Estimating soil macronutrient status using voting based machine learning approach

Uditendu Sarkar[1] Muskan Sharma[2] Shasank Dixit[2] Jyotirmoy Das[2]
Gouravmoy Banerjee[2] Indrajit Ghosh[2#]

[1]National Informatics Centre, Ministry of Electronics & Information Technology, Government of India, Jalpaiguri, West Bengal, India-735101.
[2] Dept. of Computer Science, Ananda Chandra College, Jalpaiguri, West Bengal, India-735101.
ighosh2002@gmail.com

Abstract. The fertility status of agricultural soil plays a crucial role in the growth of a healthy crop. Therefore, prior to the cultivation of a crop, it is necessary to assess the nutrient contents of the agricultural soil for the selection of suitable fertilizers. Currently, in India this facility is only available in soil testing laboratories which are located in prime urban settlements away from the rural agricultural lands and the technical expertise required for such analysis is often scarce and costly. Therefore, it is very difficult for a marginal farmer to get his soil assessed in an unhindered manner.

To address this problem several research works involving modern machine learning models have been suggested in the literature. However, to overcome the drawbacks of the existing systems, this paper proposes a hybrid voting based machine learning model to determine the fertility status of three macronutrients (N, P, and K) in agricultural soil. For better analysis, both hard and soft voting methods were used in building the models. The models were trained and tested using publicly available soil health card (SHC) data provided by Govt. of India. The evaluation of the system was carried out by means of five well established classification metrics: *Balanced Accuracy Score (Ba), Weighted F-score (F1), Weighted Precision (Pr), Weighted Recall (Rc), and Cohen's Kappa Score (κ)*. A comparative analysis with existing models reveals that the voting classifiers are more effective in determining the fertility status of the macronutrients.

Keywords: Voting classifier, Machine learning, Soil fertility, Macronutrient estimation.

1 Introduction

Agriculture is the major means of livelihood in the rural areas of India with about 58% of the population directly or indirectly dependent on agriculture. In the fiscal year 2020-21, the share of the agriculture sector into India's Gross Value Added (GVA) was recorded at 20.2% [1]. About 60% of the geographical area of India amounting to 200.2 million hectares is used for cultivating various crops [1]. As per the 4th advance estimates of 2019-20 published by the Department of Economics and Statistics under the Ministry of Agriculture and Farmers Welfare, Government of India, the total food grain production in the country is calculated to be 296.65 million tonnes which was calculated to be nine times higher than previous five years [2]. The value of total agricultural export of 2020-21 was recorded at $41.25 billion [3].

Despite such massive amounts of cultivable land, higher yield of crops and flourishing agricultural exports, Indian agriculture is challenged by various problems. Soil fertility declination, improper use of fertilizers, untimely pest attacks, crop diseases, use of poor-quality seeds, shortage of water supply, lack of mechanization, diminishing cultivable lands, and low sale prices are just some of the few difficulties encountered by the marginal farmers [4].

Out of these problem areas, soil fertility and nutrient management are of major concern for a marginal farmer, as they have a direct impact on crop yield and product quality. Providing the plants with appropriate nutrients in the right dose at the right time is the key to a successful crop production enterprise. To achieve this objective, it is essential to monitor soil nutrient levels through yearly soil tests. Post-harvest soil tests are necessary to assist the farmers in determining the status of soil organic matter, pH, electrical conductivity, levels of important macro-nutrients, and micronutrients [5]. As the primary macronutrients like Nitrogen, Potassium, and Phosphorus play a significant role in the growth of a healthy crop, their suboptimal supply can cause severe adverse effects on crop yield [6]. Hence, before cultivation starts, proper estimation of the existing soil nutrients is a vital requirement.

In general, soil nutrients are estimated by chemical analysis in the soil testing laboratories spread across the country. However, the expenditure incurred on the chemicals used is very high and expert technicians are unavailable in most cases. Moreover, these laboratories are located in urban areas away from rural agricultural settlements and are limited in number as compared to the requirement. As a result, it becomes difficult for the marginal farmers to get their land fertility assessed quickly and at a low cost.

Several alternative systems based on modern artificial intelligence techniques have been designed to address such problems, some of which are mentioned here. Jia et al., used a transfer learning technology to grade agricultural soil fertility with a pre-trained Bayesian network [7]. Helfer et al. attempted to evaluate agricultural soil fertility using a partial least squares regression model considering soil pH, temperature, rainfall, humidity, electrical conductivity and some other atmospheric parameters as inputs [8]. Terhoeven-Urselmans et al. also constructed a partial least squares regression model to estimate several soil chemical and textural properties through soil mid-infrared diffuse reflectance spectroscopy [9]. Bhaskar et al. employed data mining techniques to predict agricultural soil fertility using different soil chemical properties and nutrient contents as inputs. Sirsat et al. compared the performances of twenty different machine learning models to classify several agricultural soil parameters such as organic carbon, phosphorous, Nitrogen, Manganese, Iron, Potassium, and soil pH [10]. The comparative performance of the machine learning models was assessed using Cohen's Kappa value. In a similar type of dataset, Suchitra and Pai employed a fast-learning classification technique known as Extreme Learning Machine (ELM) to estimate the available nutrient status of agricultural soil samples [11]. Different activation functions like Gaussian radial basis, sine-squared, hyperbolic tangent, triangular basis, and hard limit were used to determine the most optimum model.

Some drawbacks specifically of the models proposed by Sirsat et al. [10] and Suchitra and Pai [11] have been observed. The current research work has been designed to overcome these pitfalls. This research work proposes a hybrid voting based machine learning model to estimate the status of three macronutrients N, P and K in agricultural soil. The proposed model was developed in two phases; a first phase involving the selection of standalone machine learning models for the hybrid voting model and a second phase to test two types of voting hybrids for better performance. All the models were evaluated based on five popular classification metrics like *Balanced Accuracy Score (Ba), Weighted F-score (F1), Weighted Precision (Pr), Weighted Recall (Rc), and Cohen's Kappa Score (κ).*

The following section of materials and methods highlights the data collection, proposed system architecture, and experimental setup. Finally, the results and conclusion are represented in the later sections.

1.1 Data collection

West Bengal is predominantly an agrarian state covering about 2.7% of the geographical area of India. Nearly 8% of India's population resides in this state and about 71.23 lakh families in West Bengal are directly or indirectly connected to agriculture. Almost 96% of the farmers of West Bengal belong to the marginal category owning less than 1 hectare of agricultural land [12]. The food grains production recorded in the state as per 2019-20 statistics is about 18.26 million tons within 6.41 million hectares of agricultural land [1]. Although West Bengal is one of the most fertile regions within India, declining soil health is a major problem in the state. As per the state soil health survey of 2015-16 and 2016-17 cycle, about 52% of the soil samples collected throughout the state has been found to be deficient in Nitrogen, 7% in Phosphorous, and 25% in Potassium [13]. This is an alarming situation and immediate mitigation efforts must be taken up to address this issue.

Table 1. Inputs for each problem, the total number of samples, and number of samples per category for the classification problems

Classification Problem	Inputs Considered	Total number of samples	Categories				
			Very Low	Low	Medium	High	Very High
Nitrogen	pH, OC, EC, P, K, S, Zn, Fe, Cu, Mn, B	332	59	76	82	63	52
Potassium	pH, OC, EC, N, P, S, Zn, Fe, Cu, Mn, B	261	48	54	78	66	15
Phosphorus	pH, OC, EC, N, K, S, Zn, Fe, Cu, Mn, B	290	41	52	76	67	54

For the current research, we have taken the soil nutrient data of Purulia, Bankura, and West Midnapore districts of West Bengal due to their high cropping intensity [14].

The agricultural soil data of these districts were obtained from the publicly available soil health card (SHC) repository [13]. The dataset obtained consists of a collection of randomly selected SHC samples of agricultural soil from the three districts which includes physical parameters such as soil pH (pH) measured using the H2O method

(1:2.5), organic carbon content (OC) measured in percentage by the Walkley-Black method, and electrical conductivity (EC) measured in Deci Siemens per meter (dS/m). Additionally, three macronutrients Nitrogen (N), Phosphorous (P), and Potassium (K) measured in Kg/ha, and secondary nutrient like Sulphur (S), and lastly, micro-nutrients such as Zinc (Zn), Iron (Fe), Boron (B), Manganese (Mn) and Copper (Cu) measured in ppm (parts per million) are included in the dataset as well. The output class of the individual macronutrients i.e., Nitrogen, Potassium, and Phosphorus were grouped into five categories: Very Low (1), Low (2), Medium (3), High (4), Very High (5) according to its level of presence in the soil. The details of each of the classification problems along with the total number of samples considered and category-wise output classes are presented in table 1.

1.2 Proposed System Architecture

Initially, the entire dataset for each classification problem was randomly divided in a 60:40 ratio for training and test purposes respectively. A set of 14 classifier models were employed to establish the fitness to purpose as given in table 2. The 60% data in the training set was additionally used for tuning the hyperparameters of the machine learning models. The hyperparameters of each of the machine learning models used were optimized using a 5-fold cross-validation grid search strategy. Only the hyperparameters providing the best results were selected for the final evaluation. In the next step, the top three learner models exhibiting comparatively high accuracy than others were selected for constructing the voting classifier.

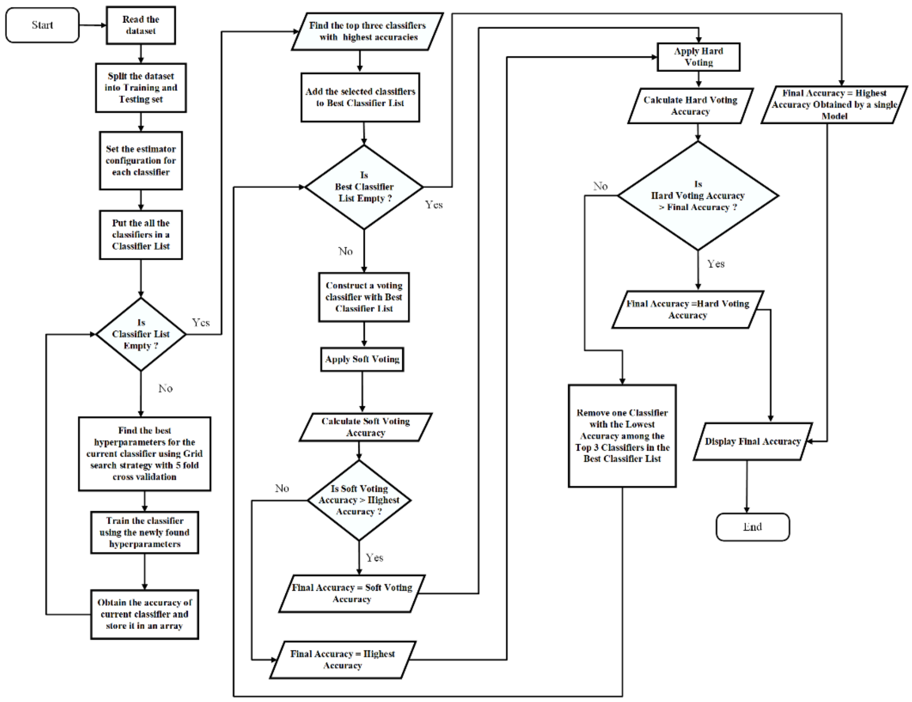
Table 2. List of machine learning models used and their hyperparameters`

Sl. No.	Machine Learning Model	Tunable Hyperparameters
1	Gradient Boost Classifier – Tree-Based (GBC)	No. of estimators
2	Light Gradient Boost Classifier (LGB)	No. of estimators
3	Extreme Gradient Boost Classifier (XGB)	No. of estimators
4	Multi-layer Perceptron Neural Network (MLP)	Hidden layer size, Activation function, Algorithm
5	K-Nearest Neighbour Classifier (KNN)	No. of neighbors
6	Decision Tree classifier (DTC)	Pruning cost, maximum depth
7	Ridge Classifier (RIC)	Regularization, Solver algorithm, Tolerance
8	Logistic Regression (LGR)	Penalty, Solver algorithm, Tolerance
9	Passive Aggressive classifier (PAC)	Regularization, Tolerance
10	Gaussian Naïve bayes classifier (GNB)	Adjustable variance
11	Bagging Ensemble – Tree-based (BDT)	No. of estimators
12	Extremely Randomized Trees classifier (EXT)	No. of estimators
13	Random Forest Classifier (RFC)	No. of estimators
14	Adaboost Classifier (ADA)	No. of estimators

There are two types of voting classifiers: hard voting classifier (HVC) and soft voting classifier (SVC). In hard voting (also known as majority voting), every individual machine learning model votes for a particular class, and thereafter the class with the majority of votes is selected as the final output. On the other hand, in the soft voting strategy, every individual machine learning model provides a probability value for a specific data point belonging to a particular target class rather than the class itself. Subsequently, the class with the highest probability is considered as the final output.

For complete and better assessment both of the voting strategies were applied in the current research. The model with the highest accuracy is finally chosen as the final accuracy for the proposed system. The flowchart for the proposed system is given below.

Flowchart:



1.3 **Performance Analysis**

Several metrics have been suggested in the literature to evaluate the performances of machine learning models [15]. However, for the current research work, the accuracy metrics that were used for analyzing the performances of the models are the *Balanced Accuracy Score (Ba), Weighted F-score (F1), Weighted Precision (Pr), Weighted Recall (Rc), and Cohen's Kappa Score (κ).* Due to the unbalanced nature of the dataset, weighted averaged versions of these metrics have been used. This strategy is useful to obtain the correct values of the metrics weighted by the number of instances for each class and overcomes the drawbacks of the previous research works [10,11]. In the case of machine learning model evaluation, true positive (*TP)*, false positive (*FP)*, false negative (*FN*) and true negative (*TN)* counts are of prime importance. These values correspond to the number of correct and incorrect predictions made by a machine learning model. The working formula of the used metrics are given below:

Balanced Accuracy Score [16]*:*

$$Balanced\ Accuracy = \frac{1}{2}\left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right] (SEQ\ "equation"\ \backslash n\ \backslash *\ \ MERGEFORMAT\ 1)$$

Weighted F1 Score [17]:

$$Weighted\ F1\ Score\ = \frac{\sum_{i=1}^{m}\left|Y_i\right|\frac{2TP_i}{2TP_i+FP_i+FN_i}}{\sum_{i=1}^{m}\left|Y_i\right|}\ (\ SEQ\ "equation"\ \backslash n\ \backslash\ *\ \ MERGEFORMAT\ 2)$$

Weighted Precision [17]*:*

$$Weighted\ Precision\ = \frac{\sum_{i=1}^{m}\left|Y_i\right|\frac{TP_i}{TP_i+FP_i}}{\sum_{i=1}^{m}\left|Y_i\right|}\ (3)$$

Weighted Recall [17]*:*

$$Weighted\ Recall\ = \frac{\sum_{i=1}^{m}\left|Y_i\right|\frac{TP_i}{TP_i+FN_i}}{\sum_{i=1}^{m}\left|Y_i\right|}\ (4)$$

The suffix *i* denotes the corresponding *FP, FN, TN* and *TP* values for the i^{th} class. m is the total number of classes and $|Y_i|$ denotes the number of instances belonging to a class.

Cohen's Kappa [18]:

For multi-class classification, Cohen's Kappa value is defined as:

$$K\ = \frac{c \times s - \sum_{k}^{K} p_k \times t_k}{s^2 - \sum_{k}^{K} p_k \times t_k}\ (5)$$

Where,

$c\ = \sum_{k}^{K} C_{kk}$ is the total number of elements correctly predicted instances,

$s\ = \sum_{i}^{K}\sum_{j}^{K} C_{ij}$ total number of elements,

$p_k\ = \sum_{i}^{K} C_{ki}$ is the number of times that class *k* was predicted (column total), and

$t_k\ = \sum_{i}^{K} C_{ik}$the number of times that class *k* truly occurs (row total)

2 Results and Discussions

The values of each of the metrics considered for this research work for each of the three macronutrient classification problems using 14 standalone machine learning models and two hybrid voting based models are provided in table 3. The noteworthy outcomes are provided in bold letters.

The experimental outcomes provided in table 3, points to the fact that using a voting based model improves the accuracy of classifying all the three macronutrients (N, P, K) significantly. For prediction of Nitrogen and Potassium fertility status in the agricultural soil, a hard voting classifier constructed with Bagging Ensemble (BDT), Extremely Randomized Trees (EXT), and Random Forest classifier (RFC) models were selected due to their higher accuracy. It is observed that this voting classifier exhibits better performance with an accuracy of 67% and 74% respectively for N and K classification problems. Contrary to this, for Phosphorus status prediction, a soft voting classifier constructed with Bagging Ensemble (BDT) and Extremely Randomized Trees (EXT) classifier models yielded the highest values across all metrics. It is also visible from table 3 that the Passive aggressive classifier model is not suitable from N, P, and K classification due to its relatively poor performance as compared to the other models.

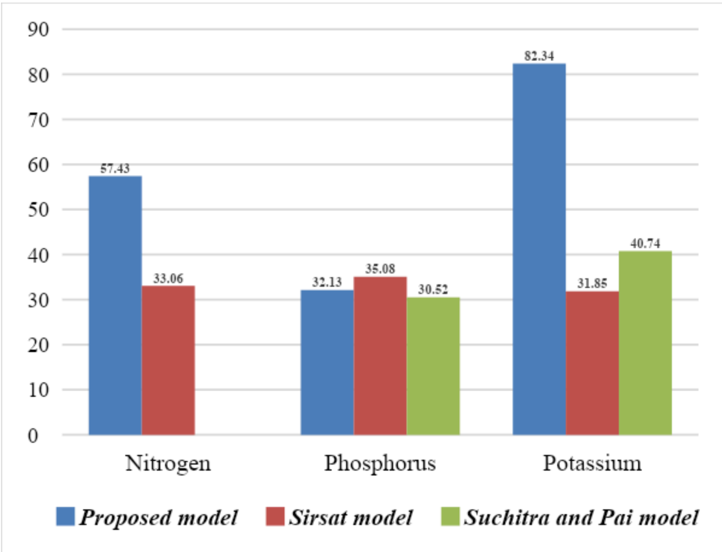


Fig. 1. Performance comparison of the proposed model with other existing models.

To establish the fitness of purpose of the proposed model, another comparative analysis was taken up with models proposed in pre-existing research works [10,11]. The comparative analysis as presented in fig. 1, reveals that out of three classification problems the proposed model performs best in two cases of N and K. In the other case the model exhibits second-best performance.

	Nitrogen					Phosphorous					Potassium				
Models	**Ba**	**F**	**Pr**	**Re**	**κ**	**Ba**	**F**	**Pr**	**Re**	**κ**	**Ba**	**F**	**Pr**	**Re**	**κ**
GBC:	0.6139	0.6054	0.6135	0.6165	0.5146	0.4155	0.4191	0.4329	0.4138	0.2538	0.6438	0.7618	0.7488	0.781	0.7113
LGB:	0.6364	0.6311	0.6348	0.6316	0.5348	0.4174	0.4089	0.4201	0.4052	0.2433	0.6788	0.7795	0.7793	0.7905	0.7249
XGB:	0.6277	0.6188	0.6312	0.6241	0.5251	0.3984	0.3841	0.3979	0.3793	0.2138	0.6951	0.7955	0.7973	0.8095	0.7491
MLP:	0.2845	0.2823	0.4118	0.3008	0.104	0.2918	0.2625	0.2416	0.2931	0.0984	0.5376	0.6336	0.6186	0.6667	0.5538
KNN:	0.2845	0.2823	0.4118	0.3008	0.104	0.2706	0.2446	0.3224	0.2586	0.0779	0.4334	0.5176	0.5024	0.5429	0.3935
DTC:	0.6302	0.6266	0.6341	0.6316	0.5333	0.3092	0.3079	0.3160	0.3103	0.1263	0.575	0.6737	0.6652	0.6952	0.5984
RIC:	0.4272	0.4086	0.409	0.4211	0.2726	0.3530	0.3644	0.4164	0.3707	0.1865	0.5204	0.6167	0.6141	0.6571	0.539
LGR:	0.4738	0.4524	0.4596	0.4662	0.3324	0.3153	0.3239	0.3674	0.3276	0.1347	0.545	0.6507	0.6333	0.6857	0.5804
PAC:	0.2136	0.1053	0.1001	0.203	0.0159	0.2655	0.1065	0.0753	0.2241	0.0617	0.2065	0.0744	0.3283	0.1905	0.0103
GNB:	0.3939	0.3447	0.4283	0.3684	0.2205	0.4259	0.3769	0.4396	0.4052	0.2558	0.603	0.698	0.7545	0.7238	0.6302
BDT:	0.6619	0.6424	0.6551	0.6541	0.5642	0.4658	0.4685	0.4775	0.4655	0.3195	0.6784	0.7733	0.7773	0.781	0.7126
EXT:	0.6517	0.6432	0.6459	0.6466	0.5547	0.4430	0.4434	0.4508	0.4397	0.2873	0.713	0.8159	0.8472	0.8381	0.785
RFC:	0.6447	0.6244	0.6272	0.6316	0.5364	0.4248	0.4080	0.4140	0.4052	0.2481	0.7372	0.8407	0.8641	0.8571	0.811
ADA:	0.3653	0.2677	0.2172	0.3609	0.1872	0.3576	0.3185	0.3437	0.3362	0.1749	0.4006	0.4111	0.5148	0.4476	0.3014
SVC:	0.6641	0.6425	0.6555	0.6541	0.5645	**0.4717**	**0.4701**	**0.4787**	**0.4655**	**0.3213**	0.7341	0.8387	0.8655	0.8571	0.8106
HVC:	**0.6731**	**0.6496**	**0.6674**	**0.6617**	**0.5743**	0.4488	0.4446	0.4529	0.4397	0.2877	**0.7446**	**0.8497**	**0.8749**	**0.8667**	**0.8234**

Table 3. Comparative analysis of the experimental results

3 Conclusion

Agricultural soil health has a significant impact on crop quality and quantity. Thus, proper estimation of the soil nutrients is very vital to maintain good soil health. Normally, the estimation of agricultural soil nutrients is done using chemical analysis in the established soil testing laboratories. This estimation process is quite expensive, furthermore, such laboratories are located in key urban settlements away from rural agricultural plots making it tedious for a marginal farmer to get his soil sample tested. Additionally, the specialized technicians required for such tests are scarce and often unavailable.

As an alternative, this paper concentrates on estimating three primary macronutrients (N, P, and K) of agricultural soil samples using voting based machine learning techniques. Thus, aid the marginal farmers by avoiding the costly manual process of estimating soil nutrients in laboratories.

In this paper, initially, a comparative study of 14 different machine learning models has been conducted. Each of the classifiers was fitted with real field agricultural soil content data. The top three learners from the set of 14 were selected for constructing the voting constituency. Both hard and soft voting methods were used for comparative assessment. The comparative metrics used for assessment were chosen in a way to overcome the drawbacks of previously conducted similar type of research [10,11]. The results show that three machine learning models giving the best results for each macronutrient when taken together through hard as well as soft voting provides better accuracy in comparison to the individual models. Furthermore, a comparison with existing works [10,11] reveals that the proposed model performs best in two (N and K classification) of three cases and the second best in the third case (P classification).

Hence, the proposed voting based model in this paper may be further employed to treat similar problems in the agriculture domain.

References

1. Pocket Book of Agricultural Statistics 2020, Technical report, Directorate of Economics & Statistics, Department of Agriculture, Cooperation & Farmers Welfare, Ministry of Agriculture & Farmers Welfare, Government of India, https://eands.dacnet.nic.in/PDF/Pocket%202020-%20Final%20web%20file.pdf, last accessed 10/12/2021.
2. Annual Report 2020-21, Technical report, Department of Agriculture, Cooperation & Farmers Welfare, Ministry of Agriculture & Farmers Welfare, Government of India, https://agricoop.nic.in/sites/default/files/Web%20copy%20of%20AR%20%28Eng%29_3.pdf last accessed 10/12/ 2021.
3. Press Information Bureau, Govt. of India, Press release, https://pib.gov.in/PressReleasePage.aspx?PRID=1725891 last accessed 10 December 2021
4. Dev, S.M.: Challenges for Revival of Indian Agriculture. National Centre for Agricultural Economics and Policy Research, ICAR, New Delhi (2008).
5. Tripathi D. K., Singh V. P., Chauhan D. K., Prasad S. M., Dubey N. K.: Role of Macronutrients in Plant Growth and Acclimation: Recent Advances and Future

Prospective. Improvement of Crops in the Era of Climatic Changes, vol 2, pp. 197-216. Springer New York, New York (2014).

6. Damar L., Lenin S.C., Lenin Y.V.: Molecular and Genetic Basis of Plant Macronutrient Use Efficiency: Concepts, Opportunities and Challenges. In: Hossain, M. A., Kamiya, T., Burritt, D. J., Tran, L. S. P., Fujiwara, T. (eds) Plant macronutrient use efficiency: molecular and genomic perspectives in crop plants. Academic Press (2017).
7. Jia H.Y., Chen J., Yu H.L., Liu D.Y.: Soil fertility grading with Bayesian Network transfer learning. In Proceedings of International Conference on Machine Learning and Cybernetics vol. 3, pp. 1159-1163. IEEE, China (2010).
8. Helfer G. A., Barbosa J.L., Dos-Santos R., Da-Costa A.B.: A computational model for soil fertility prediction in ubiquitous agriculture. Computers and Electronics in Agriculture 175, 105602 (2020).
9. Terhoeven-Urselmans T., Vagen T.G,. Spaargaren O., Shepherd K.D.: Prediction of soil fertility properties from a globally distributed soil mid‐infrared spectral library. Soil Science Society of America Journal 74(5),1792-9 (2010).
10. Sirsat M.S., Cernadas E., Fernández-Delgado M., Khan R.: Classification of agricultural soil parameters in India. Computers and electronics in agriculture 135, 269-79 (2017).
11. Suchithra M.S., Pai M.L.: Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. Information processing in Agriculture 7(1), 72-82 (2020).
12. Dept. of Agriculture, Govt. of West Bengal, https://wb.gov.in/departments-details.aspx?id=D170907140022669&page=Agriculture, last accessed 9/5/2021
13. Soil Health Card portal, https://soilhealth.dac.gov.in/, last accessed 9/5/2021
14. District wise estimated cultivable area, net cropped area, gross cropped area and cropping intensity of West Bengal for the year 2018-19. Technical report. Govt. of West Bengal, http://matirkatha.net/wp-content/uploads/2016/07/West-Bengal-District-wise-Culti-vated-Area-Gross-cropped-are-cropping-intensity-2018-19.pdf, last accessed 9/5/2021.
15. Hossin, M., Sulaiman, M. N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1-11 (2015).
16. Chicco D., Tötsch N., Jurman G.: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14(1), 1-22 (2021).
17. Behera B., Kumaravelan G., Kumar P.: Performance evaluation of deep learning algorithms in biomedical document classification. In 11th International Conference on Advanced Computing (ICoAC) pp. 220-224, IEEE, Malta (2019).
18. Cohen J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37-46 (1960).