ISSN: xxxx-xxxx (ONLINE)
DOI:

MACHINE LEARNING IMPLEMENTATION ON AGRICULTURAL DATASETS FOR SMART FARM ENHANCEMENT TO IMPROVE YIELD BY PREDICTING PLANT DISEASE AND SOIL QUALITY

G. Kavitha, M. Ganthimathi, K. Sudha and M. Ramya
Department of Computer Science and Engineering, Muthayammal Engineering College, India

Abstract

As agriculture struggles to support the rapidly growing global population, plant disease reduces the production and quality of food, fiber and biofuel crops and farmers are also not aware about the crop which suits their soil quality, soil nutrients and soil composition. The purpose of this review is to present the application of machine learning in plant resistance genes discovery and plant diseases classification and helps the system focuses on checking the soil quality to predict the crop suitable for cultivation according to their soil type which maximize the crop yield depending on the analysis done based on machine learning approach.

Keywords:

Crop Yield, Soil Quality, Plant Disease, Machine Learning

1. INTRODUCTION

Machine learning methods focus on data themselves, it mainly concentrates on four areas based on the problems to be solved: 1) detection 2) classification; 3) quantification; 4) prediction. Machine learning is divided into two categories: 1) supervised learning and 2) unsupervised learning. In the recent days, machine learning used in various disciplines, such as computing, classification, bioinformatics, marketing, medical diagnosis, game playing, healthcare and industry. Machine learning algorithms are frequently used in researches, such as Naïve Bayes Classifier, K Means clustering, support vector machine (SVM), artificial neural networks (ANN), decision trees and random forests. Machine learning includes data collection, dataset preparation, feature extraction, preprocessing, feature selection, choosing and applying machine learning algorithms and performance evaluation [10]. In this system machine learning methods are mainly applied to predict molecular biology and agriculture related to plant diseases [12].

1.1 CROP DISEASE

Crop diseases cause varies depending upon factors such as atmosphere, host health and susceptibility, and pathogen biology. There are many options for managing disease development and spread. The effectiveness of management techniques begins with proper detection of the disease and/or causal organism.

1.2 PATHOGENS

A plant disease is any physiological or structural abnormality that is caused by living organism. Organisms that cause disease are referred to as 'pathogens,' and affected plants are referred to as 'hosts' [1]. Many organisms rely on other species for sources of nutrients or as a means of survival, but are not always harmful to the host. For example, saprophytic organisms obtain nutrients from dead organic material and are a vital part of many ecosystems. Disease causing organisms include fungi, oomycetes (fungus-like organisms called water molds), bacteria, viruses, nematodes, phytoplasmas, and parasitic seed plants [14]. Once a pathogen infects a host, Symptoms are the outward changes in the physical appearance of plants.

1.3 FUNGI

Fungi are the most abundant group of plant pathogens. There are thousands of fungi capable of causing plant diseases. These multicellular organisms are typically microscopic [1]. The 'body' of a fungus is composed of filament-like threads called 'hyphae.' Masses of hyphae are called 'mycelia'. When large enough, these masses can be seen without the aid of a microscope. Powdery mildew is one example of a disease in which fungal mycelia is visible. Spores vary in color, shape, size, and function, and this variation can often be used by diagnosticians as a means to identify pathogens [15]. Some fungi produce spores within one fruiting structures (ascocarps, pustules, mushrooms pycnidia, acervuli). Other types of fungi produce exposed or unprotected spores that are not enclosed in structures [5]. These spore types are more sensitive to environmental conditions than enclosed spores. Upon infection, fungi utilize nutrients from their hosts [16]. Common symptoms caused by fungi include leaf spots, wilts, blights, cankers, fruit rots, and dieback [17]. The Fig.1 shows Septoria leaf spot of tomato is a disease familiar to many gardeners.

Fig.1. Septoria leaf spot of tomato

1.4 OOMYCETES (WATER MOLDS)

As the name implies, water is essential for survival, reproduction, infection, and spread of oomycetes. Water molds were once considered true fungi, but they are now classified as fungus-like organisms. Water molds and fungi are similar in appearance, as the 'body' is composed of hyphae that mass together to form mycelia. Infective propagules are spread via water, soil, infected plants and weeds, as well as by wind and wind-driven rain. Survival structures produced by water mold pathogens have the ability to persist in water and soil for several

years. Common symptoms caused by water molds include leaf spots, blights, cankers, root rots, wilt, damping off, and dieback.

1.5 BACTERIA

Bacteria are microscopic organisms typically composed of single cells. About 200 types of bacteria are known to cause plant diseases. Due to their small size, a high-magnification microscope is required to observe bacteria [9]. Occasionally, when a large number of cells are present, plants may be observed 'oozing' bacteria and other organic byproducts. Bacteria are capable of rapid reproduction through a process known as binary fission. In this process, one cell divides to become two, then two divide to become four cells, and so on. Within a few hours one bacterial cell can become thousands, and under ideal conditions, populations can double in as little as 20 minutes.

Unlike fungi and water molds, bacteria are not able to penetrate plant tissue directly. They must infect via wounds or natural plant openings such as stomata. Free water is required for infection. Once inside plants, bacteria begin to reproduce immediately. Some types of bacteria produce toxins or enzymes that degrade plant tissue, and the tissue is then utilized as a food source. Some bacteria can colonize vascular systems of plants, which results in restriction of water movement. Bacteria spread by water/splashing rain, wind, or insects, and then move across plant tissues in surface water to reach wounds or natural openings. Some can survive for five or more years in soil, as well as in plant debris and cankers. Common symptoms caused by bacteria include leaf spots, blights, cankers, galls, wilt, dieback, and soft rots. The Fig.2 shows Common disease that is known to infect numerous landscape and garden plants [6].

Fig.2. Common disease

Unlike developing countries, developed countries have monitoring and management mechanisms in place to mitigate the consequences of harmful diseases of crop plants: safety nets to support those most affected; food reserves that limit the risk of famine; research ability and technical support services to manage diseases; and warning systems that allow prompt application of control measures. Over the past decades, continuous studies have been performed to reveal the interactions between plant immune respond and pathogens.

Large amount of data has been generated from those researches due to the tremendous advances in genomics and proteomics. But now, the development of machine learning algorithms, which are a collection of analytic methods that automate model building process and iteratively learn from data to gain insights without explicitly programming, provides more powerful and efficient tools to not only identify genes/proteins involved in plant-pathogen interactions, but also classified plant diseases from images of infected leaves.

We here present a review of studies that utilize machine learning regarding the plant-pathogen interactions and plant disease identifications.

2. LITERATURE SURVEY

Fabrizio Balducci et al. [4] presented in this work introduces practical, cheap, and easy-to-develop tasks that are useful to increase the productivity of an agricultural company, deepening the study of the smart farm model; the technological progress in a field that needs control and optimization can really contribute to save environmental resources, respect the business and international laws, satisfy the consumer needs, and pursue economic profits. The three different data sources, with a special eye for the IoT sensors dataset, have been exploited using machine learning techniques and the more standard statistical ones. The first task shows that the forecast of apple and pear total crops on the Istat dataset could be reached with a neural network model with a success rates close to 90%, while in the second task, it emerges that for the CNR scientific data, polynomial predictive and regression models are more suited considering the nature of the dataset. Tasks 3 and 4 present the same goal faced with different machine learning methods on a pure IoT sensors dataset, showing that the decision tree model works very well; that there are specific environmental factors coming from sensors hardware that affect the model performances; and, moreover, that short-term future values with few past data can be predicted using statistical regressions [7]. It cannot be left out, however, that in cases where there are very few data statistical models such as linear or polynomial that still maintain the best predictive performances; moreover, the detection of faulty monitoring stations in Task 5 successfully employs a clustering of the stations based on their geographic location useful to detect hardware faults. In real cases highlight the need for integrating management and data scientists, in fact, IoT systems require engineering and diffusion investments that only a wise and visionary management can favor in smart/medium industries; moreover, the necessity to invest in skills and knowledge to profitably employ the IoT paradigm at higher levels emerges [8]. The main reason for the proposed tasks using different machine learning techniques is that an exploratory and highly experimental work has been employed; the Information Fusion together with the related optimization of methods and results is expected in future work, where new experiments and tasks exploit other sensor types and datasets will be designed and performed to meet the great heterogeneity of agri-companies and of the hardware sensor market. The intelligent systems developed with machine learning algorithms (supervised and non) have to manage fault tolerance and hardware malfunction prediction, and, in this way, they require designing of integrated tools, user-interfaces, and machines that easily adapt to a contexts subjected to natural events not as easily predictable as the agricultural one. Finally, smart systems that provide real-time suggestions and make long-term forecasts based on user choices and preferences must be studied and tested.

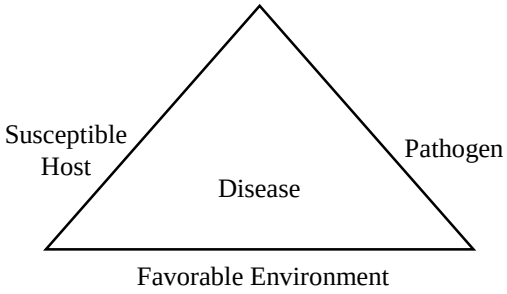
R. Ghadge et al. [2] used a system that consists of a supervised and unsupervised machine learning algorithms and gives best result based on accuracy. The results of the two algorithms will be compared and the one giving the best and accurate output will be selected. Thus the system will help reduce the difficulties faced

by the farmers and stop them from attempting suicides. It will act as a medium to provide the farmers efficient information required to get high yield and thus maximize profits which in turn will reduce the suicide rates and lessen his difficulties.

M.A. Adejumobi et al. [3] observed among the three crops planted in the study area in fig.2, that rice had the highest yield of 4.0 ton/ha followed by maize with 1.1 ton/ha and okra with 3.1 ton/ha for phase II while garden egg had the highest yield of 5.20 ton/ha followed by watermelon with 3.91 ton/ha while pepper had the least yield of 3.1 ton/ha for phase III. Rice yield decreased from 4.0 ton/ha in 2015 to 3.6 ton/ha in 2016 and subsequently decreased to 3.3 ton/ha in 2017 for phase II and garden egg yield decreased from 5.2 ton/ha in 2015 to 4.60 ton/ha in 2016 and then decreased to 4.00 ton/ha in 2017 for phase III. The decrease in the crop yield between 2015 and 2017 may be due to the increase in exchangeable bases, in the soil such as phosphorous, magnesium, potassium and sodium, in the soil. Maize and okro (phase II) and watermelon and pepper (phase III) yields decreased slightly from 2015 to 2017 as compared to rice (phase II) and pepper (phase III) yields respectively. This indicate that there is not much effect of chemical properties on maize and okro as compared to rice which has been affected by increased in chemical content of the soil in phase II, there is a slight negative effect due to chemical properties on the yield of both watermelon and pepper in phase III. Comparing this with FAO crop yield standards the yields for the crop in both phase II and phase III are lower than the FAO crop yield standards for each respective crops (Fig.1). However the one way analysis of variance of the crop yields and FAO standard as presented in Table.3, which indicates that the yields of maize, pepper, garden egg and watermelon are significant while yields of rice and okro are insignificant of 95% level of significance with respect to the FAO standard.

3. MACHINE LEARNING AND PREDICTIONS OF PLANT OR PATHOGEN RELATED MOLECULES

Due to the large number of plant resistance genes families, a high-throughput method is needed to identify genes involved in the resistance to pathogens. Although machine learning techniques have been used for various subjects, only a few studies have been conducted to predict plant pathogens related genes/ proteins. One of the major interests in plant-pathogen interactions is to identify the plant resistance genes. A good example of applying machine learning in plant-pathogen interaction research that Support Vector Machine (SVM), which was used to predict plant resistance proteins. Among the different machine learning methods, Naïve Bayes achieved highest performance of prediction. Later, based on SVM to predict effectors sub cellular localizations, which provide critical clues about the functions of effectors in plant cells. Machine learning is promising in the several ways to discover new insights and knowledge in molecules involved in plant-pathogen interactions. In addition to plant resistance genes/proteins, plant susceptible genes/proteins identifications are often ignored but are also important. Meanwhile, protein partners that physically interact with resistance genes are critical to elucidate the plant defense pathways and should be investigated using machine learning methods. Another good question for pathologists is whether a specific pathogens effectors target a variety of plant molecules. Machine learning tools can be potentially used to predict and understand how different plants genes/proteins response to pathogen effectors.



(a) Disease Triangle - Plant disease results when there is a susceptible host, viable pathogen and favorable environment

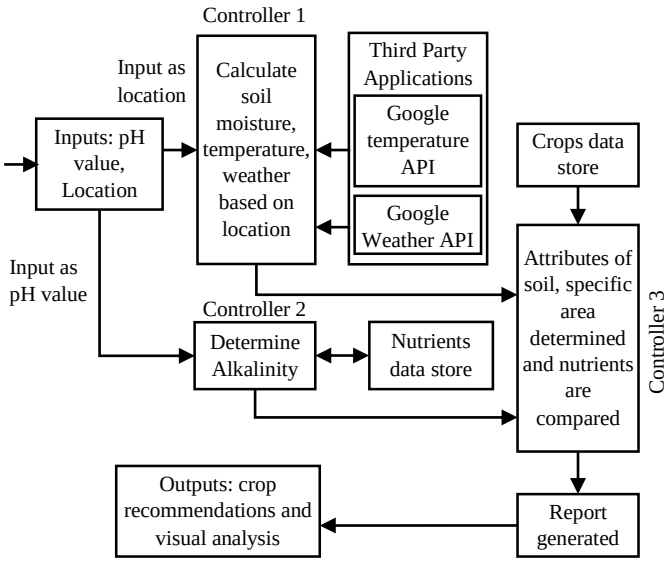


Fig.3. Block diagram of crop analysis process

3.1 CROP YIELD

The system aims to help farmers to cultivate proper crop for better yield production. To be precise and accurate in predicting crops, the project analyze the nutrients present in the soil and the crop productivity based on location. It can be achieved using unsupervised and supervised learning algorithms, like Kohonen Self Organizing Map and BPN (Back Propagation Network) [2]. Dataset will then trained by learning networks. It compares the accuracy obtained by different network learning techniques and the most accurate result will be delivered to the end user. Along with this, the end user is provided with proper recommendations about fertilizers suitable for every particular crop.

The proposed system will check soil quality and predict the cop yield accordingly along with it provide fertilizer recommendation if needed depending upon the quality of soil [3]. The functionality of the architecture in Fig.3 is as follows: The system takes inputs pH value (based on percentage of nutrient) and location from the user. Result processing is done by two controllers. Location is used as an input to controller 1, along with the use of third party applications like APIs for weather and temperature, type of soil, nutrient value of the soil in that region,

amount of rainfall in the region, soil composition can be determined.

The pH value is given as an input to controller 2, from which alkalinity of the soil is determined. Along with it, percentage of nutrients like Nitrogen (N), Phosphorous (P), Potassium (K), Sulphur(S), Magnesium (Mg), Calcium (Ca), Iron (Fe), Manganese, Boron and Zinc and Organic matter can be obtained. The result of the controller 1 and controller 2 are compared with a predefined "nutrients" data store. These compared results are supplied to controller 3 wherein the combination of the above results and the predefined data set present in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage wherein the combination of the above results and the predefined data set present in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage predefined data set present in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage of controller 3, wherein the combination of the above results and the predefined data set present in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage. Predefined data set present in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage.

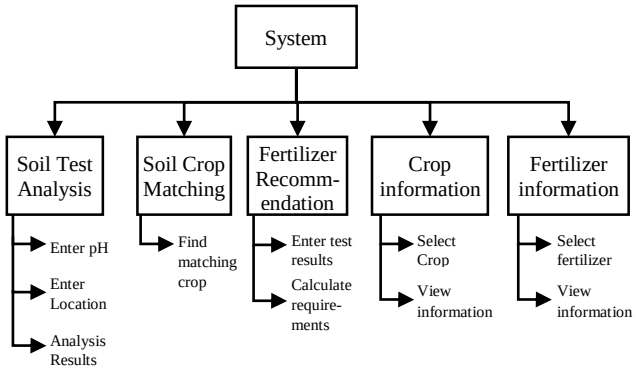


Fig.4. Modular Diagram

3.2 CHEMICAL ANALYSIS

The results of the selected average soil chemical properties of phase II and III are presented in Table.2. The soil pH on the field was moderately to slightly acidic for phase II and moderately acidic for phase III, as given by [9] ranged from 5.30 to 6.87 and from 5.72 to 5.88 for phase II and phase III respectively. According to United State Department of Agriculture (USDA), too high or too low soil pH leads to deficiency of many nutrients, decline in microbial activities, decrease in crop yield and deterioration of soil health. Therefore, the soil of the study area is thus suitable for crop growth. Organic matter content, parent material and degree of weathering. However, the soil at phase II and phase III will be good for crop that requires much phosphorus. From 19.296 to 30.242 mg/l. The level of phosphorous content of the soil is high and may be due to moderate. The system has five modules as depicted in Fig.4.

In soil test analysis user enters pH and location. Output of this module is analysis result of the percentage of nutrients in that soil. Soil crop matching module finds the matching crop that could be grown in that soil by comparison with the crop database. In fertilizer recommendation module, user is recommended with fertilizer that will give the highest crop yield. In the crop information module, user can select a crop and view information about it.

Table.1. Prediction Error over various cultures

Culture	Prediction Error (%)		
	NN	LR	Polynomial
Artichokes	139	101.63	2.70
Pear	1779.38	81.80	10
Pacciamata Eggplant	933.1	564.89	6.26

Table.2. Average soil chemical properties of phase II & phase III

Parameters	Phase II			Phase III		
	0-20 cm	0-20 cm	20-60 cm	0-20 cm	0-20 cm	20-60 cm
Ph	6.52	5.3	6.87	5.72	5.77	5.88
Mg++ (me/l)	1.36	1.16	2.26	2.06	1.29	2.78
Ca++ (me/l)	5.25	4.36	6.22	4.29	4.99	6.05
Na+ (me/l)	0.25	0.16	0.65	0.18	0.17	0.21
N (me/l)	1.04	0.18	1.15	0.22	1.07	0.32
K (me/l)	0.23	1.26	0.18	1.06	0.29	1.15
OC (%)	0.23	0.14	0.27	0.12	0.1	0.09
OM (%)	0.17	0.16	0.13	0.2	0.19	0.15
P (mg/l)	28.34	26.06	20.28	30.24	24.07	19.29
ESP (%)	10	9.11	5.9	9.02	8.01	6.8
CEC (me/l)	5.34	4.76	5.52	4.75	4.83	4.62
SAR (meg/l)	0.14	0.1	0.32	0.1	0.09	0.1

Table.3. Prediction Error of Apple and Pears between NN & LR

Italian Province	Prediction Error - Apple (%)		Prediction Error - Pears (%)	
	NN	NN	LR	LR
Udine	6.10	25.50	3.53	14.19
Gorizia	12.72	45.56	6.64	16.33
Trieste	21.80	21.25	9.83	21.25
Pordenone	12.04	38.47	154.79	153.12
L'Aquila	0.05	0.06	0.04	0.08
Teramo	2.52	2.53	1.52	13.03
Pescara	3.74	5.45	10.52	19.17
Chieti	3.65	10.54	2.26	2.73
Cosenza	22.68	63.79	16.57	20.62
Catanzaro	8.23	21.12	2.97	55.93
Reggio Calabria	11.38	42.60	6.14	13.08
Crotone	7.00	95.57	7.46	133.60
Vibo Valentia	7.50	27.55	29.40	45.31
Mean	**9.19**	**30.77**	**19.36**	**39.11**

In fertilizer information module, user can select a fertilizer and display information about it. The organic carbon gives a direct measure of available nitrogen of the soil. The average organic carbon in phase II ranged from 0.142-0.267% and phase III ranged from 0.0860-0.1151% of the entire soil nutrients relating to soil fertility [10]. Stated that organic carbon for the soil is considered high if it is within the range of 0.96-1.08%. It is observed from Table.3 that the level of organic carbon has decreased from surface depth (0-20cm) to the second depth (20-60cm) and sharply increased from second depth (20-60cm) to third depth (60-100cm) of the soil for phase II and that the level of organic carbon has generally decreased. Phosphorus is an essential element classified as a macro-nutrient because of the relatively large amount required by plants [11]. Available phosphorous content of the soil at phase II is high ranged from 20.276 to 28.342mg/kg. Available phosphorous content of the soil at phase III ranged.

4. CONCLUSION

The system uses supervised and unsupervised Machine learning algorithms and gives best result based on accuracy. Thus the system will help reduce the difficulties faced by the farmers and stop them from attempting suicides. In conclusion, machine learning provides a powerful tool to analyze tremendous amount of data. Careful selection of pre-processing data methods and machine learning tools is critical to obtain highest accuracy of classification. Meanwhile, compared to traditional methods of identifying genes involved plant pathogen interactions, methods integrating machine learning approaches are relatively scarce in the literature. Thus more machine learning based tools are needed to predict important plant resistance genes, as well as make contribution to the agriculture. With aerial imaging platforms and sensor technology, collecting field data becomes easier and more precise, which is critical for improving machine learning accuracy. More sophisticated methods such as deep learning algorithms will be applied in detecting plant diseases and discovering plant resistance genes. It will act as a medium to provide the farmers efficient information required to get high yield and thus maximize profits which in turn will reduce the suicide rates and lessen his difficulties.

The system can be enhanced further to add following functionality: Crop diseases detection using Machine learning and giving suggestions about fertilizers. The users can upload picture of diseased crop and get organic pesticides recommendations. Implementation of Smart Irrigation System to monitor weather and soil conditions, plant water usage etc. to automatically alter watering schedule.

REFERENCES

[1] X. Yang and T. Guo, "Machine Learning in Plant Disease Research", *European Journal of Biomedical Research*, Vol. 3, No. 1, pp. 6-9, 2017.

[2] R. Ghadge, J. Kulkarni, M. Pooja, N. Sachee and R.L. Priya, "Prediction of Crop Yield using Machine Learning", *International Research Journal of Engineering and Technology*, Vol. 5, No. 2, pp. 31-37, 2018.

[3] M.A. Adejumobi, H.A. Hussain and O.R. Mudi, "Physio-Chemical Properties of Soil and Its Influence on Crop Yield of Oke-Oyi Irrigation Scheme, Nigeria", *International Research Journal of Engineering and Technology*, Vol. 6, No. 4, pp. 1-8, 2019.

[4] Antonieta De Cal, Inmaculada Larena, Belen Guijarro and Paloma Melgarejo, "Use of Biofungicides for Controlling Plant Diseases to Improve Food Availability", *Agriculture*, Vol. 2, No. 2, pp. 109-124, 2012.

[5] Fabrizio Balducci, Donato Impedovo and Giuseppe.Pirlo, "Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement", *Machines*, Vol. 6, No. 3, pp. 21-38, 2018.

[6] Kimberly Leonberger, Kelly Jackson, Robbie Smith and Nicole Ward Gauthier, "Plant Diseases", Available at: http://www2.ca.uky.edu/agcomm/pubs/ppa/ppa46/ppa46.pdf.

[7] A. Lesser, "A Big Data and Big Agriculture", Available at: https://gigaom.com/report/big-data-and-big-agriculture/

[8] S.S Patil and S.A. Thorat, "Early Detection of Grapes Diseases using Machine Learning and IoT", *Proceedings of 2nd International Conference on Cognitive Computing and Information Processing*, pp. 1-5, 2016.

[9] T. Truong, A. Dinh and K. Wahid, "An IoT Environmental Data Collection System for Fungal Detection in Cropfields", *Proceedings of 30th International Conference on Electrical and Computer Engineering*, pp. 1-4, 2017.

[10] A.A. Sarangdhar and V.R. Pawar, "Machine Learning Regression Technique for Cotton Leaf Disease Detection and Controlling using IoT", *Proceedings of International Conference on Electronics, Communication and Aerospace Technology*, pp. 449-454, 2017.

[11] K.P. Satamraju, K. Shaik and N. Vellanki, "Rural Bridge: A Novel System for Smart and Co-Operative Farming using IoT Architecture", *Proceedings of International Conference on Multimedia, Signal Processing and Communication Technologies*, pp. 22-26, 2017.

[12] Shivnath Ghosh and Santanu Koley, "Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 2, No. 2, pp. 292-297,2014.

[13] Z. Hong, Z. Kalbarczyk and R.K. Iyer, "A Data Driven Approach to Soil Moisture Collection and Prediction", *Proceedings of International Conference on Smart Computing*, pp. 292-297, 2016.

[14] Sabri Arik, Tingwen Huang, Weng Kin Lai and Qingshan Liu, "Soil Property Prediction: An Extreme Learning Machine Approach", *Proceedings of International Conference on Neural Information Processing*, pp. 666-680, 2015.

[15] Vaneesbeer Singh and Abid Sarwar, "Analysis of Soil and Prediction of Crop Yield (Rice) using Machine Learning Approach", *International Journal of Advanced Research in Computer Science*, Vol. 8, No. 5, pp. 1254-1259, 2017.

[16] J.R. Okalebo, K.W. Gathna and P.L. Woomer, "Laboratory Methods for Soil and Plant Analysis. A Working Manual", 2nd Edition, Tropical Soil Biology and Fertility Institute of CIAT, 2002.

[17] Soil Testing, Available at: http://www.wikipedia.com/soil_testing, Accessed on 2009.