Environment International 153 (2021) 106529

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

# Living with arsenic in the environment: An examination of current awareness of farmers in the Bengal basin using hybrid feature selection and machine learning

Debasish Mishra [a], Bhabani S. Das [a], Tathagata Sinha [a], Jiaul M. Hoque [b], Christian Reynolds [c,d,e], M. Rafiqul Islam [b], Mahmud Hossain [b], Pinaki Sar [a], Manoj Menon [c,*]

[a] *Indian Institute of Technology Kharagpur, WB 721302, India*
[b] *Bangladesh Agricultural University, Mymensingh 2202, Bangladesh*
[c] *Department of Geography, University of Sheffield, S102TN, United Kingdom*
[d] *Centre for Food Policy, City, University of London, Myddelton Street Building, Myddelton Street, EC1R 1UW London, United Kingdom*
[e] *Barbara Hardy Institute, UniSA STEM, Mawson Lakes Blvd, Mawson Lakes, SA 5095, Australia*

ARTICLE INFO

Handling Editor: Thanh Nguyen

*Keywords:*
Arsenic
Awareness
Survey
Hybrid feature selection
Machine learning
Boruta

ABSTRACT

High levels of arsenic in drinking water and food materials continue to pose a global health challenge. Over 127 million people alone in Bangladesh (BD) and West Bengal (WB) state of India are exposed to elevated levels of arsenic in drinking water. Despite decades of research and outreach, arsenic awareness in communities continue to be low. Specifically, very few studies reported arsenic awareness among low-income farming communities. A comprehensive approach to assess arsenic awareness is a key step in identifying research and development priorities so that appropriate stakeholder engagement may be designed to tackle arsenic menace. In this study, we developed a comprehensive arsenic awareness index (CAAI) and identified key awareness drivers (KADs) of arsenic to help evaluate farmers' preferences in dealing with arsenic in the environment. The CAAI and KADs were developed using a questionnaire survey in conjunction with ten machine learning (ML) models coupled with a hybrid feature selection approach. Two questionnaire surveys comprising of 73 questions covering health, water and community, and food were conducted in arsenic-affected areas of WB and BD. Comparison of CAAIs showed that the BD farmers were generally more arsenic-aware (CAAI = 7.7) than WB farmers (CAAI = 6.8). Interestingly, the reverse was true for the awareness linked to arsenic in the food chain. Application of hybrid feature selection identified 15 KADs, which included factors related to stakeholder interventions and cropping practices instead of commonly perceived factors such as age, gender and income. Among ML algorithms, classification and regression trees and single C5.0 tree could estimate CAAIs with an average accuracy of 84%. Both communities agreed on policy changes on water testing and clean water supply. The CAAI and KADs combination revealed a contrasting arsenic awareness between the two farming communities, albeit their cultural similarities. Specifically, our study shows the need for increasing awareness of risks through the food chain in BD, whereas awareness campaigns should be strengthened to raise overall awareness in WB possibly through media channels as deemed effective in BD.

## 1. Introduction

Inorganic arsenic (iAs) is a group 1 carcinogen (IARC, 2012) and is considered as one of the top ten chemicals having significant public health concern (WHO, 2020). Specifically, the long-term exposure to iAs is known to affect almost every organ of the human body with symptoms

*Abbreviations:* iAs, inorganic arsenic; WB, West Bengal; BD, Bangladesh; CAAI, comprehensive arsenic awareness index; KAD, key awareness driver; ML, machine learning; kNN, k-nearest neighbours; NB, naïve bayes; RF, random forest; SVML, support vector machines (Linear); SVMR, support vector machines (Radial), CART, classification and regression tree; MLP, multi-layer perceptron; SGB, stochastic gradient boosting; MARS, multivariate adaptive regression splines; SCT, single C5.0 tree.

[*] Corresponding author.
*E-mail address:* m.menon@sheffield.ac.uk (M. Menon).

https://doi.org/10.1016/j.envint.2021.106529
Received 9 January 2021; Received in revised form 9 March 2021; Accepted 12 March 2021

ranging from skin lesions to cancers (Kapaj et al., 2006; Rahman et al., 2009; Mandal and Suzuki, 2002). Moreover, adverse physical health effects may also influence an individual's mental health with the possibilities of uncertainty, injustice, and isolation both within a society and in families (Hassan et al., 2005). With a spectrum of toxicity to plants, animals, and humans, ~500 million people across 108 countries (e.g., 32 countries from Asia and 31 countries from Europe among others) live with arsenic in environments where iAs concentrations exceed the 0.01 mg $L^{-1}$ limit set by the WHO (Shaji et al., 2020). With widespread occurrence and formidable health impacts, the arsenic problem is portrayed as 'a curse from God' or 'act of the devil' (Chowdhury et al., 2006).

The Ganga-Bramhaputra-Meghna (GBM) delta of the Bengal Basin is an arsenic hotspot of the world (Mukherjee and Fryar, 2008). More than 50 million people in West Bengal (WB) and 77 million people in Bangladesh (BD) are exposed to > 0.01 mg $L^{-1}$ of arsenic in drinking water (Chakraborty et al., 2015). Nine out of 23 districts of WB alone are affected by arsenic with almost 38,861 $km^2$ area identified as the highly contaminated zone in the state (Chakraborti et al., 2009; Chakraborty et al., 2015). Rahman et al. (2003) analyzed fourty eight thousand and thirty water samples from hand tubewells of North 24-Parganas and found around 53% samples having arsenic above 0.01 mg $L^{-1}$. A decade later, another similar study by Rahman et al. (2014) in Nadia district, reported around 51% of tubewell samples having arsenic above 0.01 mg $L^{-1}$, with arsenical skin lesions prevalence rate of 7.1%. In Bangladesh, arsenic contents in drinking water of 61 districts (out of the total 64) exceed the WHO's maximum permissible limit (Chakraborti et al., 2010; Saha et al., 2019). Specifically, a substantial number of sampled wells in several districts have shown arsenic contents exceeding 0.05 mg $L^{-1}$ (e.g., 90% for Chandpur; 83% for Munshiganj; 69% for Madaripur and Noakhali; and 65% for Comilla, Faridpur, and Shariatpur among others). In addition to this, an emerging body of literature focuses on iAs exposure through principal staples such as rice (Mondal et al., 2010; Srivastava, 2020; Upadhyay et al., 2019; Xu et al., 2020). For example, Joseph et al. (2015) reported both drinking water and daily foodstuff as principal sources of arsenic exposure for the affected population of the Bengal Delta Plain. In some cases, food is reported to play a dominant role over drinking water towards intake of arsenic into the human body (Signes et al., 2008; Liu et al., 2010) and, often, infants and small children are more vulnerable than adults (Carlin et al., 2016; Carey et al., 2018; Menon et al., 2020a).

Although arsenic research is rapidly progressing over the last decades (>1000 publications per year with arsenic in their title; Carlin et al., 2016), we do not know if the wider public is aware of its risks. Specifically, some of the misconceptions may still exist in rural communities. For instance, uncertainty continues with whether "(1) Safe wells will remain safe (2) National water quality standards are safe (3) Providing safe water will end arsenic poisoning and arsenic-related disease (4) Arsenic in food is less harmful (5) Skin lesions are indicative of those affected by arsenic poisoning" (WHO, 2018). Similarly, it is often believed that skin lesions are typical symptoms, although this does not mean that everyone exposed to arsenic will develop such symptoms; moreover, the health impacts can vary significantly across a population. Periodic water testing and colouring the tube wells is required to meet the local water quality standards to address 1 and 2 above; both 3 and 4 are directly related to the exposure through food because exposure is not limited to drinking water alone. Therefore, increased awareness of arsenic in both drinking water and food chain is a crucial starting point to address these uncertainties.

An epidemiological survey in the Nadia districts of WB revealed that almost 15% of 10,469 participants from 37 villages suffer from arsenicosis and people, in general, have poor arsenic awareness (Mazumder et al., 2010). Recently, Singh et al. (2018) evaluated awareness of arsenic in Bihar, another severely affected Indian state. Using questionnaire survey and machine learning (ML) approaches, these authors identified the key drivers of arsenic awareness in terms of socio-economic factors (caste, education level, and occupation), water and sanitation, and social capital and trust factors. In contrast to WB, arsenic awareness studies have been reported from Bangladesh nearly a decade and a half ago (Paul, 2004; Parvez et al., 2006). Paul (2004) surveyed in both low- and medium-risk districts of Bangladesh and observed that arsenic awareness primarily depends on the level of education, gender, age of people, and the severity of the arsenic problem (i.e., respondents from the medium-risk districts were more aware of the arsenic problem than the low-risk districts). Interestingly, most of the respondents of that study were unaware of the arsenic-induced health problems and mitigation measures available to prevent arsenic contamination. Similarly, Parvez et al. (2006) conducted an intensive survey comprising of Ca. 6,000 respondents from a small study area (25 $km^2$) of the Araihazar Upazila in Bangladesh and observed that the age, sex, occupation, type of housing, number of tube wells in *bari* (house) play significant roles in arsenic awareness.

WB shares an international border with BD and shares a similar language (Bengali) and culture. Despite these similarities, there are policy and implementation differences in how each region is dealing with the arsenic problem through infrastructure, provision to supply clean water, monitoring of water quality, international support, and awareness campaigns. For example, the National Policy for Arsenic Mitigation is in place in BD since 2004 (Government of Bangladesh, 2004). The policy states that alternative water supply options will be implemented to ensure safe water for drinking and cooking in the arsenic-contaminated areas. The policy mandates the identification of arsenicosis cases all over the country and the implementation of an effective management system. Moreover, provisions have also been made to evaluate the effects of arsenic on agriculture and develop possible management strategies. Similarly, the Planning Commission, Government of India (Planning Commission, 2007) has set up a task force in WB and proposed short-, medium- and long-term recommendations for tackling arsenic contamination. Specifically, to promote awareness, the task force recommended regular home visits by the health workers for identifying arsenicosis patients, imparting training to local clubs, NGOs, and charities to promote awareness, and distribution of posters and pamphlets. Despite such national efforts, the arsenic awareness in communities continue to be low (Singh et al., 2018).

With limited arsenic awareness data, this study was designed to comprehensively evaluate arsenic awareness among local farming communities of two neighbouring geographies of WB and BD. We conducted a structured survey in both the countries to identify a) the extent of arsenic awareness among the farming community members; b) factors that indirectly lead to such awareness build up in a person, and c) the choices a person makes in response to arsenic awareness and knowledge build-up as a part of living with arsenic in a community. We used a novel data analytic approach by combining the capabilities of hybrid feature selection and ML algorithms to process complex community responses with a mix of both numerical and categorical data. Results of this study will help develop future public-health interventions based on the current challenges, local needs, and perspectives of the communities living with arsenic in the environment.

## 2. Methods

### 2.1. Study area and questionnaire survey

Two questionnaire surveys were conducted in selected arsenic affected areas (or hotspots) of West Bengal in India and Bangladesh. This contiguous region constitutes a part of the greater Gangetic basin. The study area in West Bengal included four districts (Nadia, North 24 Parganas, South 24 Parganas, and Purba Burdwan) known to be severely affected by arsenic contamination (Planning Commission, 2007). The questionnaire survey was conducted on 181 participants covering six blocks (Haringhata, Chakdaha, Nakshipara, Bethuadohari, Krishnanagar, and Chapra) of Nadia, four blocks (Deganga, Barasat-1, Amdanga, and Guma) of North 24 Parganas, and one block each of Purba Bardhaman (Kalna-1) and South 24 Parganas (Baruipur) districts (from now on referred to as WB dataset). For the BD survey, 200 participants were selected from two arsenic affected districts (Faridpur and Chandpur) covering three Upazilas (sub-districts) such as Faridpur Sadar and Bhanga Upazila and Chandpur Sadar Upazila (from now on referred to as BD dataset). Surveys were conducted during May-July 2019 in WB and August-September 2019 in BD.

The questionnaire consisted of 73 questions (Table S1, Supplementary Material) for which the ethical approval was obtained through the Department of Geography, University of Sheffield, UK. The first 12 survey questions (SQs) centred around a person's awareness on how one's health is affected by arsenic (SQ1-4), the presence or absences of arsenic in drinking and irrigation water sources, awareness of social problems and mitigation options (SQ5-8), and arsenic linked to food chain within a community (SQ9-12). A set of 49 questions (SQ13-61) were designed to understand how a person's level of education, social position, and different community attributes may contribute to arsenic awareness. Thus, these 49 questions are indirectly linked to the arsenic awareness of an individual. Because most of our responses were categorical variables, we first coded available responses for each question (Table S3-S5 in Suppl. Materials). The remaining 12 questions (SQ62-73) were designed to explain how an affected community member responds to arsenic menace while living with arsenic in the environment. Note that we adapted some questions to fit the regional requirements (e.g., unit and measurements, currency, and some response options).

### 2.2. Development of an arsenic awareness index and hybrid feature selection

Out of 73 questions, we identified 12 questions (SQ1-12) to develop an arsenic awareness index similar to the one developed by Singh et al. (2018). Eight out of these 12 questions used in this study were also used by Singh et al. (2018), and an additional four questions (SQ9-12) were primarily focused on the awareness of arsenic in the food chain as stated before. Each question was given equal weightage, and the responses were scored as 0 (not aware) and 1 (aware). Individual scores were then summed up to create a comprehensive arsenic awareness index (CAAI). Further, we reclassified these CAAI scores into a) low awareness (score: 1–4), b) moderate awareness (5–8), and c) high awareness (9–12) categories. However, it is essential to note that there is little agreement in the previous literature on how to follow a common approach in dealing with awareness studies - every study had its way of scoring the parameters. For instance, different scores were attributed to different responses by Paul (2004), while Singh et al. (2018) used equal weightage for scoring the attributes making up the awareness index scores. Hence, this did not yield a universal system of arriving at a final set of features, and there was a need to come up with a reproducible method to evaluate awareness among the respondents. Therefore, a new hybrid feature selection procedure was adopted to identify key awareness drivers (KADs) indirectly linked to arsenic awareness using 49 questions (Table S1, SQ13-61 in Suppl. Material) of the WB dataset. Data processing and analyses were done using the caret package (Kuhn et al., 2014) in the R environment (RStudio, 2020; ver. 1.2.5033).

The hybrid feature selection procedure involved two stages of filtering (chi-square analysis) and wrapping, as shown in Fig. 1. However, chi-square ($\chi^2$) analysis alone does not exclude the possibility of random association of variables, and this can be addressed by the second (wrapping) stage using the Boruta algorithm (Kursa and Rudnicki, 2010) (explained in the next paragraph). This approach allows the selection of questions from the original data set through a series of statistical tests, and the output from the filtering stage will serve as the inputs for the wrapping stage to derive the final key drivers. Hence, variables, which are associated with a response variable by chance, should indeed be

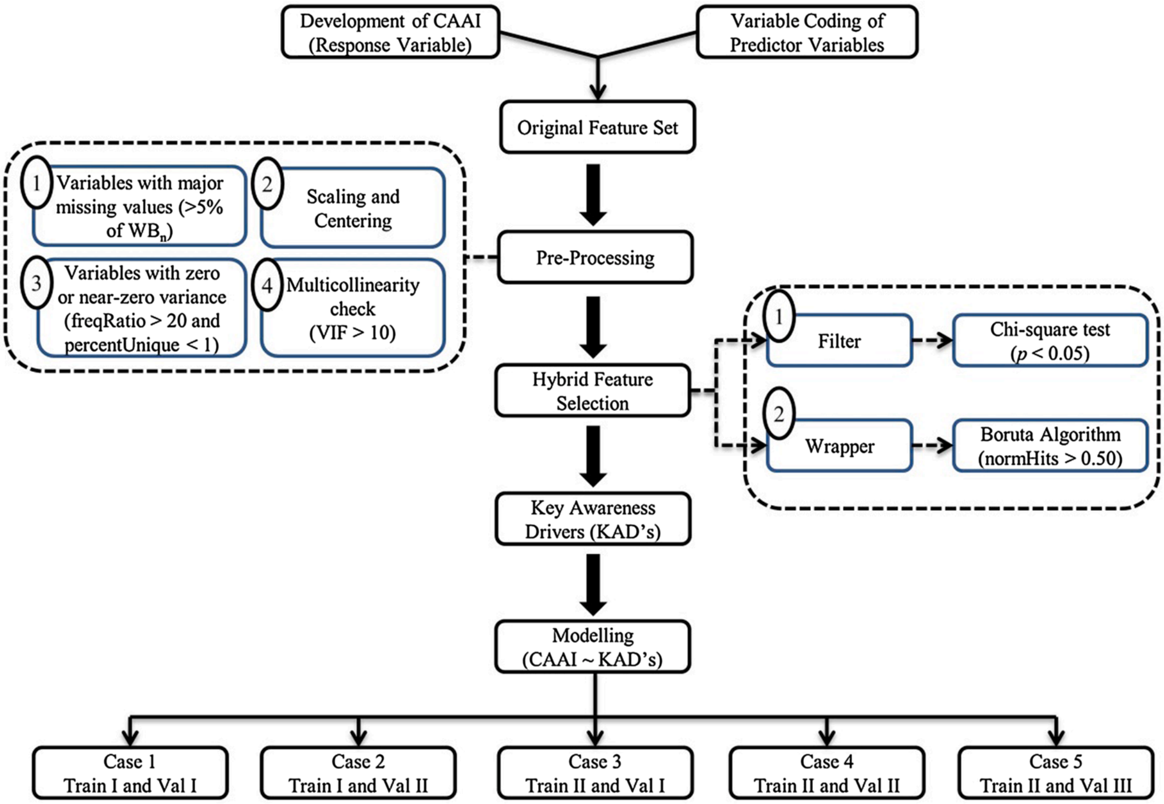
Development of CAAI (Response Variable)
Variable Coding of Predictor Variables
Original Feature Set
1 Variables with major missing values (>5% of WB$_n$)
2 Scaling and Centering
3 Variables with zero or near-zero variance (freqRatio > 20 and percentUnique < 1)
4 Multicollinearity check (VIF > 10)
Pre-Processing
Hybrid Feature Selection
1 Filter
Chi-square test ($p$ < 0.05)
2 Wrapper
Boruta Algorithm (normHits > 0.50)
Key Awareness Drivers (KAD's)
Modelling (CAAI ~ KAD's)
Case 1 Train I and Val I
Case 2 Train I and Val II
Case 3 Train II and Val I
Case 4 Train II and Val II
Case 5 Train II and Val III

**Fig. 1.** Flowchart showing the main steps linked to the newly developed Comprehensive Arsenic Awareness Index (CAAI) which includes pre-processing, hybrid feature selection procedure, derivation of key awareness drivers (KADs) and modelling under two training (Train I and Train II), and three validation (Val I, Val II, Val III) scenarios. The hybrid feature selection mainly involved two stages (filtering and wrapping) as shown in the flow chart on the right extension.

discarded from the modelling step. Data pre-processing steps (Fig. 1) helped address variables with missing values, low variance, and multicollinearity problems. Because most of the variables were categorical, no imputation was done for the missing values, and features with more than 5% of data missing were discarded. Following this, the variables with zero or small variance values were discarded using the *nearZeroVar()* function listed in the caret package. Further, features having variance inflation factor (VIF) > 10 were removed to address multi-collinearity in the system to arrive at the reduced set of features.

Following this, in the filtering stage of the hybrid feature selection procedure (Fig. 1), the coded predictors were examined for $\chi^2$ significance (95% confidence), and variables with $p < 0.05$ were retained (Table S3, S4, and S5). Subsequently, Boruta was applied on the reduced set of significant variables from the filtering stage to obtain the key awareness drivers (KADs) by setting parameters *maxRuns* (maximal number of important source runs) and the *p-value* (confidence level) at 1000 and 0.05, respectively. Boruta is built around the random forest (RF) classification algorithm and identifies *all-relevant* variables within a classification framework. For each attribute, a corresponding 'shadow' attribute is created, whose values are obtained by shuffling values of the original attribute across objects(Kursa and Rudnicki, 2010). Then, classification is performed using all attributes of this comprehensive system, and the importance of all attributes is computed using shadow attributes importance based on Z- score. The optimal feature set was thus obtained by retaining only those attributes which had Z-score more than the maximum shadow attribute's Z-score (normHits > 0.50). The CAAI was used as the dependent variable for all statistical tests.

Using the final set of KAD responses (Fig. 1) as predictors for CAAI, machine learning models were examined to classify the respondents into low, moderate, and high- awareness categories with two-fold objectives. First, we hypothesised that the key drivers selected through an elegant hybrid feature selection approach could be used for predicting the awareness category in the population. Second, we also wanted to evaluate whether simple linear models can capture the relationship between the KADs and CAAIs, notwithstanding the presence of the inherently implicit (indirect) relationships between these two sets of variables. We argue that the direct indicators of arsenic awareness are more visible ones and are expected to be manifested later than the indirect (driver) variables of arsenic awareness. Thus, if known *a priori*, KADs may lead to corrective measures for managing arsenic menace while living with arsenic in the environment.

To capture the linear and non-linear relationships between arsenic awareness and KADs, we evaluated ten machine learning algorithms (Table S2, Supplemental Materials) using a 10-fold cross-validation approach with three repetitions on both WB and BD validation datasets. The data partitioning into 70:30 training and test data was done on the optimal feature set obtained from the Boruta wrapper modelling step using the *createDataPartition()* function of the *caret* package resulting in a stratified random split. Cross-validation was conducted on the training dataset using the *trainControl()* function, which divided the training dataset into ten subsets using one subset for validation. This process was repeated three times across all ten trials to produce a reliable estimate. The final model for each trained classifier was selected based on the maximum accuracy achieved. To assess the robustness of a calibrated model, we examined the possibility to integrate both BD and WB data even though BD data had low variability using two training and three validation scenarios (Fig. 1). Training scenarios consisted of 70% of WB data alone (Train I) and a combination of 70% of WB data, with 70% of BD data (Train II). Similarly, three validation cases were 30% of WB data only (Val I), 30% of BD data alone (Val II), and a combination of 30% WB data with 30% of BD data (Val III). We estimated classification accuracy and kappa statistics for the validation datasets for all ten models.

## 3. Results

### 3.1. CAAI

Table 1 shows community awareness in WB and BD based on the coded responses to CAAI questions (SQ1-12) on the three components of health, water and community, and food. In general, there was 100% awareness for the majority of questions in BD (except for food-related questions) as compared to those of the WB site. The survey also showed that BD respondents were 100% aware of arsenic-induced health-related questions (SQ1-3) whereas, for WB, it varied from 62 to 72%. The community awareness of the medicine used for treating arsenicosis under the health component (SQ4) in both countries was minimum among the three components. Similarly, BD farming communities were fully aware (100%) of water and community-related questions (SQ5 (tube well contamination), 7–8 (social problems and mitigation measures) while WB site responses for these three questions were 12%, 58%, and 84%, respectively. Surprisingly, for SQ6 (Shallow Tube Well water contamination), awareness levels were similar for both countries. For the food-related questions (SQ9 (rice contamination), 11 (rice physiological symptoms), and 12 (methods to reduce arsenic toxicity in rice)), almost no awareness was recorded in BD whereas, awareness on the same varied from 15 to 88% for WB site. On the use of arsenic-contaminated shallow tube well water for irrigating rice crops (SQ10), we found that awareness levels were high (92–100%) for farming communities of both the countries. Furthermore, we also added the coded responses under each component to obtain component-specific awareness indices to evaluate how arsenic awareness differs between our study sites (Fig. 2A). The sum of all the 12 coded responses in the form of a CAAI was also used to describe overall arsenic awareness (Fig. 2B). These figures show the general pattern of arsenic awareness in both the communities and across the three major components of CAAI. Overall, around 39.0% of WB farmers had high awareness with 29.0% and 32.0% having low and medium awareness whereas, for BD, 99.0% of farmers were having medium awareness with the rest 1.0% falling under the high awareness category.

In general, the accumulated scores averaged over the survey population were similar across health-, water and community-, and food-related components for the WB sites (average score ranged from 2 to 2.47 out of a maximum score of 4). However, the scores for the BD sites

**Table 1**
Major components of arsenic awareness and the degree of awareness expressed in percentage in the farming community of West Bengal (WB) site ($n = 181$) and Bangladesh (BD) site ($n = 200$).

| Question related to | Questions dealing with knowledge on | WB Site (%) | BD Site (%) |
|---|---|---|---|
| Health | Arsenic toxicity in human health (SQ1) | 71.8 | 100 |
| | Visible symptoms of arsenic contamination (SQ2) | 70.7 | 100 |
| | Family member/villagers affected by arsenicosis (SQ3) | 61.9 | 100 |
| | Drugs used for treating arsenicosis (SQ4) | 0.5 | 0 |
| Water and community | Contamination status of tube well water (SQ5) | 12.1 | 100 |
| | Contamination status of shallow tube well water (SQ6) | 71.3 | 69.5 |
| | Social problems faced by people due to arsenicosis (SQ7) | 57.5 | 100 |
| | Mitigation steps for arsenicosis (SQ8) | 84 | 100 |
| Food | Contamination status of rice (SQ9) | 51.4 | 1.0 |
| | Whether contaminated STW used for irrigating rice crop (SQ10) | 92.3 | 100 |
| | Physiological symptoms of arsenic contamination in rice (SQ11) | 14.9 | 0 |
| | Methods to reduce arsenic contamination in rice (SQ12) | 88.4 | 0 |

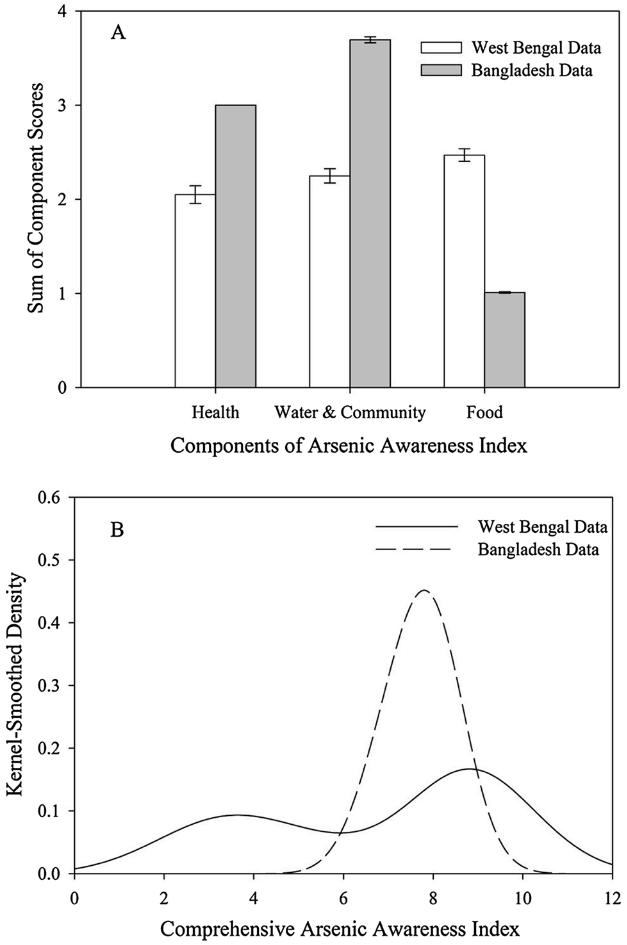

**Fig. 2.** Arsenic awareness across different components (A) and comprehensive arsenic awareness index score distribution for West Bengal and Bangladesh sites (B).

showed significantly high awareness for health-related (average score of 3.0 out of 4) and water and community-related (average score of 3.7 out of 4) components; the average awareness score for the food-related component for the BD data was the minimum at 1.0 out of a total possible score of 4.

These results show that the respondents in the BD sites are highly aware of the potential menace of arsenic in health- and water and community-related aspects of CAAI but not on the food-related arsenic awareness. Even though the average component scores for health and water in BD data were high, the average CAAI values 6.8 WB, and 7.7 for BD datasets were close to each other. Interestingly, the 25th and 75th percentiles for the overall CAAI values were 4 and 9 for WB data and 7 and 8 for the BD data, respectively. This is also reflected in the kernel-smoothed frequency distributions for the CAAI scores (Fig. 3B). A narrow and normally-distributed CAAI for the BD site further confirmed that the farming community in Bangladesh is aware of arsenic contamination in water and potential health challenges compared to mixed awareness levels for the WB site. These variations in the overall CAAI values across the study sites and across different components of the arsenic awareness provide a unique opportunity to evaluate the developed CAAI for both its capability to be used as a comprehensive awareness index and for the identification of key drivers of arsenic awareness in farming communities.

### 3.2. Factors influencing arsenic awareness

The survey questionnaire for WB and BD had 49 different questions designed to evaluate how different aspects of community living in typical farming families influenced arsenic awareness. With the CAAI developed, we used such a rich information base to identify KADs that may indirectly help build up arsenic awareness among people in the farming communities. Several of these questions were interrelated and carried information on the socio-economic and demographic conditions, water management and cropping practices, and other coping mechanisms people adopt while living with arsenic in the environment.

#### 3.2.1. Socio-economic and demographic drivers

Socio-economic and demographic data based on responses to ten different questions (Table S3) suggest that the bulk of the respondents from WB were above 41y old. In contrast, BD respondents were mostly from age groups of 41-50y and 51-60y with 27.5% each with a significant proportion of respondents from < 40y groups. There was a considerable difference in gender among those surveyed; both male (77%) and female (23%) participants were involved in WB whereas all of the respondents (100%) were male for BD. A third of the respondents in the WB site had no formal education, others had primary (24%), senior secondary (23%) education, and the remaining 20% had higher secondary school education or above. The proportion of BD respondents with no formal education was nearly the same as those in WB; the proportion of respondents with just primary (40%) and secondary school (26%) education was higher in BD than in WB. The rural population in WB state has an average literacy rate of 72.13% (Barik and Ghosal 2014) and the 33% of respondents in WB sites with no education in this survey more or less matches the nation-wide decadal census conducted in India.

The average family size in the WB site was large – almost half of the respondents (44%) having > 5-member family while the percentage of large families was 77% in BD. More than 90% of the respondents in the WB site and 87% in the BD site had their land although most of them were marginal farmers (96% in WB and 82.5% in BD) with < 1 ha of land. The majority of participants (77%) had an annual income of less than 70,000 INR (~US$ 945) with only 13% of the respondents indicated an increase in income over the last five-year period in the WB site. Most of the participants had a monthly expenditure of 5000 INR (~US$ 68) or less (63%), while 37% of the participants had a monthly expenditure of more than 5000 INR and the rest did not opt to reveal their monthly expenditures. Roughly half (51%) of the BD farmers had an annual income of less than 80,000 BDT (~US$ 945) with 99% indicating a monthly expenditure of more than 5700 BDT (~US$ 68).

#### 3.2.2. Irrigation and drinking water

With 19 questions focusing on irrigation and drinking water status in the community, 52% of the respondents from the WB sites reported that they had their tube wells installed before 1990 (Table S4). More than half of the participants (54%) had tube wells that are more than 100 m deep; however, as much as 24% of respondents drew water from less than 40 m deep tube wells. In general, farmers did not know about the marking of affected tube wells with red paint in the WB sites. Predictably, 180 out of 181 respondents at the WB site informed that they do not drink water from red-painted tube wells. A large fraction of respondents (33%) also did not know who carried out tests for arsenic on tube well water. Only 7% of the people out of 63% of participants with deep tube wells used it for collecting drinking water. This suggests that WB farming communities may not have a clear understanding of the linkage between arsenic and tube well water. About 75% of respondents were informed about such a drive in the community.

In contrast, 66% of the respondents in the BD site mentioned that they had their tube wells installed between 2001 and 2010; none replied for tube well installation before 1990. The majority (78.5%) of them drew their drinking water from tube wells of depth 40 m or less while the rest 21.5% had tube wells within the range of 41–100 m deep. Around half (48.5%) of the respondents reported that their tube wells were red painted. Nearly a third of the respondents (27%) did not know who

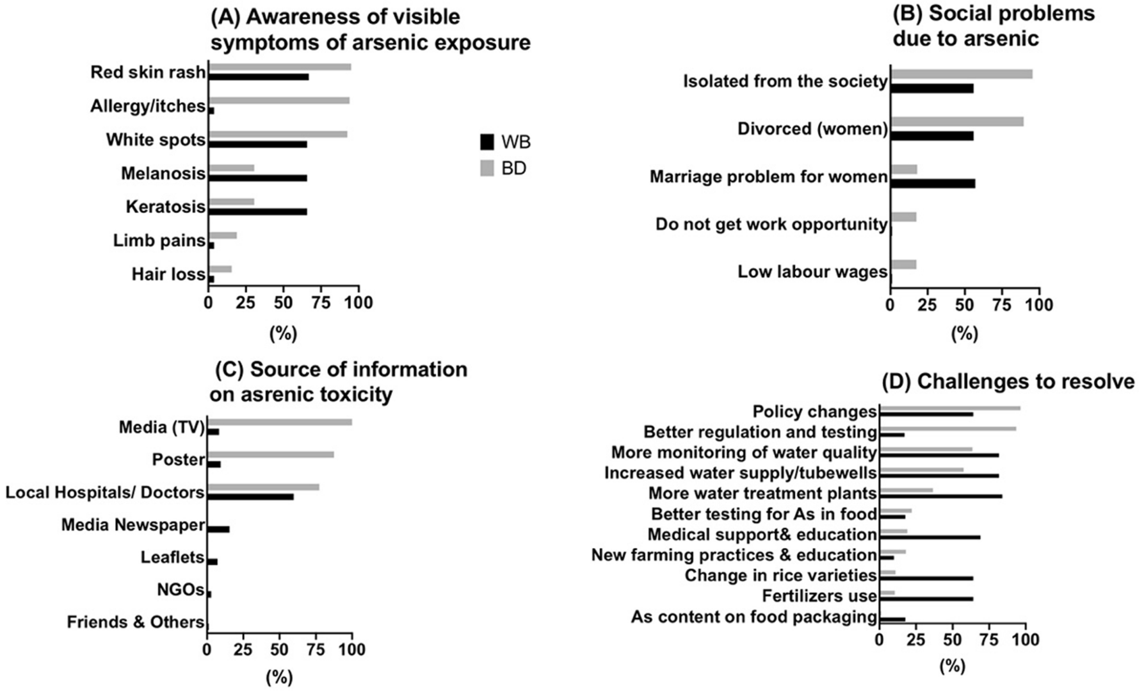

**Fig. 3.** Responses from participants from both countries (black for WB and gray for BD) for specific questions; (A) on visible symptoms of arsenic contamination; (B) social problems; (C) source of information; and (D) changes needed to resolve arsenic contamination.

tested their tube well water for arsenic contamination. All of the respondents in the study areas used tube wells (i.e. with the hand pump) and shallow tube wells (usually used for an irrigation water source) for drinking water. Participants confirmed that there was no provision of arsenic-free water supply in their villages in BD.

Almost 70% of the WB respondents agreed on the use of shallow tube wells as the primary source of irrigation for the *boro* (winter) rice. The majority of WB respondents (73%) did not think that such irrigation practice results in the red colouration of irrigation channels. However, two-third of the respondents agreed that long-term irrigation hardened their soil. Approximately one-third (35%) of the population confirmed that their fields were flooded during the monsoon whereas it was not valid for the remaining respondents. In the BD site, the whole farming community adopted shallow tube wells for irrigation purposes with 40% of farmers affirming that they were using it for growing *boro* rice. Respondents appeared to believe that such an irrigation system resulted in the red colouration of irrigation channels and hardened their field soil. Again, 31% of the BD respondents confirmed that their fields were flooded during the monsoon.

### 3.2.3. *Farming and consumption practices*

The survey also encompassed questions dealing with the cropping practices followed as well as the community preparedness on arsenic-related issues (Table S5). Although rice remains the main staple for both countries, a significant challenge in the farming community at the WB site was low productivity. Nearly a third (34%) of the respondents produced < 1000 kg, whereas the majority of them (43%) reported annual yield > 2000 kg. For the BD site, annual rice yield was < 1000 kg for 57% of participants, and only 11% having a yield > 2000 kg. Almost two-thirds (66%) of the WB farmers agreed that they had stopped *boro* rice cultivation while the rest continue to cultivate *boro* rice. Nearly all (98.3%) WB respondents denied any change in cropping pattern owing to arsenic contamination. Among the surveyed population it was found that a lot of farmers (90%) were feeding rice straw to their farm animals and many of them (83%) were also selling rice straw as feed for farm animals. In the BD site, almost all of the respondents stopped *boro* rice cultivation and using alternative crops instead and have not made any changes. All of the BD farmers were feeding rice straw to their farm cattle.

Several of our survey questions were designed to capture the rice cooking and eating aspects of the farmers in the surveyed locality. When the option was provided for the type of raw rice used for cooking between parboiled and non-parboiled, all the WB respondents mentioned parboiled rice as their preference. When farmers were asked whether they rinsed rice before cooking, most of the farmers (85%) agreed, a few (10%) declined while only 8 (4.4%) of them were unsure. When questioned on the frequency of rice consumption, a large section of the surveyed population (96%) self-reported to be consuming rice more than six times a week whereas only a handful of them (<5%) ate rice less than or up to six times a week. All BD farmers reported that they used preferred parboiled rice, rinsed rice before cooking, and consumed rice more than six times a week.

## 3.3. *Health concerns, source of information, and changes required*

When respondents were asked whether they were concerned about the arsenic contamination and its ill effects on human health, 81% of respondents in the WB site replied that they were concerned or very concerned; while 82% in the BD site were not concerned (Table S5). Almost half (50.8%) of the respondents at the WB site report that they found out the ill-impacts of arsenic on human health more than five years ago, while 44% did not know when exactly this awareness came. More than half (55%) of the respondents at the WB site reported to have their family members treated for arsenicosis. Although certified physicians made the treatment, most of the respondents (~55%) revealed that the disease was not cured except for one respondent. The rest of the respondents (45%) were not sure about the success of the treatment. For BD site, only 4 out of 200 respondents reported having their family members treated for arsenicosis, of which only one case was cured.

A number of explanatory questions with multiple-choice options were provided to capture the preferences of farmers related to arsenic-induced symptoms (Fig. 3A), social problems associated with arsenicosis (Fig. 3B), source of information (Fig. 3C), and mitigation options (Fig. 3D). The BD communities were well aware of visible symptoms of arsenic exposure such as skin rash (95%) and allergy (94%), whereas 66% of WB farming communities were aware of skin rashes, melanosis, keratosis and white spots (Fig. 3A). Interestingly, WB communities do not associate arsenic toxicity with allergies, hair loss and limb pains as is evident with a negligible response of 4% each whereas around 20% for BD population associate limb pains and hair loss with arsenic exposure. Communities in both countries (Fig. 3B) suffer from isolation from society (WB = 56% and BD = 95.5%), divorced women (WB = 56% and BD = 89.5%), and other marriage problems for women (WB = 57% and BD = 18%). The BD communities (17.5%) also expressed concerns for obtaining work opportunities and fair wages, whereas these two were of marginal importance in WB. For the BD population, both TV (100%) and posters (87.5%) constituted the major sources of information whereas these played a minimal role for WB (~8%) (Fig. 3C). Local hospitals/doctors play a predominant role in both countries (BD = 77.5% and WB = 60%). Unlike BD, newspapers (15.5%), leaflets (7%), and NGOs (3%) in WB sites were found to be providing information on arsenic. Communities from BD suggested policy changes (96.5%) and better regulation and testing (93.5%) as their top two priorities (Fig. 3D). More than 50% of the BD communities reported a preference for more work to be done in monitoring water quality and increased water supply. Interestingly, the majority (>50%) of WB respondents also agree with the above challenges. Also, WB respondents recognised additional challenges, including additional water treatment plants, medical support and education, changes in rice varieties, and fertiliser use. Nearly 20% of participants in both countries also indicated the need for better testing for food contamination. The most preferred mitigation option for WB was the installation of more water treatment plants, and in BD was bringing about new policy changes.

### 3.4. Identification of key awareness drivers (KADs) for arsenic awareness

Since BD data lacked variability in responses for SQ1-12 linked to arsenic awareness and estimated CAAI values as demonstrated in Table 1 and Fig. 2B, we used only the WB data to arrive at KADs for arsenic awareness (Fig. 1). The feature selection could only be applied on the WB dataset because of necessary variability available for such analysis was not present in the BD data, which may be because of a narrow but high level of awareness (average CAAI = 7.71, $^{first}$ quartile = 7, $^{third}$ quartile = 8) in the BD dataset. Thirty-one variables from 49 were removed as a result of the pre-processing and filtering stage of feature selection. Then, the Boruta algorithm was applied on the reduced set of 18 significant variables, resulting in the rejection of three variables (Fig. 4) selected by chi-square significance test, namely, “Age”, “Gender”, and “Annual Income”, thereby reducing the number of key drivers to 15. Fig. 4 shows the relative importance of the finalised 15 KAD questions in comparison to shadow attributes (Table S8). We also averaged the responses to each KAD question among the respondents having low, medium, and high categories CAAI values in WB and all the respondents from BD; corresponding CAAI scores for these four categories were 3.08, 6.96, 9.37, and 7.71, respectively. Table 2 shows the average KAD responses and their Pearson correlation coefficients ($\gamma$) with these average CAAI values both with ($\gamma_{WB+BD}$) and without ($\gamma_{WB}$) BD dataset. The Boruta median importance (Z value) ranged from 4.49 for KAD15 on the farmer type to 21.84 for KAD1 on the awareness of who did the test of the shallow tube well water for As. Interestingly, the $\gamma_{WB}$ were all high except for the three drivers of KAD5 on soil becoming hard because of long-term irrigation, KAD7 on whether a farmer cultivates *boro* rice using shallow tube well irrigation and KAD15 on farmer type. Additionally, despite having high Z value and high $\gamma_{WB}$ values, KAD6 and KAD9 yield very low $\gamma_{WB+BD}$ values where KAD6 reflects an As-aware person’s concern on arsenic toxicity and adverse impact on

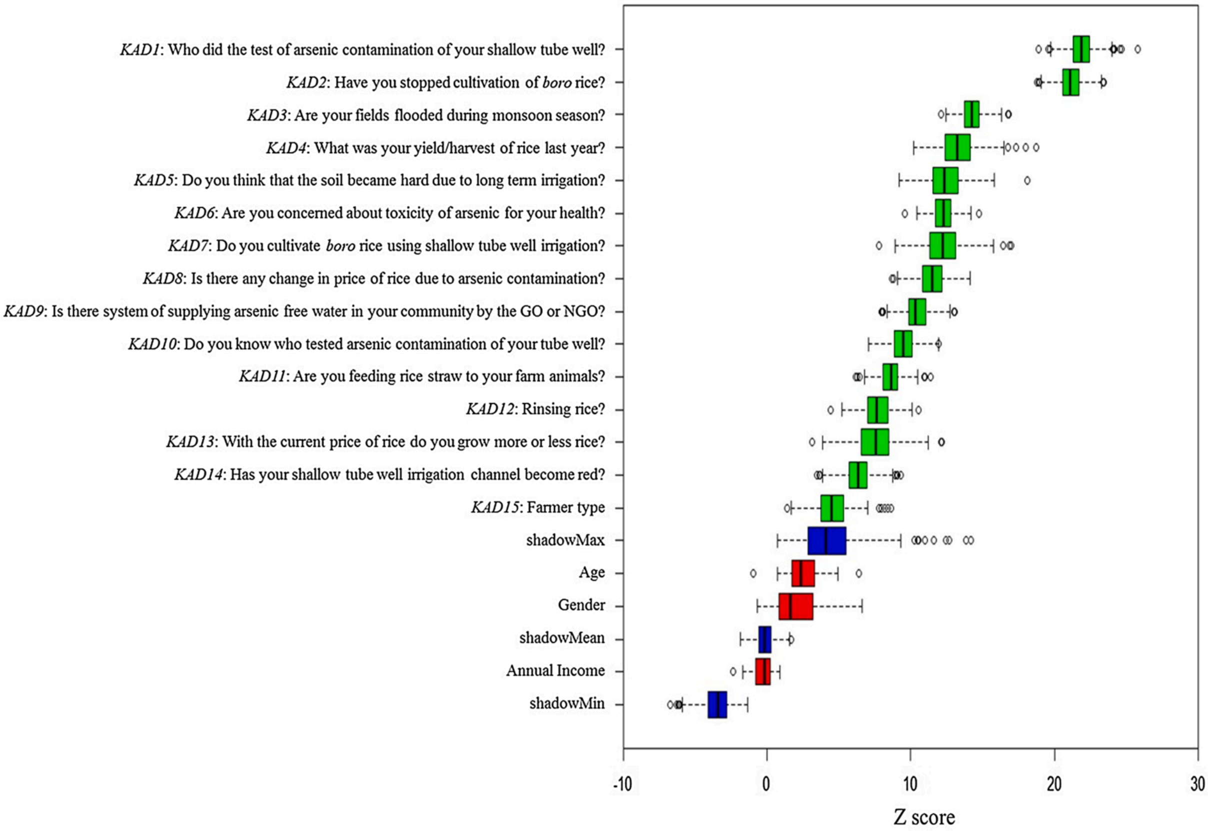

**Fig. 4.** Boruta relative feature importance box-plot. Blue box plots correspond to minimal, average and maximum Z score of a shadow attribute. Green box plots represent Z scores of confirmed attributes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Mean response to the key awareness driver (KAD) questions for respondents with low, medium and high comprehensive arsenic awareness index (CAAI) values in West Bengal (WB) and Bangladesh (BD); Pearson correlation coefficients (γ) between mean KAD response and corresponding mean CAAI score values; and Boruta median importance (Z value). KAD questions are provided in Fig. 4.

| KAD | $\mathrm{WB}_{\mathrm{low}}$ | $\mathrm{WB}_{\mathrm{medium}}$ | $\mathrm{WB}_{\mathrm{high}}$ | BD | $\gamma_{\mathrm{WB}}$ | $\gamma_{\mathrm{WB+BD}}$ | Z value | Agreement with BD data |
|---|---|---|---|---|---|---|---|---|
| KAD1 | 1.25 | 1.86 | 2.19 | 2.18 | 1.00 | 0.97 | 21.84 | High |
| KAD2 | 1.43 | 2.36 | 2.97 | 3.00 | 1.00 | 0.96 | 21.08 | High |
| KAD3 | 2.37 | 1.79 | 1.12 | 1.62 | −0.99 | −0.98 | 14.23 | High |
| KAD4 | 1.53 | 2.18 | 2.47 | 1.54 | 1.00 | 0.68 | 13.20 | High |
| KAD5 | 2.55 | 2.30 | 2.62 | 3.00 | 0.07 | 0.24 | 12.34 | High |
| KAD6 | 2.45 | 2.79 | 2.98 | 1.36 | 1.00 | 0.08 | 12.26 | Low |
| KAD7 | 2.39 | 1.96 | 2.79 | 1.79 | 0.36 | 0.11 | 12.22 | High |
| KAD8 | 1.74 | 1.54 | 1.54 | 1.00 | −0.91 | −0.49 | 11.47 | Moderate |
| KAD9 | 2.22 | 2.32 | 2.94 | 1.00 | 0.87 | 0.12 | 10.30 | Low |
| KAD10 | 1.90 | 2.21 | 2.78 | 2.46 | 0.96 | 0.96 | 9.49 | High |
| KAD11 | 2.53 | 2.96 | 2.88 | 3.00 | 0.84 | 0.83 | 8.63 | High |
| KAD12 | 2.65 | 2.55 | 2.96 | 3.00 | 0.64 | 0.62 | 7.64 | High |
| KAD13 | 2.57 | 1.96 | 1.66 | 2.51 | −1.00 | −0.72 | 7.56 | High |
| KAD14 | 1.98 | 1.61 | 1.18 | 3.00 | −0.98 | −0.19 | 6.35 | Low |
| KAD15 | 1.08 | 1.00 | 1.12 | 1.13 | 0.20 | 0.29 | 4.49 | High |

health; and, response to KAD9 reflects the administrative intervention of As-free water supply. Finally, KAD14 representing responses to the shallow tube well irrigation channels becoming red for WB and BD data had a low agreement with $\gamma_{\mathrm{WB+BD}}$ value of −0.19.

### 3.5. *Model performance for WB and BD data*

Based on the kappa values, the robustness of models can be categorised into slightly robust (0–0.2), fairly robust (0.21–0.40), moderately robust (0.41–0.60), substantially robust (0.61–0.80), and almost perfectly robust (0.81–1.0) (Landis and Koch, 1977). It can be seen in Table 3 that when only WB was used for calibration and validation, model robustness ranged from moderate to substantial (kappa = 0.46–0.67) with an accuracy of 65–78%. However, when such calibrated models were applied to the validation data (case of Train I and Val II, Table 3) from BD, we were able to obtain an accuracy that goes up to 95% (e.g. CART) but the resulting kappa values dropped substantially signifying no agreement (kappa < 0.00) to slightly robust model performance. Based on the confusion matrix obtained for all the modelling scenarios, poor kappa statistics are likely due to the lack of variability in the BD responses. When we combined 70% BD data to the 70% of WB data as training (Case of Train II in Table 3), kappa values did not dramatically change for Val I and Val II scenarios except for MLP and CART models which showed an increase of 9% each, while kNN and NB models kappa values dropped by 12% for Val I scenario. However, for Val III in which the 30% each from WB and BD data was pooled together, there was a significant improvement in the kappa values (0.66–0.78) with more than 80% classification accuracy for all the models. Although not shown in Table 3, kappa statistics for the BD dataset did not improve even when the whole of WB data was used as the calibration dataset and validated against the whole of BD data. All these results above strongly indicate that it may be possible to obtain high classification accuracy when both WB and BD data are pooled, with the kappa statistics becoming substantial to almost perfectly robust the moment BD data are added to the WB data in the validation dataset while drops when validated against BD data alone.

## 4. Discussion

Arsenic in drinking water is hailed as the largest mass poisoning of a population in history (Sen and Biswas, 2013) when the scale of the problem was unravelled in the 1990 s in both WB and BD. We did not find any studies that evaluated arsenic awareness among farming communities in WB although Singh et al. (2018) recently studied arsenic awareness in Bihar. Thus, the present study captures the community arsenic awareness in the affected districts of WB for the first time. For BD, awareness studies have been conducted as early as in 2004 (Paul, 2004) suggesting the evidence of increased attention given to the BD in comparison to WB. The newly developed CAAI evaluates arsenic awareness in a comprehensive manner because it includes arsenic exposure through food to the health and water and community-related arsenic awareness. Moreover, a notable advancement was the

**Table 3**
Overall accuracy and kappa statistics for different machine learning approaches for the West Bengal (WB) and Bangladesh (BD) datasets using two training cases (Train I: 70% of WB data alone; Train II: 70% of WB data + 70% of BD data) and three validation cases (Val I: 30% of WB data; Val II: 30% of BD data; Val III: 30% WB data + 30 of BD data).

| Model | Accuracy | | | | | Kappa statistic | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Train I | | Train II | | | Train I | | Train II | | |
| | Val I | Val II | Val I | Val II | Val III | Val I | Val II | Val I | Val II | Val III |
| kNN | 0.76 | 0.42 | 0.69 | 0.98 | 0.86 | 0.64 | −0.02 | 0.52 | 0.00 | 0.74 |
| NB | 0.71 | 0.50 | 0.63 | 0.98 | 0.82 | 0.55 | −0.02 | 0.43 | 0.00 | 0.66 |
| RF | 0.74 | 0.70 | 0.76 | 0.98 | 0.88 | 0.61 | −0.02 | 0.64 | 0.00 | 0.76 |
| SVML | 0.78 | 0.73 | 0.74 | 0.98 | 0.87 | 0.67 | −0.02 | 0.62 | 0.00 | 0.73 |
| SVMR | 0.74 | 0.37 | 0.76 | 0.98 | 0.88 | 0.61 | −0.02 | 0.64 | 0.00 | 0.76 |
| CART | 0.65 | 0.95 | 0.71 | 0.98 | 0.86 | 0.46 | −0.02 | 0.55 | 0.00 | 0.72 |
| MLP | 0.73 | 0.63 | 0.78 | 0.98 | 0.89 | 0.58 | −0.02 | 0.67 | 0.00 | 0.78 |
| SGB | 0.74 | 0.18 | 0.72 | 0.98 | 0.86 | 0.61 | −0.02 | 0.58 | 0.00 | 0.73 |
| MARS | 0.74 | 0.52 | 0.72 | 0.98 | 0.86 | 0.61 | −0.03 | 0.58 | 0.00 | 0.73 |
| SCT | 0.73 | 0.83 | 0.74 | 0.98 | 0.87 | 0.58 | −0.03 | 0.61 | 0.00 | 0.75 |

* kNN: k-Nearest Neighbours; NB: Naïve Bayes; RF: Random Forest; SVML: Support Vector Machines (Linear); SVMR: Support Vector Machines (Radial): CART: Classification and Regression Tree; MLP: Multi-Layer Perceptron; SGB: Stochastic Gradient Boosting; MARS: Multivariate Adaptive Regression Splines; SCT: Single C5.0 Tree.

incorporation of hybrid feature selection which allowed rigorous analysis of the qualitative data collected in this survey. The developed CAAI clearly showed a stark difference in community arsenic awareness in WB and BD and links this awareness to self-reported health, economic, and community impacts. Our results also offer evidence of how differences in the implementation of policy, mitigation measures, and tube well deployment have influenced communities and farmers in WB and BD.

### 4.1. *Design of arsenic awareness index for the food chain*

The newly developed CAAI includes six awareness questions (Q3, Q4, Q5, Q7, Q8, Q9) of Singh et al. (2018). Their Q1 about the general knowledge of arsenic is implicit in their Q3, which is the same question we have asked in our SQ1 (Table 1). Similarly, their question on the effectiveness of the mitigation program is also similar to our SQ8. We modified their question on arsenic sources (Q2) in the environment to capture the knowledge of arsenic in the water sources used for drinking (as in hand tube wells in SQ5), irrigation (as in shallow tube wells in SQ6), and agricultural produce (as in rice in SQ9). Thus, the awareness of arsenic source is recast to the arsenic in food and water in the newly designed CAAI. We did not consider the awareness of where to test for health hazards such as arsenicosis (Q6) in our study; instead, we considered additional questions such as awareness on whether a respondent faces any social problem because of arsenic toxicity (SQ7) and the practices of growing rice with arsenic-containing irrigation water (SQ10). We further expected that a farmer should be aware of any visible morphological changes in rice plants because of high arsenic content in the root zone and corrective measures to be adopted to reduce arsenic uptake in plants. Therefore, we considered the knowledge of arsenic effects on growing crops such as rice in SQ11 and awareness of methods to reduce arsenic toxicity in growing rice in SQ12.

The newly developed CAAI enabled us to compare the distribution of awareness across health, water and community, and food-related components in both WB and BD. For example, the distribution of kernel-smooth density curves (Fig. 2B) suggests an uneven distribution of arsenic awareness in WB whereas for BD a narrow interquartile range for the CAAI scores indicates moderate to highly-aware population. This increased awareness of BD farmers on health and water and community-related components (Fig. 2A) is possibly due to the increased influence of television, posters and health centres (Fig. 3C) in effectively enhancing the awareness in comparison to WB. In addition to the sources of information, Bangladesh Arsenic Mitigation and Water Supply Program (BAMWSP) tested arsenic in drinking water as early as 1998. Wells with arsenic levels $>$ 0.05 mg L$^{-1}$ (the BD health standard) were labelled “unsafe” and painted red, while those with arsenic levels below 0.05 mg L$^{-1}$ were labelled “safe” and painted green (Milton et al., 2012). This could have played a significant role in community awareness in BD whereas our survey showed that in the areas we surveyed in WB, farmers did not see any red paints on the existing tube wells, hence erroneously believing them to be safe for use although the majority of them (75%) were aware of such arsenic-free water supply systems. However, although the objectives laid out by BD government under National Strategy for Water Supply and Sanitation 2014 (Government of the People’s Republic of Bangladesh, 2014) for arsenic mitigation was to promote piped water supply in arsenic affected areas, in our research, we found no infringement of such activities in surveyed areas as all farmers denied prevalence of any such water supply systems.

Despite high awareness level, the BD farmers lacked knowledge of the food component of CAAI (score of 1.0 out of 4.0). This might be reflecting the slow update of awareness message on arsenic-related to the food aspect in BD. Over the last 15 years, there has been negligible testing and monitoring of arsenic in wells in BD as highlighted by Human Rights Watch (HRW, 2016). At the same time, more scientific evidence has emerged on arsenic exposure through food; this perhaps has failed to reach our targeted communities in BD. On the other hand, WB farmers were more aware of the food component with an average score of 2.7 out of 4.0. In our survey, we did not specifically ask about the source of this information on food. However, our data show that this awareness probably has led to the need for improved rice varieties and fertiliser use (Fig. 3D), which reflects their awareness of the arsenic impact on rice cultivation.

### 4.2. *Relevance of the hybrid feature selection and key awareness drivers*

Our findings suggest that in contrast to previous studies (Paul 2004; Parvez et al. 2006; Singh et al. 2018) age, education, gender and annual income of individuals are relatively less important than other factors (e.g. KAD1-15, Fig. 4). This could be due to (1) exhaustive nature of the survey questionnaire in comparison to previous studies (2) the hybrid feature selection approach followed. For instance, to select potential predictors of arsenic awareness, Singh et al. (2018) performed chi-square analysis for selecting statistically significant variables for model development. In contrast, our method goes beyond chi-square by using a wrapper Boruta which rejected the above variables initially selected by chi-square. For instance, the arsenic awareness gained by children through their schools have allowed not only children to be better educated on this, but also their friends and family members, hence rendering age as an ineffective tool in distinguishing awareness levels. This could imply that future research needs to consider a communities awareness as a sum, rather than singling out individuals within each community.

The hybrid feature selection approach was able to narrow down the number of the predictors from 49 to 15 KADs, which potentially had a higher association with CAAI for the WB farmers. KADs influence the cognitive decisions made by farmers in adapting to arsenic in the environment. Each KAD ranked based on the Z value of Boruta algorithm helped to delineate the most influential KADs among the rest. Fig. 4 clearly shows that KAD1 (Who did the arsenic test of your shallow tube well?) and KAD2 (Have you stopped cultivation of *boro* rice?) had higher median Z value (21.84 and 21.08) association with CAAI. The study by Parvez et al. (2006) in Araihazaar, BD also reported that respondents with knowledge on testing of tube wells were more aware of arsenic toxicity. Table 2 provides a concise yet apparent contradiction among the WB and BD farmers based on KADs average variability. Even with a moderate to high CAAI values, the reduction in the correlation coefficients with the addition of BD data may be reflecting ground realities of the lack of infrastructure on As-free water supply and a person’s requirement to live with the so-called ’curse of God’. Specifically, 4 of the 15 KADs (KAD 6, 8–9 and 14) chosen via hybrid feature selection for WB site, were found to have low to moderate agreements with its counterpart BD farmers based on $\gamma_{WB}$ and $\gamma_{WB+BD}$ values. This could be due to the difference between the study sites in dealing with arsenic in the environment. For example, KAD6 (Are you concerned about the toxicity of arsenic for your health) reflects that in WB site highly aware people were more concerned about arsenic toxicity but in BD, even though farmers were well aware of arsenic toxicity but 82% of the respondents were not concerned with the aftermath. This potentially could be due to the lack of alternatives as reflected in KAD9 (Is there a system of supplying arsenic free-water in your community by GO or NGO?) with BD farmers denying of any such measures. Similarly, KAD8 (Is there any change in the price of rice due to arsenic contamination?) reflects moderate agreement in both study sites suggesting the lack strong guidelines regarding price regulations on permissible arsenic concentrations in rice, when compared to the EU and other countries. The possible explanation for KAD14 (Has your shallow tube well irrigation channel become red?) is presumably due to local water quality differences. These results will have a strong bearing on how CAAI may be used to classify people based on key drivers of arsenic awareness using the machine learning approaches.

To test the prediction accuracies using KADs for CAAI, ten different (both linear and non-linear) machine learning algorithms were examined. To summarise, non-linear models such as MLP and three homogeneity-based models (i.e., CART, RF and SCT) performed better than linear models such as kNN and NB (except for SVML) under given modelling circumstances (Table 3). Overall, CART and SCT proved to be the best models with an average accuracy of 84% across all the cases considered. Singh et al. (2018) had deployed eight machine learning models against a binary arsenic awareness index, built using a set of 10 questions (10-point scale). In our analysis, we developed an index with 3-class classification on a 12-point scale. Emphasis has been laid on 3-class classification as it renders more logic in separating the low awareness group from the high awareness ones with an intermediate group of moderately aware people. Singh et al. (2018) results show that around 36% had no awareness (score of zero), whereas none belonged to no awareness category in our survey. In line with Singh et al. (2018), we found landholding size ($p = 0.17$) and household size (no. of family members) ($p = 0.36$) to be insignificant variables (Table S3), but unlike Singh et al. (2018) in our case, we found that gender ($p = 0.05$), age ($p = 0.02$) and income ($p = 0.02$) to be statistically significant whereas education to be insignificant ($p = 0.10$). Although Singh et al. (2018) reported that existing socio-economic and socio-behavioural factors play a crucial role in arsenic-exposed communities, our study finds socio-economic factors playing the least role in a person's awareness when other dynamics (such as cropping practices) are considered. However, both the studies possibly suggest a non-linear association between arsenic awareness and covariates as the performance of non-linear models were better than linear models in general with RF leading in both the studies, alongside with two other models used in our study, namely, CART and SCT.

### 4.3. *Living with arsenic in the environment*

In the 90s, it was clear that these two regions were similarly and severely affected by arsenic-contaminated groundwater (WHO, 2018). This study has shown that there have been distinct impacts of different actions taken over the past 30 years in both BD and WB by various governmental and non-governmental agencies and researchers across the world. A striking example of different policy actions over 30 years, is shown by how effective the widespread testing and painting of tube wells has led to the increasing the awareness in BD, while WB farmers were not well-aware of this policy due to lack of widespread implementation in their region. This may be because the WB government prioritised in supplying clean water to these areas instead of painting tube wells. For instance, the 2014 report on block-level awareness programme held at Rajapur Gram Panchayat with the coordination of WSSO PHED (Government. of West Bengal, 2014) mentions about three tube wells attached with Garaimari pipe water supply scheme to collect the safe drinking water; highlighting the alternative course of action adopted by WB Government to tackle the arsenic menace. Another report on Field Visit of Joint Secretary (Water) to Nadia district (2015) (Review of Arsenic Mitigation Measures, 2015) states that the WB State Government has taken up 12 (twelve) mega surface water based piped water supply schemes and 338 groundwater based piped water supply schemes, which the WB farmers were aware of. On the other hand, our BD interviewees confirmed that even though BD government had laid out plans in supplying piped water supply channels, this has not reached their areas, and policy changes were their top priorities (Fig. 3D). Due to this lack of piped water, BD communities still rely on hand tube wells and shallow tube wells for drinking and irrigation water, respectively, as evident from the survey. However, some behaviours and practice change cannot be linked to a specific policy. For instance, farmers in both regions use shallow tube wells as the primary source of irrigation but have stopped *boro* rice cultivation and/or have not made any changes in cropping patterns. We cannot determine from our survey if these changes are due to arsenic or other reasons. This warrants further investigation.

Though awareness of arsenic is trickling down slowly in WB, they are more aware of its impact on the food chain, in comparison to BD. This may indicate that the farmers are probably experiencing farming challenges and hence suggesting more actions on selecting suitable rice cultivars and fertiliser management (Fig. 3D). Interestingly, both countries prefer parboiled rice and traditionally cook in excess water. There are many scientific papers on reducing arsenic through different cooking methods to reduce arsenic exposure through rice consumption (Atiaga et al., 2020; Gray et al., 2015; Menon et al., 2020b) and these findings could be easily be communicated to the affected communities through public media, and educational/awareness/public intervention routes.

Another striking difference between these two regions is the source of information around arsenic impacts. Most WB farmers came to know about arsenic effects on human health five years ago, whereas BD farmers showed extensive awareness on this issue, as found in this, and previous studies. Notably, in BD, TV and posters (Fig. 3C) have been used effectively in increasing the awareness amongst the local population. This could be implemented in WB, for raising the awareness. Despite these differences, the communities in both regions were broadly aware a range of symptoms (Fig. 3A) and experience severe social problems (Fig. 3B) due to arsenic contamination; especially isolation from society, and welfare of women (difficulty in getting married or separation issues). One of the potential ways by which people directly come to know about arsenic is when they or their relatives/family members suffer from arsenicosis. BRAC (2000) (a non-government development organisation in Bangladesh) found that knowledge about the arsenic problem was related to a prior experience of seeing an afflicted patient. In the current study, all such variables (Table S6, Suppl. material) suffered multi-collinearity (VIF $> 10$) and were removed during the pre-processing step. Several studies have also revealed that arsenicosis has created extensive social and economic problems for the victims and their families in affected areas including social degradation, social injustice, and social isolation (Hassan et al., 2005; Argos et al., 2007; Rahman et al., 2018). Hassan et al. (2005) in one of the studies in Bangladesh reported that arsenicosis patients face difficulties in getting a job due to their disease. However, it was one of the least preferred options (Fig. 3B) by WB farmers as most of them (93%) owned their cultivable lands.

There are several suggestions by the interviewees on improving the health and well-being of these communities (Fig. 3D). There are differences in their priorities; however, they broadly agree on ensuring the supply of clean water and regular testing of wells through policy changes are needed in the future. It also appears that medical support is an urgent issue to be tackled, according to the WB participants; whereas testing or labelling for arsenic in food is comparatively less important issues for both the study sites. This could mean there is lack of awareness on the risks of arsenic exposure through food in these communities or potential economic impacts (e.g. a decline in the market value of the product if their products do not meet the required standards) if testing and quality control are introduced.

A major bottleneck of our survey was the lack of variation amongst the surveyed population in BD, which involved only male participants with a narrow spectrum of awareness spread on the index developed. Also, because the questionnaire had been adjusted to the needs of locals, it resulted in different responses, which created hindrance in variable coding while merging of both data sets onto common options in certain variables. Again even though the methodology followed in the study is robust, we are well aware that some variables concerning community health impacts like "Has any of your family member treated with arsenicosis?", have been discarded due to multi-collinearity problem, which otherwise could have had an impact on arsenic awareness.

## 5. Conclusions

This study is a first of its kind arsenic awareness survey conducted in two culturally similar farming communities of two neighbouring countries covering WB from India and Faridpur and Chandpur from BD. With a set of structured questionnaire survey, the responses from 181 farmers from WB and 200 farmers from BD were used to develop a comprehensive arsenic awareness index (CAAI) based on health, water and community and food-related factors. Our results showed that a large proportion of farmers were aware of the arsenic situation in both WB (average CAAI score of 6.8 out of 12) and BD (average CAAI score of 7.7 out of 12). Application of hybrid feature selection identified 15 KADs, which included factors related to stakeholder interventions and cropping practices instead of commonly perceived factors such as age, gender and income. Among ML algorithms, CART and SCT could estimate CAAIs with an average accuracy of 84%. Our study also showed that there is need for strengthening overall arsenic awareness in WB and, more specifically, on the risk through the food chain in BD. Mass media channels (e.g., TV and posters) have been successful in raising arsenic awareness in BD. Similar approaches may be adopted in WB.

The interviewees suggested several research and policy needs to improve arsenic awareness for improving the health and well-being of farming communities. The communities agreed on ensuring the supply of clean water and regular testing of wells through policy changes. It was also evident that the medical support is an urgent issue to be tackled in both countries whereas testing and labelling for arsenic in food is comparatively a less important issue.

Despite the pragmatic outcomes, our study was confined to arsenic-affected districts with limited number of participants in both countries. Future studies may be conducted to cover areas with varying degrees of arsenic contamination and with an increased sample size, which will provide the desired variability to validate the efficacy of developed models. Nevertheless, this study highlighted several necessary research and outreach gaps which need to be addressed in these regions in order to live with arsenic in the environment. This study addresses the UN sustainable development goals (SDGs) such as clean water and sanitation (SDG6), zero hunger (SDG2), good health and well-being (SDG3). This approach echoes the WHO's comprehensive action plan involving water testing, awareness-building campaigns, and mitigation options, including arsenic removal technologies to combat arsenic toxicity menace (Basu et al., 2015).

## CRediT authorship contribution statement

**Debasish Mishra:** Formal analysis, Methodology, Validation, Visualization, Writing - original draft. **Bhabani S. Das:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing - original draft. **Tathagata Sinha:** Data curation. **Jiaul M. Hoque:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Writing - review & editing. **Christian Reynolds:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing - review & editing. **M. Rafiqul Islam:** Conceptualization, Funding acquisition, Investigation, Writing - review & editing. **Mahmud Hossain:** Funding acquisition, Investigation, Writing - review & editing. **Pinaki Sar:** Funding acquisition, Investigation, Writing - review & editing. **Manoj Menon:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing - original draft.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The project was funded by the GCRF Global Impact Accelerator Award (Grant Reference Number: EP/S51991/1). We thank Dr. Binoy Sarkar, Dr. Sanjay Lanka, Dr. Niranjan Panigrahi, Dr. Balaram Mahopatra and Mr Budhadeb Bhattacharya for their help and support. We thank all interviewees for their participation in this study.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106529.

## References

Argos, M., Parvez, F., Chen, Y., Hussain, A.Z.M.I., Momotaj, H., Howe, G.R., Graziano, J. H., Ahsan, H., 2007. Socioeconomic status and risk for Arsenic-related skin lesions in Bangladesh. Am. J. Public Health 97, 825–831. https://doi.org/10.2105/AJPH.2005.078816.

Atiaga, O., Nunes, L.M., Otero, X.L., 2020. Effect of cooking on arsenic concentration in rice. Environ. Sci. Pollut. Res. 27, 10757–10765. https://doi.org/10.1007/s11356-019-07552-2.

Basu, A., Sen, P., Jha, A., 2015. Environmental arsenic toxicity in West Bengal, India: A brief policy review. Indian J. Public Health 59, 295. https://doi.org/10.4103/0019-557x.169659.

BRAC, 2000. Combating a deadly menace: early experiences with a community-based arsenic mitigation project in Bangladesh. Research Monograph no. 16. Dhaka.

Carey, M., Donaldson, E., Signes-Pastor, A.J., Meharg, A.A., 2018., n.d. Dilution of rice with other gluten free grains to lower inorganic arsenic in foods for young children in response to European Union regulations provides impetus to setting stricter standards.

Carlin, D.J., Naujokas, M.F., Bradham, K.D., Cowden, J., Heacock, M., Henry, H.F., Lee, J.S., Thomas, D.J., Thompson, C., Tokar, E.J., Waalkes, M.P., Birnbaum, L.S., Suk, W. A., 2016. Arsenic and environmental health: State of the science and future research opportunities. Environ. Health Perspect. https://doi.org/10.1289/ehp.1510209.

Chakraborti, D., Das, B., Rahman, M.M., Chowdhury, U.K., Biswas, B., Goswami, A.B., Nayak, B., Pal, A., Sengupta, M.K., Ahamed, S., Hossain, A., Basu, G., Roychowdhury, T., Das, D., 2009. Status of groundwater arsenic contamination in the state of West Bengal, India: A 20-year study report. Mol. Nutr. Food Res. 53, 542–551. https://doi.org/10.1002/mnfr.200700517.

Chakraborti, D., Rahman, M.M., Das, B., Murrill, M., Dey, S., Chandra Mukherjee, S., Dhar, R.K., Biswas, B.K., Chowdhury, U.K., Roy, S., Sorif, S., Selim, M., Rahman, M., Quamruzzaman, Q., 2010. Status of groundwater arsenic contamination in Bangladesh: A 14-year study report. Water Res. 44, 5789–5802. https://doi.org/10.1016/j.watres.2010.06.051.

Chakraborty, M., Mukherjee, A., Ahmed, K.M., 2015. A review of groundwater arsenic in the Bengal Basin, Bangladesh and India: from Source to Sink. Curr. Pollut. Reports. https://doi.org/10.1007/s40726-015-0022-0.

Chowdhury, M.A.I., Uddin, M.T., Ahmed, M.F., Ali, M.A., Rasul, S.M.A., Hoque, M.A., Alam, R., Sharmin, R., Uddin, S.M., Islam, M.S., 2006. Collapse of socio-economic base of Bangladesh by arsenic contamination in groundwater. Pakistan J. Biol. Sci. 9, 1617–1627. https://doi.org/10.3923/pjbs.2006.1617.1627.

Ghosal, B., 2014. Nature and dimension of disparities in the state of primary education in West Bengal, 147–164.

Government of Bangladesh, 2004. National Policy for Arsenic Mitigation (https://old.dphe.gov.bd/pdf/National-Policy-for-Arsenic-Mitigation-2004.pdf).

Government of the People's Republic of Bangladesh (https://www.who.int/globalchange/resources/wash-toolkit/national-strategy-for-water-supply-and-sanitation-bangladesh.pdf), 2014. National Strategy for Water Supply and Sanitation.

Government of West Bengal, 2014. Report on Awareness Programme on Arsenic Contamination and Remedial Measures , Treatment Prepared and Submitted By Water Sanitation Support Organisation (WSSO).

Gray, P.J., Conklin, S.D., Todorov, T.I., Kasko, S.M., 2015. Cooking rice in excess water reduces both arsenic and enriched vitamins in the cooked grain. Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess. 33, 78–85. https://doi.org/10.1080/19440049.2015.1103906.

Hassan, M.M., Atkins, P.J., Dunn, C.E., 2005. Social implications of arsenic poisoning in Bangladesh. Soc. Sci. Med. 61, 2201–2211. https://doi.org/10.1016/j.socscimed.2005.04.021.

HRW, 2016. The Failing Response to Arsenic in the Drinking Water of Bangladesh's Rural Poor | HRW [WWW Document]. hrw.org. URL https://www.hrw.org/report/2016/04/06/nepotism-and-neglect/failing-response-arsenic-drinking-water-bangladeshs-rural (accessed 12.19.20).

Humans, I.W.G. on the E. of C.R. to, 2012. A review of human carcinogens. International Agency for Research on Cancer.

Joseph, T., Dubey, B., McBean, E.A., 2015. A critical review of arsenic exposures for Bangladeshi adults. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2015.05.035.

Kapaj, S., Peterson, H., Liber, K., Bhattacharya, P., 2006. Human health effects from chronic arsenic poisoning - A review. J. Environ. Sci. Heal. - Part A Toxic/Hazardous Subst. Environ. Eng. 41, 2399–2428. https://doi.org/10.1080/10934520600873571.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R., 2014, 2014. Caret: classification and regression training. ui.adsabs.harvard.edu.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Softw. 36, 1–13. https://doi.org/10.18637/jss.v036.i11.

Landis, J.R., Koch, G.G., 1977. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics 33, 363. https://doi.org/10.2307/2529786.

Liu, W.J., Wood, B.A., Raab, A., McGrath, S.P., Zhao, F.J., Feldmann, J., 2010. Complexation of arsenite with phytochelatins reduces arsenite efflux and translocation from roots to shoots in Arabidopsis. Plant Physiol. 152, 2211–2221. https://doi.org/10.1104/pp.109.150862.

Mandal, B.K., Suzuki, K.T., 2002. Arsenic round the world: A review. Talanta. https://doi.org/10.1016/S0039-9140(02)00268-0.

Mazumder, D.N., Ghosh, A., Majumdar, K., Ghosh, N., Saha, C., Mazumder, R.N., 2010. Arsenic contamination of ground water and its health impact on population of district of Nadia, West Bengal, India. Indian J. Community Med. 35, 331–338. https://doi.org/10.4103/0970-0218.66897.

Menon, M., Dong, W., Chen, X., Hufton, J., Rhodes, E.J., 2020a. Improved rice cooking approach to maximise arsenic removal while preserving nutrient elements. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.143341.

Menon, M., Sarkar, B., Hufton, J., Reynolds, C., Reina, S.V., Young, S., 2020b. Do arsenic levels in rice pose a health risk to the UK population? Ecotoxicol. Environ. Saf. 197, 110601. https://doi.org/10.1016/j.ecoenv.2020.110601.

Milton, A.H., Hore, S.K., Hossain, M.Z., Rahman, M., 2012. Bangladesh arsenic mitigation programs: lessons from the past. Emerg. Health Threats J. 5, 7269. https://doi.org/10.3402/ehtj.v5i0.7269.

Mondal, D., Banerjee, M., Kundu, M., Banerjee, N., Bhattacharya, U., Giri, A.K., Ganguli, B., Roy, S. Sen, 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. Environ. Geochem. Health 32, 463–477. https://doi.org/10.1007/s10653-010-9319-5.

Mukherjee, A., Fryar, A.E., 2008. Deeper groundwater chemistry and geochemical modeling of the arsenic affected western Bengal basin, West Bengal, India. Appl. Geochem. 23, 863–894. https://doi.org/10.1016/j.apgeochem.2007.07.011.

Parvez, F., Chen, Y., Argos, M., Iftikhar Hussain, A.Z.M., Momotaj, H., Dhar, R., van Geen, A., Graziano, J.H., Ahsan, H., 2006. Prevalence of arsenic exposure from drinking water and awareness of its health risks in a Bangladeshi population: Results from a large population-based study. Environ. Health Perspect. 114, 355–359. https://doi.org/10.1289/ehp.7903.

Paul, B.K., 2004. Arsenic contamination awareness among the rural residents in Bangladesh. Soc. Sci. Med. 59, 1741–1755. https://doi.org/10.1016/j.socscimed.2004.01.037.

Planning Commission, G. of I., 2007. Report of the task force on formulating action plan for removal of arsenic contamination in West Bengal. Gov. India Plan. Comm. New Delhi.

Rahman, M.A., Rahman, A., Khan, M.Z.K., Renzaho, A.M.N., 2018. Human health risks and socio-economic perspectives of arsenic exposure in Bangladesh: A scoping review. Ecotoxicol. Environ. Saf. 150, 335–343. https://doi.org/10.1016/j.ecoenv.2017.12.032.

Rahman, M.M., Mandal, B.K., Chowdhury, T.R., Sengupta, M.K., Chowdhury, U.K., Lodh, D., Chanda, C.R., Basu, G.K., Mukherjee, S.C., Saha, K.C., Chakraborti, D., 2003. Arsenic Groundwater Contamination and Sufferings of People in North 24-Parganas, One of the Nine Arsenic Affected Districts of West Bengal, India. J. Environ. Sci. Heal. Part A 38, 25–59. https://doi.org/10.1081/ESE-120016658.

Rahman, M.M., Mondal, D., Das, B., Sengupta, M.K., Ahamed, S., Hossain, M.A., Samal, A.C., Saha, K.C., Mukherjee, S.C., Dutta, R.N., Chakraborti, D., 2014. Status of groundwater arsenic contamination in all 17 blocks of Nadia district in the state of West Bengal, India: A 23-year study report. J. Hydrol. 518, 363–372. https://doi.org/10.1016/j.jhydrol.2013.10.037.

Rahman, M.M., Naidu, R., Bhattacharya, P., 2009. Arsenic contamination in groundwater in the Southeast Asia region. Environ. Geochem. Health. https://doi.org/10.1007/s10653-008-9233-2.

Report on Field Visit of Joint Secretary (Water) to Nadia district of West Bengal to Review Arsenic Mitigation Measures – 11th and 12th September, 2015.

Saha, R., Dey, N.C., Rahman, M., Bhattacharya, P., Rabbani, G.H., 2019. Geogenic arsenic and microbial contamination in drinking water sources: Exposure risks to the coastal population in Bangladesh. Front. Environ. Sci. 7, 1–12. https://doi.org/10.3389/fenvs.2019.00057.

Sen, P., Biswas, T., 2013. Arsenic: the largest mass poisoning of a population in history. BMJ 346, 3–5. https://doi.org/10.1136/bmj.f3625.

Shaji, E., Santosh, M., Sarath, K.V., Prakash, P., Deepchand, V., Divya, B.V., 2020. Arsenic contamination of groundwater: A global synopsis with focus on the Indian Peninsula. Geosci. Front. https://doi.org/10.1016/j.gsf.2020.08.015.

Signes, A., Mitra, K., Burló, F., Carbonell-Barrachina, A.A., 2008. Effect of two different rice dehusking procedures on total arsenic concentration in rice. Eur. Food Res. Technol. 226, 561–567. https://doi.org/10.1007/s00217-007-0571-6.

Singh, S.K., Taylor, R.W., Mahmudur, M., 2018. Developing robust arsenic awareness prediction models using machine learning algorithms. J. Environ. Manage. 211, 125–137. https://doi.org/10.1016/j.jenvman.2018.01.044.

Srivastava, S., 2020. Arsenic in Drinking Water and Food, Arsenic in Drinking Water and Food. Springer Singapore. https://doi.org/10.1007/978-981-13-8587-2.

Upadhyay, M.K., Majumdar, A., Barla, A., Bose, S., Srivastava, S., 2019. An assessment of arsenic hazard in groundwater–soil–rice system in two villages of Nadia district, West Bengal, India. Environ. Geochem. Health 41, 2381–2395. https://doi.org/10.1007/s10653-019-00289-4.

WHO [WWW Document], 2020. . who.int. URL https://www.who.int/ipcs/assessment/public_health/chemicals_phc/en/ (accessed 12.17.20).

WHO, 2018, n.d. ARSENIC PRIMER Guidance on the Investigation & Mitigation of Arsenic Contamination. unicef.org.

Xu, L., Polya, D.A., Li, Q., Mondal, D., 2020. Association of low-level inorganic arsenic exposure from rice with age-standardized mortality risk of cardiovascular disease (CVD) in England and Wales. Sci. Total Environ. 743, 140534. https://doi.org/10.1016/j.scitotenv.2020.140534.