INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Machine Learning Model for Estimation of Village Level Soil Nutrient Index

Uditendu Sarkar[1], Gouravmoy Banerjee[2], Indrajit Ghosh[3]*

1 Scientist "F", National Informatics Centre, Ministry of Electronics & Information Technology, Government of India, Jalpaiguri, West Bengal, India-735101
2 State Aided College Teacher, Dept. of Computer Science, Ananda Chandra College, Jalpaiguri, West Bengal, India-735101
3 Associate Professor, Dept. of Computer Science, Ananda Chandra College, Jalpaiguri, West Bengal, India-735101

OPEN ACCESS

Received: 20-04-2022
Accepted: 30-08-2022
Published: 22-09-2022

Citation: Sarkar U, Banerjee G, Ghosh I (2022) A Machine Learning Model for Estimation of Village Level Soil Nutrient Index. Indian Journal of Science and Technology 15(36): 1815-1822. https://doi.org/10.17485/IJST/v15i36.851

* **Corresponding author**.

ighosh2002@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose an innovative technique for designing an efficient and adaptive machine learning model using classifier assembly for estimating village level soil nutrient index using soil datasets. **Methods:** Freely available soil datasets were collected from the concerned authority of Govt. of India. These datasets were used by the proposed machine learning model designed with a classifier assembly of fifteen diverse classifiers for nutrient class identification. The performance of each classifier was evaluated in terms of five well-accepted standard metrics. The outputs of the best performing classifier were then used for estimation of village level nutrient index using modified Parker's method**. Findings:** The model was applied for nutrient class identification, and estimation of the nutrient index of different villages using freely available benchmarked Soil health Card datasets provided by the Govt. of India. The empirical results depicted that for nutrient class identification, this proposed machine learning model overperformed the other existing models in terms of average accuracy score. In the case of Copper, it provided the highest average accuracy of classification (0.949) and estimation accuracy of 95.48%. For Sulphur, an average classification accuracy of 0.891 and an estimation accuracy of 90.66% were achieved. Similarly, for Zinc, an average classification accuracy of 0.883 and an estimation accuracy of 89.63% were observed. **Novelty:** This study suggests a novel architecture of a machine learning model using classifier assembly to estimate the village level nutrient index with the highest possible accuracy, using freely available soil datasets.

Keywords: Nutrient index; Village level soil fertility; Fertilizer management; Machine learning; Classifier assembly

1 Introduction

Depleting soil fertility is a major threat to the sustainable agricultural production system and food security. Before cultivating a crop, prior assessment of soil fertility is

indispensable for maintaining soil health and enhancing crop production [1]. The overall fertility of the soil is defined in terms of the nutrient index of each nutrient. The nutrient index quantifies the amount of a nutrient present in agricultural soil and helps to assess overall soil fertility. Before cultivating any crop, proper estimation of the nutrient index is one of the prime tasks for proper fertilizer management and maintaining good soil health. For proper estimation of the nutrient index, a large number of samples are collected from different locations of a village and are classified into three groups; low, medium, and high, based on the estimated quantity of nutrients. Proper classification of the nutrient group is a prerequisite for appropriately estimating a nutrient index.

In the traditional approach, nutrient index estimation is done using various laboratory-based chemical methods. However, these methods suffer from significant implementational limitations. In countries like India, extensive and expensive soil testing in rural villages is impractical due to the lack of accessibility and infrastructure.

As an alternative, several machine learning (ML) based models were suggested to estimate different parameters and nutrients of agricultural soils [2]. In a study, Motia and Reddy comprehensively reviewed and analyzed the potential of several ML techniques for nutrient management and fertilizer recommendations [3]. In the recent past, several ML models were suggested to assess fertility status using various soil datasets. Sheeba et al. proposed an extreme learning model that used IoT sensor datasets to assess various soil parameters in four districts of Tamil Nadu, India [4]. In the USA, Longchamps et al. suggested a random forest classifier to predict the soil fertility classes based on UV-Vis-induced fluorescence sensor datasets [5]. Zhang et al. proposed a model to estimate soil organic matter, total nitrogen, and total carbon where remote sensing data were used as inputs to a support vector machine and an artificial neural network to determine these three soil attributes [6]. To estimate the total nitrogen content of the soil, Wang et al. developed a machine learning model using Visible-near-infrared spectrum (Vis-NIR) spectroscopy. Four machine learning models, random forest, ordinary least squares regression, extreme learning machines, and convolution neural networks, were used to process the sensor data [7]. Khanal et al. suggested a model for the prediction of soil acidity (pH), cation exchange capacity (CEC), organic matter, magnesium, and potassium, where five different machine learning techniques were used with remote sensing data [8]. In Iran, Emadi et al. employed six machine learning techniques and remote sensing data to map soil organic carbon content [9]. However, these different models were designed with isolated classifiers. Other limitations of these models were that they either used costly sensors or remote sensing datasets, which are expensive and inaccessible to rural farmers.

To overcome the limitations, as an alternative to the costly sensors or remote sensing datasets, Suchitra and Pai [10] proposed a model to estimate village level nutrient index using freely available Soil Health Card (SHC) data [11] provided by the Ministry of Agriculture and Farmers Welfare, Government of India. Extreme learning machines were employed to classify soil pH, organic carbon, phosphorus, Potassium, and Boron. However, they evaluated the performance of the system in terms of unweighted metrics, which were not appropriate for the unbalanced datasets they used.

The three nutrients, Sulphur (S), Zinc (Zn), and Copper (Cu), play vital roles in maintaining the good health of the plants in agriculture [12]. S promotes plant enzyme activation, chlorophyll formation, timely maturation of leaves and seeds, and drought tolerance. It provides protection against certain plant diseases. Zn assists metabolic activities and enzyme production for plant growth and is an essential nutrient for the better production of chlorophyll and carbohydrates. Cu is a vital component of various oxidase enzymes and proteins required by plants. It has a significant role in photosynthesis and proper plant vegetative growth for a higher crop yield.

This paper suggests an innovative technique for designing a machine learning model using a classifier assembly to better estimate village level nutrient index using freely available Soil Health Card (SHC) data. Regarding the estimation of the nutrient index of three vital nutrients; Sulphur (S), Zinc (Zn), and Copper (Cu), no work has been reported so far. Our secondary objective is to estimate the nutrient indices of these three nutrients using our proposed model. The outcomes are presented as case studies.

2 Materials and Methods

2.1 Datasets

The system uses datasets obtained from the Soil Health Card (SHC) repository. The Soil Health Card scheme is a flagship program launched in February 2015 and is run by the Government of India for monitoring soil health. In the SHC scheme, uniform norms are followed across different States in India to assist site-specific fertilizer management. The scheme is managed by Integrated Nutrient Management Division in the Ministry of Agriculture and Farmers Welfare, Government of India. Soil samples collected from different locations are analyzed in several soil testing laboratories across India as per the norms provided by the authority. The results are regularly uploaded to the National Soil Health Card portal [11].

The datasets include physical parameters of soil such as pH, organic carbon content (OC), electrical conductivity (EC), and content of nutrients such as Nitrogen (N), Phosphorous (P), Potassium (K), Sulphur (S), Zinc (Zn), Iron (Fe), Boron (B), Copper (Cu), and Manganese (Mn) [13]. The raw datasets were pre-processed by eliminating the records having missing data and outliers to minimize the imbalances as much as possible. The strategy adopted to put these nutrients in the respective target class was based on the guidelines suggested by the Department of Agriculture & Cooperation, Ministry of Agriculture, Government of India [14]. S was classified as the medium for content >10 ppm and low otherwise. If Zn content > 0.6 ppm, it was classified as medium and otherwise low. Similarly, Cu content > 0.2 ppm was categorized as medium or otherwise low. The target class high for these three nutrients did not arise for the study locations.

2.2 System architecture

The proposed system was designed to estimate the nutrient index with a three-step architecture. In the first step, the content of the various nutrients and other contributory parameters of the soil of a village was collected from the SHC data repository. Based on village level datasets, a targeted nutrient was classified into three groups of samples, high, medium, and low, using a set of machine learning-based classifiers (classifier assembly). In the second step, a performance evaluator measured the performance of each classifier of the classifier assembly in terms of five performance metrics and selected the best performing one for that particular context. The outputs of the classifiers and the values of performance metrics were stored in various arrays. In the final step, the nutrient index estimator estimated the value of a nutrient index based on the outputs of the best-performing classifier. As the best performing classifier was selected, the system always achieved the highest possible accuracy. The architecture of the proposed system is presented in Figure 1. All system modules were coded using Python (Ver. 3.7) using standard library functions and can easily be deployed using a laptop or desktop.

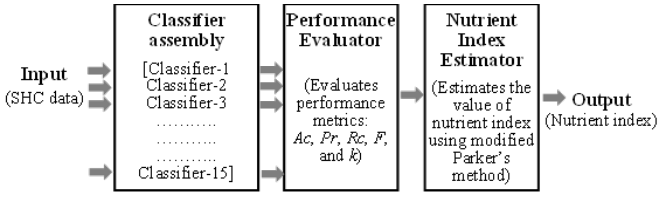

Fig 1. The architecture of the proposed system

2.2.1 Classifier assembly

The classifier assembly consisted of seven standalone classifiers and eight ensemble classifiers. The standalone classifiers are less complex and require much less computation time. In general, the ensemble classifiers perform better than a standalone classifier by amalgamating the results of several single classifiers. Each of the fifteen classifiers independently classified the target nutrient into three groups; low, medium, and high. The classifiers used in the assembly and their respective tunable hyperparameters are presented in Table 1 .

Table 1. Classifiers used and their tunable hyperparameters.

Sl. no	Classifiers	Tunable hyperparameters
1.	Logistic regression (LGR)	Penalty, Solver algorithm, Tolerance
2.	Ridge classifier (RIC)	Regularization (alpha), Solver algorithm, Tolerance
3.	Passive-Aggressive classifier (PAC)	Regularization, Tolerance
4.	Gaussian Naïve bayes classifier (GNB)	Adjustable variance
5.	Multi-layer perceptron neural network (MLP)	Hidden layer size, Activation function, Training algorithm
6.	K-nearest neighbor classifier (KNN)	Number of neighbors
7.	Decision tree classifier (DTC)	Pruning cost, maximum depth
8.	Bagging ensemble [Tree based] (BDT)	Number of trees
9.	AdaBoost classifier (ADA)	Number of estimators
10.	Gradient boost classifier [Tree based] (GBC)	Number of trees
11.	Light gradient boost classifier (LGB)	Number of estimators
12.	Categorical boost classifier (CAB)	Number of estimators
13.	Extreme gradient boost classifier (XGB)	Number of trees

Continued on next page

Table 1 continued

14.	Extremely randomized trees classifier (EXT)	Numbers of estimators
15.	Random forest classifier (RFC)	Numbers of estimators

For a targeted nutrient, the training and testing of the classifiers were done using the respective village level SHC datasets[11]. Randomly selected 60% of data was used for training, and the rest 40% was used for testing. The hyperparameters were tuned using a five-fold cross-validation grid search strategy. Only the values of the hyperparameters leading to the best results were selected for the final training and testing.

2.2.2 Performance evaluator

The performance evaluator accepted the outputs of each classifier as inputs and evaluated the performance of each classifier. Several metrics have been suggested in the literature to measure the performance of a classifier. However, due to the unbalanced nature of the present datasets, we used "weighted averaged" versions of five popular metrics. Such a strategy enabled us to get the correct values of the metrics, weighted by the number of instances for each class. The five weighted metrics used in this system were balanced accuracy (Ac), weighted precision (Pr), weighted recall (Rc), weighted F-scores (F), and Cohen's kappa (k)[15].

In supervised machine learning, measures of the classification quality are based on a confusion matrix that contains correctly and incorrectly recognized examples for each class. In the confusion matrix, t_p denotes true positive, f_p is false positive, f_n is false negative, and tn is the true negative count. Thus, the balanced accuracy (Ac), weighted precision (Pr), weighted recall (Rc), and weighted F-score (F) are defined as[15,16]:

$$\text{Balanced Accuracy } (Ac) = \frac{1}{2}\left(\frac{t_P}{t_p + f_n} + \frac{t_n}{t_n + f_P}\right) \tag{1}$$

$$\text{Weighted Precision } (Pr) = \frac{\sum_{i=1}^{m} (y_i| \frac{t_{P_i}}{t_{P_i} + f_{P_i}}}{\sum_{i=1}^{m} (y_i|} \tag{2}$$

$$\text{Weighted Recall } (Rc) = \frac{\sum_{i=1}^{m} (y_i| \frac{t_{P_i}}{t_{P_i} + f_{n_i}}}{\sum_{i=1}^{m} (y_i|} \tag{3}$$

$$\text{Weighted F} - \text{Score } (F) = \frac{\sum_{i=1}^{m} (y_i| \frac{2t_{P_i}}{2t_{P_i} + f_{p_i} + f_{n_i}}}{\sum_{i=1}^{m} (y_i|} \tag{4}$$

The suffix i denotes the corresponding f_p, f_n, $t_{n,}$ and t_p values for thei-th class. m is the total number of classes, and $|y_i|$ denotes the number of instances belonging to class i.

Cohen's kappa (k) is a standard and well-accepted measure of the accuracy of a classifier. It expresses the degree of agreement or disagreement between two instances. The generalized expression to measure the kappa (k) value for m classes is[17]:

$$\text{Cohen}'\text{s Kappa } (k) = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

Where p_0 is the total probability of agreement while p_e is the proportion of agreement expected.

The more the value of Ac, Pr, Rc, F, and k, the better the model's performance. The outputs and the values of these five metrics obtained against each classifier were stored in arrays. The performance evaluator selected the best performing classifier based on the highest average value (Avg) of these five metrics, and the output of the best classifier was selected for nutrient index estimation.

2.2.3 Nutrient index estimator

The nutrient index estimator was designed using modified Parker's method[18]. This method uses the number of samples categorized in each of the three classes, low, medium, and high, as inputs to estimate the nutrient index. The number of samples in each of the three classes is multiplied by 1, 2, and 3, respectively. The sum of the products is then divided by the total number of samples to obtain the nutrient index. The nutrient index (I_N) of a nutrient is defined as:

$$\text{Nutrient Index } (I_N) = \frac{(n_l \times 1 + n_m \times 2 + n_h \times 3)}{(n_l + n_m + n_h)} \tag{6}$$

Where n_l is the number of samples in the low group, n_m is the number of samples in the medium group, and n_h is the number of samples belonging to the high group.

2.2.4 ***Empirical case studies***

Several real field case studies were conducted to estimate the nutrient index for different villages in the state of West Bengal. West Bengal is one of India's most agriculturally productive states, producing 2856 million tons of food grains in 6.41 million hectares of land[19]. It is one of the most fertile regions in India, where the agriculture sector contributes to about 20% of the state's Gross State Value Added (GSVA)[19]. Moreover, no such studies have been reported in this state.

The case studies for six villages from three districts, Purulia, Bankura, and West Midnapur in West Bengal, are presented for illustration. These three districts in the undulating lateritic region of West Bengal were chosen because of adequately available SHC datasets and high cropping intensity[20]. The cropping intensity of Purulia, Bankura, and West Midnapore are 118, 164, and 168 percent, respectively[20]. The selected case studies on the three vital nutrients, S, Zn, and Cu, are presented because no work has been reported so far.

Datasets were collected for Balia and Chakulia villages in Purulia, Andharthaul, and Chaitali villages in Bankura, and Gopalnagar and Hariatara villages in West Midnapore districts from the SHC repository[11]. The SHC datasets contained values of required physical parameters and nutrients. For training, validation, and testing of the classifiers, the village level datasets were labeled as per the guideline of the concerned authority[14]. The total number of samples and the set of input parameters used to classify the groups (low, medium, and high) of these three nutrients are presented in Table 2.

Table 2. Summary of the used samples with input parameters.

Nutrients	Total number of samples	Input parameters
Sulphur (S)	581	pH, OC, EC, N, P, K, Zn, Fe, Cu, B, Mn
Zinc (Zn)	640	pH, OC, EC, N, P, K, S, Fe, B, Cu, Mn
Copper (Cu)	695	pH, OC, EC, N, P, K, S, Zn, Fe, B, Mn

The values of five performance metrics for each classifier against S, Zn, and Cu were evaluated by the performance evaluator using equations 1-5. The experimentally obtained values of weighted accuracy (Ac), weighted precision (Pr), weighted recall (Rc), weighted F-score (F), and Cohen's kappa (k) along with their average values (Avg) against these three nutrients; S, Zn, and Cu are presented in Table 3.

Table 3. Experimental values of Ac, Pr, Rc, F, k, and Avg obtained against the fifteen classifiers for S, Zn, and Cu.

	Metrics	LGR	RIC	PAC	GNB	MLP	KNN	DTC	BDT	ADA	GBC	LGB	CAB	XGB	EXT	RFC
Sulphur	**Ac**	0.726	0.694	0.635	0.715	0.672	0.672	0.809	0.883	0.819	0.859	0.886	0.910	0.874	0.903	0.899
	κ	0.452	0.388	0.280	0.428	0.342	0.342	0.613	0.762	0.635	0.715	0.770	0.817	0.746	0.802	0.794
	F	0.726	0.695	0.601	0.715	0.671	0.671	0.805	0.881	0.818	0.857	0.885	0.909	0.873	0.901	0.897
	Pr	0.727	0.695	0.778	0.716	0.673	0.673	0.814	0.884	0.819	0.860	0.886	0.910	0.874	0.904	0.900
	Rc	0.726	0.694	0.655	0.714	0.671	0.671	0.806	0.881	0.817	0.857	0.885	0.909	0.873	0.901	0.897
	Avg	0.672	0.633	0.590	0.657	0.606	0.606	0.769	0.858	0.782	0.830	0.862	**0.891**	0.848	0.882	0.877
Zinc	**Ac**	0.755	0.744	0.496	0.732	0.500	0.500	0.847	0.869	0.884	0.889	0.903	0.881	0.878	0.891	0.884
	κ	0.508	0.486	-0.008	0.463	0.000	0.000	0.695	0.739	0.768	0.776	0.804	0.761	0.754	0.782	0.768
	F	0.754	0.743	0.320	0.732	0.358	0.358	0.848	0.870	0.884	0.888	0.902	0.880	0.877	0.891	0.884
	Pr	0.756	0.745	0.436	0.732	0.272	0.272	0.848	0.870	0.884	0.891	0.903	0.882	0.878	0.891	0.884
	Rc	0.754	0.743	0.475	0.732	0.522	0.522	0.848	0.870	0.884	0.888	0.902	0.880	0.877	0.891	0.884
	Avg	0.705	0.692	0.344	0.678	0.330	0.330	0.817	0.843	0.861	0.866	**0.883**	0.857	0.853	0.869	0.861
Copper	**Ac**	0.874	0.874	0.829	0.863	0.884	0.884	0.870	0.923	0.926	0.937	0.958	0.937	0.944	0.937	0.937
	κ	0.747	0.747	0.656	0.726	0.768	0.768	0.740	0.846	0.853	0.874	0.916	0.874	0.888	0.874	0.874
	F	0.873	0.874	0.824	0.862	0.884	0.884	0.870	0.923	0.926	0.937	0.958	0.937	0.944	0.937	0.937
	Pr	0.876	0.876	0.861	0.877	0.887	0.887	0.878	0.923	0.926	0.937	0.958	0.937	0.944	0.937	0.937
	Rc	0.874	0.874	0.828	0.863	0.884	0.884	0.870	0.923	0.926	0.937	0.958	0.937	0.944	0.937	0.937
	Avg	0.849	0.849	0.800	0.838	0.862	0.862	0.846	0.907	0.912	0.924	**0.949**	0.924	0.933	0.924	0.924
Average Accuracy Score		**0.785**	**0.771**	**0.653**	**0.770**	**0.685**	**0.685**	**0.842**	**0.892**	**0.876**	**0.895**	**0.916**	**0.909**	**0.899**	**0.910**	**0.907**

The experimental results in Table 3 depict that the Categorical boost classifier (CAB) is the best performing one for the classification of S with Avg = 0.891. Therefore, the outputs of the CAB classifier were considered by the nutrient index estimator

to estimate the value of the nutrient index of S ($I_{N(S)}$) using equation 6. Similarly, the nutrient indexes of Zn ($I_{N(Zn)}$) and Cu ($I_{N(Cu)}$) were obtained using the outputs of the best performing Light gradient boost classifier (LGB). The actual values obtained from the concerned authority and the experimentally found values of village level nutrient index (I_N) for S, Zn, and Cu are presented in Table 4. To quantify the goodness of fit of the predicted values with the observed values, the accuracy percentages of estimation were measured and are presented in Table 3.

Table 4. The actual values, estimated values and the estimation accuracy (%) of village level nutrient index (I_N) for S, Zn, and Cu.

Nutrient Index	Villages	Actual value	Estimated value	Estimation accuracy (%)	Average estimation accuracy (%)
$I_{N(S)}$	Balia	1	0.909	90.90	**90.66**
	Chakulia	1	0.891	89.10	
	Andharthaul	1.25	1.138	91.04	
	Chaitali	1.14	1.037	90.96	
	Gopalnagar	1	0.91	91.00	
	Hariatara	2	1.819	90.95	
$I_{N(Zn)}$	Balia	1.5	1.355	90.33	**89.63**
	Chakulia	1.5	1.325	88.33	
	Andharthaul	2	1.805	90.25	
	Chaitali	2	1.766	88.30	
	Gopalnagar	1.11	1.002	90.27	
	Hariatara	1	0.903	90.30	
$I_{N(Cu)}$	Balia	2	1.898	94.90	**95.48**
	Chakulia	2	1.898	94.90	
	Andharthaul	2	1.915	95.75	
	Chaitali	2	1.915	95.75	
	Gopalnagar	1.66	1.59	95.78	
	Hariatara	2	1.916	95.80	
Overall accuracy (%)					**91.92**

3 Results and Discussion

It is revealed from Table 3 that for the estimation of S, the Categorical boost classifier (CAB) yielded the highest average accuracy of classification (Avg = 0.891). However, the Extremely randomized tree classifier (EXT) also had nearly equal performance (Avg = 0.882). For the estimation of Zn and Cu, a clear dominance of the Light gradient boost (LGB) classifier was observed. The LGB took precedence over the other classifiers, with Avg = 0.883 for Zn and Avg = 0.949 for Cu, respectively. It can be inferred from the experimental results that tree-based ensembles such as the Categorical boost (CAB) and Extremely randomized tree (EXT) classifiers performed best for nutrient index estimation of S. In contrast, the Light gradient boost (LGB) yielded the highest average classification accuracy for Zn and Cu estimation. Other tree-based models such as Categorical boost (CAB), Extremely randomized tree (EXT), Extreme gradient boost (XGB), and Random Forest (RFC) exhibited nearly equal performances for both Zn and Cu. The experimental results indicate that the tree-based ensembled classifiers with boosting and/or bagging techniques are the best choice for village level nutrient index estimation.

It is observed from Table 4 that the highest average percentage of estimation accuracy (95.48%) was achieved in the case of Cu. Nearly equal performances are observed in the case of the other two nutrients, S (90.66%) and Zn (89.63%). The overall average percentage of estimation accuracy is measured as 91.92 %. These results authenticate the fitness of the proposed model.

The classification performance of our proposed model was compared with the other four contemporary ML models proposed by Keerthan et al.[21], Chaudhari et al.[22], Suchitra and Pai[10], and Pant et al.[23]. These models were designed using different techniques for various targeted nutrients other than S, Zn, and Cu. To compare the performances of these models, model accuracy was considered as the criterion of comparison. The average accuracy score was used for better comparison because all four models used different classifiers, but the accuracy score was considered the metric for evaluating classification performance. Table 5 presents the average accuracy scores of the other four similar models and our model. The accuracy scores of the three best-performing classifiers in the classifier assembly have been considered to obtain the average accuracy score of our model. The average accuracy score is plotted against each of the five models as presented in Figure 2. Figure 2 depicts that our model

(with the highest average accuracy score of 0.911) outperformed the other models.

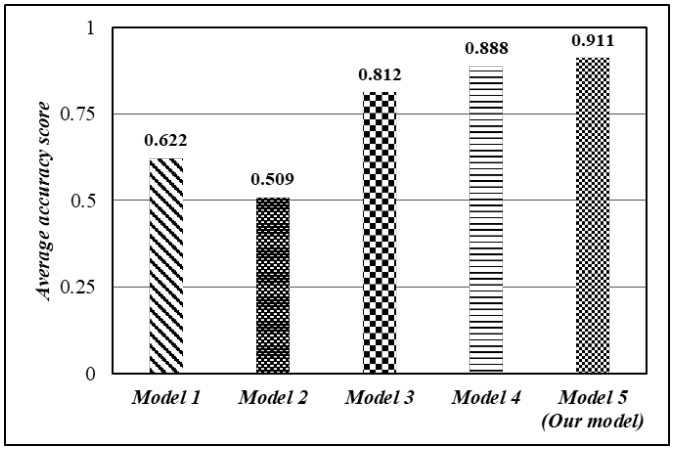

Fig 2. Comparative performances of other four similar models along with that of our model

Table 5. The average accuracy scores of the five models.

Models	Year	Classifiers used	Accuracy score	Average accuracy score
Model-1 [21]	2019	Random Forest	0.727	0.622
		Support vector machine	0.633	
		Gaussian Naïve Bayes	0.508	
Model-2 [22]	2020	Decision tree	0.600	0.509
		Support vector machine	0.446	
		k-Nearest Neighbour	0.559	
		Naïve Bayes	0.430	
Model-3 [10]	2020	ELM-tanh	0.821	0.812
		ELM-sinsq	0.782	
		ELM-tribas	0.782	
		ELM-hardlim	0.821	
		ELM-grbf	0.855	
Model-4 [23]	2021	ANN-ReLU	0.901	0.888
		ANN-tanh	0.876	
Model-5 (Our model)	2022	LGB	0.916	**0.911**
		EXT	0.910	
		CAB	0.909	

4 Conclusion

This innovative technique of applying classifier assembly serves as the base for designing more efficient and adaptive ML models. The empirical study concludes that instead of arbitrarily selected classifiers, an assembly of diverse classifiers is more efficient in designing an ML model to estimate the nutrient index better using the soil datasets. The classifier assembly always ensures the highest possible performance of the model. For example, in the case of S, the Categorical boost classifier (CAB) was selected as the best one to give the highest average accuracy, Avg = 0.891. However, in the case of Zn and Cu, the Light gradient boost (LGB) classifier is the best performing, with Avg = 0.883 for Zn and Avg = 0.949 for Cu, respectively.

The proposed model estimates the nutrient index of a nutrient using freely available soil datasets that make it an affordable alternative to costly laboratory or sensor-based systems. It will be helpful to the agricultural administration to address nutrient deficiency issues. It offers an elegant solution to rural farmers in developing countries like India.

Our future attempt is to develop a machine learning-based integrated fertilizer recommendation system (IFRS) for the marginal farmers in India using the village level nutrient index.

References

1) Barooah A, Bhattacharyya HK, Chetri KB. Assessment of Soil Fertility of Some Villages of Lahowal Block, Dibrugarh, India. *International Journal of Current Microbiology and Applied Sciences*. 2020;9(8):1438–1450. Available from: https://doi.org/10.20546/ijcmas.2020.908.165.
2) Shahare Y, Gautam V. Soil Nutrient Assessment and Crop Estimation with Machine Learning Method: A Survey. In: Cyber Intelligence and Information Retrieval;vol. 291. Springer Singapore. 2022;p. 253–266. Available from: https://doi.org/10.1007/978-981-16-4284_22.
3) Motia S, Reddy SR. Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. *Journal of Physics: Conference Series*. 2021;1950(1):012037–012037. Available from: https://doi.org/10.1088/1742-6596/1950/1/012037.
4) Sheeba B, Anand T, Manohar LD, Selvan G, Wilfred S, Muthukumar CB, et al. Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms. *Journal of Nanomaterials*. 2022;2022. Available from: https://doi.org/10.1155/2022/5343965.
5) Longchamps L, Mandal D, Khosla R. Assessment of Soil Fertility Using Induced Fluorescence and Machine Learning. *Sensors*. 2022;22(12):4644–4644. Available from: https://doi.org/10.3390/s22124644.
6) Zhang S, Lu X, Zhang Y, Nie G, Li Y. Estimation of Soil Organic Matter, Total Nitrogen and Total Carbon in Sustainable Coastal Wetlands. *Sustainability*. 2019;11(3):667–667. Available from: https://doi.org/10.3390/su11030667.
7) Wang Y, Li M, Ji R, Wang M, Zheng L. Comparison of Soil Total Nitrogen Content Prediction Models Based on Vis-NIR Spectroscopy. *Sensors*. 2020;20(24):7078–7078. Available from: https://doi.org/10.3390/s20247078.
8) Khanal S, Fulton J, Klopfenstein A, Douridas N, Shearer S. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture*. 2018;153:213–225. Available from: https://doi.org/10.1016/j.compag.2018.07.016.
9) Emadi M, Taghizadeh-Mehrjardi R, Cherati A, Danesh M, Mosavi A, Scholten T. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing*. 2020;7(1):72–82. Available from: https://doi.org/10.3390/rs12142234.
10) Suchithra MS, Pai ML. Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. *Information Processing in Agriculture*. 2020;7(1):72–82. Available from: https://doi.org/10.1016/j.inpa.2019.05.003.
11) Ministry of Agriculture and Farmers Welfare, Government of India [Internet]. Soil health card. . Available from: https://soilhealth.dac.gov.in/.
12) Fageria NK. The Use of Nutrients in Crop Plants. Boca Raton, FL. CRC Press. 2016.
13) Sharma RP, Singh SK, Chandran P, Chattaraj S. Development of soil health card (SHC) using GIS technique. *Indian Farming*. 2020;70:25–28.
14) Department of Agriculture & Cooperation, Ministry of Agriculture, Government of India. Methods Manual:Soil Testing in India. . Available from: https://agriculture.uk.gov.in/files/Soil_Testing_Method_by_Govt_of_India.pdf.
15) 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*. 2021;14(1):1–22. Available from: https://doi.org/10.1186/s13040-021-00244-z.
16) Behera B, Kumaravelan G, B PK. Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification. *2019 11th International Conference on Advanced Computing (ICoAC)*. 2019. Available from: https://doi.org/10.1109/ICoAC48765.2019.246843.
17) Vergni L, Todisco F, Lena BD. Evaluation of the similarity between drought indices by correlation analysis and Cohen's Kappa test in a Mediterranean area. *Natural Hazards*. 2021;108(2):2187–2209. Available from: https://doi.org/10.1007/s11069-021-04775-w.
18) Gopan GMV, Hasan A, Thomas T, David AA, Reddy IS. Correlation of Physico-chemical Parameters of Soil and Soil Nutrient Index Status of Kollam District. *International Journal of Plant & Soil Science*. 2022;34(20):270–276. Available from: http://doi.org/10.9734/IJPSS/2022/v34i2031151.
19) Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare [Internet]. Agricultural Statistics at a glance 2019 . . Available from: https://eands.dacnet.nic.in/PDF/At%20a%20Glance%202019%20Eng.pdf.
20) Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India [Internet]. Agriculture Contingency Plan . 2021.
21) Kumar K, Shubha TG, Sushma C, A S. Random forest algorithm for soil fertility prediction and grading using machine learning. *International Journal of Innovative Technology and Exploring Engineering*. 2019;9(1):1301–1304. Available from: http://doi.org/10.35940/ijitee.L3609.119119.
22) Chaudhari R, Chaudhari S, Shaikh A, Chiloba R, Khadtare T. Soil fertility prediction using data mining techniques. *Mukt Shabd J*. 2020;9:2347–315.
23) Pant J, Pant P, Pant RP, Bhatt A, Pant D, Juyal A. Soil Quality Prediction for Determining Soil Fertility in Bhimtal Block of Uttarakhand (India) Using Machine Learning. *International Journal of Analysis and Applications*. 2021;19(1):91–109.