# *SOIL MOISTURE PREDICTION USING MACHINE LEARNING*

Shikha Prakash
Electronics and Communication
Birla Institute of technology
Mesra ,Ranchi
Email:shikhaprakash04@gmail.com

Animesh Sharma
Electronics and Communication
Birla Institute of Technology
Mesra,Ranchi
Email:animeshsharma97@gmail.com

Sitanshu Shekhar Sahu
Electronics and Communication
Birla Institute of technology
Mesra, Ranchi
Email:sssahu@bitmesra.ac.in

***Abstract:*****Prediction of soil moisture in advance is useful to the farmers in the field of agriculture. In this paper we have used machine learning techniques such as multiple linear regression, support vector regression and recurrent neural networks for prediction of soil moisture for 1 day, 2 days and 7 days ahead. These techniques were applied on three different datasets collected from different online repositories. The performance of the predictor is evaluated on the basis of mean squared error(MSE) and coefficient of determination ($R^2$ ).The comparison result shows that multiple linear regression is superior providing MSE and $R^2$ of 0.14 and 0.975 for 1 day ahead, 0.353 and 0.939 for 2 days ahead, 1.59 and 0.786 for 7 days ahead.**

***Keywords:Agriculture, machine learning, multiple linear regression,prediction,recurrent neural network, support vector regression.***

## I. Introduction

India is a country where majority of the population is dependent on agriculture for their livelihood. Indian soils are less fertile especially in case of micronutrients.In recent years, it has been seen that soil health is somehow related with the sustainability in the field of agriculture and also the current crop yield levels can be improved by maintaining the fertility of the soil. Agriculture needs decision support system in variety of ways such as type of crop to be cultivated [1].By monitoring soil moisture, water usage can be optimized to a large extent as the water table is lowering day by day. Soil moisture is beneficial for the production of crops so, the processes which are involved for the growth of the crops can be more enhanced if we successfully predict the soil moisture content of any area or location. By knowing the soil moisture content farmers can get information about what could be the best time of sowing and cultivating the crops, infiltration of the soil is proper or not, if enough water has been provided to the roots of the crops for growth or not. Data mining techniques plays a significant role in the field of agriculture.Good crop yield prediction results have been achieved by many researchers after applying data mining techniques under different climatic scenario [2, 3].Data mining techniques were used in order to estimate tea yield analysis of four regions of Assam using the multiple linear regression [4].Nowadays machine learning is one of the state of art techniques for predicting unknown values. This paper deals with the prediction of soil moisture using machine learning. The multiple linear regression is used for various applications like stock market prediction [5].Experimental analysis of the ε-insensitive support vector regression technique to soil moisture content estimation from remotely sensed data at field/basin scaleis done which will be useful for satellites or real time applications [6].A multiple regression evaluation method based on correlation analysis can also be used in the field of wireless sensor network [7].Neural networks is used in the prediction of the stock exchange of Thailand[8].Recurrent neural networks is a powerful tool for learning sequential data, like time series data, natural language data, etc. It predicts the output not just on the basis of present inputs but also remembers previous inputs and outputs to better learn the interdependence of inputs.In recurrent neural network an echo state network for the prediction of seed moisture content is used and compared with elman network [9].Other neural network techniques can also be used for the prediction purpose.Overall the idea

given to the farmers could help them in their agricultural practices and even increase their productivity every year. In future along with soil moisture,more parameters such as soil temperature, soil pH could be used in soil health monitoring.

## II. Methodology

### *A. Datasets*

The three different datasets are used for the prediction of soil moisture.

The first dataset used is the Braggs Farm data located in Alabama. It comes under the Natural Resources Conservation Service of the United States Department of Agriculture.The samples are taken from June 2015 to December 2016.In total, there are 569 samples. The samples contain the daily recorded soil moisture at a depth of two inches collected on an hourly basis. For the measurement of soil moisture and soil temperature a hydra probe soil sensor (2.5 volt) having an accuracy of +/− 0.03 wfv ($m^3 \cdot m^{-3}$) and +/− 0.6 degrees celsius (from −10˚C to 36˚C) respectively[10].More details can be seen fromhttps://soilmoisture.tamu.edu [11].

Then our second dataset is taken from TAMU North American soil moisture database which is located in Texas US. The samples are taken from January 2000 to September 2012.In total there are 4749 samples.The samples contain the daily recorded soil moisture a depth of two inches for complete day.More details are available on https:// soilmoisture.tamu.edu [12].

Similarly the third dataset is taken from theOz Net Hydrological Monitoring Network which is an Australian Monitoring Network. It contains the soil moisture data of six different regions. In this paper we have considered the Kyeamba region. The samples are collected for every 20 minutes. For soil moisture measurement, hydra probe sensor was used. Samples are taken from March 2016 toMay 2016.In total, there are 92 samples. More information can be viewed from the website [13].

The complete flow graph of the methodology is shown in fig. 1.


Fig. 1. Flow graph of the methodology

The collected data is normalized using standard scaling technique defined as:

$$X_i = \frac{x_i - \text{mean}(X)}{\sigma^2} \quad (1)$$

All the available dataset is divided into a window of seven, then its mean and standard deviation is calculated .In total there are nine features used as input to the prediction model. Out of these, 80% is used for training the model and rest 20% is used for independent test. Three machine learning techniques such as multiple linear regression, support vector regression and recurrent neural network is used for model development. The prediction is done for 1 day ahead, 2 days ahead and 7 days ahead. The prediction techniques used are listed below.

### *B.Multiple linear regression*

It is a commonly used regression technique in a wide variety of problems. Multiple linear regression assumes linear relationship between the variables and tries to fit all the given data points with a straight line minimizing the residual error. Itis an extended version of the linear regression. As it takes all the data points into account while finding the optimal line, it is more prone to outliers.

It can be explained as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots . \beta_n x_n + e \quad (2)$$

Here the term ε is a random variable whose mean is 0 and variance is $\sigma^2$ and $\beta_0, \beta_1, \beta_2 \ldots . \beta_n$ are the parameters which is needed to estimate.

### *C. Support Vector Regression*

The support vector regression is based on the concept of vectors and since it depends on some data points it is less prone to outliners. For regression task, the loss function is defined such that it ignores the error for the data points which are within a certain distance from the true values. It is called ε-insensitive loss function and is shown in fig. 2.

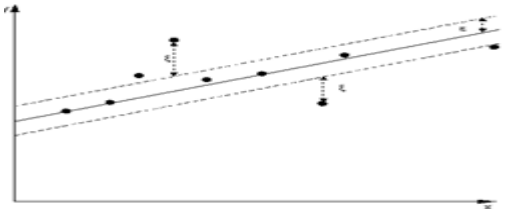

Fig. 2.One dimensional support vector regression with epsilon insensitive region.

The cost is zero for data points lying in that specified region and cost is calculated for points lying outside that region. A linear model is then constructed using the formula:

$$f(x)=w^T x+b \quad (3)$$

Where, w is the weight and b is bias which is needed to be updated whenever we perform the analysis and the formula for its calculation is**:**

$$w= \sum_{i=1}^{n} \beta_i x_i \quad (4)$$

$$b= -1/2(w^T(x_r+x_s)) \quad (5)$$

Where, w is the weight vector and b is the bias vector.

Sometimes during the training process, the hyper-plane formed is not a straight line and hence is called non-linear. The kernel trick concept comes into picture which transforms the non-linear data into linear data. It maps the dataset into the higher dimensional space and then finds the new margin in the feature space.

$$Margin=1/||w|| \quad (6)$$

Where, w is the weight vector.

The support vector regression uses various loss functions. Five among them which are popularly used are Quadratic, Laplace, Huber, ε-insensitive and Quadratic ε-insensitive. Quadratic is a conventional loss function and is not being used nowadays. Laplace is less sensitive to outliers. Huber has no sparseness in it. Considering all loss functions, ε-sensitive was used in the prediction of soil moisture whose formula is given below.

$$L_\epsilon = \begin{cases} 0 & \text{for } |f(x)-y|<\epsilon \\ |f(x)-y|-\epsilon & \text{otherwise} \end{cases} \quad (7)$$

β can then be determined using:

$$\beta=\arg\min_\beta \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\,\beta_j x_i^T x_j - \sum_{i=1}^{n}\beta_i y_i \quad (8)$$

Such that $-C_i<\beta<+C_i$ $i=1,2,\ldots N$ and $\sum_{i=1}^{n}\beta_i = 0$,

WhereC is known as the inverse regularization parameter.

Here the parameters i.e. C (inverse of regularization) and γ (gamma) are varied according to need. C is varied to avoid over-fitting of the model and for γ controls the influence of a single training data. Both C and ε are related with the model complexity but both of them effects in a different way.

*D.*Recurrent neural network (RNN)

These are neural networks which are suitable for modeling sequence of inputs, like speech data, natural language dataetc. They have the same structure as that of a feed-forward neural network or artificial neural network (ANN), but unlike the acyclic nature of ANN, RNN has cyclic connections also. The neurons in a layer can be connected to one another and can even be connected to itself which was not allowed in ANN. Because of this cyclic nature, previous inputs are used to compute outputs at each step and thus, RNNs have a memory of previous events by using which it can make further predictions. The cyclic connections allow previous information to affect the decision-making process. The structure of recurrent neural network is shown in fig.3.

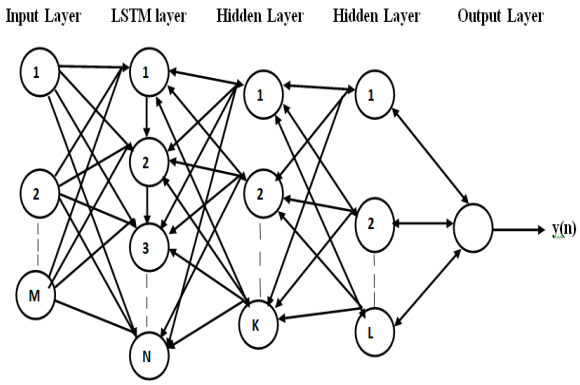


Fig. 3.Structure of Recurrent Neural Network

The input layer has total of 9 neurons as 9 features are there in our dataset which is denoted by M and M=9. The LSTM layers has 50 neurons in it denoted by N and so N=50.In the LSTM layer each neurons are connected to each other. Then comes the first hidden layer which has 25 layer ANN(artificial neural network) denoted by K and hence K=50.Similarly the second hidden layer denoted by L which has total of 12 neurons in it and therefore L=12.Finally all the connections goes to the output layer of the neuron denoted by y(n).

The formulas for calculating the hidden state of LSTM:

$$i= \sigma(x_t U^i + s_{t-1} W^i) \quad (9)$$

$$f = \sigma(x_t U^f + s_{t-1} W^f) \quad (10)$$

$$o = \sigma(x_t U^o + s_{t-1} W^o) \quad (11)$$

$$g = \tanh(x_t U^g + s_{t-1} W^g) \quad (12)$$

$$h_t = h_{t-1} * f + g * i \quad (13)$$

$$s_t = \tanh(h_t) * o \quad (14)$$

Where i,f,o are called the input, forget and output gates respectively.g and h are the hidden states. U and W are the weights of the LSTM. tanh and σ are the activations functions used.

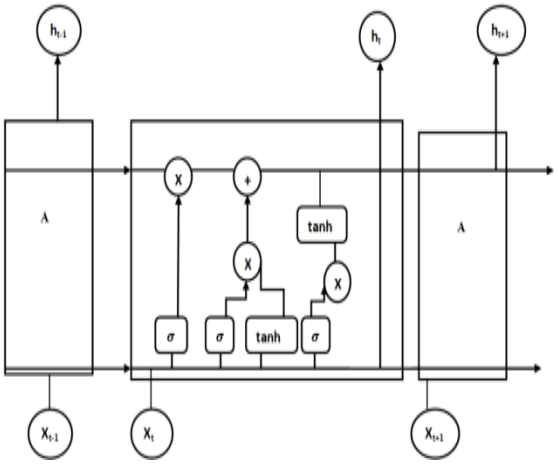


Fig. 4. Structure of LSTMs

The internal structure of LSTM is shown in fig.4.The cyclic connections allow previous information to affect the decision-making process. This model can be better understood by unrolling the neuron in time. But the conventional RNNs have problems of exploding and vanishing gradients which is meticulously countered by a modified version of RNN called the long short term memory (LSTM) networks. They usually work better than the vanilla RNN model and are rapidly being used to achieve excellent results in many complex tasks like image captioning, speech recognition, language modeling, etc.

LSTMs are better suited for long term dependencies and for the purpose of predicting the soil moisture, we will also use LSTMs to build a regression model rather than using vanilla RNN.

*E. Prediction Parameters*

Mean Squared Error (MSE) is the mean of square of errors, i.e. the difference between the true values and the predicted values. It is one of the most commonly used prediction parameter to compare various regression models.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (15)$$

where ,

n = total number of data points.

$y_i$ = true values

$\bar{y}_i$= predicted values

Co-efficient of determination($R^2$)is also a very widely used predition parameter that tells us about the goodness of fit. The value of $R^2 = 1$ tells us that the model perfectly fits the data and $R^2 = 0$ tells that the model unfits the data. It is calculated using the formula :

$$R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}}\right) \quad (16)$$

$$SS_{res} = \sum_i (y_i - \overline{y_i})^2 \quad (17)$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (18)$$

where,

$SS_{res}$ = residual sum of squares

$SS_{tot}$= total sum of squares

$y_i$ = true values , $\bar{y}$ = mean of actual values

## II. Results

The regression techniques used for predicting of soil moisture were MLR (multiple linear regression),SVR(support vector regression) and RNN(recurrent neural network).The comparison results of all machine learning models is listed in table I forBragg's farm data,in table II for Tamu_Namsd data and in table III for Kyeamba data.

Table I. MSE and $R^2$ results of Braggs farm data(Alabama) on test data:

| S. no. | Techniques | MSE(1 day ahead) | MSE(2 days ahead) | MSE(7 days ahead) | $R^2$(1 day ahead) | $R^2$(2 days ahead) | $R^2$(7 days ahead) |
|---|---|---|---|---|---|---|---|
| 1. | Multiple Linear Regression | 0.15 | 0.40 | 1.2 | 0.96 | 0.90 | 0.713 |
| 2. | Support Vector Regression | 0.14 | 0.42 | 1.3 | 0.96 | 0.90 | 0.68 |
| 3. | Recurrent Neural Network | 1.26 | 1.8 | 3.2 | 0.84 | 0.80 | 0.76 |

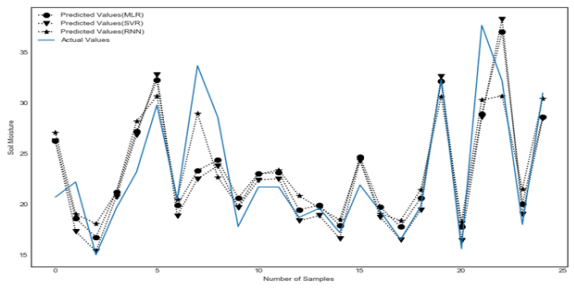


Fig. 5.Performance results of Bragg's farm (Alabama) on the test data.

It can be observed that the MLR has MSE of 0.15 for one day ahead, 0.40 MSE for two days ahead and 1.2 MSE for seven days ahead. Similarly the $R^2$ for one day ahead is 0.96, for two days ahead is 0.90 and for seven days ahead is 0.713.In the same way SVR has the MSE 0.14 for one day ahead,0.42 for two days and 1.3 for seven days ahead and for $r^2$ of 0.96 for one day ahead,0.90 for two days ahead and 0.68 for seven days ahead. In RNN the MSE for one day ahead is 1.26, for two days ahead is 1.8 and for seven days ahead is 3.2 for $R^2$ it is 0.96, 0.922 and 0.76 respectively. Hence it is observed from the tabulated results that multiple linear regression shows better predictive capability in comparison to the other two methods. It is also observed that recurrent neural network is not a good predictive model as its results are inferior to both multiple linear regression and support vector regression.

The tabulation results of other two datasets are listed below.

Table II. MSE and $R^2$ results of Tamu data(North American Soil Moisture Database) on test data:

| S. no. | Techniques | MSE(1 day ahead) | MSE (2 days ahead) | MSE(7 days ahead) | $R^2$ (1 day ahead) | $R^2$ (2days ahead) | $R^2$ (7days ahead) |
|---|---|---|---|---|---|---|---|
| 1. | Multiple Linear Regression | 0.12 | 0.17 | 0.87 | 0.983 | 0.976 | 0.877 |
| 2. | Support Vector Regression | 0.16 | 0.19 | 0.874 | 0.983 | 0.973 | 0.877 |
| 3. | Recurrent Neural Network | 0.89 | 0.952 | 1.2 | 0.88 | 0.876 | 0.8 |

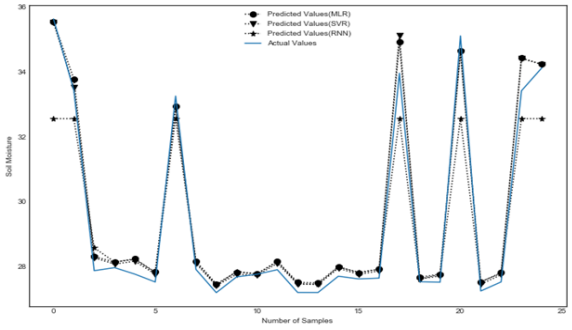


Fig. 6. Performance results of Tamu data(North American soil moisture database) on the test data.

Table III.MSE and $R^2$ results of Kyeamba data(OzNet Hydrological Monitoring Network) on test data:

| S. no. | Techniques | MSE (1 day ahead) | MSE( 2 days ahead) | MSE (7 days ahead) | $R^2$ (1 day ahead) | $R^2$ (2days ahead) | $R^2$ (7days ahead) |
|---|---|---|---|---|---|---|---|
| 1. | Multiple Linear Regression | 0.15 | 0.49 | 2.7 | 0.983 | 0.943 | 0.77 |
| 2. | Support Vector Regression | 0.065 | 0.16 | 3.8 | 0.983 | 0.981 | 0.685 |
| 3. | Recurrent Neural Network | 0.17 | 1.3 | 1.7 | 0.98 | 0.88 | 0.82 |

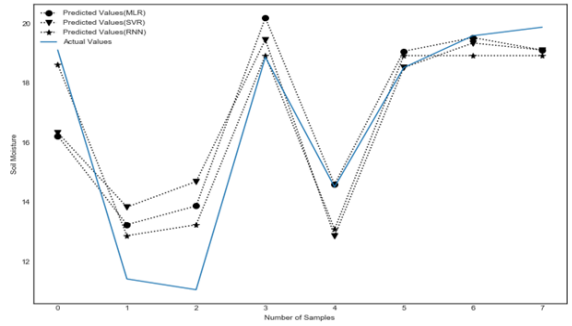


Fig. 7.Performance results of Kyeamba data(OzNet hydrological monitoring network)on the test data.

## IV. Conclusion

In this paper, we introduced machine learning techniques for prediction of soil moisture in advance. We have used multiple linear regression, support vector regression and recurrent neural network for the prediction. From the results it is concluded that multiple linear regression is superior to the support vector regression and recurrent neural network. Although the prediction results are pretty good for 1 day and 2 days ahead but we can try to improve more the 7 days ahead results by applying some other techniques. This will help the farmers to adjust their management strategy beforehand.

## REFERENCES


[1]. Hemageetha N, "A survey on application of data mining techniques to analyze the soil for agricultural purpose,"Computing for Sustainable Global Development (INDIAcom), 3rd International Conference on IEEE, 2016.

[2]. A. Raorane and R. Kulkarni, "Data Mining: An effective tool for yield estimation in the agricultural sector", International Journal of Emerging Trends and Technology in Computer Science, vol. 1, no. 2, pp. 75-79, 2012.

[3]. D. Ramesh and B. Vardhan, "Data mining techniques and applications to agricultural yield data", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 9,pp. 3477-3480, 2013.

[4]. Nivetha, R. Yamini, and C. Dhaya, "Developing a Prediction Model for Stock Analysis,"Technical Advancements in Computers and Communications (ICTACC), 2017 International Conference on IEEE, 2017.

[5]. Rupanjali D. Baruah,R.M. Bhagat,Sudipta Roy ,L.N. Sethi, "Use of data mining technique for prediction of tea yield in the face of climate change of Assam,"India Information Technology (ICIT), 2016 International Conference on IEEE, 2016.

[6]. Luca Pasolli, Claudia Notarnicola, Lorenzo Bruzzone, "Estimating soil moisture with the support vector regression technique,"IEEE geoscience and remote sensing letters vol. 8, no. 6, November 2011.

[7]. Yan, Xiaozhen, Hong Xie, and Wang Tong, "A multiple linear regression data predicting method using correlation analysis for wireless sensor networks," Cross strait quad-regional radio science and wireless technology conference , 2011. Vol. 2. IEEE, 2011.

[8]. Chaigusin, Suchira, ChaiyapornChirathamjaree, and Judy Clayden, "The use of neural networks in the prediction of the stock exchange of Thailand (SET) Index,"Computationalintelligence for modelling control & automation, International Conference on IEEE, 2008.

[9].Elliott, Daniel L, and Russell E. Valentine, "Recurrent neural networks for moisture content prediction in seed corn dryer buildings," Tools with Artificial Intelligence (ICTAI),23rd IEEE International Conference on IEEE,2011.

[10]. Bushra Zaman, Mac McKee,"Spatio-temporal prediction of root zone soil moisture using multivariate relevance vector machines,"Open Journal of Modern Hydrology, 2014.

[11]. https://wcc.sc.egov.usda.gov/nwcc(Braggs Farm dataset).

[12]. https:// soilmoisture.tamu.edu (Tamu _North American Soil Moisture Dataset).

[13]. https:// www.oznet.org.au (Kyeamba Dataset, OzNet Hydrological Monitoring Network).