

## TECHNICAL REPORT

## Environmental Models, Modules, and Datasets

# Determination of bioavailable arsenic threshold and validation of modeled permissible total arsenic in paddy soil using machine learning

Jajati Mandal<sup>1,2</sup>  | Vinay Jain<sup>3</sup> | Sudip Sengupta<sup>4,5</sup> | Md. Aminur Rahman<sup>6,7</sup> | Kallol Bhattacharyya<sup>4</sup> | Mohammad Mahmudur Rahman<sup>6,8</sup> | Debasis Golui<sup>9,10</sup>  | Michael D. Wood<sup>1</sup>  | Debapriya Mondal<sup>11</sup> 

<sup>1</sup>School of Science, Engineering and Environment, University of Salford, Salford, UK

<sup>2</sup>CSIRO, Land and Water, Waite Campus, Urrbrae, SA, Australia

<sup>3</sup>Centre of Excellence, Agilent Technologies (International) Pvt. Ltd, Manesar, Haryana, India

<sup>4</sup>Department of Agricultural Chemistry and Soil Science, Bidhan Chandra Krishi Viswavidyalaya, Nadia, West Bengal, India

<sup>5</sup>School of Agriculture, Swami Vivekananda University, Barrackpore, West Bengal, India

<sup>6</sup>Global Centre for Environmental Remediation (GCER), College of Engineering, Science and Environment, The University of Newcastle, Callaghan, New South Wales, Australia

<sup>7</sup>Department of Public Health Engineering (DPHE), Zonal Laboratory, Jashore, Khulna, Bangladesh

<sup>8</sup>Department of General Educational Development, Faculty of Science & Information Technology, Daffodil International University, Savar, Dhaka, Bangladesh

<sup>9</sup>Department of Civil, Construction and Environmental Engineering, North Dakota State University, Fargo, North Dakota, USA

<sup>10</sup>Division of Soil Science and Agricultural Chemistry, ICAR-Indian Agricultural Research Institute, New Delhi, India

<sup>11</sup>Department of Population Health, Faculty of Epidemiology and Population Health, School of Hygiene & Tropical Medicine, London, England, UK

## Correspondence

Jajati Mandal, School of Science, Engineering and Environment, University of Salford, Salford, UK.  
Email: [J.Mandal1@edu.salford.ac.uk](mailto:J.Mandal1@edu.salford.ac.uk)

Assigned to Associate Editor Ying Ouyang.

## Funding information

Netaji Subhas—ICAR International Fellowship

## Abstract

Minimizing arsenic intake from food consumption is a key aspect of the public health response in arsenic (As)-contaminated regions. In many of these regions, rice is the predominant staple food. Here, we present a validated maximum allowable concentration of total As in paddy soil and provide the first derivation of a maximum allowable soil concentration for bioavailable As. We have previously used meta-analysis to predict the maximum allowable total As in soil based on decision tree (DT) and logistic regression (LR) models. The models were defined using the maximum tolerable concentration (MTC) of As in rice grains as per the codex recommendation.

**Abbreviations:** AUC, area under curve; Cov, covariance; DT, decision tree; FN, false negative; FP, false positive; GBM, gradient boost machine; ICE, individual conditional expectation; LR, logistic regression; MCC, Mathew correlation coefficient; MTC, maximum tolerable arsenic concentration; PDP, partial dependence plots; PPV, positive predictive value; RF, random forest; ROC, receiver operating characteristic; TN, true negative; TNR, true negative rate; TP, true positive; TPR, true positive rate; VIF, variance inflation factor.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of Environmental Quality* published by Wiley Periodicals LLC on behalf of American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

In the present study, we validated these models using three test data sets derived from purposely collected field data. The DT model performed better than the LR in terms of accuracy and Matthews correlation coefficient (MCC). Therefore, the DT estimated maximum allowable total As in paddy soil of  $14 \text{ mg kg}^{-1}$  could confidently be used as an appropriate guideline value. We further used the purposely collected field data to predict the concentration of bioavailable As in the paddy soil with the help of random forest (RF), gradient boosting machine (GBM), and LR models. The category of grain As ( $<\text{MTC}$  and  $>\text{MTC}$ ) was considered as the dependent variable; bioavailable As (BAs), total As (TAs), pH, organic carbon (OC), available phosphorus (AvP), and available iron (AvFe) were the predictor variables. LR performed better than RF and GBM in terms of accuracy, sensitivity, specificity, kappa, precision, log loss, F1score, and MCC. From the better-performing LR model, bioavailable As (BAs), TAs, AvFe, and OC were significant variables for grain As. From the partial dependence plots (PDP) and individual conditional expectation (ICE) of the LR model,  $5.70 \text{ mg kg}^{-1}$  was estimated to be the limit for BAs in soil.

## 1 | INTRODUCTION

Arsenic (As) exposure has emerged as a major public health concern over the past few decades. It is now a well-established fact that not only drinking water, but also food crops cultivated using As-contaminated irrigation water is an important exposure pathway to human through water–soil–rice transfer (Mandal et al., 2021). Rice is a principal food for almost half of the world's population, particularly in Asia, Africa, and Latin America (Majumder & Banik, 2019). Milled rice consumption is substantial in India and Bangladesh that are about 103 and 268 kg per capita annually, respectively (FAO, 2017). Rice accounts for roughly 73% of calorific intake in Bangladesh and 30% in India (GriSP, 2013). It is high in dietary fiber as well as nutrients such as carbohydrates, proteins, vitamins, and minerals (Dipti et al., 2012; Mwale et al., 2018). On the other hand, the consumption of rice could be a substantial source of As exposure (Mondal & Polya, 2008; Mondal et al., 2010, 2020). Since As intake from rice has become a worldwide concern, some governments and regulatory bodies have established maximum tolerable As concentrations in rice grains (Schmidt, 2015). According to the Joint FAO–WHO Codex Alimentarius Commission (JECFA, 2017), inorganic As levels should not exceed  $0.2 \text{ mg kg}^{-1}$  in polished rice and  $0.35 \text{ mg kg}^{-1}$  in husked rice. Using a machine learning approach, Mandal et al. (2021) predicted the soil As concentrations above which the rice grains cultivated in Asian paddy fields may exceed the Codex maximum tolerable concentrations (MTC). From the logistic regression (LR) model, the maximum concentration of total As in soil was found to be  $11.75 \text{ mg kg}^{-1}$ , whereas the better performing decision tree (DT) model predicted the concentration of

total As to be  $14 \text{ mg kg}^{-1}$  above which the rice grain As concentration would exceed the MTC. The study was a meta-analysis using data that were published in 26 selected studies from Asia. Although the inclusion criteria were restricted only to field-based studies for reducing heterogeneity, it could not be fully eliminated. Furthermore, purposely collected field-based data were required to validate these models. Hence, in this study, we aimed to validate both the LR and DT models and test the efficacy of our model predictability using three different purposely collected field data sets from different rice cultivation practices: rainfed and groundwater irrigated from As-contaminated sites of West Bengal, India.

Arsenic in soil is present in both the solution and solid phases. The As may be in the form of organic and inorganic complexes (present in soil solution), adsorbed ions and compounds (clay and organic colloids), bound to secondary minerals and precipitated oxides of iron (Fe) and manganese (Mn), carbonates, and phosphates or complexed with organic matter and free ions (Raj et al., 2021). Total elemental concentrations within the soil offer little insight into the potential bioavailability of the elements (such as As), which may cause metal(loid) sequestration and recycling within the soil environment under the influence of various soil parameters (Kumari et al., 2021). The fraction of the total concentration of an element being reactive or labile is not only related to its source but also the soil properties. The most inert phase, which is contained in the crystal lattices of minerals or occluded by particles (total elemental concentration), is not potentially available for the biota; instead, only the reactive concentration is (Groenenberg et al., 2017). The potential bioavailable or bioaccessible metal(loid) fraction in soils may be a strong indicator of recent metal(loid)

depositions, as in the case of As when the field is irrigated with contaminated irrigation water (Sengupta et al., 2021). The bioavailable As is often used as a key indicator to estimate the dissolution behavior of As derived from the geochemical fractions in soils (Bari et al., 2021; ISO, 2016). In this study, we modeled the maximum tolerable available As concentrations of paddy soil above which rice grain As may exceed MTC using the collected field data. The bioavailability of As in soil is governed primarily by pH, organic carbon (OC), available phosphorus (P), and available iron (Fe) in the rice ecosystem (Hussain et al., 2021; Kumari et al., 2021). In another study, Yao et al. (2021) developed a predictive model for rice grain As in relation to bioavailable As along with soil characteristics (pH, EC, organic matter, total P, N, and As) with multiple linear regression. Iron is usually high in the Bengal delta in the groundwater as well as soil and phosphate-based fertilizers are widely used in rice that may impact As bioavailability, and hence, these two parameters may provide a useful insight within the modeling framework. Previously as reported by Tan et al. (2020) Fe and P proved to be the most important parameter in governing the groundwater (drinking purpose) As content in Bangladesh. From the comprehensive study of 26 published articles, it was observed that the most determinant soil properties for As bioavailability in typical Asian paddy soils were pH, OC, and available P and Fe (Mandal et al., 2021). We predicted the threshold for bioavailable As and also investigated the behavior of these soil parameters (pH, OC, available P and Fe) both on the bioavailability of As and also the grain As content with the help of individual conditional expectation (ICE) and partial dependence plots (PDP) using the random forest (RF), gradient boosting machine (GBM), and LR models.

## 2 | MATERIALS AND METHODS

### 2.1 | Testing of LR and DT models

Three individual test data sets composed of paired rice grain and soil's total As concentrations were used for the purpose. The test sets were selected in a way that there is a difference in terms of sample size, site, and system of rice cultivation. Test set 1 ( $n = 50$ ) was collected from three As-contaminated districts (Nadia, Murshidabad, and N-24 Parganas) of West Bengal, India under the rainfed rice system. Test set 2 ( $n = 28$ ) and test set 3 ( $n = 132$ ) were collected from the Maldah and Nadia districts of West Bengal, India, respectively, from irrigated rice systems using As-contaminated water. Total As in rice and soil samples were analyzed following the established protocols as outlined in Table S1.

The grain As content was converted to categorical variables (<MTC and >MTC) as per the methods outlined in Mandal et al. (2021). The model testing was performed using R-Studio

#### Core ideas

- Decision tree (DT) and logistic regression (LR) models are tested with field data.
- For rice cultivation, the better performing DT model predicts 14 mg kg<sup>-1</sup> total As as the soil limit.
- Both LR and random forest models identified available Fe, P, and organic carbon as important variables governing bioavailable As.
- From LR model, 5.70 mg kg<sup>-1</sup> is the threshold limit for soil bioavailable As for rice.

(version 1.3.1093 2.3.1). The “caret” package (version 6.0–86) was used to conduct prediction with LR and DT models.

### 2.2 | Predicting grain As with RF, GBM, and LR

Random forest is a supervised machine learning algorithm used for classification and regression-based problems. It is based on the principle of recursive partitioning (Breiman, 2001) and is independent of the assumption of functional relationships between the response and predictor variables. Gradient boost machine integrates the predictions from various decision trees to generate the final estimate (Friedman, 2001). Logistic regression predicts a binary outcome, based on previous explanations of a data set. It predicts a dependent variable by examining the connection between one or more existing independent variables (James et al., 2013).

For predicting rice grain As alongside the impact of other soil parameters with the RF and LR models, a compiled data set ( $n = 233$ ) of both irrigated and rainfed rice was used. The details of the data set and analysis protocols followed have been depicted in Table S2. The whole data set was randomized and split into two. Overall, 80% of the data were used as the training set and the remaining 20% formed the testing set. After this, the testing set was kept aside and the training set was subjected to repeated cross-validation. The category of grain As (<MTC and >MTC) was considered as the dependent variable, whereas bioavailable As (BAs), total As (TAs), pH, organic carbon (OC), available phosphorus (AvP), and available iron (AvFe) were considered as the predictor variables. Basically, the training set was used to generate multiple splits of the training and validation sets to reduce overfitting of the model. The “caret” package (version 6.0–86) was used to train the model with 10-fold cross-validation repeated five times. For RF model accuracy of 0.89 and kappa of 0.345 was used to select the final model using the value at  $mtry = 4$  after repeated cross-validation (Figure S1).

Similarly, after repeated cross-validation the final GBM model was selected at an accuracy of 0.86 with  $n.tree = 450$ ,  $interaction.depth = 8$ ,  $shrinkage = 0.1$ ,  $n.minobsinnode = 10$ , and  $kappa = 0.32$ . For LR model, the accuracy of 0.89 at  $kappa = 0.424$  was considered as the final model after repeated cross-validation.

The PDP shows the marginal effect that one or two features have on the predicted outcome of a machine-learning algorithm (Friedman, 2001). The correspondent to a PDP for specific data occasions is ICE plot (Goldstein et al. 2015). An ICE plot envisions the dependency of the prediction on a variable for each occurrence separately, resulting in one line per case, compared to one line in general in PDPs. The PDP and ICE plots from the RF, GBM, and RF models were prepared using the *pdp* (version 0.7.0) package. One of the assumptions for PDPs is that a variable for which the partial dependence is computed is not correlated with other variables. The RF model is highly robust against problems such as multicollinearity among the variables (Sarkar et al., 2022). For LR model, the presence of multicollinearity may undermine the assumptions for PDPs and hence the severity of multicollinearity for each variable was tested with variance inflation factor (VIF). The presence of collinearity raises the variances of parameter estimations and might result in mistaken inferences about the relationship between dependent and independent variables (Midi et al., 2010). VIF measures the multicollinearity of predictor variables in a regression analysis (Franke, 2010). As per Franke (2010), if  $VIF > 10$ , then multicollinearity is high. In our study, the VIFs were 1.15 for pH, 1.45 for OC, 1.32 for BAs, 1.31 for AvFe, 1.34 for AvP, and 1.15 for TAs.

### 2.3 | Model performance parameters

As the predictive scores are binary (usually represented as zeros and ones), there is just a single confusion matrix to analyze, to be informative, each category of the confusion matrix (true positive [TP], true negative [TN], false positive [FP], and false negative [FN]), must not be evaluated independently, but rather with respect to the other ones. The model performance parameters are accuracy (Equation 1) sensitivity (Equation 2), specificity (Equation 3), and precision (Equation 4). True positive rate is also called recall or sensitivity. True negative rate is also known as specificity. Positive predictive value is also called precision. The F1 score and the MCC were calculated by the formulae as outlined in Equations 5 and 6.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{True positive rate (TPR)/Recall} = TP / (TP + FN) \quad (2)$$

$$\text{True negative rate (TNR)/Specificity} = TN / (TN + FP) \quad (3)$$

$$\text{Positive predictive value (PPV)/Precision} = TP / (TP + FP) \quad (4)$$

$$\text{F1 score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

$$\text{MCC} = \text{Cov}(c, l) / \sigma_c \sigma_l, \quad (6)$$

where  $\text{Cov}(c, l)$  is the covariance of the true classes  $c$  and predicted labels  $l$ , whereas  $\sigma_c$  and  $\sigma_l$  are the standard deviations, respectively.

The receiver operating characteristic (ROC) curve was used to calculate the magnitude of the predicted class for a specified data that was close to the true class for that data set. The area under the curve (AUC) measured a classifier's overall performance across all possible thresholds (James et al., 2013).

## 3 | RESULTS AND DISCUSSION

### 3.1 | Testing of LR and DT models

Table S3 depicts the total As concentrations in the paired rice grain and soil samples in the three testing sets. The mean rice grain As content was  $322 \pm 166 \mu\text{g kg}^{-1}$ ,  $257 \pm 262 \mu\text{g kg}^{-1}$ , and  $337 \pm 139.2 \mu\text{g kg}^{-1}$  for test set 1, 2, and 3, respectively, while the mean soil total As concentration was  $13.8 \pm 6.9 \text{ mg kg}^{-1}$ ,  $7.4 \pm 4.0 \text{ mg kg}^{-1}$ , and  $11.6 \pm 3.6 \text{ mg kg}^{-1}$ , respectively. The three test sets have a variation in terms of rice grain and soil total As content, and hence they serve as the ideal sets for testing the models. A positive correlation was also observed between soil As and grain As, test set 1 (0.031,  $p > 0.05$ ), test set 2 (0.6,  $p < 0.05$ ), and test set 3 (0.46,  $p < 0.05$ ), as can be observed from Figure S2. A machine learning model's evaluation is just as crucial as its construction (James et al., 2013). So, testing the LR and DT models on these new and unexplored data sets have led to a complete and comprehensive review for both the models published previously (Mandal et al., 2021).

The TP, FP, TN, and FN values for the three sets with LR and DT is presented in Table 1, showing the model performance matrices for both the models. The most important metric for a classification model is accuracy. From the performance metrics of the LR and DT models, it was observed that in terms of accuracy for all three test sets, the DT outperformed LR and vice versa for misclassification. The ratio between the number of correctly classified samples and the

**TABLE 1** Confusion matrix of the testing data sets

| Model: LR                    |         |         | Model: DT |         |
|------------------------------|---------|---------|-----------|---------|
| Test set 1 ( <i>n</i> = 50)  | Actual  |         | Actual    |         |
| Predicted                    | <MTC    | >MTC    | <MTC      | >MTC    |
| <MTC                         | 19 (TP) | 1 (FP)  | 36 (TP)   | 2 (FP)  |
| >MTC                         | 27(FN)  | 3(TN)   | 10(FN)    | 2 (TN)  |
| Accuracy (%)                 | 44      |         | 76        |         |
| Sensitivity/recall           | 41.30   |         | 78.26     |         |
| Specificity                  | 75.0    |         | 50.0      |         |
| Precision                    | 95.0    |         | 94.73     |         |
| F1 Score                     | 57.57   |         | 85.71     |         |
| MCC                          | 0.10    |         | 0.18      |         |
| Test set 2 ( <i>n</i> = 28)  |         |         |           |         |
| <MTC                         | 22 (TP) | 1 (FP)  | 25 (TP)   | 2 (FP)  |
| >MTC                         | 3 (FN)  | 2 (TN)  | 0 (FN)    | 1 (TN)  |
| Accuracy (%)                 | 85.71   |         | 92.86     |         |
| Sensitivity/recall           | 88.0    |         | 100.0     |         |
| Specificity                  | 66.67   |         | 33.33     |         |
| Precision                    | 95.65   |         | 92.59     |         |
| F1 score                     | 91.66   |         | 96.15     |         |
| MCC                          | 0.40    |         | 0.54      |         |
| Test set 3 ( <i>n</i> = 132) |         |         |           |         |
| <MTC                         | 44 (TP) | 3 (FP)  | 95 (TP)   | 9 (FP)  |
| >MTC                         | 68 (FN) | 17 (TN) | 17 (FN)   | 11 (TN) |
| Accuracy (%)                 | 46.21   |         | 80.30     |         |
| Sensitivity/recall           | 39.29   |         | 84.82     |         |
| Specificity                  | 85.0    |         | 55.0      |         |
| Precision                    | 93.62   |         | 91.35     |         |
| F1 score                     | 49.15   |         | 87.96     |         |
| MCC                          | 0.18    |         | 0.41      |         |

Abbreviations: MCC, Matthews correlation coefficient; MTC, maximum tolerable concentration; TN, true negative; FN, false negative; TP, true positive; FP, false positive.

total number of samples is the most appropriate performance metric (Wang et al., 2007). This is referred to as accuracy, and it works when there are more than two labels (multiclass case). After converting the grain As to categorical variables (<MTC and >MTC) based on the codex recommendation, it was observed that for test set 1, the number of samples was 46 for <MTC and 4 for >MTC; for test set 2, it was 25 for <MTC and 3 for >MTC; and for test set 3, it was 112 for <MTC and 20 for >MTC. When the data set is unbalanced, as in our case (the number of samples in one class is far greater than the number of samples in the other classes), accuracy is no longer a reliable measure since it provides an overoptimistic estimate of the classifier's skill on the majority class (Akosa, 2017; Sokolova et al., 2006). The specificity was the highest in LR compared to DT for all the test sets, whereas the DT had a higher sensitivity compared to LR. The wise rates—true positive rate (or sensitivity, or recall) and true negative rate (or specificity)—is computed for all the possible confu-

sion matrix thresholds. These different combinations of these two metrics give rise to other measures: among them, ROC and AUC are the most important. A higher AUC value indicates a better-performing model (Mandal et al., 2021). From Figure S3, AUC for LR was 58.2%, 77.3%, and 62.1% for test sets 1, 2, and 3, respectively, and for DT the AUC was 64.1%, 66.7%, and 69.9% for test sets 1, 2, and 3, respectively. One of the three test sets (test set 2) have a higher AUC for LR compared to DT. However, ROC and AUC present several flaws (Lobo et al., 2008), and it is sensitive to class imbalance (Hanczar et al., 2010). So F1 score and MCC can be considered as the important model metrics to determine the efficacy of a model. The F1 score for LR model was 57.57, 91.66, and 49.15 for test sets 1, 2, and 3, respectively. For DT, the F1 score was 96.15 for test set 2 followed by 87.96 and 85.71 for test sets 3 and 1, respectively. The MCC for LR was 0.10, 0.40, and 0.18, and for DT, it was 0.18, 0.54, and 0.41 for test sets 1, 2, and 3, respectively. The DT model has an edge over

the LR in terms of both F1 score and MCC. The F1 score is widely used not only in the binary scenario, but also in multiclass events. In multiclass events, researchers can use the F1 micro/macro averaging method (Pillai et al., 2017). The MCC creates a high score only if the classifier correctly predicted most of the positive data cases and most of the negative data cases, and if most of its positive predictions and most of its negative predictions are correct. In fact, regarding MCC and F1, Dubey and Tarar (2018) stated that these two measures “provide more realistic estimates of real-world model performance.” In binary classification tasks, accuracy and F1 score derived on confusion matrices have been (and continue to be) among the most often used measures. However, for unbalanced data sets, these statistical techniques sometimes can produce dangerously overoptimistic inflated outcomes as they fail to reflect the ratio among positive and negative elements (Chicco et al., 2021) and MCC creates a more explanatory and honest parameter in evaluating binary classifications than accuracy and F1 score. The principle of MCC is instinctive and upfront: to get a high-quality score, the classifier must make accurate predictions independently of most of the negative and positive cases of their ratios in the overall data set. In our case, the DT model performs better than LR not only in terms of F1 score but also in terms of MCC. Hence, the DT is a better classifier as compared to LR and 14 mg kg<sup>-1</sup> total As in soil will be an appropriate guideline value for rice crop. For total As, our findings are in agreement with findings of Rahman et al. (2007) who reported that rice grain is not safe for consumption when the soil As is above  $14.5 \pm 0.1$  mg kg<sup>-1</sup>. Similarly, the recommended limit of soil As for safe cultivation of rice as proposed by the Ministry of Environment, Government of Japan is 15 mg kg<sup>-1</sup> (Punshon et al., 2017).

The boxplot in Figure 1 shows a comparison between the three test data sets in terms of category of grain As (<MTC and >MTC) with respect to soil total As and the limits predicted by the LR and DT model. The blue points below the red line (representing 14 mg kg<sup>-1</sup> of total soil As from DT) represent the instances at which the rice grain As was >MTC. These particular instances are due to the fact that in addition to total As in soil the bioavailable or bioaccessible fractions may be playing a significant role, leading to a high uptake of As in rice grain. This warrants further investigation considering the other soil parameters such as pH, OC, available Fe, and P that leads to the next part of the analysis.

### 3.2 | Confusion matrix and performance of RF, GBM, and LR models for bioavailable As

The performance of the RF, GBM, and LR models over the testing and training phase can be observed in Table 2. From the confusion matrix of the RF, GBM and LR model, it was observed that over the training set the model prediction accu-

racy was more in RF and GBM (100) compared LR (91.15) but over the testing set LR (90.24) have an edge over both RF and GBM (87.80). From the ROC in Figure S4 for RF and GBM, the AUC was 100% for training set and 89.0% and 82.4% for testing set, respectively. In case of LR, the AUC for training set was 89.60%, and for testing set, it was 85.2%. Although the AUC followed the order of RF > LR > GBM the accuracy and MCC matrices followed the order of LR > RF  $\cong$  GBM. The log loss for GBM was minimum over the training set followed by RF and LR; however, over the test set it followed the order GBM > LR  $\cong$  RF. The log loss shows how closely the prediction probability resembles the relevant true or real value (0 or 1 in case of binary classification). The higher the log loss number, the more the predicted probability deviates from the actual value (Vovk, 2015). Hence, a lower log loss value means better predictability of the model. From the accuracy, recall, precision, F1 score, and MCC of the test set, it can be concluded that the performance of the LR model was better as compared to both the RF and GBM model. As the test data set was imbalanced (<MTC = 33 and >MTC = 3) from the MCC, it can be concluded that the LR model has an edge over the RF model in terms of correctly predicting both the classes as previously recommended (Chicco & Jurman, 2020). Data collection is prone to errors leading to flaws in the data set. Noise is the name given to the errors. Machine learning algorithms that read data noise as a pattern may start generalizing from it, which might lead to issues. In general, the performance of LR is improved when the number of noise variables is less than or equal to the number of clarifying variables and RF has a higher true and false positive rate as the number of clarifying variables surges in a data set (Kirasich et al., 2018). LR having higher classification accuracy than RF has also been reported by Geng et al. (2006) in predicting colon cancer. Similarly in a financial study by Hao et al. (2016) in predicting “past-due amount,” it was reported that LR was effective in terms of predictive accuracy compared to the RF in case of big and noisy data. Although GBM and RF are excellent, they are not flawless; for instance, in comparison to logistic regression models, gradient boosting techniques typically have poor probability calibration (Niculescu-Mizi & Caruana, 2005). Additionally, certain models are intrinsically more data demanding, so perhaps the data set is simply insufficiently expressive (van der Ploeg et al., 2014) and hence results in a better performance of the LR model compared to RF and GBM.

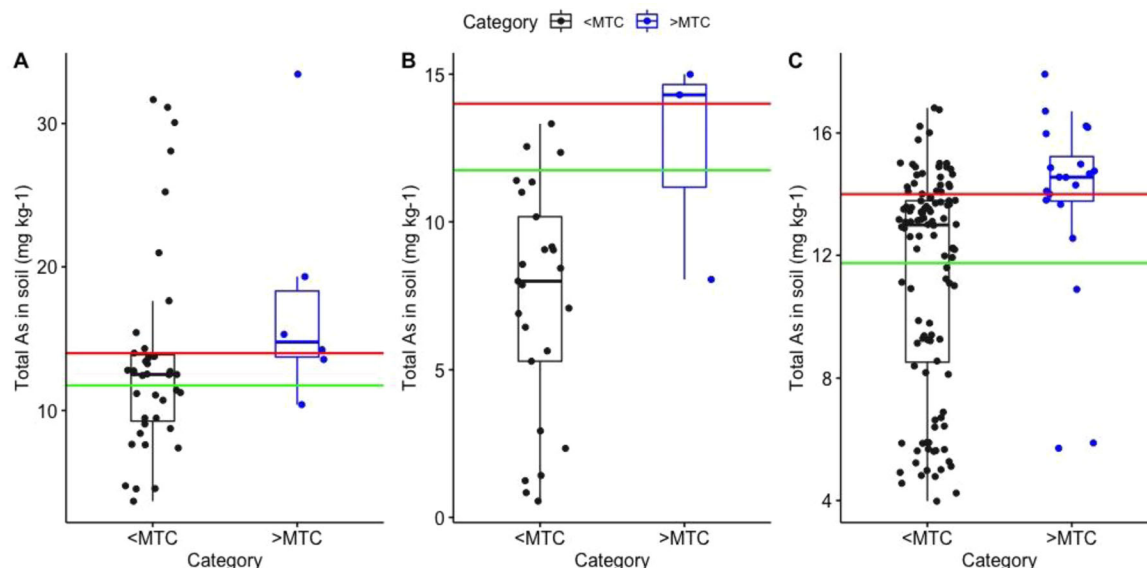
### 3.3 | Variable importance and partial dependence of the variables from better performing RF and LR model

The LR model predicted probability (< MTC | > MTC) =  $-16.82 + 4.79\text{OC} + 0.48\text{AvFe} + 1.22\text{BAs} + 0.15\text{TAs}$

**TABLE 2** Confusion matrix of random forest (RF), gradient boost machine (GBM), and logistic regression (LR) model and model parameters over training and testing phase for bioavailable arsenic (As)

|                        | Training set ( <i>n</i> = 192) |                  | Testing set ( <i>n</i> = 41) |        |
|------------------------|--------------------------------|------------------|------------------------------|--------|
| Random forest          |                                |                  |                              |        |
|                        | Actual                         |                  | Actual                       |        |
| Predicted              | <MTC                           | >MTC             | <MTC                         | >MTC   |
| <MTC                   | 166 (TP)                       | 0 (FP)           | 33 (TP)                      | 3(FP)  |
| >MTC                   | 0 (FN)                         | 26 (TN)          | 2(FN)                        | 3(TN)  |
| Accuracy (%)           | 100                            | 87.80            |                              |        |
| 95% CI                 | (0.981, 1)                     | (0.738, 0.9592)  |                              |        |
| Kappa                  | 1                              | 0.48             |                              |        |
| Sensitivity/recall     | 100                            | 94.29            |                              |        |
| Specificity            | 100                            | 50.00            |                              |        |
| Precision              | 100                            | 91.67            |                              |        |
| Log loss               | 0.074                          | 0.29             |                              |        |
| F1 score               | 100                            | 92.95            |                              |        |
| MCC                    | 1.0                            | 0.47             |                              |        |
| Gradient boost machine |                                |                  |                              |        |
|                        | Actual                         |                  | Actual                       |        |
| Predicted              | <MTC                           | >MTC             | <MTC                         | >MTC   |
| <MTC                   | 166 (TP)                       | 0 (FP)           | 33 (TP)                      | 3(FP)  |
| >MTC                   | 0 (FN)                         | 26 (TN)          | 2(FN)                        | 3(TN)  |
| Accuracy (%)           | 100                            | 87.80            |                              |        |
| 95% CI                 | (0.981, 1)                     | (0.738, 0.9592)  |                              |        |
| Kappa                  | 1                              | 0.48             |                              |        |
| Sensitivity/recall     | 100                            | 94.29            |                              |        |
| Specificity            | 100                            | 50.00            |                              |        |
| Precision              | 100                            | 91.67            |                              |        |
| Log loss               | 0.0009                         | 1.28             |                              |        |
| F1 score               | 100                            | 92.95            |                              |        |
| MCC                    | 1.0                            | 0.47             |                              |        |
| Logistic regression    |                                |                  |                              |        |
|                        | Actual                         |                  | Actual                       |        |
| Predicted              | <MTC                           | >MTC             | <MTC                         | >MTC   |
| <MTC                   | 163 (TP)                       | 14 (FP)          | 34 (TP)                      | 3 (FP) |
| >MTC                   | 3 (FN)                         | 12 (TN)          | 1 (FN)                       | 3 (TN) |
| Accuracy (%)           | 91.15                          | 90.24            |                              |        |
| 95% CI                 | (0.862, 0.9476)                | (0.7687, 0.9728) |                              |        |
| Kappa                  | 0.54                           | 0.54             |                              |        |
| Sensitivity/recall     | 98.19                          | 97.14            |                              |        |
| Specificity            | 46.15                          | 50.00            |                              |        |
| Precision              | 92.09                          | 91.89            |                              |        |
| Log loss               | 0.25                           | 0.31             |                              |        |
| F1 score               | 95.04                          | 94.34            |                              |        |
| MCC                    | 0.56                           | 0.56             |                              |        |

Abbreviations: MCC, Matthews correlation coefficient; MTC, maximum tolerable concentration; TN, true negative; FN, false negative; TP, true positive; FP, false positive.



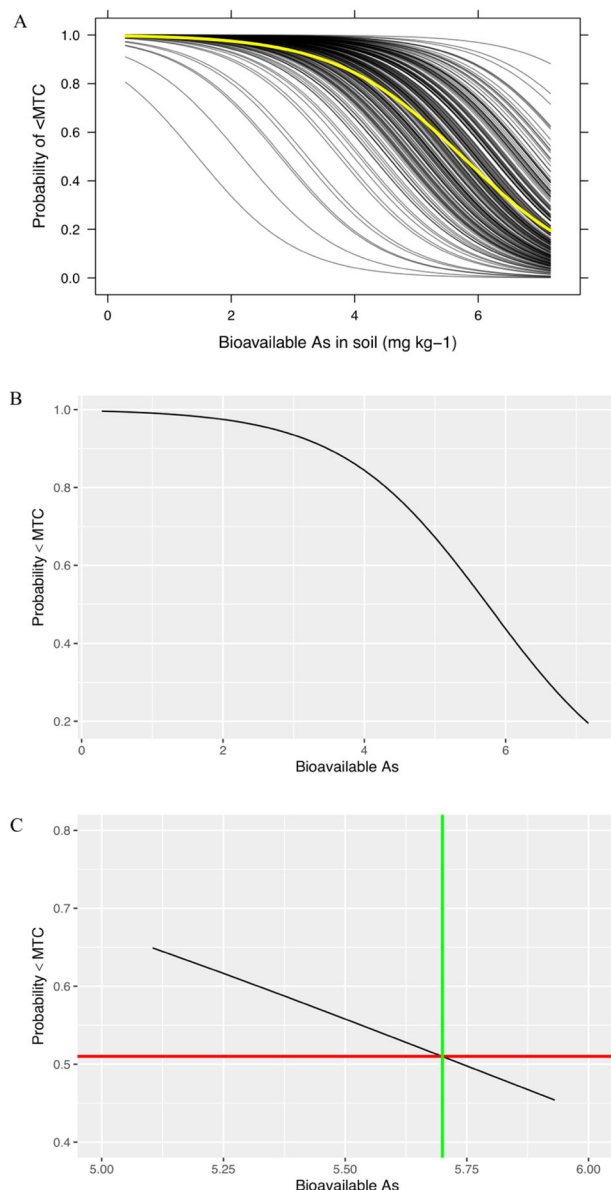
**FIGURE 1** Boxplots of total arsenic (As) in soil ( $\text{mg kg}^{-1}$ ) with respect to category of grain As concentration (<MTC and >MTC, where MTC is maximum tolerable concentration) of three testing data sets. (A: Test set 1 ( $n = 50$ ), B: test set 2 ( $n = 28$ ), C: test set 3 ( $n = 132$ )). The horizontal red line indicates the limit of soil As ( $14 \text{ mg kg}^{-1}$ ) predicted by decision tree, and the green line indicates the limit of soil As ( $11.75 \text{ mg kg}^{-1}$ ) predicted by logistic regression

(AIC = 106.92). The OC, AvFe, Bas, and TAs coefficients significantly ( $p < 0.05$ ) predicted the grain As content. When AvP and pH were considered for the model, the coefficients were statistically nonsignificant ( $p > 0.05$ ) and AIC increased to 110.25. Hence, from the LR model, the BAs, TAs, OC, and AvFe were the most important predictor variables of grain As content. From the variable importance plot of RF model shown in Figure S5, it can be observed that for predicting the category of grain As (<MTC and >MTC), the predictor variables followed the order BAs, TAs, AvFe, OC, and AvP. The pH was not an important predictor variable in either model. The importance of the BAs was 100 followed by TAs (41.75), AvFe (25.77), OC (2.52), and AvP (0.84). The soil As has the highest importance followed by pH, OC, and available soil P concentration on grain As content was previously reported by Sengupta et al. (2021) with RF regression model.

From the ICE (A) and the PDP (B and C) of the LR model shown in Figure 2, it can be observed that at cut-off probability of 0.51 (Figure S6) the limit of soil As to classify grain As < MTC was  $5.70 \text{ mg kg}^{-1}$ . Similarly, at the cut-off probability of 0.62 (Figure S5) from the RF model, it was observed that the limit of soil As to classify grain As < MTC was  $5.72 \text{ mg kg}^{-1}$  (Figure S7) above which the probability of grain As < MTC decreases. The PDP was used to show the marginal influence that one or two features have on the predicted outcome of the LR and RF model. One line per instance is displayed in an ICE plot, illustrating how the instance's prediction alters as a feature changes. The PDP does not focus on each instance but rather an overall average. A partial dependence plot can display whether the relationship between the dependent variable and an independent variable

is linear, monotonic, or more complex. The PDP in Figure 3 and Figure S8 from LR and RF model shows the probability of grain As category (<MTC) with respect to BAs (most important variable) along with other variables for LR (TAs, AvFe, OC) and RF (TAs, AvFe, OC, and AvP). It can be observed that at BAs less than  $5.70 \text{ mg kg}^{-1}$  (from LR) and  $5.72 \text{ mg kg}^{-1}$  (from RF) and TAs less than  $14 \text{ mg kg}^{-1}$  (predicted from the DT model), the probability of grain As < MTC was maximum (1.0–0.8). At BAs of  $5.72 \text{ mg kg}^{-1}$  from RF and  $5.70 \text{ mg kg}^{-1}$  from LR, the AvFe below  $8 \text{ mg kg}^{-1}$  was observed to be effective in keeping the probability of <MTC higher. The OC content between 0.6% and 0.8% was effective in keeping higher probability of grain As < MTC at BAs of  $5.72 \text{ mg kg}^{-1}$  from RF and  $5.70 \text{ mg kg}^{-1}$  from LR. For available P, it was observed that at BAs above  $5.72 \text{ mg kg}^{-1}$  from RF, the AvP was not effective in increasing the probability of grain As < MTC. However, at lower levels of As, AvP was effective for a high probability of <MTC.

To our knowledge, this is the first attempt at predicting the limit of soil bioavailable As using the PDP with respect to the cut-off probability from the models. The threshold or cut-off in a binary classification represents the probability at which most of the predictions are true. It represents the trade-off between false positives and false negatives (Sarkar et al., 2022). Neither a very rigorous nor a very slack threshold limit should be used. As because neither India nor South and South-East Asia as a whole has the luxury of cultivable land sufficient enough to feed the population, nor would a free acceptable limit help to adequately protect human health from As hazards. So, model accuracy was considered as the parameter for determining the cut-off probability rather than



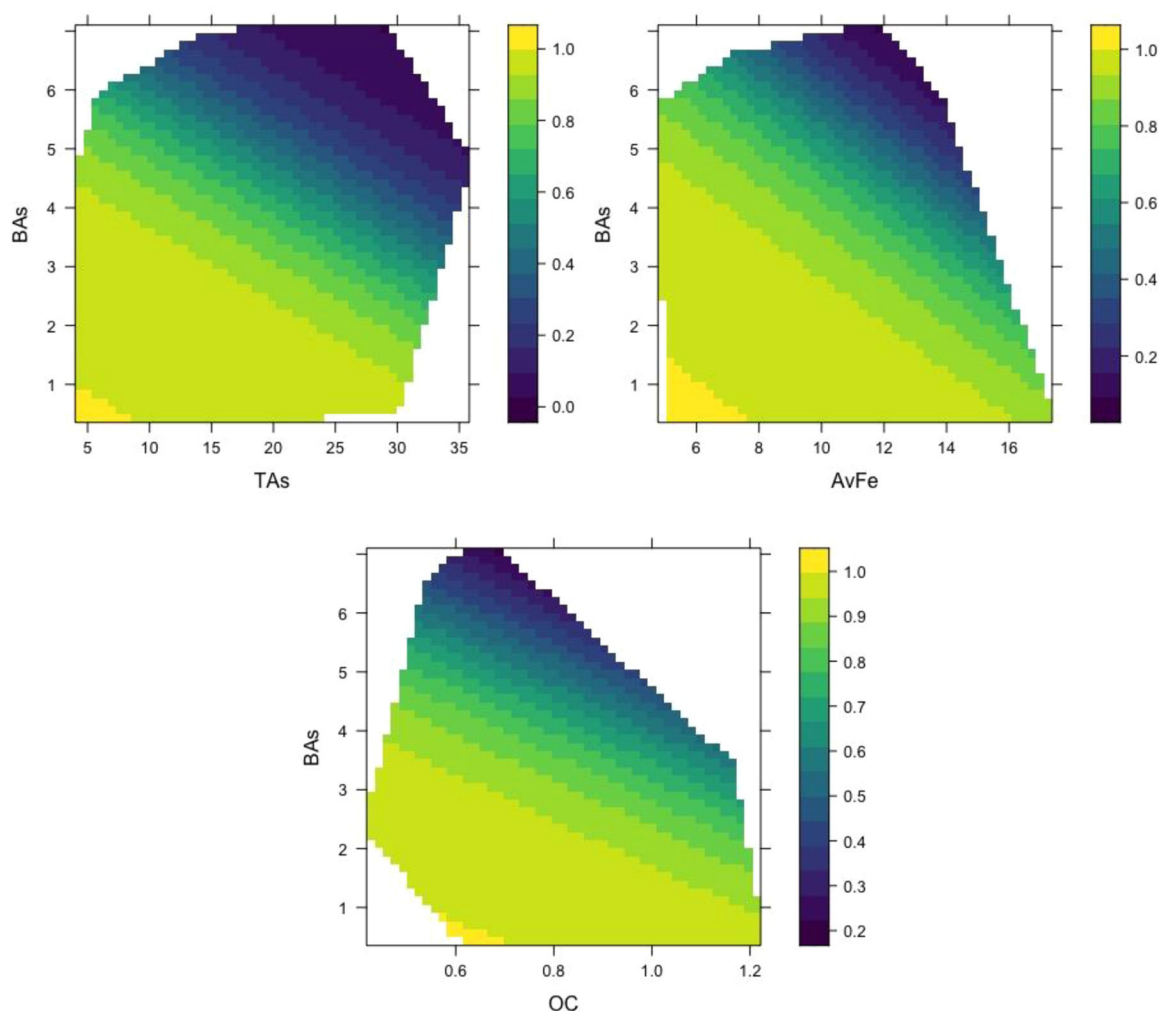
**FIGURE 2** Individual conditional expectation (ICE) and partial dependence plots (PDP) of available As ( $\text{mg kg}^{-1}$ ) from logistic regression model with respect to probability of grain As < MTC (maximum tolerable concentration)

sensitivity-specificity (Figure S9). Previous attempts in determining the safe limit of As in soil ( $\text{NaHCO}_3$  extractable As) in relation to dietary exposure from consumption of rice was undertaken by Golui et al. (2017) only for one specific area (Maldah district, West Bengal, India) and considering only two variables pH and OC over a limited data set. Although from both the models the predicted limit for available As was very close to each other, the LR model performed better and we, therefore, suggest that  $5.70 \text{ mg kg}^{-1}$  should be considered as the limit of soil available As. Previously PDP from boosted regression trees and RF were used to predict the probability of As exceedance in groundwater on the important variables (Fe

and P) by Tan et al. (2020). The visualization of two variables at once with BAs through PDP gives us an insight into the effect of changes in the variables on the probability of grain As. The PDP of BAs and TAs on grain As reveals that below the predicted limit of available As ( $5.70 \text{ mg kg}^{-1}$  from this study) and total As ( $14 \text{ mg kg}^{-1}$  from Mandal et al., 2021 and tested with field data in this study), the probability of grain As < MTC was maximum. The relationship between BAs and AvFe revealed that Fe aids in the reduction of As absorption in rice. Previously, it has been reported the use of Fe causes the formation of oxides of Fe in form of Fe plaques close to rice plant roots, which reduces As uptake, and increases co-precipitation of Fe and As (Lee et al., 2012). Metallic Fe and Fe-oxide have been observed to decline As accretion in rice by 51% and 47% (Matsumoto et al., 2015). BAs and OC relationship revealed the fact that the presence of organic matter within the soil can restrict the availability of As and its uptake by rice. Soil organic fractions that comprise humic acid (HA) and fulvic acid (FA) behave as an active binder of As through metal-humate complexes of variable stability (Kumar et al., 2021; Sengupta et al., 2022). The application of organic amendments reducing the As uptake in rice has been reported from the field experiments conducted by Sengupta et al. (2021). Phosphorous competes with arsenate (AsV) for similar adsorption sites both in the soil and on the Fe plaques mainly by ligand exchange that is a key characteristic in the rice field for bioavailability of As and uptake by roots (Peryea et al., 1995). This explains the relationship of BAs with AvP. Lee et al. (2016) proposed key factors prompting movement of As in soil and its uptake by rice: antagonism between As and P for adsorption sites and during transport in rice roots, lastly role of P in transfer of As from root to shoot. Thus, levels of AvFe, OC, and AvP (as shown in PDPs) in the soil at which the BAs will be below the projected limit would aid in the development of acceptable management techniques to mitigate As buildup in rice.

## 4 | CONCLUSIONS

Based on the model metrics, the DT model has an edge over the LR model and hence  $14 \text{ mg kg}^{-1}$  of total As in soil will be a proper guideline value below which rice cultivated in fields will not surpass the Codex recommendation. From the better-performing LR model, it was observed that BAs, TAs, AvFe, and OC were the most important variables for grain As. The PDPs of the LR model predicted the limit for bioavailable As to be  $5.70 \text{ mg kg}^{-1}$ . It is well known that Fe, P, and organic matter are used as amendments for reducing the As accumulation in crops. Thus, levels of AvFe, OC, and AvP (as shown in PDPs) in the soil at which the BAs will be below the limit would aid in the development of acceptable management techniques to mitigate As buildup in rice. In future



**FIGURE 3** Partial dependence plots (PDP) of two variables, bioavailable As (BAs) ( $\text{mg kg}^{-1}$ ) with other significant variables total As (TAs), available iron (AvFe), ( $\text{mg kg}^{-1}$ ) and organic carbon (OC) (%) from logistic regression (LR) model. Probability of <MTC (maximum tolerable concentration) is depicted in terms of color intensities

studies, manganese can also be considered as a covariate of the bioavailability of As. In spite of the uncertainties and inherent limitations of the models brought on by the lack of appropriate field data, this is a novel way of predicting the grain As content. Despite collecting paired soil and rice grain samples during different seasons and from different sites, data imbalance was observed. The efficacy of a model depends on its predictability of different types of data (balanced or imbalanced). So, from the MCC, it was observed that the LR model (predicting BAs) has an edge over the RF. Hence, the model can predict both balanced and imbalanced data sets. As the models have been developed using a specific set of data from a specific geographical region, it would be naïve to think that they could be applied to all contaminated rice growing sites globally. However, testing and fine-tuning the models with more field data will enhance their applicability and will serve as a protocol to derive site-specific regulatory limits.

## AUTHOR CONTRIBUTIONS

**Jajati Mandal:** Field work; analysis; writing—original draft; reviewing. **Vinay Jain:** Analysis. Sudip Sengupta and Debasish Golui: Field work; analysis. **Md. Aminur Rahman:** Analysis; Kallol Bhattacharyya: Supervision. **Mohammad Mahmudur Rahman:** Analysis; editing; reviewing. **Michael D. Wood:** Supervision; reviewing; editing. **Debapriya Mondal:** Conceptualization; supervision; reviewing; editing.

## ACKNOWLEDGMENTS

The authors also acknowledge the Netaji Subhas—ICAR International Fellowship provided to Jajati Mandal. Laboratory facilities and staff support provided by University of Salford, Manchester, Centre of Excellence, Agilent Technologies (International) Pvt. Ltd, Manesar and GCER, the University of Newcastle is highly appreciated.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data will be made available on request.

## ORCID

Jajati Mandal  <https://orcid.org/0000-0003-0814-0369>

Debasis Golui  <https://orcid.org/0000-0002-8554-1820>

Michael D. Wood  <https://orcid.org/0000-0002-0635-2387>

Debapriya Mondal  <https://orcid.org/0000-0002-5144-626X>

## REFERENCES

- Akosa, J. S. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. *Proceedings of the SAS Global Forum 2017 Conference*. SAS Institute Inc.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(13), 1–22.
- Dipti, S. S., Bergman, C., Indrasari, S. D., Herath, T., Hall, R., Lee, H., Habibi, F., Bassinello, P. Z., Graterol, E., Ferraz, J. P., & Fitzgerald, M. (2012). The potential of rice to offer solutions for malnutrition and chronic diseases. *Rice (N Y)*, 5(16), 16–18. <https://doi.org/10.1186/1939-8433-5-16>
- Dubey, A., & Tarar, S. (2018). Evaluation of approximate rank-order clustering using Matthews correlation coefficient. *International Journal of Engineering and Advanced Technology*, 8(2), 106–113.
- FAO. (2017). Rice market monitor. <https://www.fao.org/3/19243EN/i9243en.pdf>
- Fazle Bari, A. S. M., Lamb, D., Choppala, G., Seshadri, B., Islam, M. d. R., Sanderson, P., & Rahman, M. M. (2021). Arsenic bioaccessibility and fractionation in abandoned mine soils from selected sites in New South Wales, Australia and human health risk assessment. *Ecotoxicology and Environmental Safety*, 223, 112611. <https://doi.org/10.1016/j.ecoenv.2021.112611>
- Franke, G. R. (2010). *Multicollinearity*. Wiley international encyclopedia of marketing (pp. 197–198). Wiley-Blackwell. <https://doi.org/10.1002/9781444316568.wiem02066>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Geng, & Ming (2006). *A comparison of logistic regression to random forests for exploring differences in risk factors associated with stage at diagnosis between black and white colon cancer patients* [Unpublished master's thesis]. Graduate School of Public Health, University of Pittsburgh.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. <https://doi.org/10.1080/10618600.2014.907095>
- Golui, D., Guha Mazumder, D. N., Sanyal, S. K., Datta, S. P., Ray, P., Patra, P. K., Sarkar, S., & Bhattacharya, K. (2017). Safe limit of arsenic in soil in relation to dietary exposure of arsenicosis patients from Malda district, West Bengal—A case study. *Ecotoxicology and Environmental Safety*, 144, 227–235. <https://doi.org/10.1016/j.ecoenv.2017.06.027>
- GRiSP (Global Rice Science Partnership). (2013). *Rice almanac* (4th ed). International Rice Research Institute.
- Groenenberg, J. E., Romkens, P. F. A. M., Zomeren, A. V., Rodrigues, S. M., & Comans, R. N. J. (2017). Evaluation of the single dilute (0.43 M) nitric acid extraction to determine geochemically reactive elements in soil. *Environmental Science & Technology*, 51, 2246–2253.
- Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, 26(6), 822–830. <https://doi.org/10.1093/bioinformatics/btq037>
- Hao, J., & Priestley, J. L. (2016). *A comparison of machine learning techniques and logistic regression method for the prediction of pasture amount*. Grey literature from PhD candidates.
- Hussain, M. M., Bibi, I., Niazi, N. K., Shahid, M., Iqbal, J., Shakoar, M. B., Ahmad, A., Shah, N. S., Bhattacharya, P., Mao, K., Bundschuh, J., Ok, Y. S., & Zhang, H. (2021). Arsenic biogeochemical cycling in paddy soil-rice system: Interaction with various factors, amendments and mineral nutrients. *Science of the Total Environment*, 773, 145040. <https://doi.org/10.1016/j.scitotenv.2021.145040>
- ISO. (2016). *ISO/DIS 17586 Soil quality—Extraction of trace elements using dilute nitric acid*. Alterra, Wageningen-UR.
- Jackson, M. L. (1973). *Soil chemical analysis* (p. 498). Prentice Hall India Pvt. Ltd.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.
- JECFA. (2017). *Report of the eleventh session of the codex committee on contaminants in foods*. Joint FAO/WHO Food Standards Programme Codex Alimentarius Commission.
- Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3), 9.
- Kumar, S. N., Mishra, B. B., Kumar, S., & Mandal, J. (2021). Organo-arsenic complexation studies explaining the reduction of uptake of arsenic in wheat grown with contaminated irrigation water and organic amendments. *Water Air and Soil Pollution*, 232, 118. <https://doi.org/10.1007/s11270-021-05065-8>
- Kumari, P. B., Singh, Y. K., Mandal, J., Shambhavi, S., Sadhu, S. K., Kumar, R., Ghosh, M., Raj, A., & Singh, M. (2021). Determination of safe limit for arsenic contaminated irrigation water using solubility free ion activity model (FIAM) and tobit regression model. *Chemosphere*, 270, 128630. <https://doi.org/10.1016/j.chemosphere.2020.128630>
- Lee, C.-H., Hsieh, Y. i.-C., Lin, T.-H., & Lee, D.-Y. (2012). Iron plaque formation and its effect on arsenic uptake by different genotypes of paddy rice. *Plant and Soil*, 363, 231–241. <https://doi.org/10.1007/s11104-012-1308-2>

- Lee, C.-H., Wu, C.-H., Syu, C.-H., Jiang, P.-Y., Huang, C. - C., & Lee, D. - Y. (2016). Effects of phosphorous application on arsenic toxicity to and uptake by rice seedlings in As-contaminated paddy soils. *Geoderma*, 270, 60–67. <https://doi.org/10.1016/j.geoderma.2016.01.003>
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), 145–51. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Majumder, S., & Banik, P. (2019). Geographical variation of arsenic distribution in paddy soil, rice and rice-based products: A meta-analytic approach and implications to human health. *Journal of Environmental Management*, 233, 184–199. <https://doi.org/10.1016/j.jenvman.2018.12.034>
- Mandal, J., Sengupta, S., Sarkar, S., Mukherjee, A., Wood, M. D., Hutchinson, S. M., & Mondal, D. (2021). Meta-analysis enables prediction of the maximum permissible arsenic concentration in Asian paddy soil. *Frontiers in Environmental Science*, 9, 760125. <https://doi.org/10.3389/fenvs.2021.760125>
- Matsumoto, S., Kasuga, J., Taiki, N., Makino, T., & Arao, T. (2015). Inhibition of arsenic accumulation in Japanese rice by the application of iron and silicate materials. *Catena*, 135, 328–335. <https://doi.org/10.1016/j.catena.2015.07.004>
- Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253–267. <https://doi.org/10.1080/09720502.2010.10700699>
- Mondal, D., Banerjee, M., Kundu, M., Banerjee, N., Bhattacharya, U., Giri, A. K., Ganguli, B., Sen Roy, S., & Polya, D. A. (2010). Comparison of drinking water, raw rice and cooking of rice as arsenic exposure routes in three contrasting areas of West Bengal, India. *Environmental Geochemistry and Health*, 32, 463–477. <https://doi.org/10.1007/s10653-010-9319-5>
- Mondal, D., Periche, R., Tineo, B., Bermejo, L. A., Rahman, M. M., Siddique, A. B., Rahman, M. d. A., Solis, J. L., & Cruz, G. J. F. (2020). Arsenic in peruvian rice cultivated in the major rice growing region of Tumbes river basin. *Chemosphere*, 241, 125070. <https://doi.org/10.1016/j.chemosphere.2019.125070>
- Mondal, D., & Polya, D. A. (2008). Rice is a major exposure route for arsenic in Chakdaha block, Nadia district, West Bengal, India: A probabilistic risk assessment. *Applied Geochemistry*, 23, 2987–2998. <https://doi.org/10.1016/j.apgeochem.2008.06.025>
- Mwale, T., Rahman, M., & Mondal, D. (2018). Risk and benefit of different cooking methods on essential elements and arsenic in rice. *International Journal of Environmental Research and Public Health*, 15(6), 1056. <https://doi.org/10.3390/ijerph15061056>
- Niculescu-Mizi, A., & Caruana, R. A. (2005). Obtaining calibrated probabilities from boosting. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI2005)*. <https://doi.org/10.48550/arXiv.1207.1403>
- Olsen, S. R., Cole, C. V., Watanabe, F. S., & Dean, L. A. (1954). *Estimation of available phosphorus in soils by extraction with sodium bicarbonate* (Vol. 939, pp. 1–19). Government. Printing Office Washington DC, USDA Circular.
- Peryea, F. J., & Kammereck, R. (1995). Phosphate-enhanced movement of arsenic out of lead arsenate-contaminated top soil and through uncontaminated sub soil. *Water Air & Soil Pollution*, 93, 243–254.
- Pillai, I., Fumera, G., & Roli, F. (2017). Designing multi-label classifiers that maximize F measures: State of the art. *Pattern Recognition*, 61, 394–404. <https://doi.org/10.1016/j.patcog.2016.08.008>
- Punshon, T., Jackson, B. P., Meharg, A. A., Warczack, T., Scheckel, K., & Guerinot, M. L. (2017). Understanding arsenic dynamics in agro-economic systems to predict and prevent uptake by crop plants. *Science of the Total Environment*, 581–582, 209–220. <https://doi.org/10.1016/j.scitotenv.2016.12.111>
- Rahman, M. A., Hasegawa, H., Rahman, M. M., Rahman, M. A., & Miah, M. A. M. (2007). Accumulation of arsenic in tissues of rice plant (*Oryza sativa* L.) and its distribution in fractions of rice grain. *Chemosphere*, 69(6), 942–948. <https://doi.org/10.1016/j.chemosphere.2007.05.044>
- Raj, A., Mandal, J., Golui, D., Sihi, D., Dari, B., Kumari, P. B., Ghosh, M., & Ganguly, P. (2021). Determination of suitable extractant for estimating plant available arsenic in relation to soil properties and predictability by solubility-FIAM. *Water Air & Soil Pollution*, 232–247.
- Sarkar, S., Mukherjee, A., Gupta, S. D., Bhanja, S. N., & Bhattacharya, A. (2022). Predicting regional-scale elevated groundwater nitrate contamination risk using machine learning on natural and human-induced factors. *ACS ES&T Engineering*, 2, 689–702.
- Schmidt, C. W. (2015). In search of “Just Right”: The challenge of regulating arsenic in rice. *Environment Health Perspectives*, 123(1), A16–A19. <https://doi.org/10.1289/ehp.123-A16>
- Sengupta, S., Bhattacharyya, K., Mandal, J., Bhattacharya, P., Halder, S., & Pari, A. (2021). Deficit irrigation and organic amendments can reduce dietary arsenic risk from rice: Introducing machine learning-based prediction models from field data. *Agriculture Ecosystem and Environment*, 319, 107516. <https://doi.org/10.1016/j.agee.2021.107516>
- Sengupta, S., Bhattacharyya, K., Mandal, J., & Chattopadhyay, A. P. (2022). Complexation, retention and release pattern of arsenic from humic/fulvic acid extracted from zinc and iron enriched vermicompost. *Journal of Environmental Management*, 318, 115531. <https://doi.org/10.1016/j.jenvman.2022.115531>
- Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *Proceedings of Advances in Artificial Intelligence (AI 2006), Lecture Notes in Computer Science*, (Vol. 4304, pp. 1015–1021). Springer.
- Tan, Z., Yang, Q., & Zheng, Y. (2020). Machine learning models of groundwater arsenic spatial distribution in Bangladesh: Influence of holocene sediment depositional history. *Environmental Science and Technology*, 54, 9454–9463. <https://doi.org/10.1021/acs.est.0c03617>
- Van Der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14, 137. <https://doi.org/10.1186/1471-2288-14-137>
- Vovk, V. (2015). The fundamental nature of the log loss function. In L. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, & W. Schulte (Eds.), *Fields of logic and computation II. Lecture notes in*

- computer science*, (Vol. 9300). Springer. [https://doi.org/10.1007/978-3-319-23534-9\\_20](https://doi.org/10.1007/978-3-319-23534-9_20)
- Wang, L., Chu, F., & Xie, W. (2007). Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1), 40–53. <https://doi.org/10.1109/TCBB.2007.1006>
- Yao, B.-M., Chen, P., Zhang, H.-M., & Sun, G.-X. (2021). A predictive model for arsenic accumulation in rice grains based on bioavailable arsenic and soil characteristics. *Journal of Hazardous Materials*, 412, 125131. <https://doi.org/10.1016/j.jhazmat.2021.125131>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Mandal, J., Jain, V., Sengupta, S., Rahman, M. A., Bhattacharyya, K., Rahman, M. M., Golui, D., Wood, M. D., & Mondal, D. (2023). Determination of bioavailable arsenic threshold and validation of modeled permissible total arsenic in paddy soil using machine learning. *Journal of Environmental Quality*, 52, 315–327. <https://doi.org/10.1002/jeq2.20452>