Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series

Sk Al Zaminur Rahman
Computer Science and Engineering Discipline
Khulna University
Khulna, Bangladesh
E-mail: alzaminurr@gmail.com

Kaushik Chandra Mitra
Computer Science and Engineering Discipline
Khulna University
Khulna, Bangladesh
E-mail: kaushik30208@gmail.com

S.M. Mohidul Islam
Computer Science and Engineering Discipline
Khulna University
Khulna, Bangladesh
E-mail: mohid@cse.ku.ac.bd

***Abstract*—Soil is an important ingredient of agriculture. There are several kinds of soil. Each type of soil can have different kinds of features and different kinds of crops grow on different types of soils. We need to know the features and characteristics of various soil types to understand which crops grow better in certain soil types. Machine learning techniques can be helpful in this case. In recent years, it is progressed a lot. Machine learning is still an emerging and challenging research field in agricultural data analysis. In this paper, we have proposed a model that can predict soil series with land type and according to prediction it can suggest suitable crops. Several machine learning algorithms such as weighted *k*-Nearest Neighbor (*k*-NN), Bagged Trees, and Gaussian kernel based Support Vector Machines (SVM) are used for soil classification. Experimental results show that the proposed SVM based method performs better than many existing methods.**

***Keywords*—Soil series, Land type, Chemical feature, Geographical attribute, machine learning**

I. INTRODUCTION

Data mining means identifying hidden patterns from large datasets and establishing a relationship among them to solve the problem through data analysis [1]. Introduction of data mining in agricultural field has made benefits in research field. Classification is very important in any field of science to establish the fundamentals. It can help finding the diversity between the objects and concepts. It also provides necessary information through which research can be made in a systematic manner [2]. Soil is one of the key components in agricultural field for yielding crops. Soil classification philosophies follow the existence knowledge and practical circumstances. On the land surfaces of earth, classification of soil creates a link between soil samples and various kinds of natural entity [3].

There are about 500 soil series in Bangladesh which is identified by Soil Resources Development Institute (SRDI). Soil series means group of soils which is formed from the same kind of parent materials and remains under the similar conditions of drainage, vegetation time and climate. It also has the same patterns of soil horizons with differentiating properties. Each of the soil has different names and it is named after its locality (e.g. Barisal series, Sara series, Isshwardi series, etc). In Bangladesh, it is a starting point of soil classification to create a platform for its correlation with international soil classification systems (FAO or USDA-United States Department of Agriculture) [4]. Soil series are given names after their locality for having a convenience so that a particular series can be differentiated with another. For instance, 'Barisal Series' does not mean that the category of that series only found in the region of Barisal city, Upazilla or district. It defines the properties of 'Barisal Series' soils and it is applied to all soils those have same kinds of properties. Here the series names are entirely a label [5].

Ramesh et al. [6] used dataset, collected from Soil Science & Agricultural department, Kanchipuram and National Informatics Centre, Tamil Nadu, India. They used random forest, Bayes Net, Naïve Bayes, J48. The dataset consists of 2045 samples of ten types of soil. Gholap et al. [7] used soil datasets from three regions (Khed, Bhor and Velhe) of Pune district, India. Dataset has total 1988 instances with 9 attributes. They focus on applying various algorithms such as Naïve Bayes, JRip, J48 (which is an open source java implementation of the C4.5 decision tree algorithm) for classification task. Devi et al. [8] used several classification algorithms like *k*-means, Random Tree and Apriori, for classifying the soil and predict the suitable crops. This paper collects datasets from the Agriculture University of Coimbatore district of India which has 250 samples of 32 different places. They used clustering technique to group data, and then they classified the data by the order of soil and places with Random Tree algorithm. Then they have applied apriori Mining process to generate an association rule for finding suitable crops for the specific soil. Ramesh, D. et al. [9] used the process of applying data mining techniques on the agricultural data of East Godavari district of Andhra Pradesh in India. This dataset has four input variables and they are Year, Rainfall, Area of Sowing and Production. The datasets have been collected from Indian Meteorological Department, Statistical Institution, and Agriculture department. They have used *k*-means clustering to form a table which has divided into 4 clusters ranging from 1 to 5 cm from the total number of raining years. Then they have taken minimum, mean and maximum values of those 4 clusters. Then they compared the results of the dataset of sowing and average production by applying *k*-means clustering and Multiple Linear Regression (MLR) techniques.

The main purpose of the proposed work is to create a suitable model for classifying various kinds of soil series data along with suitable crops suggestion for certain areas of certain Upazilla of Bangladesh. We have worked with soil series of six upazillas of Khulna district, Bangladesh. Soil

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

series are recognized by machine learning methods using various chemical features and possible crops for that soil series are suggested using geographical attributes.

The rest of the paper is organized as follows. Section II describes the proposed methodology in details. In section III, experimental resulr are outlined. Consclusion is drwan in section IV.

II. Proposed Methodology

The System architecture of the proosed model is shown in fig. 1.

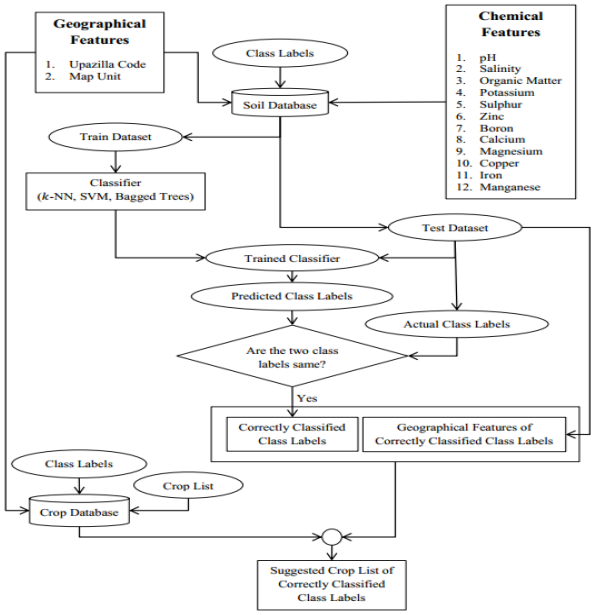

Fig. 1: Proposed System Architecture

The method invoves two phases: training phase and testing phase. Two datasete are used: Soil dataset and crop dataset. Soil dataset contains class labeled chemical features of soil. Table I shows the details of the 12 chemical attributes of soil, used in our method.

TABLE I. CHEMICAL ATTRIBUTES

Attribute	Details
pH	pH value of soil
Salinity	Ds/meter
Organic Matter %	Percentage
Potassium	Mili equivalent/100 gram soil
Sulphur	Microgram/per gram soil
Zinc	Microgram/per gram soil
Boron	Microgram/per gram soil
Calcium	Mili equivalent/100 gram soil
Magnesium	Mili equivalent/100 gram soil
Copper	Microgram/per gram soil
Iron	Microgram/per gram soil
Manganese	Microgram/per gram soil

Soil series and land type combinely reprsents the soil class in the database. The machine learning methods are used to find the soil class (i.e. soil series and land type). Three diffrent methods are used: weighted K-NN, Gaussian Kernel based SVM, and Bagged Tree.

A. Weighted K-NN

A refinement of the *k*-NN classification algorithm is to weigh the contribution of each of the *k* neighbors according to their distance to the query point x_q , giving greater weight w_i to closer neighbors [10]. It is given by

$$F(x_q) = \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i} \quad (1)$$

Where the weight is,

$$w_i = \frac{1}{d(x_q, x_i)^2}$$

In case x_q exactly matches one of x_i so that the denominator becomes zero, we assign $F(x_q)$ equals $f(x_i)$ in this case. It makes sense to use all training examples not just *k* if weighting is used, the algorithm then becomes a global one. The only disadvantage is that the algorithm will run more slowly.

B. SVM

SVM is a supervised machine learning algorithm which works based on the concept of decision planes that defines decision boundaries. A decision boundary separates the objects of one class from the object of another class [11]. Support vectors are the data points which are nearest to the hyper-plane. Kernel function is used to separate non-linear data by transforming input to a higher dimensional space. Gaussian radial basis function kernel is used in our proposed method.

$$K(X_i, X_j) = e^{-\|X_i, X_j\|^2 / 2\varphi^2} \quad (2)$$

Where, $K(X_{i,} X_j) =$ Feature vectors in input space, $\| X_i, X_j \|^2 =$ High dimensional space of X and Y co-ordinate, and σ is a free parameter.

C. Bagged Tree

We have used a bagged decision tree ensemble classifier (consisting of 30 trees). Bagging generates a set of models each trained on a random sampling of the data (Bootstrap resampling) [12]. The predictions from those models are aggregated to produce the final prediction using averaging (shown in fig. 2).

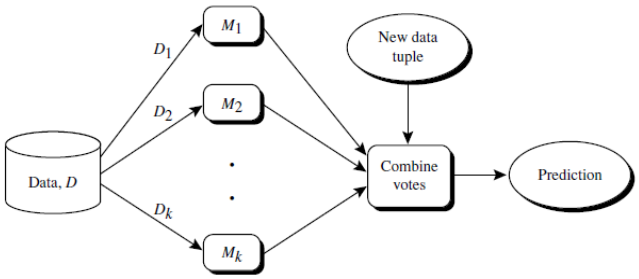

Fig. 2. Ensemble tree

The soil classes that are used in the proposed method of these upazillas are shown in table II.

TABLE II. SOIL CLASSES

Soil Class (Soil Series name with Land Type)	Class Label
Gopalpur High Land	1
Isshwardi Mid High Land	2
Ghior Mid High Land	3
Ghior Mid Low Land	4
Bajoya Mid High Land	5
Barisal Mid High Land	6
Barisal Mid Low Land	7
Harta Mid Low Land	8
Jhalokathi Mid High Land	9
Dumuria Mid High Land	10
Komolkathi Mid High Land	11

Two-third of the samples are used to train the model(s), rest one-third of the samples are used to test the models. For correctly classified samples, crop is suggeted for the corresponding map unit of corresponding upazilla. These two geogrphical features and suggested crop-list makes the crop dataset. The total map units of six upazillas of Khulna district, Bangladesh is shown in table III.

TABLE III. TOTAL MAP UNITS OF VARIOUS UPAZILLAS

Upazilla Name	Upazilla Code	Total Map Units
Rupsha	100	7
Dighalia	200	10
Fultola	300	12
Koyra	400	4
Dakop	500	8
Terokhada	600	9

III. EXPERIMENTAL RESULTS

A. Database creation and attribute selection

In this research, we have worked with soils series of six upazillas of Khulna district, Bangladesh. Upazillas are: 'Rupsha', 'Dighalia', 'Fultola', 'Koyra', 'Dakop', 'Terokhada'. There are total 15 soil series in this 6 Upazillas. In our work, we have worked with 9 soil series; they are 'Gopalpur', 'Isshwardi', 'Ghior', 'Bajoya', 'Barisal', 'Harta', 'Jhalokathi', 'Dumuria', 'Komolkathi'. Combining the land type, we get 23 classes which have 438 samples but we have taken 11 of them (shown in table II), which have 383 samples. We have excluded other classes because they have less than or equal 10 samples while other has large number of samples. That is, class imbalance problem is occurred. To remove this problem, we augmented our dataset by replicating the samples of classes which were under sampled and randomly chose the samples of classes which were over sampled. After doing that, there were 45 samples of each class and total 495 samples of 11 classes. Now by merging the geographical data (from Table III), chemical data (from Table I) and classes (from Table II), we have made our Soil database.

We have created the crop database by taking the Upazilla code, map unit and class label. Different types of crops grow well in different type of lands and soil series of each Upazilla. We have gathered all the information together and sort them by Upazilla code, map unit and class labels. Thus crop database is created. Table IV shows some crop samples that are suitable for various soil classes of specific map units.

TABLE IV. SOME CROP SAMPLES FOR VARIOUS CLASSES, MAP UNITS, AND UPAZILLAS

Class Label	Map Unit	Upazilla Code	Crop List
11	10	200	Rice (Boro,ufshi), Rice (Boro, deshi)
11	9	600	Rice (mixed Bona Aush & Amon)
9	6	600	Seasme (mixed)/Rice (Aush) & Rice (Bona Amon)
17	2	400	Rice (Ropa Amon, local)
11	7	600	Rice (mixed Bona Aush & Amon), Rice (Bona Amon)

B. Result analysis

The proposed method is based on these two databases described above. Several machine learning methods are applied separately to recognize the soil class of test sample. From experimental results, we see that though K-NN and Bagged tree shoes comparative accuracy, but SVM shows better accuracy than other methods used here. The classification accuracy is shown in table V of the following.

TABLE V. RESULT OF THE PROPSED METHOD

	Method	Accurcy (%)
Ramesh, V., et al. [6]	J48	91.90
Gholap, Jay, et al. [7]	J48	92.3
Proposed Work	SVM	94.95

After classifying the soil series, the crops, those are suitable for that series for the given map unit of corresponding upazilla are suggested. Fig. 3 shows some randomly taken correctly classified testing samples.

Sample No.	1	2	3	4	5
Upazilla Label	100	500	300	600	400
Map Unit	1	1	8	3	1
pH Value	7.1	7.9	6.4	6.6	7.8
Salinity	1.3	19	1.3	1.26	1.86
Organic Matter	2.61	1.63	4.48	4.98	1.63
Potassium	0.15	0.29	0.42	0.28	0.32
Sulphur	26.95	350	29.1	34	85.5
Zinc	0.21	0.92	0.44	2.1	0.1
Boron	1.24	1.67	1.09	1.22	0.79
Calcium	18.5	13.8	21.5	26.5	13.5
Magnesium	8.25	5.3	3.75	6.33	6
Copper	4.85	7	7.27	8.8	4.5
Iron	49.12	59	84.2	15	71.8
Manganese	7.55	32	16.39	27.7	80.8
Class Label	1	8	6	7	8

Fig. 3. Randomly taken 5 correctly classified samples of testing set

The relevant crops suggestion for the correctly classified testing sample of ***Gopalpur high land*** class and ***map unit number 1*** of ***Rupsha*** upazilla is shown in fig. 4

Detailed Croplist Suggestion for Goplapur Hoghland									
Upazila Code		100							
Upazilla Name		Rupsha							
Map Unit		1							
Class Label		1							
Class Name		Gopalpur High Land							
Appropriate Crop (Without Irrigation)	Rabi Crops	1.Barley	2.Corn	3.Cotton	4.Tomato	5.Brinjal	6.Onion	7.Green Chilli	8.Sweet Potato
	Kharif Crop 1	1.Rice (Bona Aush)	2.Corn	3.Bitter Gourd	4.String Beans	5.Pointed Gourd			
	Kharif Crop 2	1.Rice (Ropa Amon)	2.Pumpkin	3.Red Amaranth	4.Spinach				
A.Yearly Crop		A.	1.Banana	2.Papaya	3.Ginger	4.Turmeric	5.Sugarcane		
B.Long lived Crop		B.	1.Lichi	2.Guava	3.Mango	4.Lemon			
Appropriate Crop (With Irrigation)	Rabi Crops	1.Corn	2.Potato	3.Cotton	4.Okra	5.Tomato	6.Kalabash	7.Cabbage	8.Cauliflower
	Kharif Crop 1	1.Corn	2.Snake Gourd	3.Ridge Gourd	4.Bitter Gourd	5.Okra	6.Brinjal		
	Kharif Crop 2	1. Rice (Ropa Amon)							
A.Yearly Crop		A.	1.Banana	2.Papaya	3.Sugarcane				
B.Long lived Crop		B.	1.Lemon						

Fig. 4. Detailed crop list of Gopalpur high land class and map unit no. 1 of Rupsha upazilla

There is multiple distribution of crop suggestion under the condition of without irrigation and with irrigation. Rabi crops, Kharif crops 1, Kharif crops 2 are one kind of distribution of crops. Yearly crops and long lived crops are another type of distribution. Similar type of detailed crop suggestion of each class with specific map units and Upazilla are exists. Table VI of the following shows comparison result with some existing methods for soil classification. It shows that the proposed SVM based methods outperform than other methods mentioned here.

TABLE VI. PERFORMANCE COMPARISON

Classification Model	Accuracy (%)	Average Accuracy (%)
Gaussian SVM	94.95	92.93
Weighted k-NN	92.93	
Bagged trees	90.91	

C. *More findings*

We have calculated the mean and variance for twelve chemical attributes of various samples of soil database in order to find the characteristics of data. We have done that with respect of each class. We have found that for different kinds of soil, the number of independent attributes does not match with each other. Thus we can conclude that it cannot normally find any linear feature that can classify the soil series without using any machine learning algorithm.

IV. Conclusion

A model is proposed for predicting soil series and providing suitable crop yield suggestion for that specific soil. The research has been done on soil datasets of six upazillas of Khulna region. The model has been tested by applying different kinds of machine learning algorithm. Bagged tree and K-NN shows good accuracy but among all the classifiers, SVM has given the highest accuracy in soil classification. The proposed model is justified by a properly made dataset and machine learning algorithms. The soil classification accuracy and also the recommendation of crops for specific soil are more appropriate than many existing methods. In future, providing fertilizer recommendation is our concern, also data of other districts will be added to make this model more reliable and accurate.

Acknowledgement

We are grateful to Soil Resource Development Institute (SRDI), Government of the people's republic of Bangladesh, for providing valuable soil data of various upazillas of Khulna district, which helps us to conduct our experiment.

References

[1] https://www.techopedia.com/definition/1181/data-mining. [Accessed date:19th August, 2018]

[2] http://www.library.arizona.edu/exhibits/swetc/azso/body.1_div.6.html . [Accessed date:19th August, 2018]

[3] Eswaran, H., Ahrens, R., Rice, T. J., & Stewart, B. A. (2002). *Soil classification: a global desk reference*. CRC Press..

[4] http://en.banglapedia.org/index.php?title=Bangladesh_Soil. Last Accessed date: 25th August, 2018.

[5] http://en.banglapedia.org/index.php?title=Soil_Series. Last Accessed date: 25th August, 2018

[6] Ramesh, V. and Ramar, K., 2011. Classification of agricultural land soils: a data mining approach. *Agricultural Journal*, *6*(3), pp.82-86..

[7] Gholap, J., Ingole, A., Gohil, J., Gargade, S. and Attar, V., 2012. Soil data analysis using classification techniques and soil attribute prediction. *arXiv preprint arXiv:1206.1557*.

[8] Devi, M. P. K., Anthiyur, U., & Shenbagavadivu, M. S. (2016). Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining. *International Journal of Modern Computer Science*, *4*(6).

[9] Ramesh, D., & Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. *International journal of advanced research in computer and communication engineering*, *2*(9), 3477-3480.

[10] KNN algorithm, http://www.data-machine.com/nmtutorial/ distanceweightedknnalgorithm.htm.

[11] SVM, https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html.

[12] Bagging, https://medium.com/@harishkandan95/bagging-the-skill-of-bagging-bootstrap-aggregating-83c18dcabdf1.