6, Issue: 8, pp: 1812-1816, 2018.

XII. Guwahati.Assam.Link:https://www.google.co.in/maps/place/Guwahati,+Assam/data=!4m2!3m1!1s0x375a5a287f9133ff:0x2bbd1332436bde32?sa=X&ved=2ahUKEwj4jYCo1_rcAhVFKo8KHWMoB3AQ8gEwAHoECAQQAQ

XIII. Kumar, V., Vimal, B., Kumar, R., Kumar, R., & Kumar, M, "Determination of soil pH by using digital image processing technique". Journal of Applied and Natural Science, Vol.: 6, Issue: 1, pp: 14-18, 2014.

XIV. Matei, O., Rusu, T., Petrovan, A., Mihuţ G., "A Data Mining System for Real Time Soil Moisture Prediction", Procedia Engineering, Vol.: 181, pp: 837-844, 2017

XV. Mohan, R.R., Mridula S., Mohanan P., "Artificial Neural Network Model for Soil Moisture Estimation At Microwave Frequency", Progress In Electromagnetics Research M, Vol.: 43, pp: 175–181, 2015.

XVI. Pandey, A., Jha, S.K., Srivastava, J.K., Prasad R., "Artificial neural network for the estimation of soil moisture and surface roughness", Russ. Agricult. Sci. Vol.: 36, pp: 428-432, 2010.

XVII. Riccardi, M., Mele, G., Pulvento, C., lavini A., D'AndriaS R., Jacobsen E., " Non-destructive evaluation of chlorophyll content in quinoa and amaranth leaves by simple and multiple regression analysis of RGB image components". Photosynth Res. Vol.: 120, pp: 263–72, 2014.

XVIII. Rigon, J.P.G., Capuani, S., Fernandes, D.M., Guimarães, T. M., 2016. "A novel method for the estimation of soybean chlorophyll content using asmartphone and image analysis", Photosynthetica, Vol.:54, pp: 559–566, 2016.

XIX. Ruiz, N. L., Curto, V.F., Erenas, M. M., Lopez, F. B., Diamond, D., Lopez, A. J. P, Valley, A.F.C., "Smartphone-Based Simultaneous pH and Nitrite Colorimetric Determination for Paper Microfluidic Devices". Analytical Chemistry. Vol: 86, Issue: 19, pp: 1-23, 2014.

XX. Sagar, S, Debjeet, B, Advait, L,Mishra, N., "Moisture And pH Detection Using Sensors And Automatic Irrigation System Using Raspberry Pi Based Image Processing", International Journal of Engineering Technologies and Management Research, Vol.: 5, issue: 2, pp: 153-157, 2018.

XXI. Soil pH. Link: www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051574.pdf

XXII. Swapna, U.C., Prapulla, Kumar. "Measurement of Soil PH Value Using HSV Color Space Value of Image". International Journal of Innovative Research and Advanced Studies, Vol.: 3, Issue: 6, pp: 1-4, 2016.

XXIII. Taheri-Garavand, A., Meda, V., Naderloo, L., "Artificial neural Network−Genetic algorithm modeling for moisture content prediction of savory leaves drying process in different drying conditions". Engineering in Agriculture, Environment and Food. Vol.: 11, Issue: 4, pp: 232-238. 2018.

XXIV. Tenpe, A., Kaur, S., "Artificial neural network modeling for predicting compaction parameters based on index properties of soil", Int J Sci Res (IJSR), Vol.: 4, issue: 7, pp: 1198–1202, 2015.

XXV. Utai, K., Nagle, M., Hämmerle, S., Spreer, W., Mahayothee, B., Müller, J., "Mass estimation of mango fruits (Mangifera indica L., cv. 'Nam Dokmai') by linking image processing and artificial neural network", Engineering in Agriculture, Environment and Food., Vol.: 12, Issue: 1, pp:103-110, 2019.