Knowledge based Soil Classification Towards Relevant Crop Production

Waleej Haider[1, *], Aqeel ur Rehman[3]
Department of Computing
Hamdard University
Karachi, Pakistan

M. Nouman Durrani[2]
Department of Computer Science
FAST-NUCES
Karachi, Pakistan

Sadiq ur Rehman[4]
Department of Computing
Hamdard University
Karachi, Pakistan

Abstract—**Pakistan's economy is strongly associated with agriculture sector. For a country having 25 % of GDP contributed through agriculture, there is a need to modernize the agriculture by acclimatizing contemporary approaches. Unfortunately, it has become a common trend among farmers to cultivate crops, being used in food items or which can easily be sold out in the market without using knowledge about the suitability or relevancy of crops to the soil environment. Consequently, the farmers face financial losses. Many researchers have proposed soil classification methods for various soils related researches, but they have very little contribution towards guidance of the farmers to select most suitable crops for cultivation at a particular soil type. Without the use of technology and computer-assisted approaches, the process of classifying soil environments could not help the farmers in taking decisions regarding appropriate crop selection in their respective fields. In this paper, an effective knowledge-oriented approach for soil classification in Pakistan has been presented using crowd sourced data obtained from 1557 users regarding 103 agricultural zones. The data were also obtained from AIMS (Govt. of Punjab) and Ministry of National Food Security & Research. In this work, random forest classifier has been used for processing and predicting complex tiered relationship among soil types belonging to agricultural zones and major suitable crops for improving yield production. The proposed model helps in computing the degree of relevancy of crop to agricultural region that help former selecting suitable crops for their cultivated lands.**

Keywords—Knowledge creation; agriculture; soil classification; random forest; knowledge distribution; crop relevancy

I. INTRODUCTION

The agriculture has a significant impact on the overall GDP of agricultural countries. As an agricultural country, 25% of Pakistan's GDP is contributed through agricultural revenue. As 2% per year increase in the country's population has given birth to new requirements of: (i) advanced crop production methods and (ii) growing new crop categories over the existing land beside traditional agriculture. Knowing suitable crop or seed type for a particular soil type, may increase the yield. In this work, through crowd sourcing, the knowledge of agricultural experts has been used for the assignment of a crop (or a seed type) to a soil type for helping farmer growing new crops over the existing agricultural lands.

Detailed understanding of soil nature can help in improving overall efficiency of crop productions. Unfortunately, farmers are exploiting customary farming ways which do not tend to meet upcoming agricultural requirements. In this regard, the role of soil environment management is very critical. Further, weather, climate impact, time and type of crop are also very important. Cultivation without having knowledge of appropriateness of context is perilous. In connection to this, current research [1], [2] indicates that researchers, farmers and government officials need to work together for the adequate solutions related to said issues. Moreover, training sessions for the farmers and upholding their preceding experiences related to the crops can help them making efficient decisions for the subsequent crops.

Researchers have proposed methods of classifying soil based on environmental conditions i.e. temperature, moisture, humidity, pH, fertility of soil, mode of formation or structure/texture of soil but these methods could not sufficiently paid attention towards the suitable crop selection criteria [3]–[5].

Modern trends also tend to apply artificial intelligence, machine learning and deep learning mechanisms for the solutions of agricultural problems. One of such solutions expels the use of ML based classifications of soil environment. An effective approach has been proposed for classification of soil environment in Pakistan integrating crowd sourced data, machine learning with local knowledge of experts to take knowledge-based decisions for suitable crop selection.

This paper presents a generic approach for classifying soil types of Pakistan using crowd sourced data. Crop data have been obtained from various regions of Pakistan and worked on crop to soil mapping rules using proposed model, which are further verified by the experts. This model also generates knowledge of soil-crop mapping for those agricultural regions where relevant data is partially available. This knowledge-based classification of soil, mapping to suitable crops, results verification by the experts and crowd sourcing have enhanced the accuracy of the proposed model, provide new directions to the farmers and help in increasing crop production. In this paper, we've also worked on calculating the crop relevancy to different soils that helps in creating knowledge about crop production at various soils and best suitable business-oriented crop selection.

Remaining of the paper has been organized as: literature review of multiple dimensions has presented in the next Section II. Section III contains general categories of soil in Pakistan using different parameters and analytical study about

---

*Corresponding Author

cultivation of crops on the soil types. Experimental setup is given in Section IV, which is interlinked with Sections V and VI. These sections cover the implementation of the algorithm and discusses the results of different experiments respectively.

II. Existing Systems

The concept of soil classification for different reasons is not new but the use of technology has altogether changed the meaning of classifying soil. It changed the previous ways based on testing soil samples of an agricultural land in the laboratories for examining its fertility, mode of its formation, its textures or existence of different chemicals in soil by which it can be classified. Although, these samples could not represent the whole land, so, the classification could not be performed on the huge landscape using these methods. These approaches based on sample testing are time consuming, have a high frequency of errors and hard to cover maximum agricultural region to collect soil samples to be examined. Evolution of agricultural decision support systems is not ended at computer-assisted solutions, but improved by knowledge-based computerized approached. A comprehensive study covering all the aspects has been presented in this section.

A. Existing Crop Selection Approaches

Researchers have proposed methods for selecting the best crops to be cultivated in the fields. Valipour et al. have presented a model in [6] to evaluate potential evapotranspiration in a specific climate or in a particular agricultural region. Four climate types including arid, mediterranean, semiarid and lastly very humid have been considered. A multidimensional and chronological study of ETo has been done and climatological data of around 50 years (1961 to 2010) of 18 agricultural regions of Iran with several climates on a monthly basis were collected. To estimate ETo, temperature, radiation, mass transfer bases models have selected to check, which model is performing better in a particular climate. It is analyzed that radiation based models were easily adjusted better with climate change [6]. Gornott et al. have tested three regression models for the impact of climate variation on the change of crop yields production. They have focused two major crops, wheat and maize for this analysis [7].

B. Classification of Soil

The soil classification has been done by the researchers using different approaches like i) environmental conditions i.e. temperature, moisture, humidity and pH, were considered, ii) fertility of soil [2], [8], [9] was used as the basic measure for the soil classification, iii) mode of formation of soil iv) structure/texture of soil [5], [10], [11] v) regional or land cover basis [12]–[15], [14], [10], [16], [17] or based on vi) existence of chemicals in the soil [8], [9] were used for the soil classification.

Current computer-assisted methods have been used for classification in the agricultural domain using various types of statistical, AI and ML algorithms like decision trees, random forest and Convolutional Neural Networks (CNN) [18] [19]–[21]. An approach has been presented to modernize large scale, local soil maps using three sourced data. i) used old soil shape classifications, ii) obtained data from visible near infrared classifications spectroscopy, and iii) digital soil class mapping. Using the combined data, Australian Soil Classification maps have been updated. The total error rate was 55.6% after testing the model on independent validation set [3]. Random forest classifier integrated with multivariate regression for classification and to quantify some quality parameters of the soil [11]. Digital soil mapping has been used to update the national map of the soil Drenthe province without extra fieldwork. Multinomial logistic regression has been used on a legacy soil data to measure the association between auxiliary variables and different soil group [13].

Researchers have encouraged towards the design of knowledge oriented methods [22], [23]. This paper deliberates the local knowledge of regular farmers to grow crops, according to soil quality. Interviews of farmers were conducted and most of them have two or more soil types in their agricultural lands. Leigh A et al. have conducted survey and interviewed to realize how farmers living in the remote areas of southeastern Costa Rica are using knowledge of soil for proper allocation and management of crops and to discover how to integrate native and technical knowledge to select sustainable agricultural approach.

III. Context of Soil Type used for Production

Soil can be classified on the basis of their 1) structures and texture 2) regional basis 3) climate 4) vegetation 5) Mode of formation [13], [24] as sown in Fig. 1. In this section, an empirical study about soil types, their existence in different regions, commonly growing crops and the impact on yield obtained has been presented.

1) Variances in structure and texture: There are five main types of soils in Pakistan on the basis of variance in structure and texture as a) clay, b) sand and gravel, c) silt, d) loam and e) organic soil which is less common.

a) Clay soil: Clay soil, composed of smooth and gluey particles and available red, bluish gray, yellow or almost black colors. Since humid nature, it warm and dry up slowly and batter in storing reserves of nutrients it requires regular irrigation and in case of irregular watering, it become very hard during dry and warm weather [11]. In this situation, water absorption ability of the soil become very low and the sprinkling method cannot fulfill the sufficient requirements. Only furrow method can be used in this condition. Cultivators are excavating large amount of organic matters like well-rotted manure compost, peat moss or simply Gypsum in the clay soil to cover the hardness before the sowing stage. This soil is useful for growing wheat, which is one of the major crops being used for food products, gram and paddy crops that is largely cultivated in Pakistan. Clay exists in most of the regions of Punjab, Sindh and KPK provinces where major crops like wheat, rice and sugarcane are cultivated. Unlike other regions, Thatta, Badin zones are not focusing on wheat, which is more suitable for clayey soil. Some regions of Punjab and Sindh are getting good production of cotton and maize. On the other side, according to crowd sources, these zones like Thurparker, NoderoFeroz, Nawabshah, Mirpur Khas, etc. are being used for caster, guar and millat which are not most

suitable for Clay soil nor these give sufficient business to the farmers.in the same way some agricultural zones of the KPK province like Peshawar, Noshehra, Mardan and Charsada are growing pears, sugar beet and plum in clayey soil but they can produce tobacco and millet to get more business.

b) Sand and gravel soils: Sand and gravel soils are dominating particles in this type. Cultivation process is very easy on this soil, but it becomes warm and dry up and scattered very quickly [25]. So, its ability to absorb heat may damage vegetables. Pure sandy and gravelly soils have less nutrients and fertility, but mostly these types contain sufficient clay particles which make the soil suitable for cultivation and receptive to fertilizers [3].

Moreover, the quality of this soil can be improved by taking out the larger pebbles and stones and by mixing animal manure or some plant waste with this soil in the fields.

These soils are present in southern and northern irrigated plains and sandy deserts of Sindh and Punjab provinces. Sandy soil is useful in growing vegetables, which thrives in the soil. Wheat, millet, guar are being cultivated in sandy deserts of Sind. Castor is grown in some regions which is less profitable. They can grow melons, maize or oilseeds as in the northern irrigated plains of Punjab province. Same crops are being practiced in sandy deserts of Punjab along with cotton and sugarcane. Sandy and gravel soil do not provide good production of sugarcane but people are traditionally growing this crop.

c) Silt Soils: Silt soils have a transitional size amongst clay and sand with small size gritty particles which make it very hard [25]. Silt present in the soil in a variety of colors like gray, red tan, and yellow. It's usually not very fertile. There is a thick layer of clay at the top, so normally it is less fertile and have low ability of water absorption. This thick layer should be broken by mixing a sufficient amount of peat moss, rotted sawdust, compost, and wood shavings and afterwards, adding organic matter can recover the construction and fertility of this soil type. It is useful for root vegetables and pulses and berseem according to experts from the crowd. This soil exists in southern and northern irrigated plains where mustard, and berseem are being cultivated with sorghum. Regions in the Indus delta like Thatta and Badin are using silty soil for banana fruit and berseem. Pulses and many other crops can give more business to farmers. Wet mountains are used for growing apples with other major crops and people of Barani land near the Punjab province are harvesting fodder crops for animal feed. Mittet, pulses and oilseeds are also being grown there with other major crops. Some farmers are cultivation rice in silty soil which is not relevant to this soil and could not provide more yield. Agronomists from the dry western plateau are growing melons and sorghum as major crops, but some are focusing millet which is less beneficial

d) Loam Soils: Loam contains clay, sand and silt in the different proportions along with organic matters [26]. Various proportions of clay, silt, sand, and organic matter; the magnitudes of these defines the quality, productivity and behavior of the soil towards the cultivation. This soil is best for major crops and vegetables. Pure loamy soil exists in southern

irrigated plains i.e. From Hyderabad to Jacobabad where agriculturists are growing sugar cane, sorghum and mustard with other major crops, sandy desert of Sind (Tharparkar to Cholistan) and Punjab (Muzaffar Gharh to Khushab) provinces where wheat, guar and millet are majorly cultivated and some farmers are growing gram and sugar cane in Punjab whereas caster in Sindh. Cotton, mustard, maize and vegetable are also suitable crops for loamy soil. Farmers can earn growing these crops instead of caster and berseem. Regions except Indus delta, northern irrigated plains and dry mountains and Suleiman piedmont, have mixture of loamy and other soils which are fertile and useful for most of the major crops.

e) Organic Soil: Organic Soil is dark in color composed of big amount of peat moss and or leaf mold. It warms up gradually due its moisture retaining capability, but is easy to water, weed and work [11]. It is micronutrient deficient, but the use of special fertilizers cover-up the scarcities.

To analyze the yield production of crops on various soil types and the degree of irrelevancy of crops to the soil environment, knowledge has been created by processing data using statistical models and algorithms have been proposed to calculate the degree of irrelevancy of target crops to the available soil types.

2) Modes of formation: Drought is fundamentally affecting the soil in Pakistan as it is major characteristics of climate which result in shortage of soil moisture. Soil in Pakistan is composed of two parent materials[10] , a) Alluvium, Loess and wind reworked sands, b) Residual material gained from enduring of underlying rocks. The soils of Pakistan have attained different features from the parental material and based on their way of formation. The river-laid residues have established into Muddy Soils. The desert sands have also revolved into a diverse type of soils. The mountains, hills and other highlands have formed Residual Soils with spots of Grainy, loss and other material of the soils. Consequently, the soils of Pakistan have been classified into six types on the basis of its ways of construction or modes of formation as shown in Fig. 1.

3) Soil types according to spatially placed regions: Some fundamental materials may transfer from one place to another through rivers and wind and placed over the surface are called alluvium and Aeolian Soils respectively. This provides a base to another soil categorization according to spatially placed regions.

The Indus plain soil is made-up with the shifting of alluvium by the Indus river and its branches. Rivers are dropping the soil on the earth's surface in the form of layers for the last thousands of years. This soil has sufficient calcium carbonate, lower level of organic matter and further divided into three types. Bangor soils are present over the vast area of the Indus plain like many agricultural portions of the Punjab, Mardan, Kachhi plain, Peshawar and Bannu along with a major part of Sindh province. These soils somehow provide better production. Khaddar soils are present in the nearby regions of the rivers. A layer of salt clay is dropped on the surface and due to this yearly change, this soil has less gratified of salt and

organic material. Indus Delta *Soils* present across the river Indus Delta. It maintains its level from Hyderabad and exists in the southern coastal areas. These areas are normally affected by flood so it developed under this floodwater and most of the soil is clay by structure, hence due to its ability to absorb water, this is majorly suitable for cultivation of rice crop. Northern and western zones of Pakistan are highlands, so known as *mountain soil* [25]. The climate of northern mountainous lands is moist and these have large amount of organic matter contents. On the other hand, soils of western mountainous lands are deficient in the organic matter due to dry or semi-arid climate and have a large amount of calcium carbonate. The soils of Pothwar hills have great sea green content. If the sufficient amount of water is provided, this soil become very productive.

Sandy Desert Soils exists in the Cholistan, western side of Balochistan and the desert of Thar. Its structure contains a layer of sands elements and has the medium amount of calcium carbonate. Dry and semi-arid climate [15] effect this type of lands and forcefully disposition the layers of sandy soil. Consequently, this soil type is structured. This is the reason of less economic activities in the deserts.

An effective analysis of three target crops on all available soil types in three target regions has been presented in Table I. An algorithm to compute the degree of crop relevancy has been developed and the results are discussed in detail in Section VII, case-II.
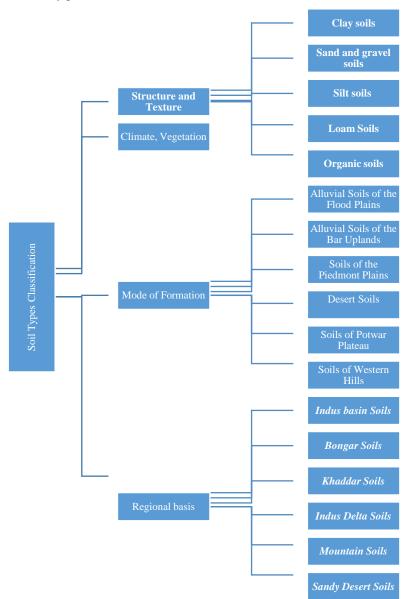

Fig. 1.   Hierarchical Structure of Soil Types Classification in Pakistan.

TABLE. I.   KNOWLEDGE RESULTANT FROM CROWD SOURCING ABOUT CROP PRODUCTION, INCOME AND RELEVANCY OF CROP

Targeted Regions	Soil Type	Texture & Structure	Water Retention Capability	Target Crops	Business Impact	Degree of Irrelevancy (in %)
Faisalabad	Clay	Smooth , sticky , rock and gluey	High	Rice	Moderate	77.25
				Maize	High	0
				Cotton	Moderate	72.70
	Sandy	Coarse	Low	Rice	low	79.50
				Maize	High	0
				Cotton	Moderate	67.84
	Loam	Mixed	Moderate	Rice	Moderate	72.88
				Maize	High	0
				Cotton	Moderate	67.30
	Silt	Shallow residual and Gritty	Moderate-high	Rice	Moderate	66.73
				Maize	High	20.47
				Cotton	High	0
	Organic	Peat moss or leaf mold	Moderate	Rice	Moderate	85.26
				Maize	High	0
				Cotton	Moderate	74.93
Multan	Clay	Smooth and gluey	High	Rice	Low	93.98
				Maize	Low	93.17
				Cotton	High	0
	Sandy	Gritty	Low	Rice	Low	93.65
				Maize	Low	92.87
				Cotton	High	0
	Loam	Mixed (variable)	Moderate	Rice	Low	90.44
				Maize	Low	92.57
				Cotton	High	0
	Silt	Gritty	Moderate-high	Rice	Low	95.87
				Maize	Low	98.85
				Cotton	High	0
	Organic	Peat moss or leaf mold	Moderate	Rice	Low	95.13
				Maize	Low	97.40
				Cotton	High	0
Bahawalpur	Clay	smooth and gluey	High	Rice	Low	97.84
				Maize	Low	94.58
				Cotton	High	0
	Sandy	Gritty	Low	Rice	Low	98.06
				Maize	Low	94.32
				Cotton	High	0
	Loam	Mixed (variable)	Moderate	Rice	Low	97.50
				Maize	Low	94.29
				Cotton	High	0
	Silt	Shallow residual and Gritty	Moderate-high	Rice	Low	99.69
				Maize	Low	99.46
				Cotton	High	0
	Organic	Peat moss or leaf mold	Moderate	Rice	Low	99.54
				Maize	Low	96.71
				Cotton	High	0

IV. MATERIALS AND METHODS

Before performing experiments, data from the crowed and online sources has been obtained and the algorithm has been selected for data processing as discussed in the next sections. Workflow of the whole classification process is shown in Fig. 2.

A. Data Acquisition

Data from volunteer users [27] belonging to all over the Pakistan has been obtained and organized. This dataset contains landscape, soil types, agricultural zones or regions and most suitable crops being cultivated in the zones. Around 1600 users has shared the data from 103 different agricultural zones of Pakistan based on four features as shown in the Table II. Another dataset was acquired from the Institute of Soil & Environmental Sciences, University of Agriculture, Faisalabad.

B. Preparation of Data Set

Input data are further processed by testing relativity by correlation, sampling and data cleaning by testing the comprehensiveness and setting missing values in data records. Observations containing missing values have been removed from training dataset. The data has been managed to meet the algorithm requirements as columns of the Table II shows the features of data and rows represents the records. Environmental factors of the soil are characterized by predictors of the environment of any agricultural region R. Finally, the output of the whole process is a function of defined environmental predictors as:

$$\text{Output (R)} = f(st[r], c[r], mc[r], l[r], az[r], ct[r]) \qquad (1)$$

For any agricultural region R, representation in the above equation are soil type (st), climate (c), major crops (mc), landscape (l), Agricultural zones (az) and context (ct). Random forest has been used to map four of these factors to output. The data has been divided into two training sets. Two predictor variables from the factors of soil environment initially used in experimentation are, i) soil types and agricultural zones in the first training data and ii) other two predictors, soil types and major crops in the second training datasets which plays a vital role in the classification as presented in Tables III and IV along with statistics of the training data set. Sandy, Clayey exists in 73 agricultural zones which is 16.25 % of the total soil types of the regions. Calcareous, Sandy, Loamy, Silty and Sandy, Loamy are less common in Pakistan. While obtaining the crowd sourced data, there are possibilities of sharing inaccurate, mixed or incomplete data by users. To remove these insufficiencies, this data has been sent to expert community where domain experts are available to rectify the data as depicted in Fig. 2.

This process makes the data cleaner and ultimately improves the accuracy of the results. The sample data of 20 users are shown in the Table II. To produce the clarity in the results, two random forests have been driven using these sets of predictors as discussed in Section VI.

TABLE. II. CROWED SOURCED DATA OF SOIL ENVIRONMENT OF PAKISTAN

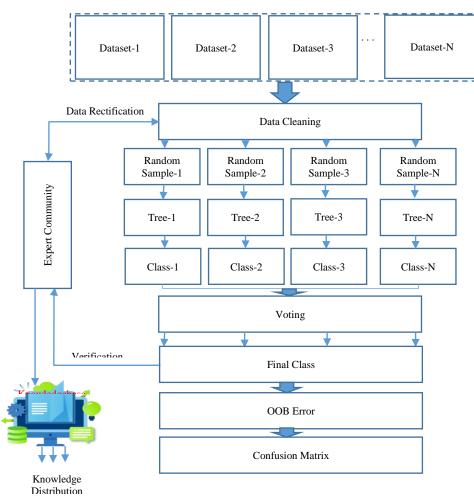| User ID	Landscape	Agricultural Zones	Soil Types	Major Crops
1	Western Dry Mountains	Pishin	Calcareous loamy	Apples, Peaches
2	Northern Irrigated Plains (a)	Jhang	Calcareous, Silt-loam	Millet, Maize
3	Southern Irrigated Plains	R.Y. Khan	Calcareous, Loamy	Sorghum, Berseem
4	Wet Mountains	Battagram	Silt loam, Silty clays	Wheat, Apples
5	Sandy Desert (a)	Cholistan	Clayey and Loamy	Guar, Castor
6	Sandy Desert (a)	Nawabshah	Clayey and Loamy	Wheat, Castor
7	Barani Lands	Bhakkar	Silt loam, Silty clay loam	Millet, Fodder crops
8	Western Dry Mountains	Jaffarabad	Calcareous loamy	Peaches, Plums
9	Southern Irrigated Plains	Shikarpur	Silty, Clayey	Berseem, Rice
10	Dry Western Plateau	Chagai	Strongly Calcareous, Silt loams	Wheat, Sorghum
11	Sandy Desert (b)	Layyah	Sandy, Loamy	Cotton, Guar
12	Barani Lands	Lakki Marwat	Silty clay loam, Clay loam	Maize, Oil seeds
13	Northern Irrigated Plains(b)	Charsadda	Clayey, Moderately Calcareous	Maize, Gram
14	Dry Western Plateau	Lasbella	Strongly Calcareous, Silt loams	Wheat, Sorghum
15	Southern Irrigated Plains	Sanghar	Calcareous, Silty	Berseem, Rice
16	Dry Western Plateau	Khuzdar	Strongly Calcareous, Silt loams	Wheat, Sorghum
17	Sandy Desert (b)	Khushab	Calcareous, Loamy	Gram, Millet
18	Sandy Desert (a)	Tharparkar	Sandy, Clayey	Wheat, Castor
19	Barani Lands	Bannu	Silty clay loam, Clay loam	Maize, Oil seeds
20	Western Dry Mountains	Quetta	Calcareous loamy	Apples, Peaches

Fig. 2. Workflow of Proposed Approach for Classification of Soil Environment of Pakistan.

TABLE. III. TREE STRUCTURE PREDICTORS, SOIL TYPES AND AGRICULTURAL ZONES

Nodes	Objects	%	Test statistic	p-value	Purity	DF	Split variable	Values	Parent node	Sons	Predicted values
Node 1	449	100.00%	1677.500	< 0.0001	4.68%	174				2; 3; 4; 5; 6; 7; 8	Badin
Node 2	41	9.13%			51.22%		Soil Types	Clayey and Silty	1		Badin
Node 3	70	15.59%	15.556	0.049	14.29%	8	Soil Types	Calcareous, Loamy; Clayey and Sandy	1	9; 10	Hyderabad
Node 4	85	18.93%	31.481	0.000	17.65%	8	Soil Types	Silty, Clayey; Loamy, Silty; Calcareous, Silty	1	11; 12	Sanghar
Node 5	73	16.26%			6.85%		Soil Types	Sandy, Clayey	1		Tharparkar
Node 6	60	13.36%			16.67%		Soil Types	Clayey and Loamy; Sandy and Loamy	1		Tharparkar
Node 7	40	8.91%			25.00%		Soil Types	Calcareous, Sandy; Sandy, Loamy	1		Muzaffaragarh
Node 8	80	17.82%			12.50%		Soil Types	Clayey, Calcareous; Calcareous, Silt-loam	1		Bahawalnagar
Node 9	45	10.02%			11.11%		Soil Types	Calcareous, Loamy	3		Hyderabad
Node 10	25	5.57%			20.00%		Soil Types	Clayey and Sandy	3		Hyderabad
Node 11	45	10.02%			11.11%		Soil Types	Silty, Clayey	4		Hyderabad
Node 12	40	8.91%			25.00%		Soil Types	Loamy, Silty; Calcareous, Silty	4		Sanghar

TABLE. IV. TREE STRUCTURE PREDICTORS, SOIL TYPES AND MAJOR CROPS

Nodes	Objects	%	Test statistic	p-value	Purity	DF	Split variable	Values	Parent node	Sons	Predicted values
Node 1	199	100.00%	447.503	< 0.0001	22.61%	24				2; 3; 4; 5	Silty, Clayey
Node 2	41	20.60%			100.00%		Major Crops	Rice, Pulses; Berseem, Banana;Pulses , Sugarcane; Pulses, Banana; Rice,Banana; Berseem, Banana;Pulses , Sugarcane; Pulses,Banana; Rice, Banana; Pulses,Banana	1		Clayey and Silty
Node 3	98	49.25%	13.237	0.153	28.57%		Major Crops	Cotton, Wheat; Rice, Wheat; Rice, Sugarcane; Mustard, Sorghum; Sorghum, Berseem	1		Silty, Clayey
Node 4	37	18.59%	21.281	0.007	45.95%	4	Major Crops	Sorghum, Cotton; Sugarcane, Sorghum; Berseem, Rice; Mustard, Wheat; Sugarcane, Berseem	1	6; 7; 8	Silty, Clayey
Node 5	23	11.56%	6.469		43.48%		Major Crops	Guar, Millet; Wheat, Guar; Wheat, Castor; Millet, Wheat; Guar, Castor; Guar, Wheat; Castor, Wheat	1		Sandy, Clayey
Node 6	9	4.52%			100.00%		Major Crops	Sorghum, Cotton; Sugarcane, Sorghum	4		Silty, Clayey
Node 7	24	12.06%			33.33%		Major Crops	Berseem, Rice; Mustard, Wheat	4		Loamy, Silty
Node 8	4	2.01%			100.00%		Major Crops	Sugarcane, Berseem	4		Calcareous, Silty

C. Selection of Algorithm

Various algorithms have been applied to the data e.g., decision trees, K- means, but the random forest from the classification and regression trees seems the most suitable for classification of soil environment of Pakistan. Keeping in view the data set, this algorithm has been selected for classification due to following reasons.

- The random forest algorithm or the random forest classifier can successfully work on both classification and the regression task.

- When the crowd sourcing is used for data acquisition, there may be a chance of missing values in the data [27]. The Random forest classifier can easily handle the missing values in the data.

- As the algorithm builds many decision trees using random samples to be a part of forest, random forest classifier in this situation, won't overfit the applied model.

- Most significantly, in case of classifying agricultural region wise soil environment and mapping these to suitable crops, the data values seem categorical in the sense, the variable have more than one level. The random forest classifier can model for categorical values in this situation.

D. Working of Algorithm

The dataset D containing 1557 observations on 04 useful features was divided into a training set by using the concept of random sampling.

The random forest algorithm can be divided into two stages.

- Creation of random forest.

- To make some predictions from the generated random forest classifier.

First, the process of random forest creation has been discussed.

Random forest is primarily an ensemble of K no. of trees, let say {$\mathbf{t}_1(j),…,\mathbf{t}_k(j)$}, where j =$\mathbf{j}_1,…,\mathbf{j}_i$ is an $\mathbf{i}$-dimensional vector of interacting variables (also called predictors) which symbolize the soil environmental factors of Pakistan. The ensemble after processing, provides K outputs as {Output$_1$ = $\mathbf{t}_1(j),…..,$ Output$_K$= $\mathbf{t}_k(j)$}, where for OutputK, k = 1,…,K, is the classification of the Soil Environment of Pakistan by the $\mathbf{k}_{th}$ tree. Now to produce a random forest, Outputs of all generated trees are formerly aggregated to perform final classification output, by majority vote from entirely generated trees. Now for prediction from a set of given training data, T = {($\mathbf{j}_1$, Obs$_1$),… , ($\mathbf{j}_n$, Obs$_n$)}, where $\mathbf{j}_i$, $\mathbf{i}$ = 1,.…,n, is a vector of interacting variables or predictors and Obs$_i$ is the representing corresponding soil classes as results, hence, training of the current random forest continues according to following steps:

1) Initially, from training rectified data (, K bootstrap samples have been drawn. Each bootstrap provided the foundation for one of the trees which is part of the forest.

2) Afterwards, a classification tree has been grown for every bootstrap sample without any pruning, to originate the classification of soil Environment as Output.

3) Instead of selecting the best split at every node between all predictors to produce a forest, randomly sampled *m* no. of predictors and selected the best split among them. While growing the forest, the value of *m* is detained constant.

4) The above steps are repeated so that K trees have been grown. These K no. of trees have created a forest.

5) Now for every tree, the data which is not a part of the bootstrap sample has been predicted (this is also called out-of-bag data) consuming the grown-up of the tree with the bootstrap sample (also called in-the-bag data).

6) After aggregating the predictions which are out-of-bag, and the predicted Output values of processed data are then compared with the ultimate observed values, Obs, of each unit present in the out-of-bag (OOB) sample. Finally, classification error rate (ER) has been calculated as:

$$\sum_{i=1}^{n} I(Output_i^{OOB} \neq Obs_i^{OOB})/n \qquad (2)$$

Where equation I represents the indicator function while n is representing the no of data which is out of the bag. On the basis of error rate, confusion matrix for both cases have been generated which shows the performance of algorithms. Each row signifies instances occurred in the predicted class. Whereas each column of the matrix characterizes the instances in a real class. Confusion matrices of both training sets have shown minimal error rate.

V. VERIFICATION OF CLASS ASSIGNMENT

To improve the accuracy of the traditional RF algorithm, experience of domain experts has been used.

Classes obtained after majority voting along with generated rules have been sent to expert community where many domain experts who are already playing their role in data rectification resides are shown in Fig. 2. They have reviewed and verified the classification results and decision rules using their experience. Some rules have been revised as discussed in Section VII and saved in the knowledge-base. This integration of generated results and experience of domain experts has created new knowledge about the soil environment which is used for the improvement of existing decision support systems of the domain. Consequently, the results have been improved after the involvement of expert experience in the system. This created knowledge will be placed on our online system for further distribution among other registered users.

VI. ALGORITHM IMPLEMENTATION

After data cleaning and management, the crowd sourced data seem categorical in a sense. A random forest algorithm has been implemented by splitting the data set into two portions as by taking soil type as a primary variable. Data have been trained as discussed in the previous section using three predictors, soil type, major crops and agricultural zones. Statistics of the training data are available in Tables V and VI. Soil categories have been counted in the current training data set and the frequency of each type has been calculated. Sandy, Clayey exists in 73 agricultural zones which is in association 16.25 % of the total soil types of the regions besides combination with other soil types. Calcareous, Sandy, Loamy, Silty and Sandy, Loamy are separately 4.45% of the overall

soil types hence, less common in Pakistan as shown in Fig. 3. Similarly, Cotton, Wheat, Rice, Sugarcane, Mustard, Sorghum and Berseem are common crop types in Pakistan. During the training and processing of generation decision trees for every random sample, a random with replacement was used in bagging process. The parameter was set as the minimum node size=2, minimum son size=1 and the maximum depth of the tree is 20 while running the random forest algorithm. Further, sample size is set 249 observations for the subset comprising soil types and major crops, maximum trees=100 and stopping condition has been set at 300. After setting these parametric values, the algorithm has run to generate two random forests as: i) to map soil type's classes and agricultural zones, ii) for a subset of predictors, soil types and most suitable major crop for cultivation in that agricultural region. This has provided better representation of results. The tree structure of the data subset i) and ii) is shown in Tables III and IV, respectively. In both trees, no of objects used at each node have been presented along with their frequencies (in %) which shows the existence of predictor variable at each node. Selection of split points on predictor variables which aims to predict membership in various classes of dependent variables. In the both tree structures, initializing the split from the root node, continuing the splits by determining best terminal node in the tree choosing best predictor variable to execute the split. For this, P-values are calculated for each terminal node to examine the significance of association of class membership along with the levels of every predictor variable. After multiple comparisons, if the computed smallest p-value is smaller than the predefined threshold p-value, then the predictor variable which produces smallest p-value has been selected to split the consistent node. Following this practice, node 1, 3 and 4 have been split in the Table III and the node 1 and 4 has performed the split using p-values as presented in Table IV. Then the degree of freedom (DF) has been calculated at each node by computing the (no. of levels of defined categorical variables – 1).

TABLE. V. SUMMARY STATISTICS OF TRAINING DATA SET OF SOIL TYPES

Variable	Categories	Counts	Frequencies	%
Soil Types	Calcareous, Loamy	45	45	10.022
	Calcareous, Sandy	20	20	4.454
	Calcareous, Silt-loam	40	40	8.909
	Calcareous, Silty	20	20	4.454
	Clayey and Loamy	30	30	6.682
	Clayey and Sandy	25	25	5.568
	Clayey and Silty	41	41	9.131
	Clayey, Calcareous	40	40	8.909
	Loamy, Silty	20	20	4.454
	Sandy and Loamy	30	30	6.682
	Sandy, Clayey	73	73	16.258
	Sandy, Loamy	20	20	4.454
	Silty, Clayey	45	45	10.022

TABLE. VI.   SUMMARY STATISTICS OF TRAINING DATA SET OF MAJOR CROPS

Variable	Categories	Counts	Frequencies	%
Major Crops	Berseem, Banana	4	4	2.010
	Pulses, Banana	2	2	1.005
	Pulses,Banana	2	2	1.005
	Pulses, Sugarcane	5	5	2.513
	Berseem, Banana	5	5	2.513
	Berseem, Rice	12	12	6.030
	Castor, Wheat	1	1	0.503
	Cotton, Wheat	20	20	10.050
	Guar, Castor	4	4	2.010
	Guar, Millet	3	3	1.508
	Guar, Wheat	2	2	1.005
	Millet, Wheat	5	5	2.513
	Mustard, Sorghum	23	23	11.558
	Mustard, Wheat	12	12	6.030
	Pulses,Banana	2	2	1.005
	Pulses, Sugarcane	4	4	2.010
	Rice, Banana	3	3	1.508
	Rice, Pulses	10	10	5.025
	Rice, Sugarcane	27	27	13.568
	Rice,Banana	4	4	2.010
	Rice, Wheat	5	5	2.513
	Sorghum, Berseem	23	23	11.558
	Sorghum, Cotton	5	5	2.513
	Sugarcane, Berseem	4	4	2.010
	Sugarcane, Sorghum	4	4	2.010
	Wheat, Castor	5	5	2.513
	Wheat, Guar	3	3	1.508

Fig. 3.   Existence of Soil Type in no. of Zones as Related Frequencies.

VII. RESULTS AND DISCUSSIONS

This section comprises of two use cases: Case-I contains a) classification of soil environments of Pakistan based on agricultural regions, b) classification crops being cultivated in various soil types and c) mapping of crop(s) to suitable soil types. Case-II encompasses the results of knowledge management using statistical model regarding i) Productions of targeted crops at all available soil types of selected regions, ii) income and business impact of these crops to the farmers and iii) degree of irrelevancy of targeted crops to predict best business-oriented crop(s) for cultivation.

Case-I: Classification of Soil Types and Crop Mapping.

Data set comprising 1557 observations and 4 features has been obtained from various regions of Pakistan.

To achieve final prediction from entire dataset, the training data set has been divided into two parts according to interacting variables (predictors) as i) soil types with major crops taking 199 out of 1557 observations as training sample ii) soil types with agricultural zones having 449 out of 1557 observations as training sample and run the algorithm to generate two random forests using parametric values K=8, m=4, soil type (st=9), climate (c=4), major crops (mc=37), landscape (l=12), Agricultural zones (az=103). To predict output at every new location, the algorithm measured the ration of votes for every class to generate a final class. This is also known as the approximation of probabilities of each class. In the classification, this calculation also gives a measure of confidence. At the end, two random forests have been generated and compared using predictors. For case i), soil type with agricultural zones. Tree structure is available in the Table III. Node 1 has 449 objects with 4.68% purity p-value< 0.0001. Forest is structured with 12 nodes, which split at node 3 and 4. Similarly for case ii) soil type with major crops, tree structure with all statistics is depicted in Table IV which shows that the node 1 have 199 objects, node 2 contains 41 and the last node 8 have only 4 objects. Node 2, 6 and 8 have 100% purity.

As a result, two random forest have been produced along with decision rules for both cases. 11 rules have been made for mapping soil types to agricultural zones as presented in Table VII. These rules are useful to guide the growers to analyze existence of soil types in various agricultural zones. In the same way, 7 decision rules have been generated for mapping suitable major crops to various soil types as shown in Table VIII.

These rules representing different classes have then sent to agricultural community for further verification and knowledge creation of this domain. Four domain experts have participated in the verification process and revised 2 out of eleven rules of case i) and 1 out of seven rules from the case ii), respectively. From Table VII, they have compared rule for the nodes 9 and 10 which are generated for Hyderabad zone and updated the rule no. 10. Also, the rule for node no. 4 and 12 have been compared, reviewed and updated the rule no. 12 where Geographical Coverage is representing agricultural zones. Similarly, after comparing rules of node 4 and 6, the rule 6 has been revised in Table VIII. This integration of expert's experience to the existing decisions obtained using algorithm has generated knowledge of this domain. This knowledge-based decision will improve the DSSs as in this case 16.6% accuracy has been increases using expert's experience. With the use of large training data set and large number of generated rules, more accuracy can be achieved using the proposed approach. These knowledge-based decisions will then place on an online platform for further distribution among many registered users to be utilized in their fields for selection of suitable crops to achieve more production.

TABLE. VII. DECISION RULES OF MAPPING SOIL TYPES TO AGRICULTURAL ZONES

Nodes	Geographical Coverage(Pred)	Rules
Node 1	Badin	
Node 2	Badin	If Soil Types in [Clayey and Silty] then Geographical Coverage = Badin in 9.1% of cases
Node 3	Hyderabad	If Soil Types in [Calcareous, Loamy; Clayey and Sandy] then Geographical Coverage = Hyderabad in 15.6% of cases
Node 4	Sanghar	If Soil Types in [Silty, Clayey; Loamy, Silty; Calcareous, Silty] then Geographical Coverage = Sanghar in 18.9% of cases
Node 5	Tharparkar	If Soil Types in [Sandy, Clayey] then Geographical Coverage = Tharparkar in 16.3% of cases
Node 6	Tharparkar	If Soil Types in [Clayey and Loamy; Sandy and Loamy] then Geographical Coverage = Tharparkar in 13.4% of cases
Node 7	Muzaffaragarh	If Soil Types in [Calcareous, Sandy; Sandy, Loamy] then Geographical Coverage = Muzaffaragarh in 8.9% of cases
Node 8	Bahawalnagar	If Soil Types in [Clayey, Calcareous; Calcareous, Silt-loam] then Geographical Coverage = Bahawalnagar in 17.8% of cases
Node 9	Hyderabad	If Soil Types in [Calcareous, Loamy; Clayey and Sandy] and Soil Types in [Calcareous, Loamy] then Geographical Coverage = Hyderabad in 10.0% of cases
Node 10	Hyderabad	If Soil Types in [Calcareous, Loamy; Clayey and Sandy] and Soil Types in [Clayey and Sandy] then Geographical Coverage = Hyderabad in 5.6% of cases
Node 11	Hyderabad	If Soil Types in [Silty, Clayey; Loamy, Silty; Calcareous, Silty] and Soil Types in [Silty, Clayey] then Geographical Coverage = Hyderabad in 10.0% of cases
Node 12	Sanghar	If Soil Types in [Silty, Clayey; Loamy, Silty; Calcareous, Silty] and Soil Types in [Loamy, Silty; Calcareous, Silty] then Geographical Coverage = Sanghar in 8.9% of cases

TABLE. VIII. DECISION RULES OF MAPPING MAJOR CROPS TO SOIL TYPES

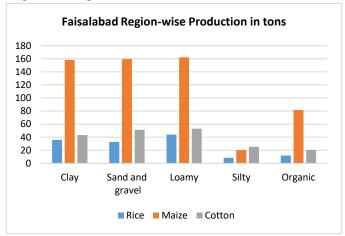| Nodes	Soil Types(Pred)	Rules
Node 1	Silty, Clayey	
Node 2	Clayey and Silty	If Major Crops in [Rice, Pulses; Berseem, Banana; Pulses, Sugarcane; Pulses, Banana; Rice,Banana; Berseem, Banana; Pulses, Sugarcane; Pulses,Banana; Rice, Banana; Pulses,Banana] then Soil Types = Clayey and Silty in 20.6% of cases
Node 3	Silty, Clayey	If Major Crops in [Cotton, Wheat; Rice, Wheat; Rice, Sugarcane; Mustard, Sorghum; Sorghum, Berseem] then Soil Types = Silty, Clayey in 49.2% of cases
Node 4	Silty, Clayey	If Major Crops in [Sorghum, Cotton; Sugarcane, Sorghum; Berseem, Rice; Mustard, Wheat; Sugarcane, Berseem] then Soil Types = Silty, Clayey in 18.6% of cases
Node 5	Sandy, Clayey	If Major Crops in [Guar, Millet; Wheat, Guar; Wheat, Castor; Millet, Wheat; Guar, Castor; Guar, Wheat; Castor, Wheat] then Soil Types = Sandy, Clayey in 11.6% of cases
Node 6	Silty, Clayey	If Major Crops in [Sorghum, Cotton; Sugarcane, Sorghum; Berseem, Rice; Mustard, Wheat; Sugarcane, Berseem] and Major Crops in [Sorghum, Cotton; Sugarcane, Sorghum] then Soil Types = Silty, Clayey in 4.5% of cases
Node 7	Loamy, Silty	If Major Crops in [Sorghum, Cotton; Sugarcane, Sorghum; Berseem, Rice; Mustard, Wheat; Sugarcane, Berseem] and Major Crops in [Berseem, Rice; Mustard, Wheat] then Soil Types = Loamy, Silty in 12.1% of cases
Node 8	Calcareous, Silty	If Major Crops in [Sorghum, Cotton; Sugarcane, Sorghum; Berseem, Rice; Mustard, Wheat; Sugarcane, Berseem] and Major Crops in [Sugarcane, Berseem] then Soil Types = Calcareous, Silty in 2.0% of cases

A maximum tree depth=3 with significance level of 5% and merge threshold is 5% was achieved. The algorithm used the Chi-square Automatic Interaction Detector (CHAID) to control the association among variables. CHAID analysis figures out a predictive model, or possibly a tree, to estimate how independent, interconnecting and dependent variables best merge produce the best output.
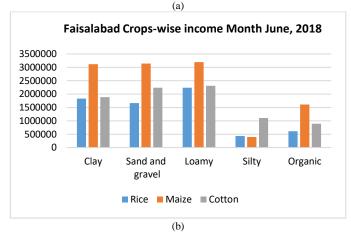
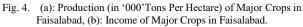Case-II: Region Wise Crop Production in Different Soil Environments.

In order to create knowledge and utilize the advantages of created knowledge, three major cash crops of Pakistan of Khareef season, Cotton, maize and rice have been selected for experimentation. Three agricultural regions, Faisalabad, Multan and Bahawalpur have been used as these have different climate conditions, almost same variety of soil types and have cultivated similar crops. 1000 farmers and experts have participated to provide data about region wise soil types, cultivated crops and verification of soil crop mapping.

The process of knowledge creation has been divided into three phases. In phase-i) The data were obtained about region wise crop cultivation and respective production from online sources e.g. Agriculture Marketing Information Service and ministry of national food security & research. A data related to soil wise crop cultivation has been obtained by crowd sources and data sets have been generated. Knowledge about the existence of soil types in different regions was created using crowd sourcing. Phase ii) covers region wise income of crops at all available soil types to find most business-related crops. Phase iii) encompasses the degree of irrelevancy of crops grown in the targeted regions to avoid growing these crops as presented in Table I.

a) Region-wise Crop Production at different Soil types:
First, Knowledge has been created and managed about production of rice crop on available soil types in the Faisalabad region as shown in Fig. 4(a). The covered area was 23.88 (in '000' hectare) and the farmers have achieved the maximum 44000 tons per hectare at loamy soil and lowest yields 7000 tons/ hectare at silt soil. This knowledge has been verified using crowd sourcing to improve the quality of results. Two other crops of 'Khareef' season have been selected and knowledge have been created for cotton and maize crops in this region at same soil environments. More production of all target crops has been observed in loamy soil. Maize produced 162.1 (in '000' tons/hectare) and cotton gives 53 ('000'bales per hectare). Sandy soil has also given good results for maize and cotton, but pure clay has less contributed in rice production. It only produced 32.7 (in '000' tons/hectare). Largely, the yield of maize has achieved the maximum level of production in the Faisalabad region as 162.1 in '000' tons per hectare. Silt soil has given a response in cultivation of cotton as compared to other two crops. It produced 25.4 in '000' bales per hectare. This knowledge created by crowd sources is helpful for farmers in choosing best crop for growing in all available soil types and to predict the combination of soil type and suitable crops for more production.
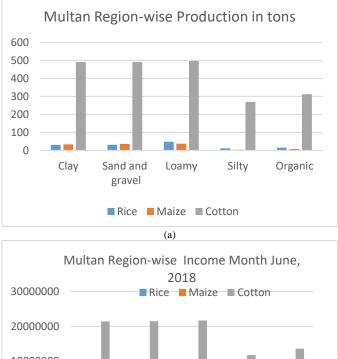
Similarly, the dataset about same crops cultivated in two other regions, Multan and Bahawalpur has been obtained. These regions have almost similar soil types and crops, but different climate conditions as compared to Faisalabad region. Results of Multan region show maximum production of 496.8 bales per hectare (in '000' bales/hectare). In the same way, rice yield= 27.76, maize= 63.55 ('000' tons/hectare) and cotton= 1114.65 (bales/hectare) were recorded in loamy soil as shown in Fig. 5(a). Where I bale=170 KG.

b) Region wise income of crops at all available soil types: The has also been obtained from online sources i.e. 'index mundi.com' about prices of the targeted crops in the month of June 2018 and processed the data using statistical models to create the knowledge about most business-oriented yields of Faisalabad regions on various soil types. The knowledge-based results show the maize gives more income to the farmers as compared to cotton and rice in Faisalabad as shown in Fig. 4(b). Maize produced 3.19 million income per hectare (excluding other expenses) to the farmers at loamy soil and a minimum 0.39 million per hectare at silt soil. Loam soil has also given good response to rice and cotton as it has given 2.24 and 2.31 million per hectare to the farmers respectively as depicted in Fig. 5(b).

(a)

(b)

Fig. 4.   (a): Production (in '000'Tons Per Hectare) of Major Crops in Faisalabad, (b): Income of Major Crops in Faisalabad.

(a)

(b)

Fig. 5.   (a): Production (in '000'Tons Per Hectare) of Major Crops in Multan Region; (b): Analysis of Income of Major Crops in Multan Region.

c) Towards Knowledge creation about Crop Irrelevancy: In order to predict which crop is not relevant to an agricultural region based on soil types, the data have been obtained from the Agricultural Marketing Information Service of Punjab Government and Pakistan Bureau of Statistics. The data have been processed using statistical models. Also, the data about yield prices has been used. To calculate the degree of crop irrelevancy to an agricultural region based on soil types, a set of steps has been discussed as under:

Consider region as R={ x,y,z…}, soil type as S= {l,m,n…} and crop type as C = {a,b,c…}

For a selected region R_i, the degree of crop C_i relevancy DR_i for a soil type S_i can be calculated as:

$$\text{Degree of Relevancy } DRi(Si) = \frac{Production(Ci)}{Production \ in \ \forall \ Ci \ values} \quad (3)$$

$$\text{Degree of irrelevancy } DIR_i.(S_i) = (1 - DR_i) \ x100 \quad (4)$$
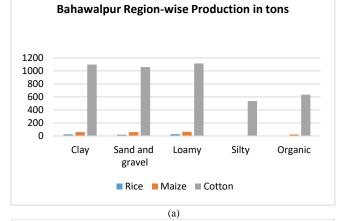
Set the threshold value for business impact of Crop $C_i(R_i)$

The degree of relevancy in Table I has been used to calculate the results of crop relevancies in all available soil types of targeted regions. Results have been presented in the Table I. Results show empirical substantiation that at same soil type of two distinct regions, a particular crop gives a completely different response. For example, at clay and loamy soil of Faisalabad and Bahawalpur, Maize and cotton have given different yield production. This shows the importance of knowledge-based results. We could now say that clay or loam soil will always respond to cotton or any particular crop. Similarly, cotton and rice have given different production at the sandy soil in Faisalabad and Multan regions.

The algorithm is generic in nature and can be used for a variety of crops of different agricultural regions where the required data is available. Knowledge verification from crowd sources have not only enhanced the quality of knowledge, but improved the accuracy of results.

VIII. REPRESENTATION OF RESULTS

Decision rules have been generated for both forests to guide the farmers in taking decisions for a) which soil type their agricultural zone have b) Which major crops they may have to cultivate in their respective lands to earn through enhancing yield production. Decision rules for mapping soil types to agricultural zones are presented in Table VII and to select most suitable crop for cultivation on particular soil types is given in Table VIII which have been further elaborated in Section VII case-I. Results of knowledge Creation about crop production at available soil types in the selected regions and regions wise income of target crops at different soil type have been presented in Fig. 4(a) and (b), Fig. 5(a) and (b) and Fig. 6(a) and b. results of presented methodology for crop relevancy to the soil types of any region have been shown in Table I.
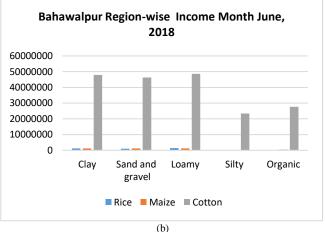
(a)

(b)

Fig. 6. (a): Production (in '000'Tons Per Hectare) of Major Crops in Bahawalpur Region. (b): Analysis of Income of Major Crops in Bahawalpur Region.

IX. CONCLUSION

In this paper, a knowledge-based classification approach has been presented to support farmers in taking decisions regarding selection of suitable crops for cultivation in relevant agricultural zones. Moreover, a procedure to calculate the degree of irrelevancy of a crop has also been discussed. The data rectification and verification of classification results from the domain experts have witnessed the improvement towards knowledge-based decisions related to the crops. In this work, a random forest classifier has been used for processing and predicting the multifarious relationship among soil types belonging to agricultural zones and major suitable crops for improving yield production.

REFERENCES

[1] L. Klerkx and A. Proctor, "Land Use Policy Beyond fragmentation and disconnect : Networks for knowledge exchange in the English land management advisory system," Land use policy, vol. 30, no. 1, pp. 13–24, 2013.

[2] S. Rajeswari and K. Suthendran, "C5. 0 : Advanced Decision Tree (ADT) classi fi cation model for agricultural data analysis on cloud," Comput. Electron. Agric., vol. 156, no. December 2018, pp. 530–539, 2019.

[3] H. Teng, R. A. Viscarra, Z. Shi, and T. Behrens, "Catena Updating a national soil classi fi cation with spectroscopic predictions and digital soil mapping," Catena, vol. 164, no. October 2017, pp. 125–134, 2018.

[4] S. Sinha, A. Norouzi, A. Pradhan, X. Yu, and D. Seo, "A Field Soil Moisture Study Using Time Domain Reflectometry (TDR) and Time Domain Transmissivity (TDT) Sensors," no. Ictim, 2017.

[5] P. Han, D. Dong, X. Zhao, L. Jiao, and Y. Lang, "Original papers A smartphone-based soil color sensor : For soil type classification," Comput. Electron. Agric., vol. 123, pp. 232–241, 2016.

[6] M. Valipour, M. Ali, G. Sefidkouhi, and M. R. Sarjaz, "Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events," Agric. Water Manag., vol. 180, pp. 50–60, 2017.

[7] C. Gornott and F. Wechsung, "Agricultural and Forest Meteorology Statistical regression models for assessing climate impacts on crop yields : A validation study for winter wheat and silage maize in Germany," Agric. For. Meteorol., vol. 217, pp. 89–100, 2016.

[8] V. Triantafyllidis, A. Kontogeorgos, C. Kosma, and A. Patakas, "An Assessment of the Soil Quality Index in a Mediterranean Agro Ecosystem," vol. 30, no. 12, pp. 1042–1050, 2018.

[9] K. J. Romanowicz, Z. B. Freedman, R. A. Upchurch, W. A. Argiroff, and D. R. Zak, "Active microorganisms in forest soils differ from the total community yet are shaped by the same environmental factors : the influence of pH and soil moisture," no. July, pp. 1–9, 2016.

[10] B. Heung, C. E. Bulmer, and M. G. Schmidt, "Geoderma Predictive soil parent material mapping at a regional-scale : A Random Forest approach," Geoderma, vol. 214–215, pp. 141–154, 2014.

[11] F. Bachion, D. Santana, A. Marcelo, D. Souza, and R. Jesus, "Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters," Spectrochim. Acta Part A Mol. Biomol. Spectrosc., vol. 191, pp. 454–462, 2018.

[12] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "ISPRS Journal of Photogrammetry and Remote Sensing A review of supervised object-based land-cover image classification," ISPRS J. Photogramm. Remote Sens., vol. 130, pp. 277–293, 2017.

[13] B. Kempen, D. J. Brus, G. B. M. Heuvelink, and J. J. Stoorvogel, "Geoderma Updating the 1 : 50 , 000 Dutch soil map using legacy soil data : A multinomial logistic regression approach," Geoderma, vol. 151, no. 3–4, pp. 311–326, 2009.

[14] R. E. Banfield, S. Member, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," vol. 29, no. 1, pp. 173–180, 2007.

[15] V. F. Rodriguez-galiano, B. Ghimire, J. Rogan, M. Chica-olmo, and J. P. Rigol-sanchez, "ISPRS Journal of Photogrammetry and Remote Sensing An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS J. Photogramm. Remote Sens., vol. 67, pp. 93–104, 2012.

[16] L. Pásztor, A. Laborczi, J. Szabó, and G. Illés, "Geoderma Compilation of a national soil-type map for Hungary by sequential classi fi cation methods," no. April, 2017.

[17] B. Melville, A. Lucieer, and J. Aryal, "Int J Appl Earth Obs Geoinformation Object-based random forest classi fi cation of Landsat ETM + and WorldView- 2 satellite imagery for mapping lowland native grassland communities in," Int J Appl Earth Obs Geoinf., vol. 66, no. November 2017, pp. 46–55, 2018.

[18] D. T. Basics, M. Notation, B. Definitions, R. Trained, D. Trees, and M. Properties, "Decision Forests : A Unified Framework for Classification , Regression , Density Estimation , Manifold Learning and Semi-Supervised Learning By Antonio Criminisi , Jamie Shotton , and Ender Konukoglu," vol. 7, no. 2011, pp. 81–227, 2012.

[19] R. L. Lawrence, S. D. Wood, and R. L. Sheley, "Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest)," vol. 100, pp. 356–362, 2006.

[20] A. H. Khan, A. Akbar, K. Farooq, N. M. Khan, M. Aziz, and H. Mujtaba, "Soil Classification Through Penetration Tests," vol. 9, no. Figure 1, pp. 76–86, 2011.

[21] H. Ahmed, T. A. Jilani, W. Haider, M. A. Abbasi, and S. Nand, "Establishing Standard Rules for Choosing Best KPIs for an E-Commerce Business Based on Google Analytics and Machine Learning Technique," vol. 8, no. 5, pp. 1–7, 2017.

[22] C. Rica, L. A. Winowiecki, M. P. Whelan, P. A. Mcdaniel, M. Villalobos, and E. Somarriba, "landscape-scale agricultural production and conservation efforts in Local soil knowledge and its use in crop allocation : Implications for landscape-scale agricultural production and conservation efforts in Talamanca , Costa Rica," no. March, 2014.

[23] K. Saito, B. Linquist, and B. Keobualapha, "Farmers ' knowledge of soils in relation to cropping practices : A case study of farmers in upland rice based slash-and-burn systems of northern Laos," vol. 136, pp. 64–74, 2006.

[24] S. Rajeswari and K. Suthendran, "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud," Comput. Electron. Agric., vol. 156, no. December 2018, pp. 530–539, 2019.

[25] M. Nafees, M. Rasul, J. Hizbullah, and A. Ali, "STATUS OF SOIL Texture and Required Associated Soil Conservation Measure of River Swat Catchments Area , NWFP ," vol. 24, no. 2, 2008.

[26] A. Arooj, "Evaluation of Predictive Data Mining Algorithms in Soil Data Classification for Optimized Crop Recommendation."

[27] Durrani, Muhammad Nouman, and Jawwad A. Shamsi. "Volunteer computing: requirements, challenges, and solutions." Journal of Network and Computer Applications 39 (2014): 369-380.