Deriving arsenic concentration guideline values for soil and irrigation water for rice cultivation

Jajati Mandal

Submitted in Partial Fulfilment of the Requirements of the Degree of

Doctor of Philosophy

School of Science Engineering and Environment

University of Salford, Greater Manchester, United Kingdom

2023

Table of Contents

List of Figures

List of Tables

List of Plates

Acknowledgement

Mere words are inadequate to express the sense of gratitude and indebtedness to those whose assistance were indispensable for the completion of my work.

First and foremost, I am grateful to my first supervisor (presently my external supervisor) Dr. Debapriya Mondal, London School of Hygiene & Tropical Medicine and present supervisor Prof. Mike Wood, University of Salford and co-supervisor Dr. Simon Hutchinson, University of Salford for their invaluable guidance, encouragement, and patience.

I would like to express my gratitude to Dr. Debapriya Mondal for her kind support, cooperation, valuable suggestions during my application for Netaji Subhas Fellowship of Indian Council of Agricultural Research (ICAR) and for application for PhD at University of Salford. It is primarily because of her able guidance, keen interest, constant encouragement, and valuable discussions during this investigation during study that the work has moulded to an appropriate shape.

I feel highly privileged to work under the guidance and supervision of Prof. Mike Wood, for his kind cooperation, valuable suggestion and for always maintaining a cheerful and helpful attitude throughout the entire course of work. I am thankful to, for his incessant advice, inexhaustible encouragement, prudent and relentless guidance, tireless support, and all sorts of cooperation since inception and till completion of this research pursuit.

I am grateful to Amy Evans, Teaching Fellow & Specialist Technician, University of Salford, who provided technical support and training on laboratory equipment and procedures.

I am highly indebted to my external supervisor Dr. Prashant Srivastava, Senior Research Scientist, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Environment Unit, Adelaide, Australia for providing me the opportunity to work at CSIRO for six months. His constant encouragement in improving the novelty of the study by incorporating nuances of chemistry like adsorption and desorption studies with biochar and introducing me to research on emerging pollutants is acknowledged. I am incredibly grateful to Dr. Jason Kirby for providing me the opportunity, the resources and support throughout my research work at CSIRO.

My acknowledgement and regards are also due to Dr. Mohammad Mahmudur Rahman, The University of Newcastle, Australia for his valuable guidance during laboratory analysis of samples, noble advice, timely suggestions, and constant inspiration during the work.

I would also like to convey my earnest gratefulness and regards to my teachers and collaborators like Prof. Kallol Bhattacharyya, Bidhan Chandra Krishi Viswavidyalaya, Dr. Debasis Golui, Indian Agricultural Research Institute, Dr. Vinay Jain, Agilent Technologies and Dr. Pradip Bhattacharyya, Indian Statistical Institute for valuable suggestions and all sorts of help throughout this research pursuit.

I would also like to bestow my sincere gratitude to Prof. Nilanjan Chattopadhyaya, Dr. Anshuman Kohli, of Bihar Agricultural University and Prof. Ashok Ghosh of Mahavir Cancer Sansthan, Patna, Bihar for their enormous support and consistent encouragement in each step of my work.

I also owe a big gratitude to Dr. Bilash Chandra Das, Assistant Director Agriculture for providing me all facilities to perform the field work in the arsenic contaminated, Murshidabad, West Bengal. He along with associated workforce ensured proper field work even in Covid pandemic period.

A special chalice of gratitude must be extended to Biswanath da, my field and laboratory helping hand in India also a friend, philosopher and guide for his tireless cooperation and incessant association with the work leaving aside his constant hurdles of life through a cheerful attitude.

I am deeply beholden to Dr Sudip Sengupta, who helped me as a brother through the nuances of laboratory incubation studies, preparation of manuscripts for advancement of the current research findings.

The financial assistance from ICAR-Netaji Subhas Fellowship throughout the research pursuit is highly acknowledged. The financial support from my cousin brother Mr. Amit Kumar Bal, my friend Mr. Sandiapan Samanta and Dr. Biswanath Dari is deeply acknowledged.

A major share of the acknowledgement should also be provided to my friends like, Helen, Piyawan (Jay), Feyisara (Feyi), Kenneth, Somdutta, Huong, Trinh, Mick, Sash and others who in spite of their busy schedule have always been associated with every ups and downs of my life and provided motivation by inestimable means.

Last but not the least I would like to express my deepest appreciation to my wife Rajrani Mandal and my daughter Riddhima Mandal (Jini) for their unwavering support encouragement and sacrifice throughout my academic journey. Their love, guidance and have been my constant source of strength and inspiration. Lastly, I want to acknowledge the role of my extended family in creating a supportive and nurturing environment for me to pursue my academic goals.

In lieu of all these, few might not have been mentioned, but none of them has been forgotten. All of you are highly endeared to me.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except in the text and acknowledgements.

List of Abbreviations, Acronyms and Symbols used

ABBREVIATIONS		
g	=	Gram
ha	=	Hectare
m	=	Metre
cm	=	Centimetre
mg	=	Milligram
mg kg^{-1}	=	Milligram per kilogram
µg	=	Microgram
µg kg^{-1}	=	Microgram per kilogram
mL	=	Milliliter
Kg	=	Kilogram
kg ha^{-1}	=	Kilogram per hectare
t ha^{-1}	=	Tonne per hectare
Av.	=	Available
ACRONYMS		
et al.	=	Other authors/and others
EC	=	Electrical Conductivity
i.e.	=	Id est (that is)
e.g.	=	Exempli gratia (for example)
SD	=	Standard deviation
SE	=	Standard Error
OC	=	Organic Carbon
BAs	=	Bioavailable Arsenic
AvFe	=	Available Iron
AvP	=	Available Phosphorus
AUC	=	Area Under Curve
DT	=	Decision Tree
FN	=	False Negative
FP	=	False Positive
GBM	=	Gradient Boost Machine
ICE	=	Individual Conditional Expectation
LR	=	Logistic Regression
MCC	=	Matthew Correlation Coefficient
MTC	=	Maximum Tolerable Concentration
PDP	=	Partial Dependence Plot
PPV	=	Positive Predictive Value
REM	=	Random Effect Model
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
SYMBOLS		
^{0}C	=	Degree Celsius
@	=	At the rate of
%	=	Percentage
&	=	And
Fe	=	Iron
P	=	Phosphorus
Al	=	Aluminum
Fe	=	Iron
Si	=	Silicon
As	=	Arsenic
As(V)	=	Arsenate
As(III)	=	Arsenite
≅	=	Equivalent

Abstract

Arsenic (As) is a naturally occurring, toxic trace element that can be found in irrigation water, soil, and crops. Rice accumulates higher concentrations of As in its grains than other cereals like wheat and barley. This leads to concern over dietary As exposure, especially in areas of India, Bangladesh, Nepal, Taiwan, Vietnam, and Thailand. This present study has been undertaken to manage the risk posed by rice grown in As-contaminated areas. The objectives of the study were to determine guideline values for total and bioavailable As in soil, as well as As levels in irrigation water, using predictability models. Additionally, the efficacy of biochar as an amendment for As-contaminated soils in rice cultivation was assessed through a meta-analysis.

Meta-analysis of a database compiled from an extensive literature review was undertaken using decision tree (DT) and logistic regression (LR) machine learning models to evaluate the relationship between As concentrations in rice grain, soil, and irrigation water. Soil total As was a stronger predictor of As in rice grain than irrigation water As. Both the DT and LR models successfully predicted the soil concentrations above which As in grain would exceed the Codex recommendation. Subsequent field studies in West Bengal, India in 2021 provided validation data, which demonstrated that 14 mg kg^{-1} of total As in soil was an appropriate guideline value for the safe cultivation of rice.

The concentration of bioavailable As in paddy soil was predicted using random forest (RF), gradient boosting machine (GBM), and LR models. The LR model was the better performing, identifying bioavailable As, total As, available iron (Fe), and organic carbon as significant predictors of grain As. Based on the LR model's partial dependence plots and individual conditional expectation plots, 5.70 mg kg^{-1} was the limit for bioavailable As in soil.

An incubation study was conducted using monolithic soil columns collected from 10 As-contaminated sites. Results were analysed using linear discriminant analysis (LDA) and LR, considering the As dose, soil pH, organic carbon, clay, available Fe, phosphorus, and total As. The LR model performed best, predicting 190 μL^{-1} as the guideline value for irrigation water.

To support remediation of As-contaminated soil, biochar was evaluated as a potential soil amendment. A meta-analysis indicated that biochar could be an effective tool in the sequestration of As in soil, but further research is required under realistic field conditions.

This study has provided the first guideline values for As in soil and irrigation water and identified a potential management option (soil remediation using biochar). The findings have direct relevance to rice farmers and regulators, with the potential to deliver significant public health benefits in As contaminated regions.

Chapter 1-Introduction

1.1 Background of Research

Arsenic (As), a potentially toxic trace element, is of great environmental concern due to its presence in soil, water, plant, animal, and human continuum (Bhattacharyya and Sengupta, 2020). The origin of the word "arsenic" comes from the Persian word "zarnikh" and the Greek word "arsenikon" meaning "yellow orpiment". Arsenic is placed in group V and period 4 of the periodic table with an atomic number of 33 and molecular weight of 74.92. Arsenic is a naturally occurring element that has both metallic and non-metallic properties.

It can be present in soil, air and water as a metalloid and as chemical compounds of both inorganic and organic forms (Matschullat, 2000; Laha et al., 2021). It enters soil from mining, coal burning, sewage sludge, and pesticides, and is then absorbed by plants and enters the food chain, posing a threat to human health (Xue et al., 2017; Wan et al., 2018; Bhattacharyya et al., 2021a).

Contamination of land and aquatic systems with As is a persistent global issue, particularly in South and Southeast Asian and Latin American countries(Hussain et al., 2021). Numerous incidents of groundwater As contamination and human suffering have been documented in 20 countries around the world (including Argentina, Chile, Finland, Hungary, Mexico, Nepal, Taiwan, Bangladesh, India, and others). Around 200 million people are thought to be at danger of As poisoning, either directly by drinking As-contaminated groundwater or indirectly via consuming As-laced food crops, primarily rice (Oryza sativa L.) irrigated with As-contaminated groundwater (Chowdhury et al., 2020). In India the severity of As contamination has been reported from the Indo-Gangetic plain (Chakraborti et al., 2003). Arsenic concentration in natural waters, except groundwater is generally low, other than in the areas characterised by the geothermal water or mining activities. While drinking water is considered as the most important source for As exposure, food is equally important exposure route and most important route of exposure in areas with safe drinking water. Food gets contaminated mainly due to contaminated irrigation water resulting in soil-crop-food transfer (Mandal et al., 2019; Kumar et al.,2021; Sengupta et al., 2021). In West Bengal, India, two methods of paddy cultivation are commonly practiced: *Boro*, which involves pre-monsoon cultivation irrigated with groundwater, and *Aman*, which involves post-monsoon cultivation that is rainfed and sometimes irrigated. *Boro* rice covers a larger area of approximately 15.12 million hectares compared to *Aman* rice, which covers around 3.93 million hectares (Chowdhury et al., 2018).

Both these cultivation methods result in a significant uptake of As in the harvested rice. The findings of Chowdhury et.al. (2020) strongly support this statement.

Apart from drinking water, rice is an important route of As exposure in endemic areas (Mondal et al., 2010; Mondal & Polya, 2008). Recently elevated levels of As in rice grain from the As affected areas of West Bengal, India have been reported (Chowdhury et al.,2020; Sengupta et al.,2021). In As impacted districts of West Bengal, India, Golui et al. (2017) reported As concentrations in rice grain ranging from 2 to 1260 g kg^{-1}, with a mean value of 146 g kg^{-1}. Similar increased levels of As in rice have been recorded from As polluted areas in the Maner block of Patna District, Bihar, India (Singh & Ghosh, 2011). Initially the maximum allowable level of As in rice grain, according to the World Health Organization (WHO), was considered to be 1 mg kg^{-1} (Meharg & Rahman, 2003). In the case of the United States Department of Agriculture (0.15 mg kg^{-1}) and the European Union (0.5 mg kg^{-1}) it is stricter (Meharg & Rahman, 2003). Recently as per the Codex Alimentarius Commission a joint committee of WHO and Food and Agriculture Organization (FAO) the maximum level (ML) of inorganic As (i-As) in husked rice is 0.35 mg kg^{-1} and it is 0.2 mg kg^{-1} for polished rice (JECFA, 2017). Apart from drinking water, irrigation water contaminated with As acts as a potent source of contamination to humans through water-soil-crop food transfer. While drinking water, already having a permissible/safe limit, quantification of safe limit for irrigation water and soil for rice is also required. However, there are a few things to keep in mind. For example, the soil total As, does not account for how crucial soil physio-chemical properties affect its availability. To address this issue inclusion of bio-accessible As (plant available As) as one of the predictor variables is required. Although total elemental concentrations are regarded as a reliable indicator for determining the long-term enrichment of the soil and estimating the source of the elements, they 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.

Management solutions that are both efficient and feasible (under local conditions) for the remediation of As-contaminated soil is of importance to reduce human health risks from soil-crop-food transfer. Both phytoremediation and bioremediation of the As contaminated sites have been undertaken (Laha et al., 2021 & 2022; Upadhyaya et al., 2018 and Mondal et al., 2021). Solutions include amendments used for the remediation of As-contaminated soil such as the use of inorganic elements like phosphorus (Hossain et al., 2009),

silicon (Ma et al., 2008), iron (Ultra et al., 2009) and selenium (Wan et al., 2018); and complexation of As by the application of organic amendments such as sugarcane bagasse (Mandal et al., 2019a and 2019b) and vermicompost (Sengupta et al., 2021; 2022). One of the limitations of organic amendments, however, is that they need to be applied in large quantities. Biochar is an effective amendment in reducing the ecotoxicity of soils that are contaminated with heavy metals because it can effectively bind metal(loid)s in water and immobilise them in soil (Guo et al., 2020; Ahmad et al., 2014; Tan et al., 2015; O'Connor et al., 2018). Biochar is prepared by carbonising organic materials through pyrolysis at high temperatures (between 300 and 1000 °C) with little or no oxygen (Lehman and Joseph, 2015). The surface functional groups on biochar, such as hydroxyls, carbonyls and carboxyls, serve as binding sites for metal(loid)s (Tan et al., 2015). Due to the presence of negatively charged surface functional groups, biochar can electrostatically bind heavy metal cations and adsorb them. Additionally, the electron-rich aromatic biochar surface may electrostatically draw electron-deficient metal cations to itself through donor-acceptor interactions (Vithanage et al., 2017). The encouraging results of many studies regarding the efficacy of biochar in binding contaminants have stimulated much interest in using it as a soil amendment for environmental rehabilitation (Guo et al., 2020). At present, most biochar research publications looked at its use from a technical or economic perspective in relation to soil quality and the remediation of surface-, ground-, and waste-water. An integrated understanding of the mechanisms of remediation of As-contaminated soils (specifically in the rice rhizosphere) through pristine and modified biochar to improve the functional properties of biochar could result in future larger-scale applications.

1.2. Aim and Objectives

The aim of this research is to determine the maximum permissible soil and irrigation water As concentrations for rice cultivation accounting for differences in chemical properties of the soil (pH, organic carbon, clay, iron and phosphorus) and to assess the efficacy of biochar as an amendment for As contaminated soils. To meet this, aim the present study has been planned with the following objectives.

1. To build predictability models for maximum permissible soil total As through meta-analysis.
2. To assess the predictability of the models with field data from As contaminated sites of India.

3. To assess the role of soil chemical properties on grain As content and deriving the limit for bioavailable As.
4. To propose the guideline value of As in soil and irrigation water for cultivation of rice in As contaminated sites of India.
5. To assess the efficacy of biochar as an amendment for As contaminated soils for cultivation of rice through meta-analysis.

1.3. Thesis Structure

Chapter 1 – Introduction

In Chapter 1, an overview is provided on the contamination of groundwater by As and its impact on human exposure through the transfer of As from soil to crops and food. Additionally, the chapter outlines the significance of undertaking this study, elucidates its purpose and primary objectives, and establishes the specific goals that the study aims to tackle.

Chapter 2 - Review of Literature

The Chapter 2 consists of a comprehensive analysis and synthesis of existing research works undertaken regarding the origin, distribution of As contamination in groundwater, soil and rice. It serves to provide a theoretical and conceptual framework for the research being conducted and demonstrates the existing knowledge and research gap.

Chapter 3- Materials and Methods

In Chapter 3 a detailed description of the experimental sites, collection of soil and plant samples, analytical procedures followed in the laboratory, and statistical methods used in the research study have been provided. This chapter explains how the studies were conducted and allows readers to evaluate the reliability and validity of the results.

Chapter 4 - Building predictability models for maximum permissible soil total As and irrigation water As through meta-analysis.

The Chapter 4 deals with the prediction of soil total As and irrigation water As guideline values using the meta-data from the published research papers. This chapter provides a detailed and in-depth analysis regarding the use of machine learning algorithms (logistic regression and decision tree) in developing predictability models.

Chapter 5 - Assessing the predictability of the logistic regression and decision tree models over field data.

In Chapter 5 the performance of the predictability models derived from the meta-data has been assessed with the three sets of field data. A piece of detailed information and discussion regarding the model performance matrices over those test data sets has been covered in this chapter.

Chapter 6- Determination of bioavailable As threshold by random forest, gradient boosting machine and logistic regression

Chapter 6 deals with the prediction of threshold value for soil bioavailable As to ensure safe cultivation of rice using the machine learning algorithms. This chapter also provides an in-depth information regarding the effect of soil properties on the bioavailability of As.

Chapter 7 - Predicting the limit of As concertation in irrigation water for cultivation of rice.

This chapter provides the detailed information regarding the prediction of guideline value for irrigation water As through an incubation study with monolithic soil columns collected from the As-contaminated sites.

Chapter 8 - Evaluation of biochar as an amendment for mitigation of As contamination in rice through meta-analysis.

The efficacy of biochar as an amendment to mitigate As in rice soils has been assessed in this chapter trough meta-analysis. An in-depth analysis and discussion on the effect of biochar (both pristine and modified) on rice grain As content, fractions of soil As and also on the rice plant growth attributes (height, tiller number, root and shoot biomass) has been covered in this chapter.

Chapter 9 – Summary and Conclusion

This chapter outlines a brief overview of the entire thesis to provide a clear understanding of the research objectives, methodology, key findings, and their significance. The main findings and contributions of the research has been summarized in this chapter.

Chapter 10- References

This chapter lists all the sources cited or consulted during the research process to provide the necessary information for readers to locate and verify those sources.

Chapter 2- Review of literature

2.1. Origin of Arsenic in soil and groundwater

In order to formulate the underlying reasons of the widespread groundwater As problem four main hypotheses for the mobilization and transportation of As in groundwater have been proposed (Bhattacharya *et al.* 2015): oxidation of pyrite, competitive ion exchange, reductive dissolution of iron oxyhydroxides, and reduction and re-oxidation. A large number of As bearing minerals are present in the environment including arsenical pyrite (FeAsS), realgar (AsS) and orpiment (As_2S_3) (Khosravi *et al.*, 2019). The alluvial plains of Eastern India and Bangladesh subjected to widespread contamination have pyrite oxidation as the major underlying process where excessive groundwater use for irrigation creates an oxidizing status of the aquifers (Chakraborty *et al.* 2015). Under aerobic conditions, Fe oxide-hydroxide or ferric oxyhydroxide (FeOOH) is stable and does not release As whereas anoxic conditions results in the reduction of FeOOH to ferrous oxide (Fe_2O_3) and As is mobilised. The maintenance of such anoxic conditions is proposed to be facilitated by the widespread practice of wetland rice cultivation in the As contaminated regions (Sanyal, 2017). The competitive ion exchange theory deals with the competition among the As oxyanions and phosphate ions that decipher the release of As in the aquifers (Fakhreddine *et al.* 2015). However, the most important factor of As contamination in India and Bangladesh is based on the principle of reductive dissolution of metal oxides and Fe hydroxides, which results in the release of As. The fourth hypothesis behind the As menace deals with the reduction and re-oxidation theory (mobilization of As via Fe oxyhydroxides reduction followed by pyrite re-oxidation). This combination of processes although enables As immobilization yet a reduced environment restricts such process making As bioavailable (Bhattacharya *et al.* 2015; Shukla *et al.* 2020). Overall, the background of As contamination suggests that under an anoxic condition of the underground aquifer, As mobilization from As-bearing sediments to the groundwater aquifer occurs. Wetland paddy cultivation in the affected regions promotes this anoxic state.

In West Bengal (Figure 2.1), the presence of As in groundwater in concentrations exceeding the maximum acceptable concentration was first detected in 1978, while the first case of As poisoning in humans was diagnosed at the School of Tropical Medicine in Calcutta in 1983 (Acharya, 1997). The effect of ingestion of inorganic As in drinking

water and the health effects in adults has also been well established (Guha Mazumdar *et al.,* 1998). The main focus of attention, until recently, has been exclusively on As contamination in groundwater-derived drinking water. The As loading of the groundwater which is used as irrigation source varied from 0.06 to 0.53 mg/L in Nonaghata mouza of the Haringhata block of Nadia district in West Bengal (ICAR, 2005). A high degree of such contamination was also found in different parts of the affected-belt, to name a few, Gotera and Ghentugachhi mouzas of Chakadaha block of Nadia district (ICAR, 2001) (ranging from trace to 0.89 mg/L); Ambikanagar, Chakla, Ijajpur and Chyangdana villages under Deganga block of North 24- Parganas district of West Bengal (varying from 0.05 to 0.50 mg/L), etc., by a group of researchers at the Bidhan Chandra Krishi Viswavidyalaya (Sanyal *et al.*, 2015).

Figure 2.1. Groundwater As status of West Bengal (*http://www.soesju.org*)

2.2. Chemistry and forms of As in groundwater-soil environment

Arsenic is released into the environment in both inorganic and organic forms. Arsenate (AsV) is inorganic, phyto-available forms of As in soil solution (Meharg & Hartley-Whitaker, 2002). However, microbes, which can methylate and demethylate As species in soils, may transform inorganic arsenic species to organic As species and vice-versa (Turpeinen et al.,1999) . As (V) exists in four forms in aqueous solution based on pH: H_3AsO_4, $H_2AsO_4^-$, $HAsO_4^{2-}$ and AsO_4^{3-}. Similarly, As (III) exists in five forms: $H_4AsO_3^+$, H_3AsO_3, $H_2AsO_3^-$, $HAsO_3^{2-}$ and AsO_3^{3-}. The ionic forms of As (V) dominate at pH > 3, and As (III) is neutral at pH < 9 and ionic at pH > 9. At pH 6 – 8, in most aquatic systems, both $H_2As_2O_4^-$ and $HAs_2O_4^{2-}$ ions occur in considerable proportions in an oxidized environment (redox potential, Eh = 0.2 – 0.5V), while the aqueous acid, H_3AsO_3, is the predominant species under reduced conditions (Eh = 0 – 0.1V) (Sadiq, 1997; Sanyal, 2014). The toxicity follows the order: *arsine* (-3) > *organo-arsine compounds* > *arsenites* (As^{3+}) and *oxides* (As^{3+}) > *arsenate* (As^{5+}) > *arsonium metals* (+1) > *native As metal* (0) (Ghosh et al.,2004).

2.3. Arsenic and Rice

Rice serves as a primary food source for over 50% of the world's population, particularly in Asian, African, and Latin American nations (GriSP, 2013). In Asia, rice is the basic staple food for majority of the population, including the region's 560 million poor (GRiSP, 2013). The Asian region accounts for 90% of global rice production and consumption (Sengupta et al., 2021), some of which occurs in As contaminated areas. In India and Bangladesh, daily consumption of milled rice is high (approximately 68.2 and 173.3 kg $person^{-1}$ $year^{-1}$ respectively (Sengupta et al., 2021). In Bangladesh, approximately 69.6% of calorific intake comes from rice and in India it comprises of 29.1% (Sengupta et al., 2021). Rice is a rich source of dietary fibre and nutrients, including carbohydrates, proteins, vitamins and minerals (Mwale at al., 2018). Rice accumulates higher concentration of As in grains than other cereals such as wheat and barley (Williams et al., 2007). Rice, hydrophilic in nature uses an enormous amount of water throughout it's lifecycle. The amount of water used to grow rice can be quantified. According to the Food and Agriculture Organization of the United Nations (FAO), it takes about 2,500 litres of water to produce one kilogram of rice (FAO, 2014). This means that growing rice is a very water-intensive crop. The FAO also notes that the amount of water used to grow rice can vary significantly depending on the specific circumstances. For example, in irrigated systems,

the amount of water used can be as high as 10,000 litres per kilogram of rice. In rainfed systems, the amount of water used can be much lower, but it can vary depending on the specific circumstances. In the current climate, with water scarcity becoming a more and more pressing issue, the amount of good quality water used to grow rice is a major concern. In some parts of Asia, where rice is a staple food, water resources are already stretched thin (FAO, 2016).

Most rice soils, which are commonly found in the alluvial lowlands in humid climates, especially Entisols or Inceptisols, have undergone little soil formation. A few chronological studies of the mineralogy of rice have indicated that the mineral composition of rice soils was almost the same as that of their parent materials. The conclusion of minimum chronological effects on mineral composition of rice soils was corroborated, except in the case of biotite, which was rapidly weathered by seasonal wetting and drying of rice soils. This latter process is grown on almost all of the major soils, either exclusively or in rotation with dryland crops (Brammer, 1978). Rice soils are affected by alternating reductive and oxidative conditions, resulting in damage to clay lattices and/or chloritization of expanding 2: 1 clay minerals (Brinkman, 1970; Yoshida and Itoh, 1974). This upper layer, resulting from ferrolysis, was impermeable to water percolation. Ferrolysis results in the transformation of partially Fe^{2+} -saturated clay to partial H^{+}-saturation during the period of oxidation (Brinkman, 1970; Yoshida and Itoh, 1974). Exchangeable H^{+} reacts with the clay lattice, resulting in release of Al^{3+} and partial Al^{3+} saturation. Thus, there was partial Al^{3+}-interlayering of the Fe^{2+}-saturated clay during the seasonally wet period. This Fe^{2+} - H^{+} Al^{3+} replacement induces clay destruction and change in acidity in rice soils (Brinkman, 1970; Yoshida and Itoh, 1974). Another reaction is observed in the rice soils, i.e., smectite and vermiculite, which have a higher cation-exchange capacity, resulted from the degradation of biotite in sediments which have originated from the Himalayan Mountains.

The mechanical and chemical compositions of rice soils from nine tropical Asian countries (i.e., Bangladesh, Burma, Cambodia, India, Indonesia, Malaysia, Philippines, Sri Lanka and Thailand) were studied (Yoshida and Itoh, 1974). The Bangladesh samples generally had a higher silt concentration than the other samples, according to the international grain-size limits (i.e., silt, 0.002 – 0.02 mm), due to sedimentation from the Ganges and the Brahmaputra rivers. Minerals rich in Fe, e.g., hematite (α-Fe_2O_3), goethite (α-FeOOH), lepidocrocite (γ-FeOOH), siderite ($FeCO_3$), jarosite ($KFe_3(SO_4)_2(OH)_6$) and vivianite

($Fe_3(PO_4)_2.8H_2O$), were identified in rice soils from nine Asian countries (Van Breemen, 1976). Minerals rich in Fe often have higher adsorbing capacities for AsO_4^{3-} than PO_4^{3-} (Violante and Pigna, 2002). Alternating oxidation/reduction in rice might strongly impact mobility and accessibility of As, since the As might be co-precipitated or trapped by iron oxides precipitated during the oxidation of reduced rice soils.

Again, redox conditions and pH significantly affected the availability and consequent phytotoxicity of inorganic and organic As species; however, it is only in wet land soils (rice paddies) where redox conditions are very different to non-wetland soils (Marin et al., 1993). The redox state and pH of the soil has a major influence on As speciation and solubility (Carbonell Barrachina et al., 2000).

As contaminated groundwater are used extensively to irrigate paddy rice mostly during the dry season (Meharg & Rahman, 2003). The spatial distribution of As in soil contaminated with groundwater used for irrigation is difficult to calculate, because As concentrations may be 70 mg kg^{-1} in the top soil near to the inlet of irrigation system, but may drop to nearly background levels of <5–10 mg kg^{-1} over a distance of few hundred meters form the inlet (Dittmar et al., 2010; Panaullah et al., 2009; Takahashi et al., 2004). Even so, As in the soil can quickly build up because only plant uptake or erosion can reduce As in the soil. Under anaerobic conditions, arsenite (63%) is the most abundant species, while monomethylarsonic acid (MMA) (14%), dimethylarsinic acid (DMA) (11%), and As(V) (39%) are also present (Abedin et al. 2002). Arsenic toxicity influences rice plant growth; in fact, As-toxicity causes ATP inhibition and oxidative stress, resulting in decreased yield (Panaullah et al., 2009). In soils containing >60 mg kg^{-1} total As, toxicity symptoms such as stunted growth, brown patches, and burning on rice plant leaves were seen (Zhao et al., 2010). When soil As concentrations were increased from 12 to 60 mg kg^{-1} in traditional paddy fields in Bangladesh, rice yields decreased from 7.5 to 2.5 t ha^{-1} (Duxbury and Panaullah, 2007).

2.4. Arsenic uptake mechanism in rice plant

Arsenic is taken up in rice roots by two mechanisms. First mechanism involves the phosphate (PO_4^{3-}) transport pathway using high affinity PO_4^{3-} transporters (Muchhal et al., 1996; Shin et al., 2004; Catarecha et al., 2007), which uptake As(V) from soil solution and subsequently to aerial parts of the plants (Zhao et al., 2010). The second route by which As is taken up by plants roots is through aquaporin channels, which uptake As(III) (silicic acid analogue) and

methylated As species (MMA and DMA) (Jian et al., 2008). As(III) utilises this Si transporter in rice root cells because it is comparable to silicic acid; both have a high pKa (9.3 for silicic acid and 9.2 for arsenous acid, respectively) and a tetrahedral structure of similar size. After As(III) is taken up by root cells, some of it is released into the rhizosphere right away, which is partially mediated by Lsi1 acting as a bidirectional channel (Zhao et al., 2010); the rest is sequestered in root vacuoles or translocated to the shoots, where it is distributed to various organs (Zhao et al., 2010). Methylated As species like MMA and DMA contribute very little to total As in soil. The intrinsic protein nodulin 26-like may be used to absorb MMA and DMA (Bakhat et al., 2017). Inorganic As, on the other hand, is thought to be significantly more hazardous than pentavalent methylated As species (Meharg et al.,2009). It is also clear that DMA is less hazardous than inorganic As species (Syu et al., 2015). In soil the As exists as arsenite (As^{3+}) and arsenate (As^{5+}) forms. Arsenite is often found in anaerobic soils, where it is more soluble and more easily taken up by plants whereas As^{5+} is found in aerobic soils. The prevalence of anaerobic condition during rice cultivation results in the uptake of As in As^{3+} forms. The relative abundance of As^{3+} and As^{5+} in soil depends on several factors, including the pH and redox potential (Eh) of the soil, the presence of organic matter, and the presence of other metals (Hussain et al., 2020). Due to continuous change of pH and Eh of the soil during the cultivation of rice quantifications of the inorganic form of As (As^{3+} or As^{5+}) is a challenging task (Sengupta et al., 2023). Regardless of rice variety, the accumulation of As in the root was found to be 28 and 75 times higher than in the shoot and uncooked rice grain, respectively (Rahman et al., 2007).

2.5. Factors affecting Bioavailability of As

2.5.1. Contaminated irrigation water source

The bioavailability of As for crops is dependent on various factors. One of the prime factors is contaminated irrigation water. Several authors have reported the fact that irrigation water significantly contributes towards the build-up of As in soil and in turn increasing the bioavailability (Bhattacharya et al., 2010b; Biswas et al., 2014; Golui et al., 2017; Sengupta et al., 2021).

2.5.2. pH and Redox potential of soil

The availability of As to the crops is governed by the redox potential (Eh), pH of the soil. The speciation and solubility of As is mainly governed by pH and Eh of the soil. At high Eh values As^{5+} exists as H_3AsO_4, H_2AsO^{4-}, $HAsO_4^{2-}$ and AsO_4^{3-} whereas at low Eh values the corresponding As^{3+} species is present. Arsenic solubility in soils is considerably low at neutral or slightly acidic pH and increased considerably in both strongly acidic and alkaline conditions. The soluble or rather available As level in soil should increase substantially with diminishing Eh and increasing pH as reported by (Majumdar & Sanyal, 2003). Furthermore, at a high pH, the OH^- ion concentration would increase, causing displacement of As^{3+} and As^{5+} species from their binding sites through competitive ligand exchange reactions (Bhattacharyya, 2004).

2.5.3. Soil Organic Carbon and clay content

The oxidizable organic carbon plays an important role in retention and release of As to bioavailable forms. Much research looks at the mitigation potential of soil amendments such inorganic fertiliser or organic manure application, which can immobilise, adsorb, bind, or co-precipitate As in situ. By producing metal-humate complexes (chelates) with varying degrees of stability, soil organic components such as humic acid (HA) and fulvic acid (FA) operate as effective As accumulators (Sinha & Bhattacharyya, 2011; Mandal et al., 2019a; Kumar et al., 2021). The stability constant (log K) of the complexes formed by the soil HA/FA with As in the contaminated soils suggested that organo- As complexes were quite stable, even in the presence of competing oxyanions such as phosphate and nitrate (Mukhopadhyay and Sanyal, 2004, Mandal et al., 2019a).

The clay content in soil can influence the mobility and bioavailability of As. Arsenic tends to bind strongly to soil particles, particularly those containing Fe oxides and Al hydroxides. Clay minerals, which are often composed of Al and Si, can also adsorb As, and affect its mobility in soil. In general, soils with a higher clay content tend to have higher As retention capacity, which can reduce As mobility and availability to plants. However, some studies have suggested that high clay content can also create reducing conditions that release As from soil particles, making it more available for uptake by plants (Hussain et al., 2021). Lin and Puls (2000) suggested that aging of the clay minerals affects the adsorption of As greatly with a possible understanding that long term aging results in stronger degree of bonding of As to the clay minerals as a result of increase in the levels due to increasing dehydrations and

As diffusion at the soil water interface to internal pores of the clay aggregates. Low Eh in paddy soils can cause enhanced flocculation and dispersion of clay particles, making it easier for clay to migrate to the bottom of the plough zone and produce a hard clay pan. Clay minerals cause finely structured soil with a large surface area. Iron (oxy)hydroxides are primarily co-precipitated on the surface of clay particles, which improves As retention in paddy soil and reduces rice plant uptake (Hussain et al., 2021).

2.5.4. Soil Phosphorus

Phosphorus (P) is an essential plant nutrient that is required for plant growth. As and P are both in Group Vb of the Periodic Table. Their interaction in the soil-plant system is a critical factor in As mobilisation. Indeed, it appears that these oxyanions would not be adsorbed separately in mixtures, but rather would compete for the same type of adsorption sites. (Raj et al., 2021). Several researchers found that the presence of phosphate reduced arsenate adsorption, and that the reduction was significantly greater for arsenate's competitive effects on phosphate adsorption by soil minerals, however there was a lot of variance in the degree of competition between these two oxyanions (Mukhopadhyay et al., 2002).

2.5.5 Silicon (Si), Iron (Fe), Manganese (Mn)

Silicon (Si) is the second most prevalent element in the earth's crust, and it is mostly absorbed by plant roots in the form of silicic acid (H_4SiO_4). The interaction of Si with As has received some attention in recent years (Chen et al., 2014; Saud et al., 2016). Bogdan & Schenk (2009) found that the concentration of Si in soil was inversely correlated with the As contents in rice. There is direct and indirect evidence that As is held in soils and sediments by oxides (e.g., Fe, Mn) with the development of inner-sphere complexes via the ligand exchange mechanism (Kumari et al., 2021; Woolson & Axley, 1971). Despite a high As content in the soil solution, Fe plaque proved to be quite effective at sequestering As and preventing rice from acquiring it. These findings point to a high mobility of As in the soil during floods, which is regulated by the formation of oxic/anoxic interfaces at the surface of the soil in contact with flooding water and in the rice rhizosphere, as described by Chowdhury et al. (2018). Soil Zn and Fe affecting the As in rice grain has been previously reported by Duan et al. (2013). The As pathway in the rice field (from groundwater to rice grain) is depicted in Figure 2.2.

2.5.6. Water management strategies

In India, particularly in West Bengal, two processes of paddy cultivation are followed: *Boro* (the pre-monsoon cultivation, irrigated with groundwater) and *Aman* (the post-monsoon cultivation, rainfed and irrigated sometimes). In terms of area *Boro* rice covers around 15.12 million ha whereas *Aman* rice covers around 3.93 million ha (Chowdhury et al., 2018). Both these methods of rice cultivation cause a significant uptake of As in the harvested produce. The findings of Chowdhury et al. (2020) vehemently confirm the above fact. The methods of cultivation (rainfed or irrigated) can be rationalised in terms of the level of contamination in irrigation water and the volume of water used for cultivation of rice to quantify the amount of As added to soil as explained by Chowdhury et al. (2020) as per the following equations:

Total water (L) = Water discharge rate (L/h) X Weekly watering (h)……(1)

Total As exposure (µg) = Total water (L) X Water As (µg L^{-1})……..(2)

2.6. Limits of As in irrigation water and soil

The European Union (EU) recommended that As in agricultural soil should not exceed 20 mg kg^{-1} (Hussain et al., 2021; Rahman et al., 2007). Lower and upper guideline values of 10 mg kg^{-1} and 50 mg kg^{-1} respectively have been prescribed by Finnish regulators (Finland, Ministry of the Environment, 2007; Toth et. al, 2016). However, these guideline values were not specific for paddy soils. For irrigation water, a regulatory limit of 100 µg L^{-1} for As has been adopted (FAO, 1992; Pescod, 1992). This is in line with the 100 µg L^{-1} maximum concentration recommended by Ayers and Westcot (1985) for trace elements in irrigation waters. The values recommended by the EU and the Ministry of Environment in Finland were for generic agricultural soils rather than for paddy soils. These guideline values will be useful for the policymakers and the stakeholders (farmers) to ensure effective management of the contaminated sites. Regulatory limits will help to ensure that exposure to As is kept within safe levels, which can help to prevent and protect public health. By establishing clear and consistent regulatory limits, stakeholders can have a better understanding of what is required to comply with regulations and can make informed decisions about how to manage and mitigate potential risks. It will also promote the use of safer and more sustainable practices and products; stakeholders can help to drive innovation and investment in new technologies and approaches that can benefit both the environment and the economy. These generic

agricultural values may not be appropriate for application to paddy soil conditions, which are known to enhance As bioavailability to rice roots (Meharg and Rahman, 2003).

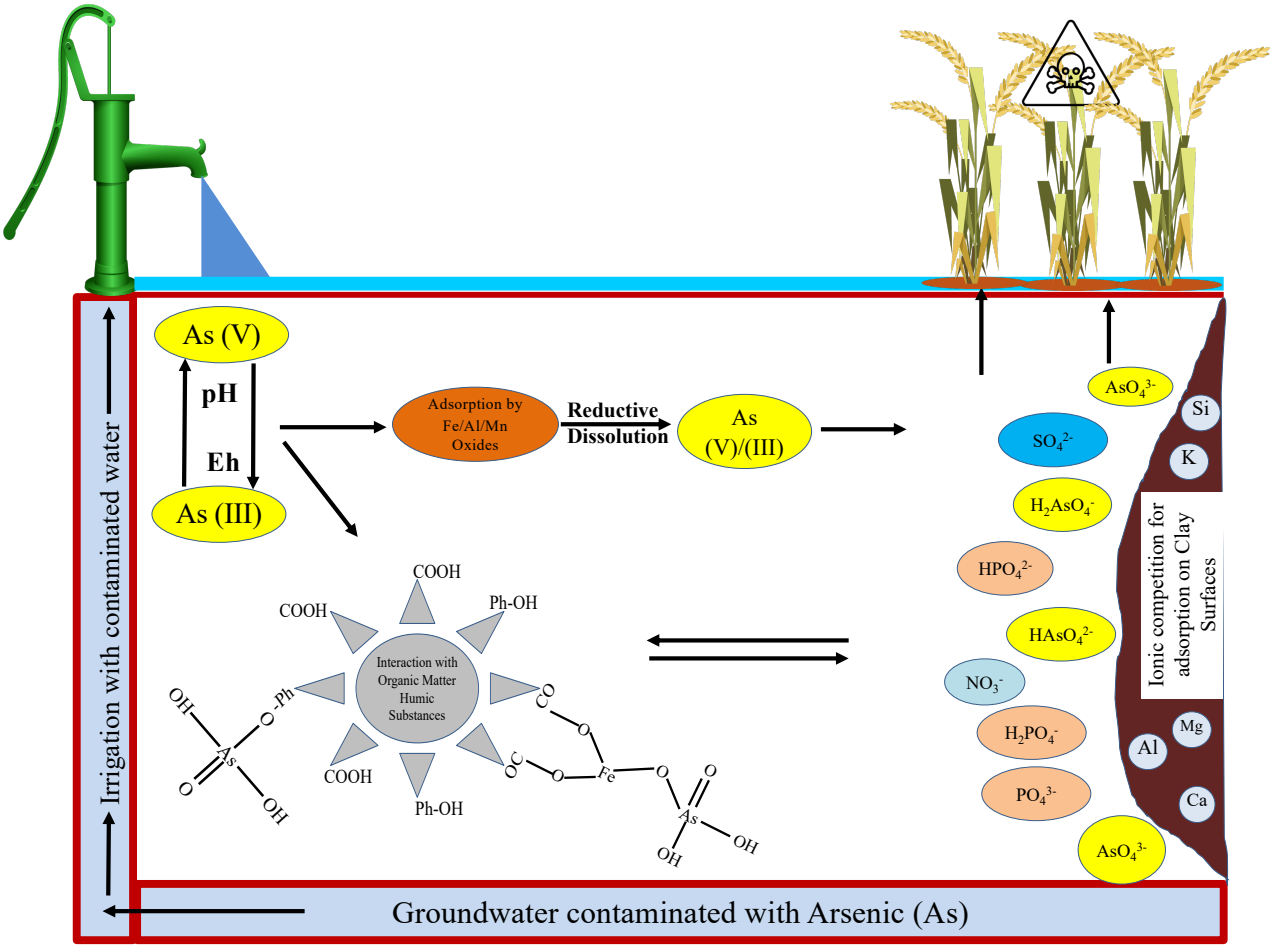


Figure 2.2. The As pathway in rice field. (Prepared by Jajati Mandal)

2.7. Biochar as an amendment for As contaminated soils

Biochar is a type of charcoal that is produced by heating organic material, such as wood, agricultural waste, or other biomass, in a low-oxygen environment. The process, known as pyrolysis, breaks down the organic material into a carbon-rich material that is highly stable and resistant to decomposition (Yaashikaa et al., 2020).

The yield and characteristics of biochar are based on thermochemical methods, operating conditions, and feedstock. It is widely known that low-temperature-generated biochars from slow pyrolysis have low hydrophobicity and aromaticity but significant surface acidity and polarity. Major biomass decomposition occurs between 200 °C and 500 °C through a series of phases that include partial hemicellulose decomposition, complete hemicellulose decomposition, full cellulose, and partial lignin decomposition (Rutherford et al., 2012). Use

of a high pyrolytic temperature (>450 °C) for preparation of biochar has been observed in all the studies considered in this review. In a meta-analysis, Arabi et al. (2021) reported that low pyrolysis temperature biochars (≤ 450 °C) did not affect As mobility in the soil, but high pyrolysis temperature biochars (> 450 °C) considerably mobilised the soil As. Biochars pyrolysed at high temperatures are more successful than those generated at low temperatures for As immobilisation, which could be attributed to the high aromaticity and porous structure, as well as the presence of mineral-phases (e.g., $CaPO_4$, $CaCO_3$) (Amen et al., 2020). The presence of acidic groups such as carboxylic, phenolic, and cationic groups on the surface of biochar produced at a shorter pyrolysis duration result in biochar with a relatively low pH (Shaaban et al., 2014). According to Beiyuan et al. (2017), biochar generated at low temperature has a greater O/C ratio than biochar produced at high temperature. These findings suggested that As stabilisation may be significantly aided by O-containing functional groups, such as carboxyl and hydroxyl groups (Shaaban et al., 2014). In contrast, a little higher pH in the higher temperature pyrolysis likely led to increase As mobilisation (Beiyuan et al., 2017; Zhao et al., 2018). The pH of the medium (i.e., soil solution) can affect the charge characteristics of the biochar surface as well as As speciation, but not the pH of biochar. For example, depending on the solution pH, various functional groups such as amine, alcohols, carboxylic, on the surface of biochar tend to be protonated, hence altering the surface charge of biochar (Vithanage et al., 2017). Contrarily, the pH of the solution affects the speciation of As into its many neutral and anionic forms such as H_3AsO_4, H_2AsO_4, $HAsO_4^{2-}$, and AsO_4^{3-}. At pH 3-6, the H_2AsO^{-4} species can dominate; however, at pH 8 and above, $HAsO_4^{2-}$ and AsO_4^{3-} species become dominant (Kumari et al., 2021; Raj et. al., 2020). Consequently, multiple species of As can be adsorbed on the surface of biochar at different pH values, making it difficult to determine the predominant species of As. The dominant As species on the surface, as well as how the As-surface complexation takes place, varies with changing solution pH. Biochar possesses various surface functional groups, such as hydroxyls, carbonyls, and carboxyls, which enable biochar to adsorb heavy metal cations through electrostatic attraction. However, immobilization of anionic metals, such as As, has been challenging, and modification or engineering of biochar is recommended to tailor its characteristics to overcome this issue.

Following the application of biochar in paddy soil, many crucial processes lead to the mechanisms of As immobilisation. The type of biochar used, and the modifications made

determine the mechanisms of As immobilisation/mobilisation. The presence of numerous functional groups like alcoholic, phenolic, and carboxylic groups associated with the biochar can play the role of electron donors regulating the reduction of As(V) to As(III) as has been seen in soils treated with biochar (Choppala et al., 2016). Anionic forms of As predominate. Therefore, the functional groups that are carboxylic and phenolic on the surface of biochar particles might not have a strong attraction for As (Irshad et al., 2020). Modification of biochar may be a solution to this problem. Rice straw biochar modified with hydroxyapatite and zeolite increased the amount of Ca in soil which sequesters As from the exchangeable fraction into insoluble Ca-bound As (Gu et al., 2018). Modification of wheat straw biochar with goethite resulted in the restricted mobility of As in paddy soil due to the presence of Fe functional groups. Applications of Fe and Mn oxide residues have been shown in numerous prior studies to minimise As uptake in rice by controlling the mobility and bioavailability of As in the soil through dissolution and mineralisation (Jindo et al., 2016). The application of maize straw biochar modified with manganese oxide decreased As(III) and As(V) mobility and availability in rice both moderately and heavily polluted rice soils (Yu et al., 2017). Thus, the application of biochar enhanced Mn content in soils as compared to the control. Fe/Mn plaque is reported to have a greater affinity for arsenate than arsenite. Reduced mobility and bioavailability are the result of As(III) and As(V) in soil due to the formation of complexes with a variety of oxides, including Fe, Mn, and Al oxides. The combination of zero-valent iron (ZVI) with oil palm fibre biochar resulted in a significant decrease in As bioavailability in rice soils (Qiao et al., 2018).

At present, most biochar research publications have focused on its use from a technical or economic perspective in relation to soil quality and the remediation of surface-, ground-, and waste-waters (Conte et al., 2021 and Yu et al., 2022). There has been little systematic, integrated research undertaken on the main properties/mechanisms of biochars that can be utilised to effectively prevent the availability and bioaccumulation of As from contaminated soils for the protection of human and animal health. The use of biochar for As remediation in India is still in its early stages, but the results of the studies that have been conducted so far are promising. Biochar has the potential to be a cost-effective and sustainable way to remediate As-contaminated soils in India.

In a real-world scenario, the selection of feedstock and charring processes for biochar production depends on several factors, including availability, cost, and local agricultural practices. The choice of feedstock may be influenced by the agricultural residues or biomass

abundant in the region. According to the Indian Ministry of New and Renewable Energy (MNRE), India produces an average of 500 million tons (Mt) of crop residue per year among which 34% comes from rice (Bhuvaneshwari et al., 2019). The specific properties of the biochar, such as surface area, pore structure, and presence of functional groups, can vary based on the charring process and feedstock used. It is challenging to determine the most effective feedstock or charring process for As immobilization in a particular location without considering site-specific factors and conducting thorough research and experimentation. Utilizing crop residues for biochar production holds great promise in managing crop residues in India and promoting agricultural sustainability. Furthermore, biochar also has the potential to address the management of As-contaminated soils when used as an amendment. Exploring this application of biochar in As-contaminated soils is necessary and can contribute to finding effective solutions.

Chapter 3-Materials and Methods

3.1. Meta analysis

The concept of meta-analysis can be traced back to the work of genealogist Karl Pearson in the early 1900s (Pearson, 1904). Pearson was interested in combining the results of multiple studies to see if there was a common underlying effect. However, it was not until the 1970s that the term "meta-analysis" was coined by statistician Gene V. Glass. Glass, along with Mary Lee Smith, popularized the use of meta-analysis in education research. They published a seminal article in 1978 that discussed the benefits of meta-analysis for summarizing the results of multiple studies and identifying patterns in the data (Glass, 1976). Their work helped establish meta-analysis as a valuable tool in social science research. Since then, the use of meta-analysis has expanded to other fields, including medicine, psychology, and environmental science. Today, meta-analysis is widely used in research to synthesize data from multiple studies and draw conclusions about the overall effect of an intervention or treatment. Meta-analysis is a statistical technique used in research to combine the results of multiple independent studies on a particular research question or topic. By combining the results of several studies, researchers can obtain a more precise estimate of the true effect size and identify patterns or inconsistencies in the results across studies. In a meta-analysis, researchers identify and collect data from multiple relevant studies, and then use statistical methods to synthesize the results from these studies into a single summary estimate. Meta-analysis typically involves a systematic review of the literature, in which researchers identify all relevant studies, assess the quality of each study, and extract data on the relevant variables.

3.1.1. Systematic Review and Data Extraction

The first step is to identify the research question or questions that the review will aim to answer. Then a set of criteria that articles must meet to be included in the review was developed. These criteria typically include factors such as study design, population, and outcomes (Higgins and Green 2011). Then a comprehensive search of multiple databases to identify all potentially relevant studies was undertaken from ISI Web of Science (https://clarivate.com/webofsciencegroup/solutions/web-of-science/) and Pub Med (https://pubmed.ncbi.nlm.nih.gov). The identified studies were screened based on the inclusion and exclusion criteria to determine which studies meet the criteria for inclusion. The

relevant data from the included studies was extracted, such as study design, population, interventions, and outcomes. The quality of each included study was assessed to determine the risk of bias and the overall strength of evidence. Finally, the data was analysed to draw conclusions about the research question(s) and determine the overall strength of evidence. Throughout the process, the established guidelines for systematic reviews that is Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines was followed (Moher et al., 2009). The process is designed to minimize bias and ensure that the review is comprehensive and transparent. The inclusion criteria for the research articles that were considered for meta-analysis have been outlined in chapter 4 (section 4.2.1) and chapters 8 (section 8.2). Further, the details of systematic review and the PRISMA diagrams are depicted in chapters 4 and 8.

3.2. Collection of soil and rice grain samples

Paired soil and rice grain samples (n=101) were collected from the rice growing As contaminated districts (Murshidabad, Nadia and North-24 Parganas) of West Bengal India (Figure 3.1) under rainfed system of rice cultivation (Plate 3.1). The soil samples were air-dried, thoroughly mixed and ground to pass through a 2-mm sieve and stored in zip-lock bags at 25°C for further analysis. Rice grain samples collected from the field were washed initially by tap water followed by dilute hydrochloric acid and finally with double distilled water. The samples were appropriately labelled, dried in a hot air-oven at 65°C for 48 hours. The dried rice grain samples were ground and stored in zip lock bags at 25°C for analysis.

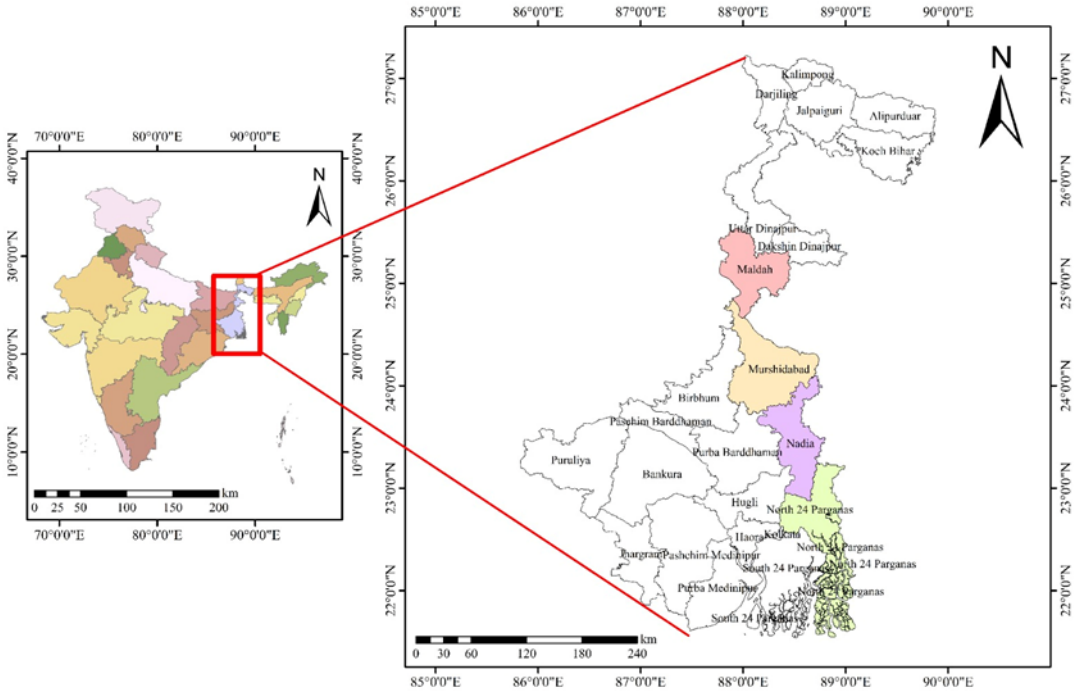


Figure 3.1 Soil and rice grain sample collection districts of West Bengal, India (Source of maps: DIVA-GIS).

Plate 3.1 Collection of soil and rice grain samples from As contaminated sites

3.3. Collection of soil columns

Monolithic soil columns were collected from Gotera, Dakshin Panchpota , Ghetugachi, Kalyani of Nadia; Jhumka, Sujapur, Beldanga, Radhavallabhpur of Murshidabad; and Baruipur, Sonarpur of South 24-Parganas comprising the As contaminated districts of West Bengal India (Figure 3.2) and Plate 3.2.

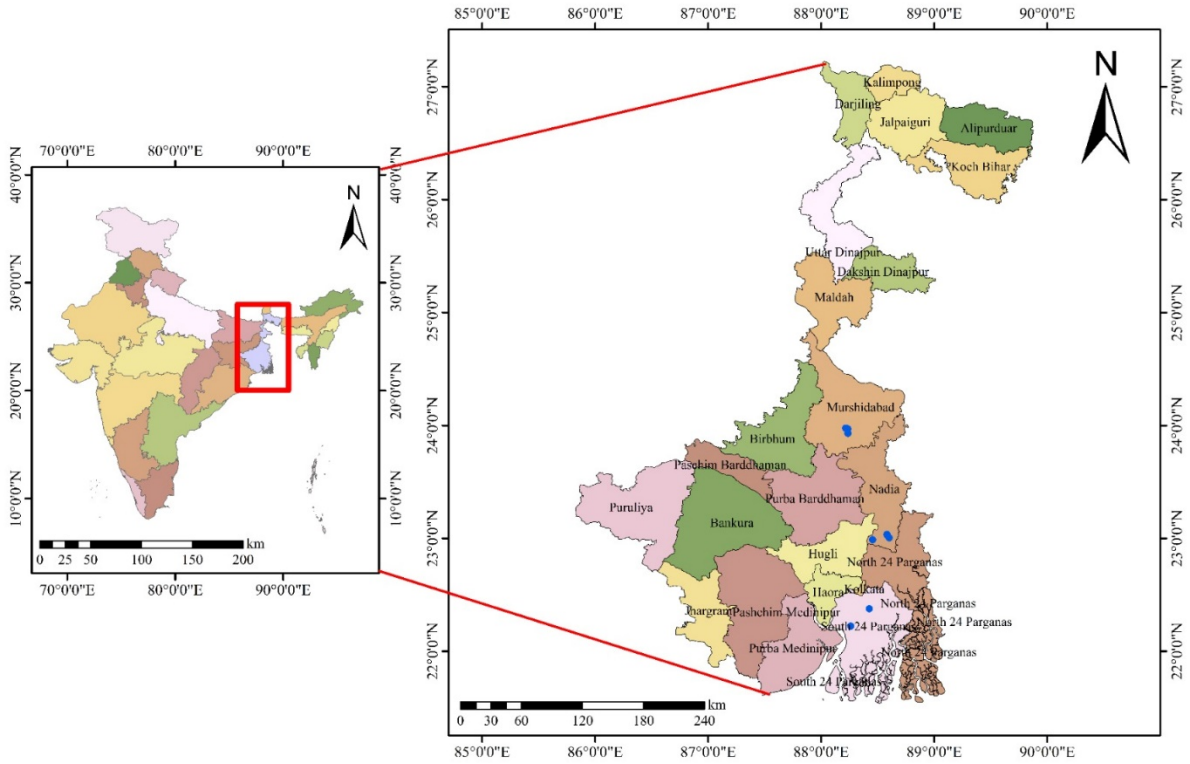


Figure 3.2 Soil columns collection sites across the As- contaminated districts of West Bengal, India (Source of maps: DIVA-GIS).

Plate 3.2 Collection soil columns from As contaminated fields

3.4. Incubation study with soil columns

As contaminated water (0, 100, 200, 300, 400, 500, 600 µ L^{-1}) was applied to the set of 210 columns as per the irrigation practices followed for cultivation of rice under rainfed conditions and remaining set of 210 columns as per the irrigated condition. After completion of 12 weeks of incubation the soil from each column was extracted, homogenised, grounded with a mortar and pestle, and sieved with 2 mm sieve. The soil samples were stored in zip-locked bags for further analysis.

Plate 3.3 Incubation study with soil columns

3.5 Analysis of soil and rice grain samples

3.5.1. Soil pH

pH of the soil sample was determined in soil suspension (soil: water :: 1:2.5) as determined by Jackson (1967) through an Eutech microprocessor based pH- EC-Ion meter.

3.5.2. Soil texture and clay content

Particle-size distribution of the soil was obtained by following the Hydrometer method (Bouyoucos, 1962), and soil textural class was determined from the percent contents of sand, silt, and clay with the help of the triangular textural diagram (Brady, 1990).

3.5.3. Oxidizable Organic Carbon

Organic carbon was determined by oxidizing the soil with 1 (N) $K_2Cr_2O_7$ in presence of concentrated H_2SO_4 and titrating back the remaining $K_2Cr_2O_7$ with ferrous ammonium sulphate (FAS) solution using diphenylamine indicator (DPA), following the wet digestion method of Walkley and Black (1934).

3.5.4. Available phosphorus (P)

The soil available phosphorus was extracted with 0.5 M $NaHCO_3$ as suggested by Olsen and Sommers (1982) and estimated through UV-VIS spectrophotometer (VARIAN CARY-50).

3.5.5. Available Iron (Fe)

Soil was analysed for available Fe by extracting with 0.005 M DTPA extracting solution (pH 7.3) following the method of Lindsay and Norvell (1978) through an Atomic Absorption Spectrophotometer (Model- Perkin Elmer AAnalyst 200).

3.5.6 Amorphous iron (AmFe) and aluminium oxides (AmAl)

Amorphous Fe and Al oxide of the soil sample was determined by extracting with 0.02 M ammonium oxalate, pH 3.0 following the method of Mckeague and Day (1966). Amorphous Fe oxide was determined using air-acetylene flame while amorphous Al using nitrous oxide-acetylene flame, using Atomic Absorption Spectrophotometer (Model- Perkin Elmer AAnalyst 200).

3.5.7. Bioavailable As

For bioavailable As the method outlined by ISO 2016 was followed (Groenberg et al., 2017). Five g of sieved air-dried soil sample was weighed into a polypropylene screw closure bottle and 50.0 ml of a 0.43 M HNO_3 solution were added. The bottles were mechanically shaken (end-over-end rotation) for 2 hours at room temperature and the extracts were filtered through vacuum driven filtration using a Millipore® filter unit and a Millipore® filter paper (0.45 µm pore size). All filtrates were kept at 4°C until further analysis. Each soil sample was extracted in triplicate. Two extraction blanks were included in each batch of 20 bottles. The samples were analysed in ICP-MS (Agilent 7850).

3.5.8. Digestion of soil samples

The EPA 3051 method was followed for sample digestion (method 3051A; USEPA 2007). The dried, finely powdered soil sample (0.2 g) was weighed into a dry, clean Teflon digestion vessel and 5 ml of aqua regia was added. The vessel was closed, placed into the rotor, and tightened. The loaded rotor was then placed into the microwave oven. A microwave digestion system (MARS 5; CEM Corp, Matthews, N C) with a rotor for 14 Teflon digestion vessels (HP 500) was

used for sample digestion and extraction. The microwave conditions for digestion were: stage 1: microwave power 1200 W, 300 PSI, and ramp for 2 min; stage 2: microwave power 1200 W, 300 PSI, and ramp for 3 min followed by a 5-min hold. After cooling for 30 min, the vessels were opened carefully. Each digestion solution was transferred to a 50-ml volumetric flask, diluted to the mark with MilliQ water, filtered through a Millipore membrane (0.45 µm), and kept in a plastic container for analysis.

3.5.9. Digestion of rice grain samples

0.5 g ground rice sample were weighed directly into a 75 mL digestion tube and 5 mL concentrated HNO_3 (trace analytical grade, 70%), obtained from Fisher Chemicals was added to it. The mixture was then allowed to stand overnight under fume hood. The following day, the digestion tubes were heated using temperature-controlled digestion block (A.I. Scientific Block Digestion System AIM 500), programmed to slowly ramp up to 140°C over 8 h and then to maintain temperature for the digestion of rice samples. Sample digestion was continued until only a small residual liquid remained in each tube. The tubes were removed from the digestion block and allowed to cool to room temperature in the fume cupboard prior to dilution (10 mL). The samples were mixed thoroughly by vertexing and filtered by a 0.45 µm syringe filter directly into plastic containers for storage prior to analysis.

3.5.10. Analysis of As in soil and rice grain samples

Rice and soil samples were analysed following the established protocols from Rahman et al. (2009a), Rahman et al. (2009b) and Alloway (2013). 7500ce series inductively coupled plasma mass spectrometer (ICP- MS) (Agilent Technologies, Tokyo, Japan), coupled with an auto-sampler (ASX-520, CETAC Technologies) and integrated samples introduction system (ISIS) was used to determine the amount of total As in soil and rice grain samples.

3.5.11. Quality Assurance and Quality Control

For quality control, Standard reference materials (SRM) from the National Institute of Standards and Technology (NIST), USA (Rice flour (SRM 1568b) and Montana soil (SRM 2711a)) were used. The CRM, blanks, duplicates, and continuing calibration verification (CCV) were included in each batch throughout the elemental analysis. Mean total recoveries from rice SRM was 268 ± 14.98 (n=5) µg kg^{-1} which indicates 94.09% recovery of the certified value of

285 ± 14 µg kg^{-1}. The mean total recovery from soil SRM was 104.27±3.13 (n=5) mg kg^{-1} which indicates 97.45% recovery of the certified value of 107 ± 5 mg kg^{-1}. Both the recovery from rice and soil SRMs confirmed accuracy of rice and soil digestion and analysis.

3.6. Other Data Sources

For prediction model training and validation, data from research collaborators in India from Bidhan Chandra Krishi Viswavidyalaya, West Bengal, and Indian Agricultural Research Institute, New Delhi was used.

3.7. Statistical Analysis

The statistical analysis of the data and prediction modelling was performed with R-Studio (*version 1.3.1093 2.3.1*). The R packages used for the purpose has been mentioned in each subsequent chapter.

3.7.1. Normality test of the data

The Shapiro-Wilks method was used for normality test of the data. The null hypothesis of these tests is that "sample distribution is normal". If the test is significant, the distribution is non-normal. A p-value > 0.05 implies that the distribution of the data is not significantly different from normal distribution. In other words, we can assume the normality. The details of the normality test of the data are depicted in Table 3.1. The statistical tests were undertaken based on the normality of the data. As the data were not normally distributed the non-parametric tests like Kruskal-Wallis and Wilcox were performed.

Table 3.1 Shapiro-Wilk normality test of the data across different chapters

Chapter Number	Variable name	p-Value	Data distribution
4	Irrigation water As	$8.261e\text{-}^{08}$	Non-normal
	Soil As	$2.2e^{-16}$	Non-normal
	Grain As	$2.126e^{-10}$	Non-normal
6	pH	$9.837e^{-10}$	Non-normal
	Organic carbon	$2.217e^{-15}$	Non-normal
	Available P	$2.202e^{-05}$	Non-normal
	Available Fe	$1.358e^{-06}$	Non-normal
	Bioavailable As	0.004169	Non-normal
	Total As	$1.454e^{-10}$	Non-normal
	Grain As	0.0002851	Non-normal
7	Total As	$2.2e\text{-}^{16}$	Non-normal
	Bioavailable As	$2.2e^{-16}$	Non-normal
	Available Fe	$5.446e^{-09}$	Non-normal
	Available P	$2.331e^{-12}$	Non-normal
	pH	0.0002379	Non-normal
	Organic carbon	$1.914e^{-07}$	Non-normal
	Clay	$2.762e^{-15}$	Non-normal
	Amorphous Fe	$2.2e^{-16}$	Non-normal
	Amorphous Al	$8.479e^{-16}$	Non-normal

3.7.2. Spearman Correlation

To examine the association between two variables and assess their monotonic relationship, Spearman's correlation coefficient was employed. This non-parametric method is particularly suitable when dealing with ordinal or ranked data, allowing for the evaluation of both linear and non-linear associations (Daniel 1990). Spearman's correlation coefficient is denoted by the symbol ρ (rho), ranges from -1 to 1. A value of -1 indicates a perfect negative monotonic relationship, where an increase in one variable corresponds to a decrease in the other. A value of 1 indicates a perfect positive monotonic relationship, where an increase in one variable corresponds to an increase in the other. A value of 0 suggests no monotonic relationship

between the variables. To calculate Spearman's correlation coefficient, the ranks of the values for each variable are used instead of the actual values. The correlation coefficient is then computed based on these ranks. If the variables have tied ranks (i.e., multiple values with the same rank), a correction factor is applied to adjust the correlation coefficient. To determine the statistical significance of the correlation coefficient, a two-tailed hypothesis test was conducted at the $\alpha = 0.01$, and $\alpha = 0.05$ significance level. The null hypothesis of no correlation was rejected if the p-value was less than 0.01 or 0.05. The details of the variables for the correlation studies are mentioned in chapter 4, 5 and 7.

3.7.3. Generalised Linear Model

A generalized linear model (GLM) is a statistical framework that extends the linear regression model to handle a broader range of response variables and error distributions. While linear regression assumes a normally distributed response variable with constant variance, GLM allows for more flexibility by accommodating different types of response variables, such as binary, count, or categorical data, as well as non-constant variance (James et al., 2013). GLM with a continuous dependent variable and continuous independent variables is a statistical modelling framework that extends linear regression to accommodate non-normal distributions and relationships between variables. GLMs are useful when the assumptions of traditional linear regression, such as normality of residuals, are violated or when the response variable is not normally distributed. The probability distribution of the response variable was chosen based on the nature of the data. For continuous data the Gaussian distribution was used. The details of the dependent and the independent variables used is mentioned in chapter 7.

3.7.4. Wilcox Test

The Wilcoxon test, also known as the Wilcoxon signed-rank test, is a non-parametric statistical test used to compare paired samples or repeated measures from the same population when the data do not meet the assumptions required for parametric tests, such as the paired t-test. The Wilcoxon test is particularly useful when the data are ordinal, skewed, or have outliers. It does not assume a specific distribution for the data, making it robust against violations of normality assumptions. The test evaluates whether the median difference between paired observations is significantly different from zero. It does this by comparing the ranks of the

absolute differences between the pairs, rather than the actual values themselves (Siegel 1956). The details of the null and alternative hypothesis have been mentioned in chapter 7.

3.7.5. Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric statistical test used to compare the medians of three or more independent groups or samples. It is an extension of the Mann-Whitney U test (Wilcoxon rank-sum test) for two groups and allows for the comparison of more than two groups. The Kruskal-Wallis test is suitable when the assumptions of parametric tests, such as the analysis of variance (ANOVA), are violated. It does not assume a specific distribution for the data and is robust against non-normality, making it applicable to a wide range of data types. The test evaluates whether the distributions of the groups differ significantly in terms of their location (median). It does this by comparing the ranks of the observations across the different groups. The details of the null and alternative hypothesis have been mentioned in chapter 7.

3.8. Machine Learning Algorithms

3.8.1. Logistic Regression (LR)

A logistic regression model or logit model as mentioned by James et al. (2013) was used to model the binary dependent variables. A probability value between 0 and 1 was allocated to each class. To identify the best fitting model both accuracy and kappa values were considered. The residuals of the LR model were checked for normality and the distribution was further confirmed from the plot. To estimate the coefficients from the data, the model could have two (X1 and X2) or more predictors. A linear relationship can be written in the mathematical form shown by equation 1, where p is the probability of the event that Y=1 and Y is the binary response variable. The quantity p (X) /(1−p(X)) is called the odds, which can take any value between 0 and ∞ and is calculated by maximum likelihood method. *β0*, *β1*and *β2* are the coefficients.

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 + \beta 1X1 + \beta 2X2 \dots\dots\dots\dots (1)$$

In terms of model complexity, it is low in logistic regression, particularly when no or few interaction terms and variable transformations are used. The dependent and independent variables used in LR is mentioned in details chapter 4, 6 and 7.

3.8.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a supervised learning model that is like logistic regression in that the outcome variable is categorical and can therefore be used for classification. It aims to project the feature space onto a lower-dimensional space while maximizing the class separability, so that the classes are linearly separable in the new feature space. LDA finds the linear combinations of features that best distinguish the classes, and the resulting transformed features can be used as inputs for a classifier to make predictions. The goal of LDA is to find the maximum ratio of between-class variance to within-class variance, which results in the best separation of classes in the transformed feature space (Fisher, 1936). The simplest form of LDA uses a linear discriminant function (LDF) as shown in equation 2 that runs through the centre points of the two groups to distinguish between them.

$$LDF = a + b1x1 + b2x2 + \cdots + bpxp \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where a is a constant, and b1 to bp are the regression coefficients for p variables. The simplest form, two-group LDA, uses a LDF that separates the two groups by passing through their centroids, or geometric centres. The boundary between the two groups can be represented as a function that is perpendicular to the LDF. There are infinite possibilities for this boundary, but the most common choice is one that is equidistant from both centroids, known as LDA with equal prior probabilities. However, if it is known beforehand that there are unequal proportions of objects in each group, the boundary can be shifted along the direction of the LDF towards one of the centroids to increase the likelihood of assigning objects to that group. The dependent and independent variables used in LDA is mentioned in detail in chapter 7.

3.8.3. Decision Tree (DT)

Classification and Regression Tree (CART) which is a non-parametric supervised learning method was proposed by Breiman (1984) and Ripley (1996). Decision trees are not black-box models, their outputs are easy to interpret, and the DT maps the behaviour or relationship

between the predictor and target variable (Dreiseitl and Ohno-Machado, 2002). The DT method has been used widely, for example, to identify heavy metals in environment by Jouanneau et al. (2011) and to streamline the mapping of soil pollution, for example in a study on rice cadmium concentration (Wang et al., 2020).

The algorithm divides the data set several times according to a criterion that maximizes data separation, resulting in a tree-like structure (Breiman, 1984). The used criterion is knowledge gain, which implies that the decrease in entropy, due to each split, is maximized. The ratio of y class elements over all elements of the leaf node that contains data item x is the estimate of P(y|x) (Dreiseitl and Ohno-Machado, 2002). The best DT model is selected based on Complexity Parameter (cp) and accuracy. The cp is used to control the size of the DT and to select the optimal tree size. If the cost of adding another variable to the DT from the current node is above the value of cp, then tree building is discontinued. The dependent and independent variables used in DT is mentioned in chapter 4.

3.8.4. Random Forest (RF)

Random Forest is a supervised machine learning algorithm for classification and regression based on the principle of recursive partitioning (Breiman, 2001), and independent of the assumption of functional relationships between the response and predictor variables. Briefly, Random Forest analysis ensembles numerous regression trees following a process called “bootstrap aggregation” or “bagging.” First, a random subset of the data space is drawn (with replacement) to grow a tree to its full length, and each node of the tree group’s observations are characterised by certain conditions on the predictor variables to produce an average prediction for the response variable. Each tree growing process uses only two-thirds of the bootstrapped data and one-third of the observations (out-of-bag data, OOB) are used for estimating the prediction errors. Second, each node split in a tree considers a random subset of predictor variables, usually a square root of the total number of predictor variables. The predictions from all the trees are averaged to make final predictions. The variable importance function within the Random Forest algorithm ranks predictor variables based on the increase in model error by randomly permuting the values of the predictor variables. The dependent and independent variables used in RF is mentioned in chapter 6.

3.8.5. Gradient Boost Machine (GBM)

Gradient Boost Machine (GBM) is a popular machine learning algorithm that is used for both regression and classification problems. It belongs to the ensemble learning family of algorithms and combines multiple decision trees to create a strong predictive model. Gradient Boost Machine integrates the predictions from various decision trees to generate the final estimate (Friedman, 2001). The algorithm works by sequentially adding decision trees to the model, with each new tree attempting to correct the errors of the previous trees. In other words, each tree is trained on the residuals (i.e., the difference between the predicted values and the actual values) of the previous trees. This process continues until the algorithm reaches a predetermined number of trees or until the error rate has been reduced to an acceptable level. The "gradient" in the name "Gradient Boosting" refers to the use of gradient descent optimization method to minimize the loss function of the model. The loss function is a measure of how well the model is performing, and the goal of gradient boosting is to minimize this function by iteratively adding new trees. However, it can be computationally intensive and requires careful tuning of its hyperparameters to avoid overfitting. The dependent and independent variables used in GBM is mentioned in chapter 6.

3.9. Model performance parameters

3.9.1. Confusion Matrix

A confusion matrix is a table that is often used to evaluate the performance of a classification model (James et al., 2013). It is a matrix of actual vs predicted values, where the predicted values are the model's predictions for each class, and the actual values are the ground truth labels for each class. The confusion matrix shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

The confusion matrix can be used to calculate various performance metrics for the classification model, such as accuracy, precision, recall, F1-score, and others. These metrics provide insight into how well the model is performing and can help in optimizing the model's performance. This parameter has been used in chapter 4, 5, 6 and 7.

3.9.2. Accuracy

Model accuracy is a performance metric that is used to evaluate how well a classification model can predict the correct class labels for the input data. It measures the percentage of

correctly predicted instances out of all the instances in the test dataset. This parameter has been used in chapter 4, 5, 6 and 7.

Accuracy= (TP+TN)/(TP+TN+FP+FN) …… (3)

3.9.3. Recall

Recall, also known as sensitivity or true positive rate (TPR), is a performance metric that measures the proportion of actual positive cases that are correctly identified by a classification model. In other words, recall measures the ability of a model to identify all positive instances in a dataset.

Recall = TP/ (TP+FN) …….. (4)

Recall is a useful metric to evaluate the performance of a classification model when the goal is to identify all positive instances in the dataset, even if it means accepting a higher number of false positives. However, a high recall value may come at the cost of lower precision (i.e., the proportion of predicted positive cases that are actually positive), which is a trade-off that needs to be considered based on the specific application of the model. This parameter has been used in chapter 5, 6 and 7.

3.9.4. True negative rate (TNR)

True negative rate, also known as specificity, is a performance metric that measures the proportion of actual negative cases that are correctly identified by a classification model. In other words, true negative rate measures the ability of a model to correctly identify all negative instances in a dataset.

True negative rate (TNR) = TN/ (TN+FP) …….. (5)

True negative rate is a useful metric to evaluate the performance of a classification model when the goal is to correctly identify all negative instances in the dataset, even if it means accepting a higher number of false negatives. However, a high true negative rate may come at the cost of lower recall or sensitivity (i.e., the proportion of positive cases that are

correctly identified), which is a trade-off that needs to be considered based on the specific application of the model. This parameter has been used in chapter 5, 6 and 7.

3.9.5. Precision

Precision is a performance metric that measures the proportion of predicted positive cases that are actually positive. In other words, precision measures the accuracy of positive predictions made by a classification model.

Precision = TP/(TP+FP) …… (6)

Precision is a useful metric to evaluate the performance of a classification model when the goal is to minimize the number of false positives. However, a high precision value may come at the cost of lower recall or sensitivity (i.e., the proportion of actual positive cases that are correctly identified), which is a trade-off that needs to be considered based on the specific application of the model. This parameter has been used in chapter 5, 6 and 7.

3.9.6. F1 Score

F1 score is a performance metric that combines precision and recall into a single value that balances both metrics. It is the harmonic mean of precision and recall.

F1 score = (2 x Precision x Recall)/ (Precision + Recall) ……. (7)

F1 score is a useful metric to evaluate the overall performance of a classification model, especially when there is an uneven distribution of positive and negative cases in the dataset. It provides a single score that balances both precision and recall and can be used to compare the performance of different models. In general, a high F1 score indicates that the model is performing well in terms of both precision and recall. However, in some cases, a higher priority may be given to either precision or recall, depending on the specific application of the model. This parameter has been used in chapter 4, 5, 6 and 7.

3.9.7. Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC) is a performance metric used to evaluate the quality of binary classification models. It considers true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) to produce a score between -1 and +1, where +1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between the prediction and actual values.

MCC = (TP * TN - FP * FN) / sqrt((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))…….(8)

MCC values range from -1 to +1, with 0 indicating no correlation and values closer to +1 indicating a strong positive correlation between the predictions and actual values. MCC is a useful metric in cases where the classes are imbalanced, as it considers the proportion of true and false positives and negatives. It is also useful when the cost of false positives and false negatives is different, as it provides a single value that balances both types of errors. In general, a higher MCC value indicates better performance of the model, and a value of zero indicates random prediction. This parameter has been used in chapter 5, 6 and 7.

3.9.8. Receiver Operating Characteristic (ROC)

ROC (Receiver Operating Characteristic) is an evaluation metric used to assess the performance of binary classification models. ROC is a graphical representation of the performance of a classification model that shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different classification thresholds. A ROC curve plots the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis for various classification thresholds. The curve is generated by moving the classification threshold from 0 to 1 and calculating the TPR and FPR at each threshold. This parameter has been used in chapter 4, 5, 6 and 7.

3.9.9. Area Under Curve (AUC)

AUC, on the other hand, is a metric that represents the overall performance of the model by calculating the area under the ROC curve. AUC ranges from 0 to 1, with 1 indicating perfect classification and 0.5 indicating random classification. A high AUC value indicates that the model can correctly classify a large proportion of positive and negative instances, while

minimizing the number of false positives and false negatives. AUC is a useful metric when the classes are imbalanced, and it provides a single value that summarizes the performance of the model. In general, a higher AUC value indicates better performance of the model, and a value of 0.5 indicates random classification. This parameter has been used in chapter 4, 5, 6 and 7.

Chapter 4

Building predictability models for maximum permissible soil total As and irrigation water As through meta-analysis

The contents of this chapter have been published as: **Mandal, J**., Sengupta, S., Sarkar, S., Mukherjee, A., Wood, M.D., Hutchinson, S.M. and 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

4.1. Introduction

Arsenic exposure, mainly through contaminated groundwater used for drinking, has widely been associated with detrimental health effects (Rahman et al., 2009). Though As exposure affects more than 200 million people worldwide (Shakoor et al., 2017), it has emerged as a major public health concern in Bangladesh and India, over the last few decades (Chakraborti et al., 2015). The World Health Organization (WHO) has established a guideline value of 10 µg L^{-1} for As in drinking water. Although contaminated irrigation water also contributes to As exposure by enhancing As concentrations in food crops (Mandal et al., 2019; Bhattacharyya et al., 2021), no WHO or international guideline value for irrigation water has been established to date.

Rice is a staple food for more than half of the global population, especially in Asian, African and Latin American countries (Majumder and Banik, 2019). In India and Bangladesh, daily consumption of milled rice is high (approximately 103 and 268 kg per capita $year^{-1}$ respectively; FAO, 2017). In Bangladesh, approximately 73% of calorific intake comes from rice (Mwale et al., 2018) and in India it comprises 30% (IRRI, Knowledge Bank). Rice is a rich source of dietary fiber and nutrients, including carbohydrates, proteins, vitamins, and minerals (Dipti et al., 2012; Mwale et al., 2018). However, rice consumption may also be a major route of As exposure (Mondal and Polya, 2008; Mondal et al., 2010; Mondal et al., 2020). Soil serves as a significant sink for As, which is highly bioavailable to rice roots under the conditions in which rice is cultivated (Kumarathilaka et al., 2018). Rice plants are major accumulators of As compared to other cereal crops (Williams et al., 2007) and irrigation of a paddy field with As contaminated water elevates As concentrations in paddy soil (Meharg and Rahman, 2003), rice straw, and grain (Panaullah et al., 2008). In Asia, rice is the basic staple food for the majority of the population, including the region's 560 million poor (GRiSP Global Rice Science

Partnership, 2013). During 2018–19, rice consumption in China was to the extent of 146.7 million tons, followed by India at 102 million tons (ICAR-NRRI Annual Report 2020). Apart from China and India, the other major rice producing countries are Bangladesh, Indonesia, Vietnam, Thailand, and Philippines. The production together accounts for more than 80% of global rice production (ICAR-NRRI Annual Report 2020) but unfortunately some of these regions are As contaminated. For example, in Bangladesh, 2.4 million out of 4 million hectares of paddy field have been found to be As contaminated (Akinbile and Haque, 2012).

The Joint FAO-WHO Codex Alimentarius Commission has recommended a maximum concentration of 0.2mg kg^{-1} for inorganic As in polished rice and 0.35mg kg^{-1} in husked rice (Codex Alimentarius Commission, 2017). However there have been limited attempts to establish paddy soil and irrigation water As concentrations above which the maximum recommended concentrations in rice may be exceeded. The usual range of total As in uncontaminated soil is 0.1–10 mg kg^{-1} (Zhao et al., 2010). The European Union (EU) recommended that As in agricultural soil should not exceed 20 mg kg^{-1} (Rahman et al., 2007; Rahaman et al., 2013; Hussain et al., 2021). Lower and upper guideline values of 10 and 50 mg kg^{-1} respectively have been prescribed by Finnish regulators (Ministry of the Environment, 2007; Toth et al., 2016). However, the values recommended by the EU and the Ministry of Environment in Finland were for generic agricultural soils rather than for paddy soils. These generic agricultural values may not be appropriate for application to paddy soil conditions, which are known to enhance As bioavailability to rice roots (Meharg and Rahman, 2003). For irrigation water, a regulatory limit of 100 µg L^{-1} for As has been adopted (Food and Agriculture Organization FAO, 1992; Pescod, 1992). This is in line with the 100 µg L^{-1} maximum concentration recommended by Ayers and Westcot (1985) for trace elements in irrigation waters but is again focused on generic agricultural production rather than rice specifically.

To our knowledge, no previous studies have derived maximum tolerable concentrations of paddy soil and irrigation water As above which rice grain As may exceed the maximum allowable concentrations set by the Joint FAO-WHO Codex Alimentarius Commission (JECFA, 2017). Using a meta- analysis approach, we attempt to determine soil and irrigation water As concentrations above which rice grains cultivated in Asian paddy fields may exceed the maximum tolerable concentrations of 200 µg kg^{-1} for inorganic As in polished rice and 350 µg kg^{-1} in husked rice.

4.2. Materials and Methods

4.2.1. Data Sources

We systematically reviewed published articles reporting As concentrations in paddy soil, irrigation water, and rice grains cultivated in Asian countries. We used Boolean operators (e.g., "OR" and "AND") to develop search terms from keywords ("arsenic," "contamination," "soil," "water," "rice," "risk"). Searching ISI Web of Science and PubMed with these terms, we identified relevant research papers published between 1980 and 2021, since 1980 onward the severity of As contamination was recognized in Asia. Studies were only included in subsequent meta-analysis if (1) the research was carried out in the field and not as pot experiments in the laboratory; (2) it was undertaken in Asian countries; (3) the As concentration data presented included total As of soil, rice grain, and irrigation water from the same study location; (4) the analysis of As was carried out using appropriate laboratory instruments rather than Field Testing Kits; and (5) details of the analytical method(s) and quality assurance procedures used for the study were provided. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flowchart can be seen in Figure 4.1. The details of the 26 selected research papers can be observed in Table 4.1.

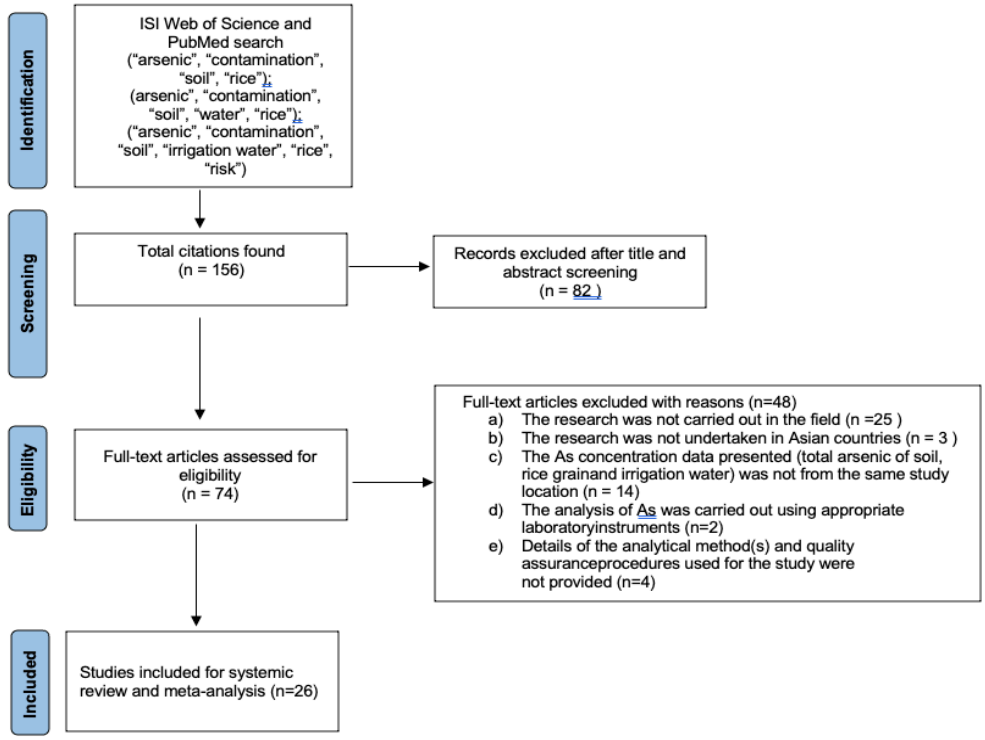


Figure 4.1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the selection of studies eligible for a meta-analysis.

Table 4.1. Characteristics of the studies included the meta-analysis

Sl. No	Author and year of publication	Location	No. of sites	Parameters Analysed (Range or mean)											Correlation (grain As vs soil As)
				t-As in grain (mg kg^{-1}) (dry mass)	SE(m)	As in Irrigation water ($\mu g\ L^{-1}$) (filtered)	As in soil (mg kg^{-1}) (dry mass)	pH	OC (%)	Texture	Redox (mv)	Fe (mg kg^{-1})	P (mg kg^{-1})	S (mg kg^{-1})	
1	Roychowdhury, 2008a	India	23	0.043-0.662	---	18-200	3.34-31.6	---	---	---	---	---	---	---	---
		India	18	0.045-0.386	---	4-82	5-95.3	---	---	---	---	---	---	---	---
2	Chowdhury et al., 2018	India	10	0.036-1.56	---	74-301	12.75-37.23	---	---	---	---	---	---	---	---
3	Roychowdhury et al., 2008b	India	8	0.045-0.386	---	2-82	5-95.3	---	---	---	---	---	---	---	---
4	Chowdhury et al., 2020	India	3	0.224-0.389	---	10-493	1.53-30.17	7.39 - 7.74	1.86 - 2.14	---	153-163	---	---	---	---
5	*Biswas et al., 2018	India	24	0.550	---	410	7.06	8.1	3.97	Silty Clay	---	14.99	6.24	---	---
6	*Bhattacharya et al., 2010a	India	18	0.160-0.230	---	530	3.34-4.6				---	---	---	---	---
		India	12	0.160-0.300	---	400	5.26-7.10				---	---	---	---	---
		India	12	0.230-0.400	---	420	7.03-9.72	7.66	0.72	Clay Loam	---	---	---	---	---
		India	12	0.240-0.580	---	400	5.31-5.82				---	---	---	---	---
		India	9	0.290-0.540	---	440	4.01-5.52				---	---	---	---	---
7	*Bhattacharya et al., 2010b	India		0.140-0.310	---	360-470	4.26-5.85	---	---	---	---	---	---	---	---
8	*Biswas et al., 2014	India	94	0.330	---	420	8.35	---	---	---	---	---	---	---	---
		India	78	0.230	---	350	6.17	---	---	---	---	---	---	---	---
9	Golui et al., 2017	India	13	0.002-1.26	---	180-570	0.196-2.33	8.06 - 8.12	0.45 - 0.61	---	---	---	---	---	0.76
10	Mukherjee et al., 2017	India	22	0.210-0.720	---	56-585	9.05-25.80	---	---	---	---	---	---	---	0.85
11	*Rahaman and Sinha, 2013	India	2	0.390-0.670	---	430-540	9.8-10.7	7.91 - 8.30	---	Silty Clay-Silty Loam	---	2.79-3.11	13-19	---	0.673
12	Sarkar et al., 2012	India	1	0.420-0.560	---	106-573	16.22-18.74	7.22	0.99	Silty Clay	---	---	32.85	---	---
13	Sinha and Bhattacharyya, 2014	India	1	0.103-0.141	---	320	2.38-3.03	---	---	---	---	---	25.23-37.04	---	---

14	Srivastava et al., 2015	India	58	0.179-0.932	---	0-312	3-35	---	2.52	Clay Loam-Clay	---	---	25.6	6.84	---
15	Talukder et al., 2011	Bangladesh	1	0.470	---	100	8.12	6.1	0.95	Sandy Clay Loam	---	68.36	5.47	2.36	---
16	*Dahal et al., 2008	Nepal	10	0.60-0.330	0.01	5-1014	6.1-16.7	8.0	---	---	---	---	---	---	0.68
17	*Hsu et al., 2012	Taiwan	1	0.290-0.660	---	26-67	11.8-112	5.6-6.5	---	Silty Clay Loam	---	5.85-13.1	---	---	---
18	Rahman et al., 2014	Bangladesh	2	0.290-0.650	---	25-419	9.12-11.23	6.8		---	---	---	---	---	---
19	Rahman et al., 2007	Bangladesh	6	0.600-0.700	---	70	14.5	7.1		Clay loam	---	---	6.8	---	---
20	Rahman et al., 2010	Bangladesh	44	0.230	---	87.30	13.0	---	---	---	---	---	---	---	---
21	*van Geen et al., 2006	Bangladesh	6	0.280-0.440	---	0-185	2.9-29	---	---	---	---	---	---	---	---
22	*Islam et al., 2017	Bangladesh	3	0.288-0.320	---	2.4-255.4	2.7-15.7	6.1-7.6	2.0-2.4	Clay-Silty Clay Loam	---	---	---	---	---
23	*Ahmed et al., 2011	Bangladesh	10	0.101-0.338	0.012-0.024	0-234	0.9-8.7	5.0-7.5	0.74 - 1.62	---	---	9.1-17.5	---	---	
24	Sharma et al.,2017	India	12	0.03-0.33	---	2.31-15.91	0.06-0.11	---	---	---	---	---	---	---	---
25	Reid et al., 2021	Vietnam	16	0.063-0.528	---	0-751	6-20	---	---	---	---	---	---	---	---
26	*Wang et al., 2019	China	5	0.039–0.084		5.5–9.1	---	---	---	---	---	---	---	---	---

* Reported As concentration for polished rice

4.2.2. Classification of data

The amount of inorganic As (i-As) in rice grain was computed using published data (18 research) from Asian countries, yielding a weighted mean of 80 percent for polished rice and 75 percent for husked rice, respectively (Table 4.2). In our meta-data, the total As (t-As) concentrations in rice grain were transformed to i-As. The rice grain concentrations were categorized into two groups: (i) "within the maximum tolerable concentration (≤ MTC)": As ≤ 350 µg kg^{-1} (husked rice) and ≤ 200 µg kg^{-1} (polished rice); (ii) "above the maximum tolerable concentration (>MTC)": > 350 µg kg^{-1} (husked rice) and > 200 µg kg^{-1} (polished rice) (based on the recommendation of JECFA, 2017). The whole data set was randomly split into two, 80% of the data were used as the training set and the remaining 20% formed the testing set (Mukherjee et al., 2021).

Table 4.2. Calculation of inorganic Arsenic (i-As) from total As in polished and husked rice

Sl.No.	Author and year of publication	Location	Sample Size	i-As (%)
Polished Rice				
1.	Mondal et al., 2021	India	29	98.4
2.	Williams et al., 2006	Bangladesh (different varities)	3	71
			3	66
			3	60
			3	83
			3	82
			3	81
			3	72
3.	Williams et al., 2005	India	15	81
		Thailand	12	74
		Taiwan	3	67
4.	Rahman et al., 2014	India	1	98
		Pakistan	1	96
5.	Torres-Escribano et al., 2008	Thailand	3	68
6.	Nookabkaew et al., 2013	Thailand (different locations)	29	61.45
			18	60
			22	63.25
			8	60.48
7.	Pal et al., 2009	India	1	95
8.	Islam et al., 2017	Bangladesh	10	92
9.	Roychowdhury, 2008a	India (different locations)	34	88.1
			38	90.4
10.	Roychowdhury et al., 2008b	India	18	90.3
11.	Chen et al., 2016	Taiwan (different locations)	15	87.8
			12	88.2
			4	82.8
			7	76.9
12.	Schoof et al., 1998	India	2	58
13.	Meharg et al., 2009	Bangladesh	15	61.53
			Weighted Average ± SD	**80.0 ± 13.13**
Husked Rice				
1.	Schoof et al., 1998	Taiwan	1	67
2.	Nookabkaew et al., 2013	Thailand (different locations)	19	57.54
			5	58
			9	52.8
			3	63.93
3.	Reid et al., 2021	Vietnam	45	84
4.	Sinha and Bhattachryya, 2014	India	4	85
5.	Chen et al., 2016	India (different locations)	2	96.5
			4	93
			7	81
			Weighted Average ± SD	**75±13.71**

4.2.3. Data Analysis

The logistic regression (LR) and decision tree (DT) algorithms were used to predict the binary variables: ≤ MTC and >MTC . The data analysis was performed using R-Studio (version 1.3.1093 2.3.1). Splitting the data into training and test data was performed using the stats (version 4.0.3) package. The Caret package (version 6.0-86) was used to conduct logistic regression and DT analysis (Kuhn, 2008). The probability graph from the logistic regression was prepared using *ggplot2* (version 3.3.3) and *tidyr* (version 1.1.3), and for ROC and AUC *pROC* (version 1.17.0.1) was used. The LR method uses only the statistically significant predicator variables in the model whereas DT uses the predictor variables in a hierarchical and recursive manner. DT have the flexibility of assigning the classes in one or more steps. One advantage of the LR is that it can be used to generate probabilities of class membership for each object whereas DT only generates average probabilities applicable to all the objects assigned to a particular group (Worth and Cronin, 2003).

4.2.4. Model limitations and assumptions

The two criteria used to assess the quality of a classification model are discrimination and calibration. Discrimination is a measure of how well the two classes in the data set are separated; calibration determines how accurate the model probability estimate is to predict the true probability (Dreiseitl and Ohno-Machado, 2002). To provide an unbiased estimate of a model's discrimination and calibration, these values should be calculated from a data set not used in the model building process. Usually, a portion of the original data set, called the test or validation set, is put aside for this purpose, since testing on a separate data set would, in an Ideal case, provide an unbiased estimation of generalisation error. In small data sets as in this study, there may not be enough data for both training and testing. For this reason, the total data set was split into training set and testing set and the training data set was used as the source of information. In this case, the whole data set was divided into k pieces, k-1 pieces are used for training, and the last piece was the test set. This process of k-fold cross-validation builds k models; the numbers reported are the averages over all k test sets (Stone, 1974; Allen, 1977). The problem of over fitting both in the logistic regression and the DT analysis was controlled by k-fold cross validation (k=10) of the training data (James et al., 2013). On the observations in the remaining fold, the number of misclassified observations was calculated.

This procedure was repeated, with each validation set consisting of a different set of observations (James et al., 2013). To quantify the extent to which the predicted response value for a given observation was close to the true response value for that observation, the receiver operating characteristic (ROC) curve was used. The overall performance of a classifier, summarized over all possible thresholds, was given by the area under the curve (AUC) (James et al., 2013).

4.3. Results

4.3.1. Relationship between rice grain As with soil and irrigation water As

Arsenic concentrations in rice grain, soil and irrigation water based on the meta data (n=134) are summarized in Table 4.3. The rice grain t-As concentration ranged from 18 to 1560 µg kg^{-1} with a mean value of 420 µg kg^{-1}. The As concentration in soil ranged from 0.06 to 112 mg kg^{-1} with a mean value of 11.74 mg kg^{-1} and the irrigation water As content ranged from 0 to 1014 µg L^{-1} with a mean value of 235.49 µg L^{-1}.

Table 4.3. Total As concentrations in rice grain, soil, and irrigation water (n=134)

Parameters	**Mean±SD**	**Median**	**Range (min- max)**	**IQR (Q_3-Q_1)**
Grain As (µg kg^{-1})	420±30	350	18-1560	540-210
Soil As (mg kg^{-1})	11.73±12.06	8.40	0.06-112	16.22-5.20
Irrigation water As (µg L^{-1})	235.49±215.48	192.00	0.0-1014	410.0-25.5

The rice grain As content was found to be positively and significantly correlated with the irrigation water As (spearman's rho= 0.46, $p<0.01$) and the soil As (spearman's rho= 0.65, $p<0.01$). The irrigation water As and soil As was also observed to have a significant positive correlation. between themselves (spearman's rho=0.32, $p<0.05$) as can be visualized from Figure 4.2. From the collated meta data, 12.68% had soil As concentrations above 20 mg kg^{-1} and 63.43 % of the data had irrigation water As above 100 µg L^{-1}; 54% of the polished rice grain i-As meta-data exceeded the concentration of 200 µg kg^{-1} and 74% of the husked rice grain i-As meta-data exceeded 350 µg kg^{-1}.

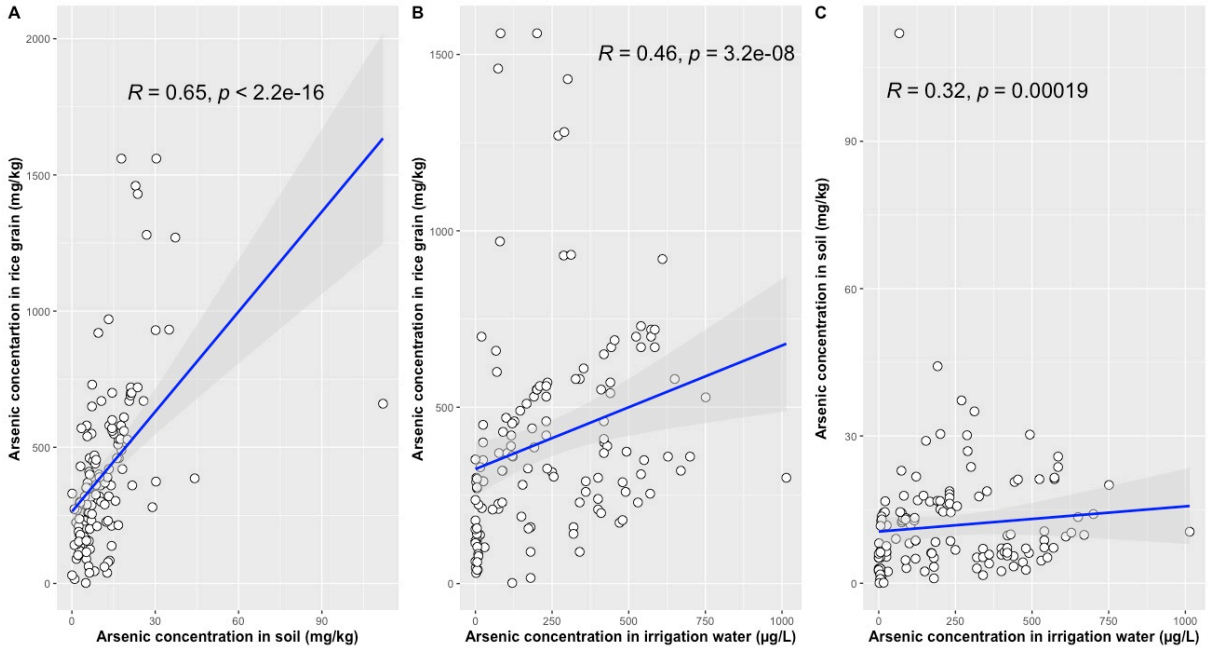


Figure 4.2. Spearman's correlation between (A) rice grain t-As and soil As, (B) rice grain t-As and irrigation water As and (C) irrigation water As and soil As (n=129).

4.3.2. Determination of the limit of As in soil and irrigation water

The prediction by the LR model was Probability (≤ MTC |> MTC) = -1.6822 + 0.1429 $Soil_{As}$ (AIC=123.68). The soil As coefficient significantly (p<0.01) explained the grain i-As content. When irrigation water As was added to the model, the coefficient was statistically non-significant (p>0.05) and AIC increased to 128.17. Soil As content was 11.75 mg kg^{-1} when probability (≤ MTC |> MTC) reached 50% as can be observed from Figure 4.3. Hence, 11.75 mg kg^{-1} may be considered as the limit of As in soil beyond which grain i-As content may exceed 200 µg kg^{-1} for polished rice and 350 µg kg^{-1} for husked rice. From DT it was observed that the soil As appeared as the primary splitting variable at 14 mg kg^{-1} as can be observed from Figure 4.4. When soil As was greater than 14 mg kg^{-1}, the probability of grain As being classed > MTC was 0.85 and 32% of the data was in this node. No further splitting of the tree and inclusion of irrigation water as a successful variable was observed. An attempt was made to predict the maximum concentration in irrigation water above which the soil As exceeded 11.75 mg kg^{-1} and 14 mg kg^{-1} using the LR and DT models respectively. With LR model the irrigation water As was observed to be non-significant and in case of DT the irrational splitting was observed which was also not suitable for pruning based on the complexity parameter.

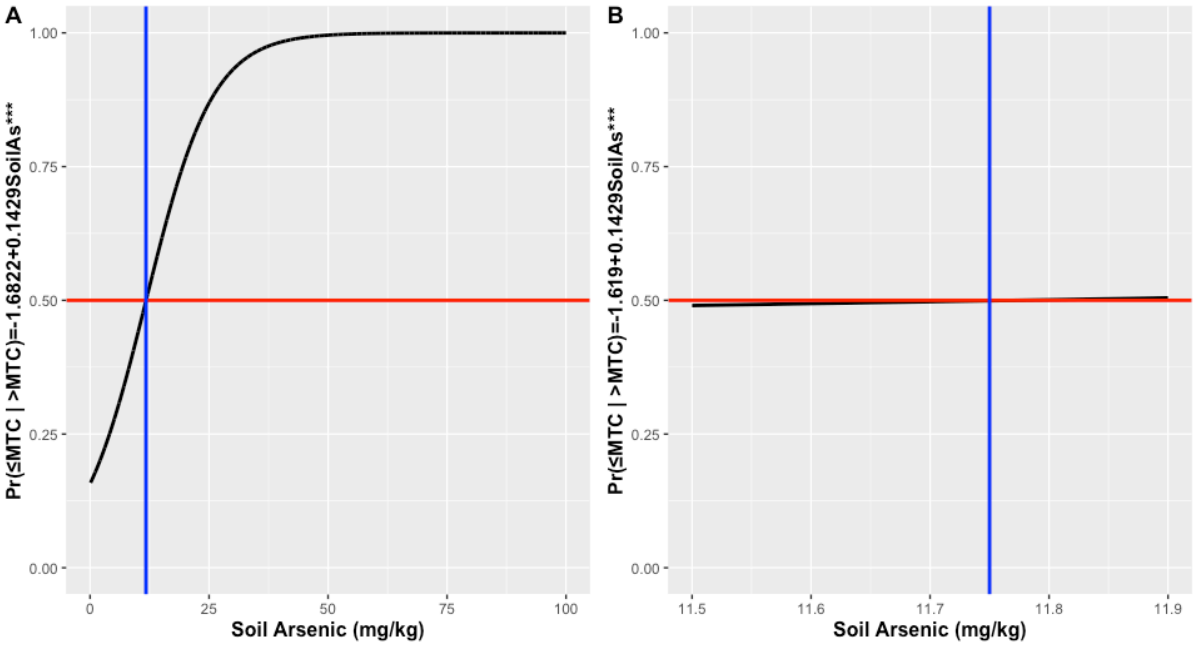


Figure 4.3. Limit of As in soil based on the probability plot of logistic regression with respect to soil As (A) and the cut-off point magnified (B).

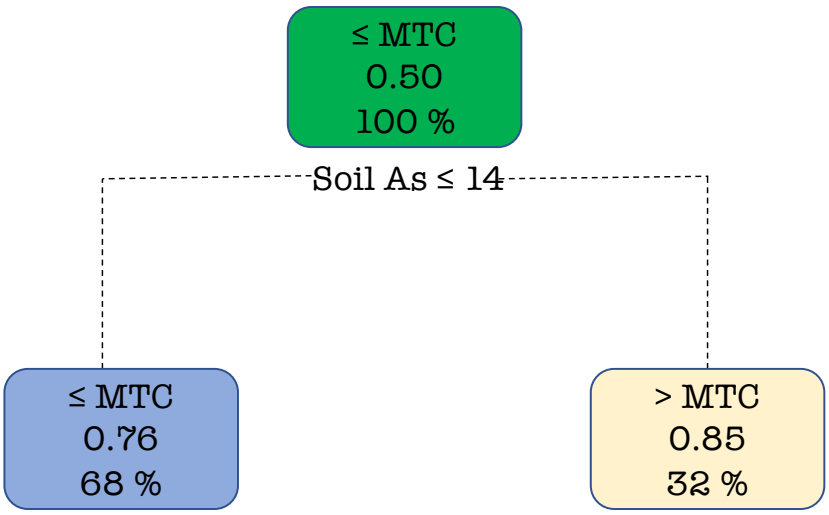


Figure 4.4. Decision Tree explaining the probability of the category (≤MTC or >MTC) of rice grain based on the As content in soil. Percentage of observations in the node and probability of class observations in the node are displayed.

4.3.3. Comparison between the two models

A comparison between the predictability models developed by Decision Tree (DT) and Logistic Regression (LR) over the training phase and the testing phase can be visualized from Table 4.4. The results revealed that the in terms of model accuracy and misclassification percentage DT have an edge over LR both over the training and the testing phase.

Table 4.4. Model performance over the training phase (n=108) and testing phase (n=26)

Model	Accuracy (%)	Misclassification (%)
Training Phase		
Decision Tree	73.15	26.85
Logistic Regression	65.74	34.26
Testing Phase		
Decision Tree	73.08	26.92
Logistic Regression	65.38	34.62

To evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions match the observed class. That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation/class. The ROC curve represents the overall performance of a classifier, summarized over all possible thresholds, is given by the AUC (James et al., 2013). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier. We expect a classifier that performs no better than chance to have an AUC of 0.5. The ROC plots confirmed that DT performed better than LR. During the training phase, DT achieved an AUC of 72.5% and LR 65.5% (Figure 4.5, A and B) and in the testing phase (Figure 4.5, C and D), the AUC for DT was 70.6% and for LR 65.5%.

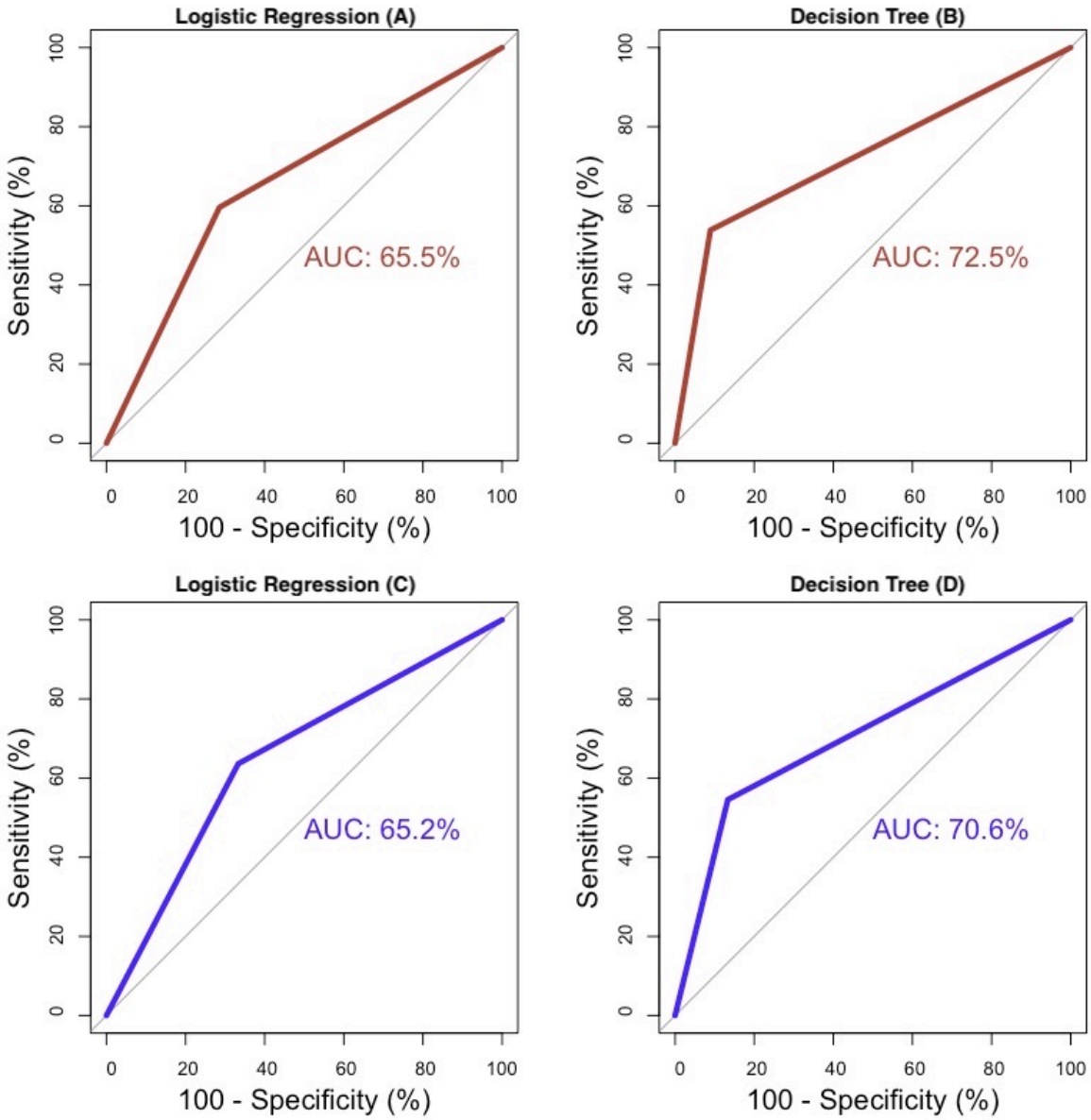


Figure 4.5. Sensitivity vs Specificity plot for Decision Tree and Logistic Regression over the training phase (A, B) and the testing phase (C, D).

4.5. Discussion

To our knowledge this is the first meta-analysis utilizing published data from Asian paddy fields on As in rice grain, soil and irrigation water to determine the relationship between them and to develop a model to estimate the maximum concentration in paddy soil and irrigation water above which Codex standards for the maximum allowable i-As in rice would be exceeded (JECFA, 2017). From the 156 papers reviewed, only 26 studies (15 from India, one from Taiwan, one from Nepal, one from Vietnam, one from China and 7 from Bangladesh) met the inclusion criteria for the meta-analysis; these studies all reported t-As concentrations in rice grain, soil, and irrigation water. There was near equal split between studies which reported t-As

concentrations in husked rice (15 studies) and those in polished rice (11 studies). After converting t-As to i-As, 43% for husked and 60% for polished rice concentrations exceeded the Codex standard.

In this meta-analysis, soil As was the main determining factor and this was confirmed by a) the stronger positive correlation between paddy soil and rice As concentrations compared to irrigation water and rice grain concentrations (Figure 4.2); b) the LR model having non-significant contribution of irrigation water As; and c) the DT model predictions taking only into account the soil As classified data. This aligns with previous studies where authors reported most significant impact of soil As (Sengupta et al., 2021) and 'modest if any' impact of irrigation water on t-As content of rice (van Geen et al., 2006). On the contrary, in a recent study, mean As concentrations in groundwater used for irrigation were strongly correlated with grain t-As (Reid et al., 2021). Regardless, many studies have suggested that soil As concentrations increase with contaminated groundwater irrigation of paddy fields (Huq et al., 2006; Panaullah et al., 2008; Dittmar et al., 2010), eventually resulting in increase of grain As (Rahman et al., 2007; Lu et al., 2009; Rahman et al., 2010). The non-significant influence of irrigation water on the grain As concentrations in this study was perhaps unsurprising given the complexity in the transfer of irrigation water As to rice grain via the soil. For example, the accumulation of As in soil from the irrigation water is dependent on several factors like the temporal variation in As concentration throughout the crop-growth period, the volume of irrigation water used, and the area of the field being irrigated (Chowdhury et al., 2018; Chowdhury et al., 2020). The complexity in the relationship between irrigation water As and grain As could be further enhanced due to irrigation practices which often include the use of both groundwater and rainwater. For example, the phase wise soil As movement and its enrichment pattern in rice due to the use of As contaminated irrigation water showed a moderate accumulation of soil As in the vegetative phase followed by a severe drop in the reproductive phase and continued buildup of As in the ripening phase (Chowdhury et al., 2018). Whereas, in rainfed rice cultivation, a moderate accumulation of As in soil in the vegetative phase followed by a rise in the reproductive phase and a decrease at the ripening stage mainly due to the dilution of the As accumulation in soil due to rainwater was noted (Chowdhury et al., 2020).

The maximum concentration of As in soil from LR model was found to be 11.75 mg kg^{-1} whereas, based on the better performing DT model the maximum concentration was 14 mg

kg^{-1} above which the As concentration in rice grain would exceed the Codex recommendation. This is in agreement with a) the study from Bangladesh, investigating the accumulation and distribution of As in rice grain, and reporting that the rice grown in soils contaminated with As of 14.5 ± 0.1 mg kg^{-1} was not safe for human consumption (Rahman et al., 2007); b) the recommendation of the maximum acceptable limit of As in soil of 20 mg kg-1 by the European Union (Rahman et al., 2007) and c) the limit of 15 mg kg^{-1} of As in paddy soils by Ministry of Environment, Government of Japan (Punshon et al., 2017). These findings suggest that, based on the scientific data currently available for rice cultivation in Asia, an As guideline value of 14 mg kg^{-1} in paddy soil may be appropriate.

4.6. Conclusion

This meta-analysis determined that the concentrations of 14 mg kg^{-1} in paddy soil may be an appropriate guideline value above which rice grains cultivated in Asian paddy fields will exceed the Codex recommended maximum allowable concentrations of 200 µg kg^{-1} for i-As in polished rice and 350 µg kg^{-1} in husked rice. Both LR and DT models predicted that soil As was the main determining factor. A guideline value for the irrigation water could not be derived using either the LR or DT models and warrants further investigation. The non-significant contribution of irrigation water was unsurprising, given that the As accumulation in soil due to contaminated irrigation water depends on several factors and the relationships governing transfer to rice grains are complex. Considering uncertainties and limitations of the available meta data and models, experimental studies collecting more appropriate soil and rice grain samples, and analysis of soil bio-accessible/available As, along with soil parameters (pH, organic carbon, available iron, available phosphorus) rather than only total As in paddy soil, is warranted to validate these findings.

Chapter 5

Assessing the predictability of the logistic regression and decision tree models over field data

The contents of this chapter have been published as: **Mandal, J**., Jain, V., Sengupta, S., Rahman, M.A., Bhattacharyya, K., Rahman, M.M., Golui, D., Wood, M.D. and Mondal, D. (2023). Determination of bioavailable arsenic threshold and validation of modelled permissible total arsenic in paddy soil using machine learning. *Journal of Environmental Quality*. https://doi.org/10.1002/jeq2.20452

5.1. Introduction

Using a meta-analysis approach Mandal et al. (2021) predicted the soil As concentrations above which rice grains cultivated in Asian paddy fields may exceed the Codex maximum tolerable concentrations (MTC) of 200 µg kg^{-1} for inorganic As in polished rice and 350 µg kg^{-1} in husked rice. The maximum concentration of As in soil from the logistic regression (LR) model was found to be 11.75 mg kg^{-1} whereas, based on the better performing decision tree (DT) model the maximum concentration was 14 mg kg^{-1} above which the As concentration in rice grain would exceed the MTC. 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 will lead towards a complete and comprehensive review for both the models published previously (Mandal et al., 2021). 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 datasets: purposely collected field-data from different rice cultivation practices, rainfed and groundwater-irrigation from As-contaminated sites of West Bengal, India.

5.2. Materials and Methods

Testing machine learning models with different datasets is crucial for assessing their generalization, performance, detecting issues like overfitting or underfitting, comparing models, and monitoring their long-term performance. It helps ensure that the models are reliable, accurate, and capable of making accurate predictions on new, unseen data (James et al.,2013). Three individual test data sets comprised of paired rice grain and soils total As concentrations were used for the purpose. The test sets were selected in a way that there is a difference in terms of site (agro-climatic zones) and system of rice cultivation (rainfed and

irrigated). The test set 1 (n=101) was collected from three As-contaminated districts (Nadia, Murshidabad and N-24 Parganas) of West Bengal, India under rainfed rice system. The test set 2 (n=28) and test set 3 (n=132) were collected from Maldah and Nadia districts of West Bengal, India respectively from irrigated rice system. All these sites from where the samples for test sets were collected are predominantly considered as the severely As contaminated belt of West Bengal, India. The Maldah district falls under the Vindhyan alluvial (old alluvial zone) agro-climatic zone. Nadia and Murshidabad distract falls under the Gangetic alluvial (new alluvial zone) of West Bengal (Mandal et al., 2022). By testing the model with different datasets, it can be identified whether the model suffers from overfitting or underfitting issues. This will help to determine if the model needs adjustments such as regularization techniques, architectural changes, or more data for training. Testing different machine learning models with diverse datasets allows to compare their performance and choose the most suitable one for the task. By evaluating multiple models on the same test datasets, one can objectively assess their predictive abilities, accuracy, robustness, and other relevant metrics. This facilitates informed decision-making and model selection (James et al., 2013). The total As analysis of the soil and rice grain samples were analysed following the protocols as outlined by Rahman et al. (2009 a, b) using ICP-MS (PerkinElmer NexION 350, USA).

The grain As content was converted to categorical variables (<MTC and >MTC) as per the methods outlined (75% of the total As in husked rice) in Mandal et al., (2021). The model testing was performed using R-Studio (*version 1.3.1093 2.3.1*). The "*caret*" package (*version 6.0–86*) was used to conduct prediction with logistic regression and decision tree models.

5.2.1. Model performance parameters

The evaluation of a model's performance involves the analysis of a single confusion matrix, which consists of four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). It is important to assess the relationships between these categories, rather than evaluating them individually, to accurately evaluate the model's performance. The model performance parameters include accuracy, sensitivity or recall, specificity or true negative rate, and precision or positive predictive value. The F1 score and the Mathews correlation coefficient (MCC) were also calculated.

The Area Under the Curve (AUC) is calculated by plotting the receiver operating characteristic (ROC) curve, which plots sensitivity against specificity at different classification

thresholds. The "*ROCR" (version 1.0-11)* and the "*pROC*" *(version 1.17.0.1)* were used to plot the ROC curves over the training and testing phase.

5.3. Results

5.3.1. Test Data Sets

Table 5.1 displays the total concentrations of As in rice grain and soil samples across the three testing sets. Positive correlations were observed between soil As and grain As in all three testing sets (test set 1: $r=0.34$, $p<0.01$; test set 2: $r=0.60$, $p<0.01$; test set 3: $r=0.46$, $p<0.01$), as depicted in Figure 5.1. After converting grain As concentrations into categorical variables (<MTC and >MTC), the sample counts were as follows: test set 1 - <MTC: 86, >MTC: 15; test set 2 - <MTC: 25, >MTC: 3; test set 3 - <MTC: 112, >MTC: 20.

Table 5.1. Total As concentrations in rice grain and soil in the testing sets

Parameters	**Mean±SD**	**Median**	**Range (min- max)**	**IQR (Q_3-Q_1)**
Test Set 1 (n=101)				
Grain As (µg kg^{-1})	294.54±162.41	255.40	84.38-870.85	360.01-175.35
Soil As (mg kg^{-1})	13.08±7.40	10.85	2.01-36.13	15.77-8.19
Test Set 2 (n= 28)				
Grain As (µg kg^{-1})	257.49±262.02	222.0	10.75-1260	343.25-102.00
Soil As (mg kg^{-1})	7.43±4.02	8.02	0.55-15.00	10.38-4.79
Test Set 3 (n=132)				
Grain As (µg kg^{-1})	327.70±139.25	330.0	10-650.21	412.50-255.00
Soil As (mg kg^{-1})	11.65±3.63	13.15	3.98-17.92	14.24-9.24

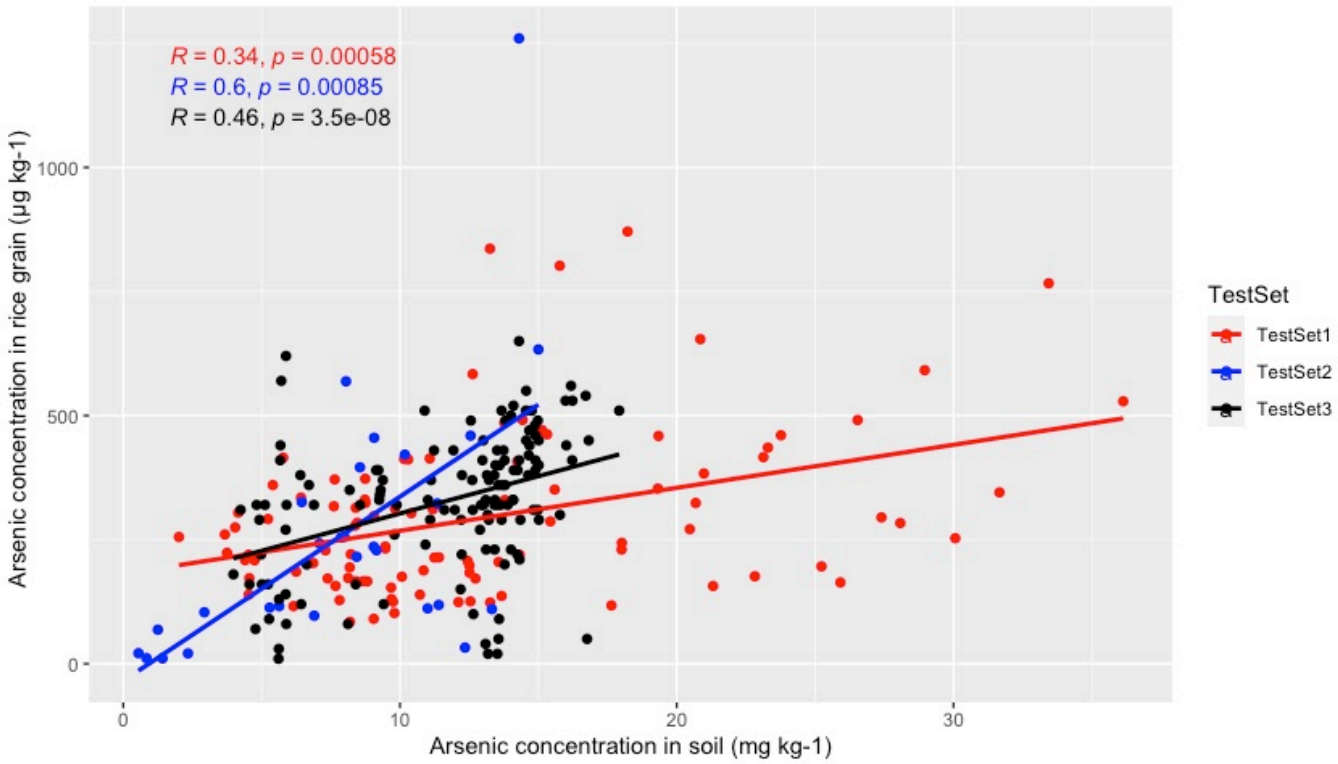


Figure 5.1. Spearman's correlation between grain As and soil total As for all the test sets ((Test Set 1, n=101), (Test Set 2, n=28) and (Test Set 3, n=132)).

5.3.2. *Confusion matrix and model parameters from testing of LR and DT*

The TP, FP, TN, and FN values for the three sets using LR and DT can be observed in Table 5.2. Model performance metrics for LR and DT are presented in Table 5.3. The accuracy of the LR model was 69.13%, 85.71%, and 46.21% for test sets 1, 2, and 3, respectively. For DT, the accuracy was 78.22%, 92.86%, and 80.30% for the corresponding test sets. DT demonstrated lower misclassification compared to LR. The F1 scores for the LR model were 78.01%, 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 85.89% for test sets 3 and 1, respectively. The MCC values for LR were 0.43, 0.40, and 0.18, and for DT, they were 0.45, 0.54, and 0.41 for test sets 1, 2, and 3, respectively.

Figure 5.2 presents a boxplot comparing the three test data sets in terms of grain category (As < MTC and As > MTC) with respect to soil As, along with the limits predicted by the LR and DT models. Figure 5.3 depicts a comparison of the receiver operating characteristic (ROC) curves for the two models across the three test sets.

Table 5.2. Confusion matrix of the testing data sets

Model: LR			**Model: DT**	
Test Set 1	*Actual*		*Actual*	
Predicted	**<MTC**	**>MTC**	**<MTC**	**>MTC**
<MTC	55 (TP)	1 (FP)	67 (TP)	3 (FP)
>MTC	31(FN)	14 (TN)	19 (FN)	12 (TN)
Test Set 2				
<MTC	22 (TP)	1 (FP)	25 (TP)	2 (FP)
>MTC	3 (FN)	2 (TN)	0 (FN)	1 (TN)
Test Set 3				
<MTC	44 (TP)	3 (FP)	95 (TP)	9 (FP)
>MTC	68 (FN)	17 (TN)	17 (FN)	11 (TN)

Table 5.3. Model parameters over the testing phase

Test Set	Parameters	Model	
		Logistic Regression (LR)	**Decision Tree (DT)**
Test Set 1 (n=101)	Accuracy (%)	69.13	78.22
	Misclassification (%)	30.69	21.78
	Sensitivity (%)	63.95	100
	Specificity (%)	93.33	80.00
	Kappa	0.345	0.402
	AUC	78.6	79.0
	Precision	100	95.71
	Recall	63.95	77.90
	F1 Score	78.01	85.89
	MCC	0.43	0.45
Test Set 2 (n=28)	Accuracy (%)	85.71	92.86
	Misclassification (%)	14.29	7.14
	Sensitivity (%)	88.00	100
	Specificity (%)	66.67	33.33
	Kappa	0.422	0.471
	AUC	77.30	66.70
	Precision	95.65	92.59
	Recall	88.00	100
	F1 Score	91.66	96.15
	MCC	0.40	0.54
Test Set 3 (n=132)	Accuracy (%)	46.21	80.30
	Misclassification (%)	53.79	19.70
	Sensitivity (%)	39.29	84.82
	Specificity (%)	85.00	55.00
	Kappa	0.104	0.342
	AUC	62.10	69.90
	Precision	93.62	91.35
	Recall	33.33	84.82
	F1 Score	49.15	87.96
	MCC	0.18	0.41

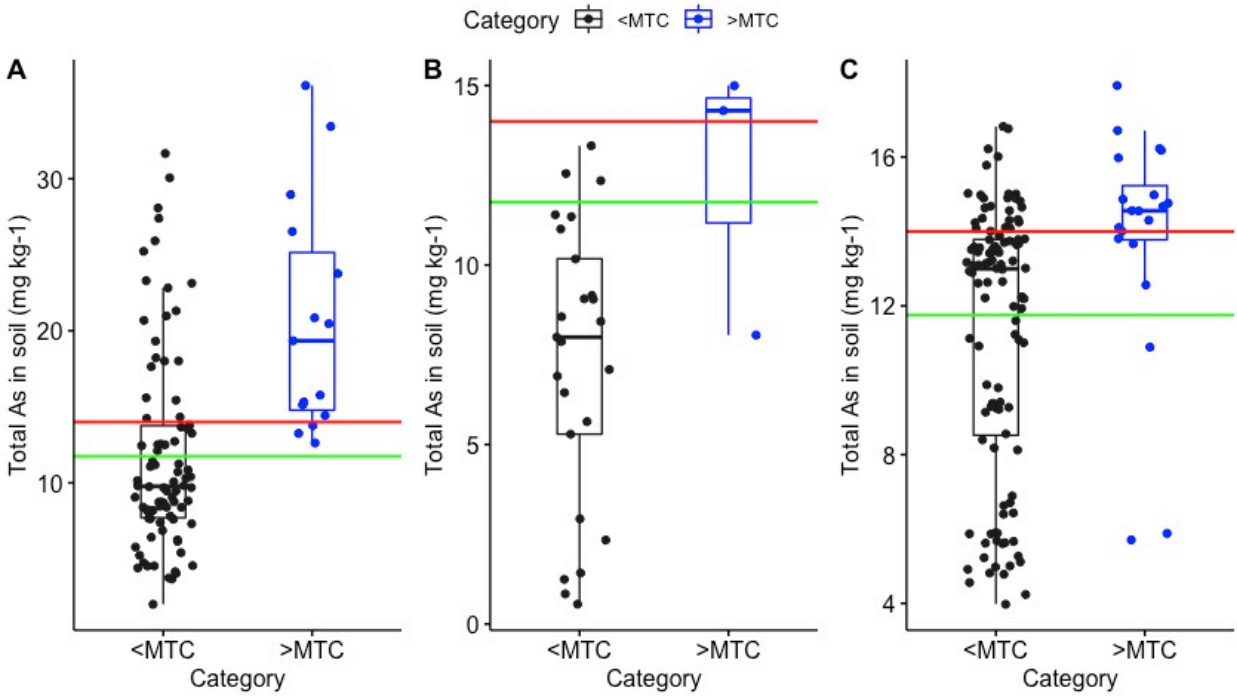


Figure 5.2. Boxplots of total As in soil (mg kg^{-1}) with respect to category of grain As concentration (<MTC and >MTC) of three testing data sets. (A: Test Set 1 (n=101), B: Test Set 2 (n= 28), C: Test Set 3 (n=132)). The horizontal red line indicates the limit of soil As (14 mg kg^{-1}) predicted by Decision Tree and the green line indicates the limit of soil As (11.75 mg kg^{-1}) predicted by Logistic Regression.

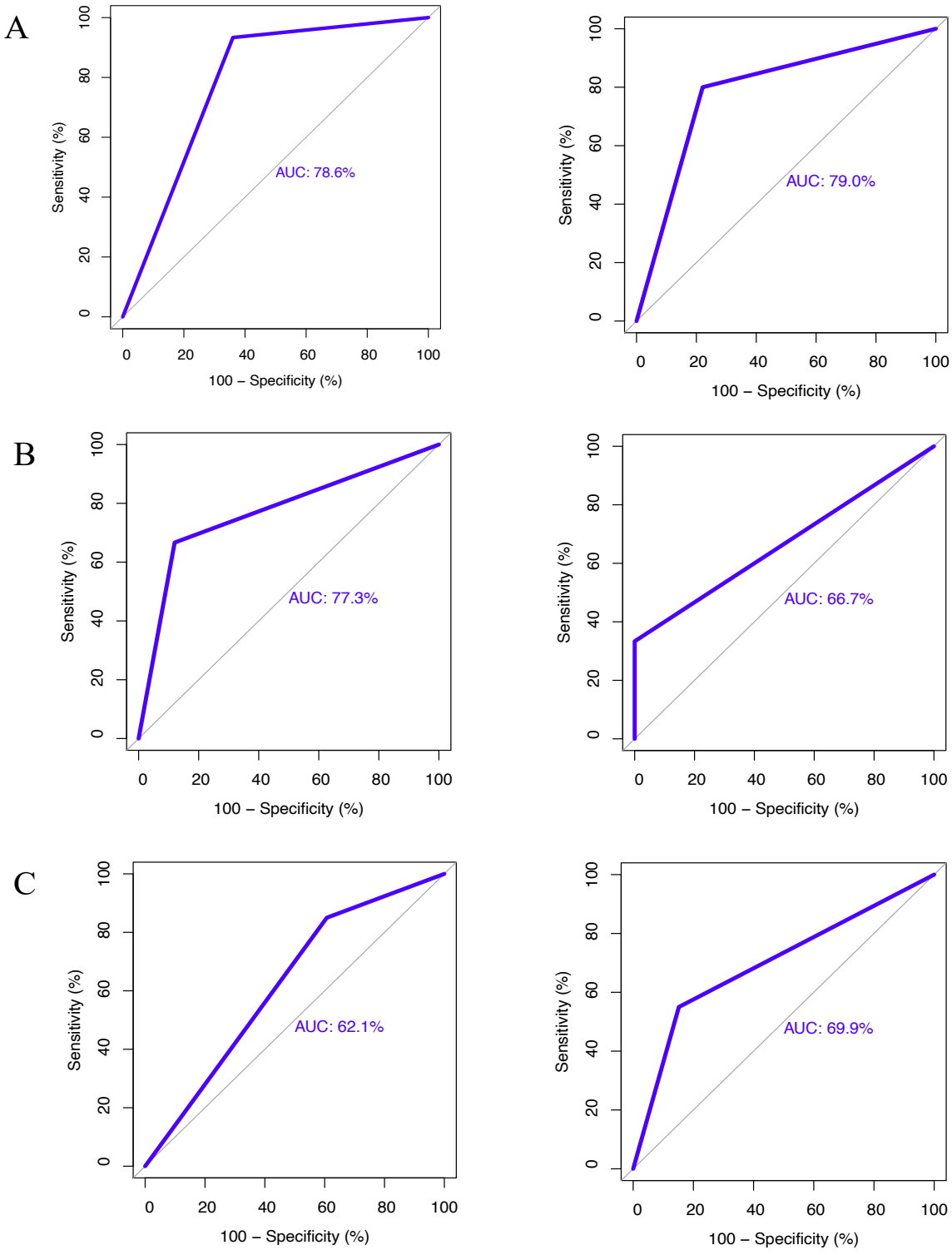


Figure 5.3. Sensitivity vs. specificity plot for logistic regression and decision tree over Test set 1 (A), 2 (B) and 3 (C).

5.4. Discussion

From the performance metrics of the LR and DT models, it was observed that DT outperformed LR in terms of accuracy for all three test sets, while LR performed better in terms of misclassification. The ratio between the number of correctly classified samples and the total number of samples, as suggested by Wang et al. (2007), is considered the most appropriate performance metric. However, when the dataset is unbalanced, as in our case where the number of samples in one class is significantly greater than the others, accuracy alone becomes unreliable. It provides an overly optimistic estimate of the classifier's skill on the majority class, as pointed out by Sokolova et al. (2006) and Akosa (2017). In our case, for test set 1 (<MTC = 86, >MTC = 15), test set 2 (<MTC = 25, >MTC = 3), and test set 3 (<MTC = 112, >MTC = 20).

Logistic regression showed higher specificity compared to DT for all test sets, while DT exhibited higher sensitivity compared to LR. Taking advantage of class-wise rates such as true positive rate (sensitivity/recall) and true negative rate (specificity), alternative measures like ROC and AUC can be derived. In our case, DT had a higher AUC than LR in two cases (test set 1 and 3) and lower AUC than LR in test set 2. However, it is worth noting that ROC and AUC have some flaws, as highlighted by Lobo et al. (2008), and are sensitive to class imbalance, as noted by Hanczar et al. (2010).

Therefore, considering F1 score and MCC as important model metrics to determine efficacy, the DT model outperformed LR. The F1 measure is widely used across various machine learning applications, not only in binary scenarios but also in multiclass cases (Pillai et al., 2017). MCC generates a high score only if the classifier correctly predicts most positive and negative instances, considering the overall dataset. Dubey and Tarar (2018) also support the use of MCC and F1 as these measures provide more realistic estimates of real-world model performance.

In binary classification tasks, accuracy and F1 score derived from confusion matrices have been widely used. However, for unbalanced datasets, these statistical techniques can produce dangerously overoptimistic outcomes since they fail to consider the ratio between positive and negative elements (Chicco et al., 2021). The MCC criterion is intuitive and straightforward: a high-quality score is achieved when the classifier makes correct predictions for both negative and positive cases, regardless of their ratios in the overall dataset. Therefore,

considering MCC, the DT model outperformed the LR model and can be considered a better classifier.

The boxplot in Figure 5.2 shows a comparison between the three test data sets in terms of category of grain As (< MTC and >MTC) with respect to soil As and the limits predicted by the LR and DT model. The blue points below the red line (representing 14 mg kg-1 of total soil As from DT) represents 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 bioavailable 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 like pH, OC, available Fe and P which leads to the next part of the thesis.

The redox potential pH of the soil influences the bioavailability of As to the crops. The pH of the soil plays a major role in determining As speciation and solubility and the level of As in soil rises significantly with declining Eh and rising pH (Majumdar & Sanyal, 2003). Both the release of As into accessible forms and retention of As depend heavily on the oxidizable organic carbon. Humic acid (HA) and fulvic acid (FA) are two examples of soil organic components that function as efficient As-humate complexes (chelates) with varying degrees of stability (Mandal et al., 2019; Kumar et al., 2021 and Sengupta et al., 2022). Arsenic mobilisation is greatly influenced by P (both being Group Vb elements in the Periodic table) and how they interact in the soil-plant system. In fact, both As and P would compete for the same kind of adsorption sites rather than being adsorbed separately in mixtures (Raj et al., 2021). Arsenic is adsorbed in soils and sediments by oxides (such as iron (Fe), and manganese (Mn)) with the creation of inner-sphere complexes via the ligand exchange process, (Kumari et al., 2021) and hence affecting the bioavailability of As.

5.5 Conclusion

Based on the model metrics the DT model has an edge over the LR model and hence 14 mg kg^{-1} of total As in paddy soil may be an appropriate guideline value above which rice grains cultivated in paddy fields will exceed the Codex recommended maximum allowable concentrations of 200 µg kg^{-1} for i-As in polished rice and 350 µg kg^{-1} in husked rice. However, it would be naive to believe that our models could be applicable to all contaminated rice growing sites worldwide, as the models have been trained with a particular set of data from a specific geographical region. A procedure for creating predictability models for other

contaminated locations throughout the world can be developed using this technique of estimating the limit for soil total As.

Developing a procedure for creating predictability models for other contaminated locations worldwide based on the research findings can have both time and resource implications. The feasibility of this endeavour depends on several factors: data collection, model training and validation, model refinement and iteration followed by implementation and deployment of the models. Considering these factors, developing predictability models for other contaminated locations worldwide based on the mentioned research findings can be a resource-intensive and time-consuming process but it is not an impossible task. It requires a collaborative effort involving domain experts, data scientists, and adequate funding for data collection, model development, and validation. Additionally, the availability of comprehensive and diverse datasets from different contaminated locations plays a crucial role in determining the feasibility and accuracy of the models. While the procedure described in the research offers a promising approach, it is important to acknowledge the practical challenges and resource implications associated with its implementation. Flexibility, adaptability, and continuous refinement of the models are key considerations to ensure their effectiveness across different geographical regions with varying soil and environmental characteristics.

Chapter 6

Determination of bioavailable arsenic threshold by random forest, gradient boosting machine and logistic regression

The contents of this chapter have been published as: **Mandal, J**., Jain, V., Sengupta, S., Rahman, M.A., Bhattacharyya, K., Rahman, M.M., Golui, D., Wood, M.D. and Mondal, D. (2023). Determination of bioavailable arsenic threshold and validation of modelled permissible total arsenic in paddy soil using machine learning. *Journal of Environmental Quality*. https://doi.org/10.1002/jeq2.20452

6.1. Introduction

Heavy metals, such as As in soil is present in both the solution and solid phases. Arsenic can be present as free ions and organic and inorganic complexes (present in soil solution) or adsorbed ions and compounds (clay and organic colloids) or as bound to secondary minerals and precipitated oxides of Fe and Mn, carbonates, and phosphates or complexed with organic matter (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 their source but also with the soil properties. The mostly 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 key indicators to estimate the dissolution behaviour of As derived from the geochemical fractions in soils (Bari et al., 2021).

The bioavailability of As in soil is governed primarily by pH, organic carbon (OC), available phosphorus (P), and available iron (Fe) in 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 Bengal delta in the groundwater

as well as soil and phosphate-based fertilizers are vastly used in rice planting which may impact As bioavailability and hence these two parameters may provide an useful insight within the modelling 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.

In this study, we modelled the maximum tolerable available As concentrations of paddy soil above which rice grain As may exceed maximum tolerable concentration (MTC) as per the Codex recommendation using the collected field data. We predicted the threshold for bioavailable As and also investigated the behaviour 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.

6.2. Materials and Methods

6.2.1. Collection and analysis of soil and rice grain samples

Paired soil and rice grain samples collected from both the irrigated and rainfed rice system (n=233) of Nadia, Murshidabad and N-24 Parganas district of West Bengal, India. The total As (TAs) analysis of the soil and rice grain samples were done following the protocols as outlined by Rahman et al. (2009 a,b) using ICP-MS (PerkinElmer NexION 350, USA). For bioaccessible As (AvAs), 0.43 M HNO_3 was used as an extracting agent followed by analysis in ICP-OES (Agilent 720) (Bari et al., 2021). Other soil parameters were done as per the following methods: pH (soil:water 1:2.5, Jackson 1973), organic carbon (OC) (oxidation with potassium dichromate ($K_2Cr_2O_7$), Walkley and Black 1934), available P (AvP) (extraction with sodium bi-carbonate ($NaHCO_3$), Olsen et al. 1954), and available Fe (AvFe) (extraction with diethylene tri-amine penta-acetic acid (DTPA), Lindsay and Norvel, 1978).

6.2.2. Predicting grain As with RF, GBM and LR

For predicting rice grain As alongside impact of other soil parameters with the RF, GBM and LR models, a compiled data set (n=233) of both irrigated and rainfed rice was used. The whole data set was randomly split into two, 80% of the data were used as the training set and the remaining 20% formed the testing set. After this the test set was kept aside and the train set was subjected to repeated cross- validation. Each model was trained through the procedure

of k-fold repeated cross-validation (k=10 and repeats =5). The remaining 20% of the test data was used for mode testing. 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) as the predictor variables. Basically, the training set was used to generate multiple splits of the training and validation sets to reduce over fitting of the model. The "*caret*" package (*version 6.0–86*) was used to train the model with 10-fold cross-validation repeated 5 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 6.1). 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 (in this case it is RF model) (Friedman, 2001). The equivalent to a PDP for individual data instances is ICE plot (Goldstein et al., 2017). An ICE plot visualizes the dependence of the prediction on a variable for each instance separately, resulting in one line per instance, compared to one line overall in partial dependence plots. 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 like multicollinearity among the variables (Sarkar et al., 2022). For LR model 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 rises the variances of parameter estimates and thus leads to inaccurate conclusions about the relationship between dependent and independent variables (Midi et al., 2010). Variance Inflation Factor measures the severity of 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 VIF 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.

6.2.3. Model performance parameters

The evaluation of a mode''s performance involves the analysis of a single confusion matrix, which consists of four categories: true positive (TP), true negative (TN), false positive (FP), and

false negative (FN). It is important to assess the relationships between these categories, rather than evaluating them individually, to accurately evaluate the mode''s performance. The model performance parameters include accuracy, sensitivity or recall, specificity or true negative rate, and precision or positive predictive value. The F1 score and the Mathews correlation coefficient (MCC) were also calculated.

The Area Under the Curve (AUC) is calculated by plotting the receiver operating characteristic (ROC) curve, which plots sensitivity against specificity at different classification thresholds. The *"ROCR" (version 1.0-11)* and the "pROC" *(version 1.17.0.1)* were used to plot the ROC curves over the training and testing phase.

6.3. Results

6.3.1. Confusion matrix and performance of RF, GBM and LR models

The performance of the RF and LR models over the testing and training phase can be observed in Table 6.1. Over the training set both the RF and GBM model performed better compared to the LR model. The accuracy was 100 for both RF and GBM model and 91.15 for LR model for training phase. The sensitivity, specificity, precision, F1 score was 100 and Kappa, MCC was 1.0 for RF and GBM model over the training set. The sensitivity, specificity, precision, F1 score were 98.19, 46.15, 92.09 and Kappa, MCC were 0.54 and 0.56 for LR model over the training set. For the testing set the sensitivity, specificity, precision, F1 score were 94.29, 50, 91.67, 92.95 respectively for both the RF and GBM. The Kappa and MCC were 0.48 and 0.47 respectively for both RF and GBM over testing phase. For LR model the sensitivity, specificity, precision, F1 score were 97.14, 50, 91.89, 94.34 respectively over the testing phase. Over the testing set the Kappa and MCC were 0.54 and 0.56 respectively for LR model. From the ROC at Figure 6.2 for RF the AUC was 100% for training set and 89.0% for testing set. In case of LR the AUC for training set was 89.60% and for testing set it was 85.2%. Although the AUC of RF>LR >GBM but the accuracy and MCC of LR> RF $\cong$ GBM. Besides 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. In terms of model performance metrics over the testing phase LR was better than RF and GBM in terms of accuracy, sensitivity, specificity, kappa, precision, log loss, F1score and MCC.

Table 6.1. Confusion matrix of RF, GBM and LR model and model parameters over training and testing phase

Training set (n=192)			**Testing set (n=41)**	
Random Forest (RF)				
	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 (GBM)				
	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 (LR)				
	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	

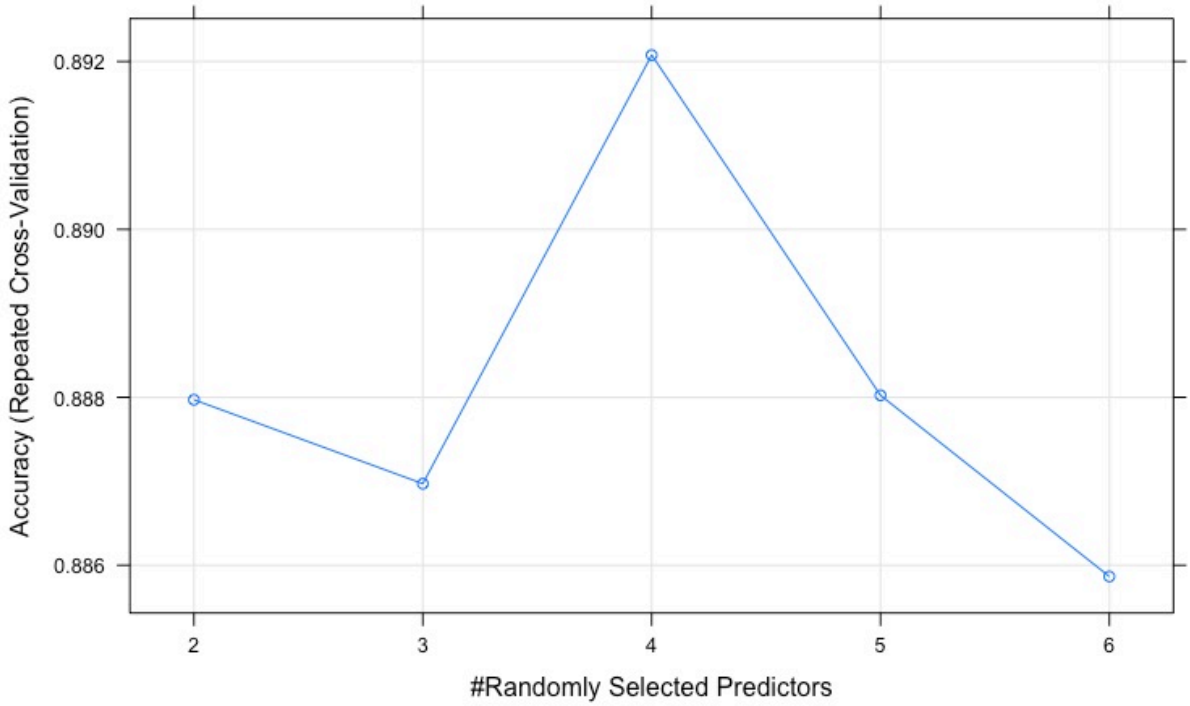


Figure 6.1. Accuracy from repeated cross-validation with randomly selected parameters plot of RF model

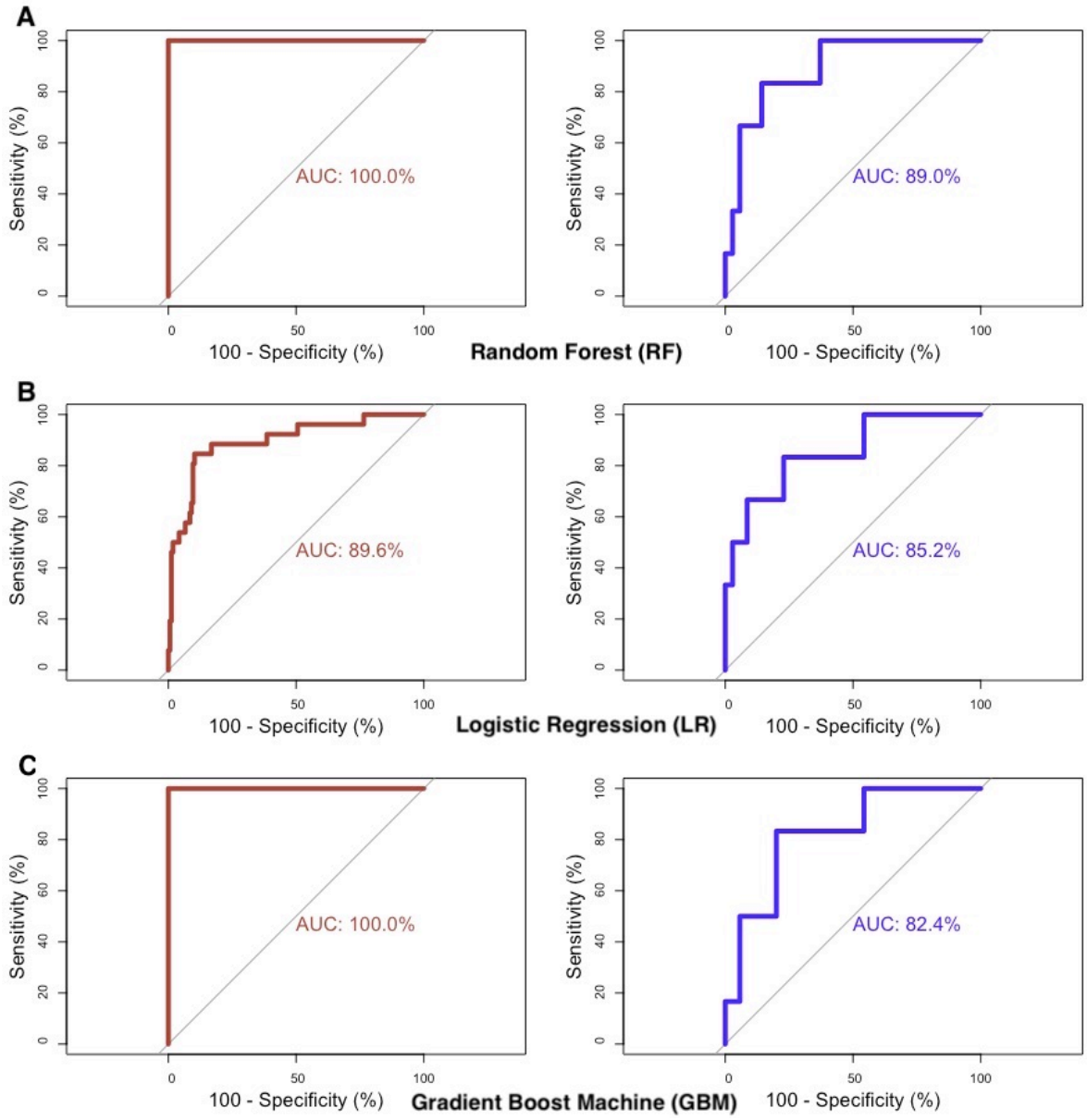
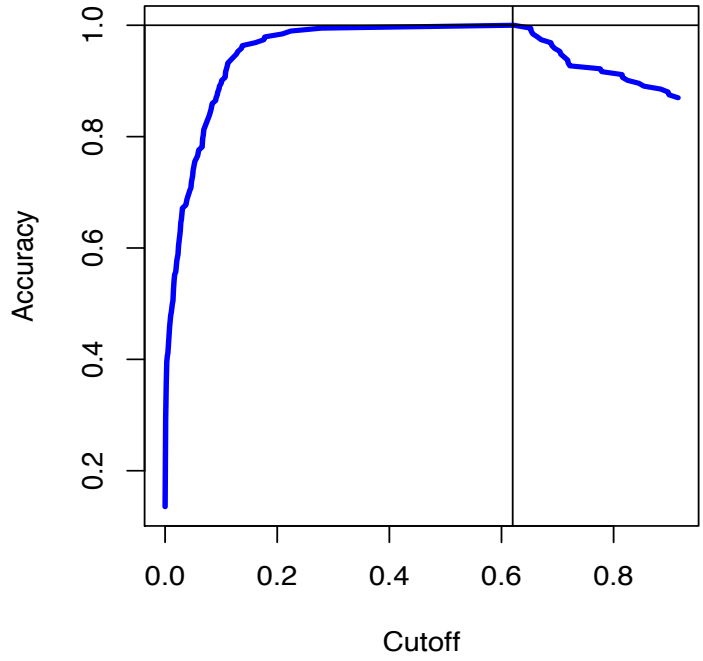


Figure 6.2. Sensitivity vs. specificity plot for random forest (A) and logistic regression model (B) and gradient boost machine (C) over the training and testing phase.

A



B

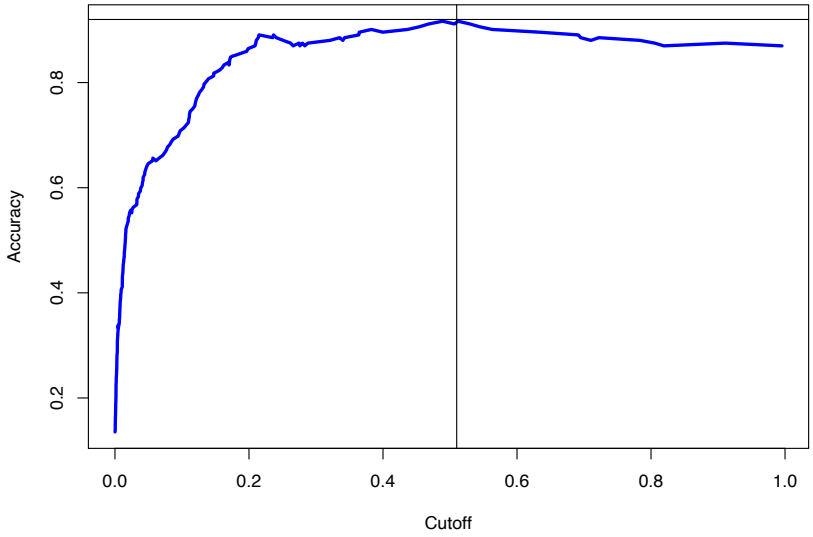


Figure 6.3. Cut-Off or threshold probability for RF at 0.62 (A) and for LR at 0.51(B) with respect to maximum accuracy.

6.3.2. Variable importance and partial dependence of the variables from better performing RF and LR model

The LR model predicted Probability (≤MTC |> MTC) = -16.82 + 4.79OC + 0.48AvFe + 1.22BAs + 0.15TAs (AIC=106.92). The OC, AvFe, BAs and TAs coefficient significantly ($p < 0.05$) explained the grain content. When AvP and pH were considered for the model, the coefficient was statistically non-significant ($p > 0.05$), 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 of Figure 6.4 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 did not come up as an important predictor variable. The importance of the BAs was 100 followed by TAs (41.75), AvFe (25.77), OC (2.52) and AvP (0.84). From the ICE (A) and the PDP (B and C) plots of the RF model at Figure 6.5 to AvAs it can be observed that the probability of class <MTC decreases when the soil bioavailable As increases. At the cut off probability of 0.62 at the highest accuracy from the RF model (Figure 6.3) it was observed that the limit of soil As to classify grain As (<MTC) was 5.72 mg kg^{-1} (Figure 5.5) above which the probability of <MTC decreases. Similarly, from the ICE (A) and the PDP (B and C) of the LR model at Figure 6.6 it can be observed that at cut-off probability of 0.51 (Figure 6.3) the limit of soil As to classify grain As (<MTC) was 5.70 mg kg^{-1}. The PDP at Figure 6.7 and 6.8 from RF and LR model shows the probability of grain As category (<MTC) with respect to BAs (most important variable) with other important variables (TAs, AvFe, OC and AvP) for RF and significant variable for LR (TAs, AvFe, OC). It can be observed that at BAs less than 5.70 mg kg^{-1} from LR and 5.72 mg kg^{-1} from RF and TAs less than 14 mg kg^{-1} (predicted from the DT model) the probability of <MTC was maximum (1.0-0.8). At BAs of 5.72 mg kg-1 from RF and 5.70 mg kg^{-1} from LR the AvFe between 12-14 mg kg^{-1} was observed to be effective in keeping the probability of <MTC higher. The OC content between 0.6-0.8% was effective in keeping higher probability of grain As <MTC at BAs of 5.72 mg kg^{-1} from RF and 5.70 mg kg^{-1} from LR. For available P it was observed that at BAs above 5.72 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 formulating high probability of <MTC.

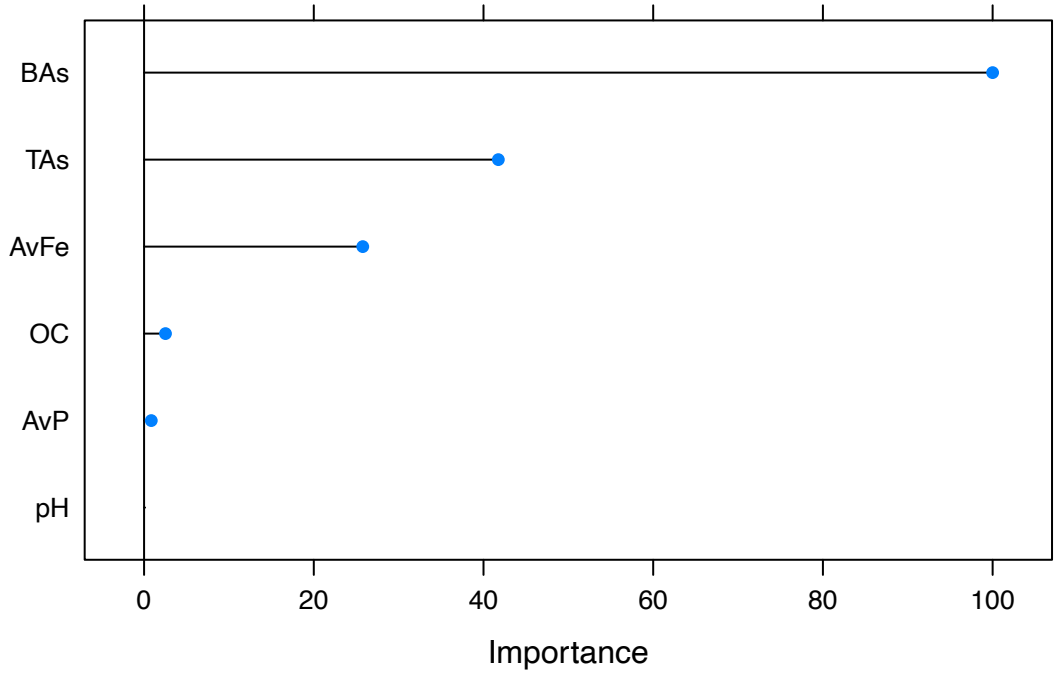


Figure 6.4. Variable importance plot from random forest model.

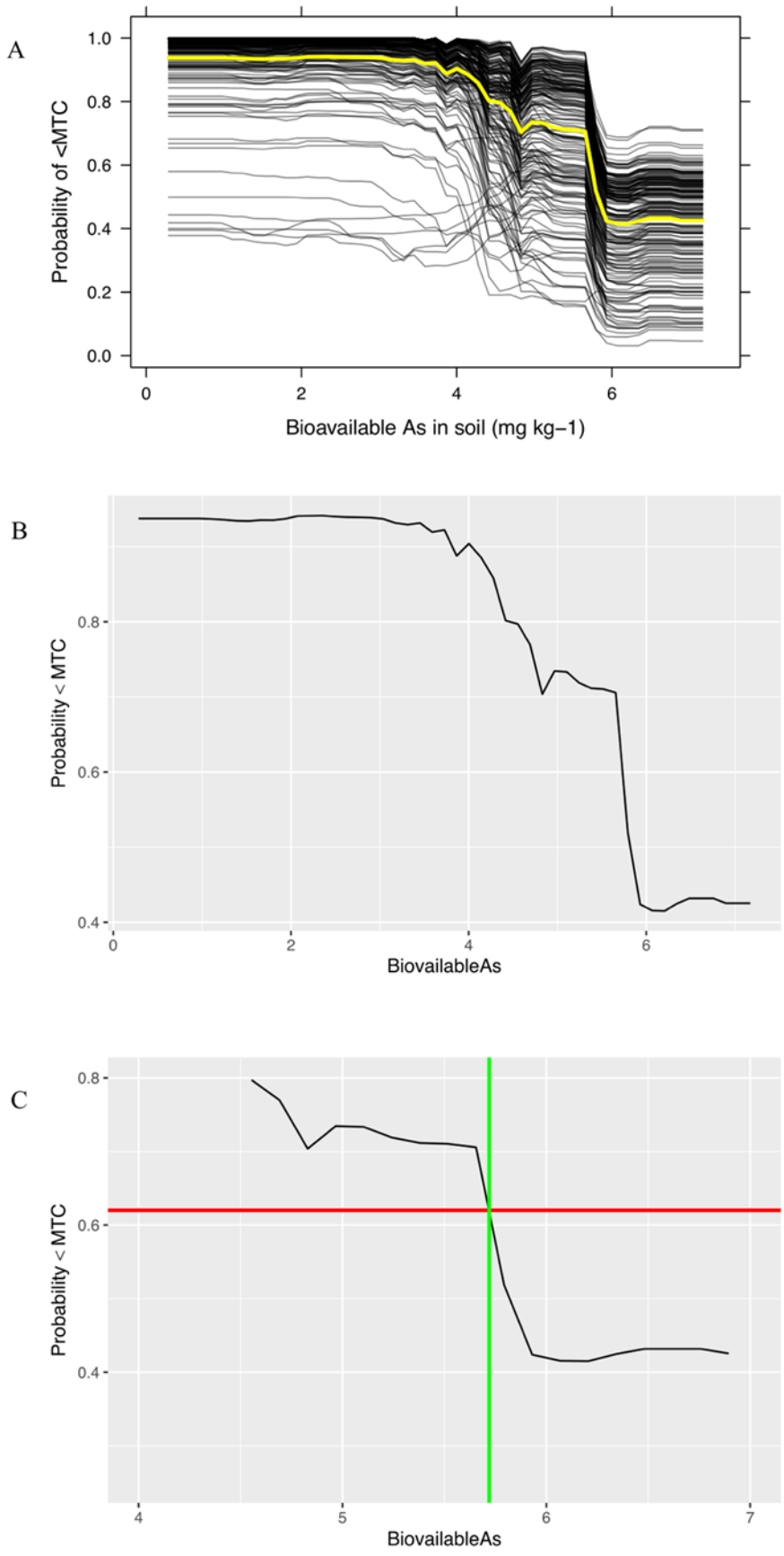


Figure 6.5. ICE and PDP of bioavailable As (mg kg^{-1}) from random forest model with respect to probability of grain As <MTC.

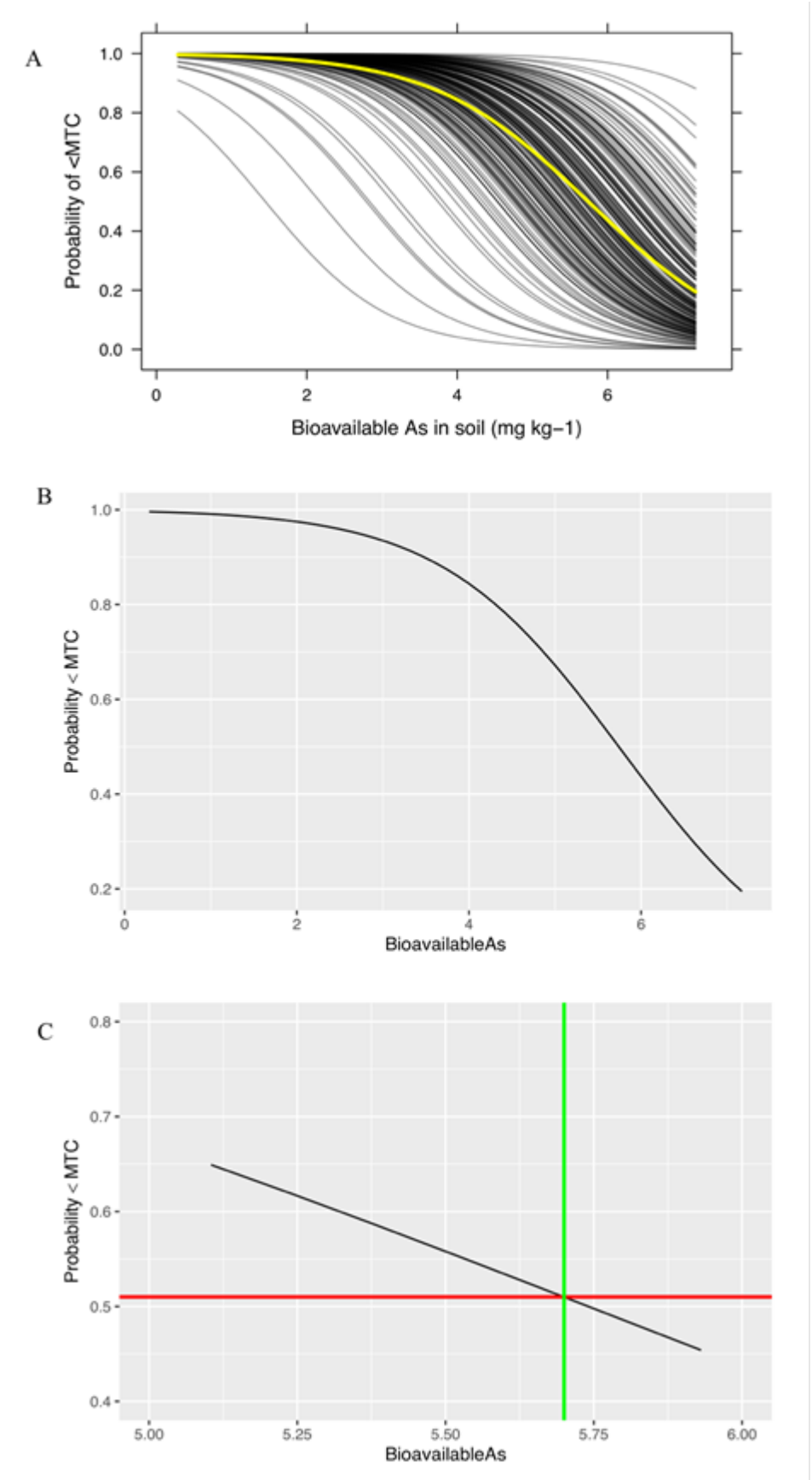


Figure 6.6. ICE and PDPs of available As (mg kg^{-1}) from logistic regression model with respect to probability of grain As <MTC.

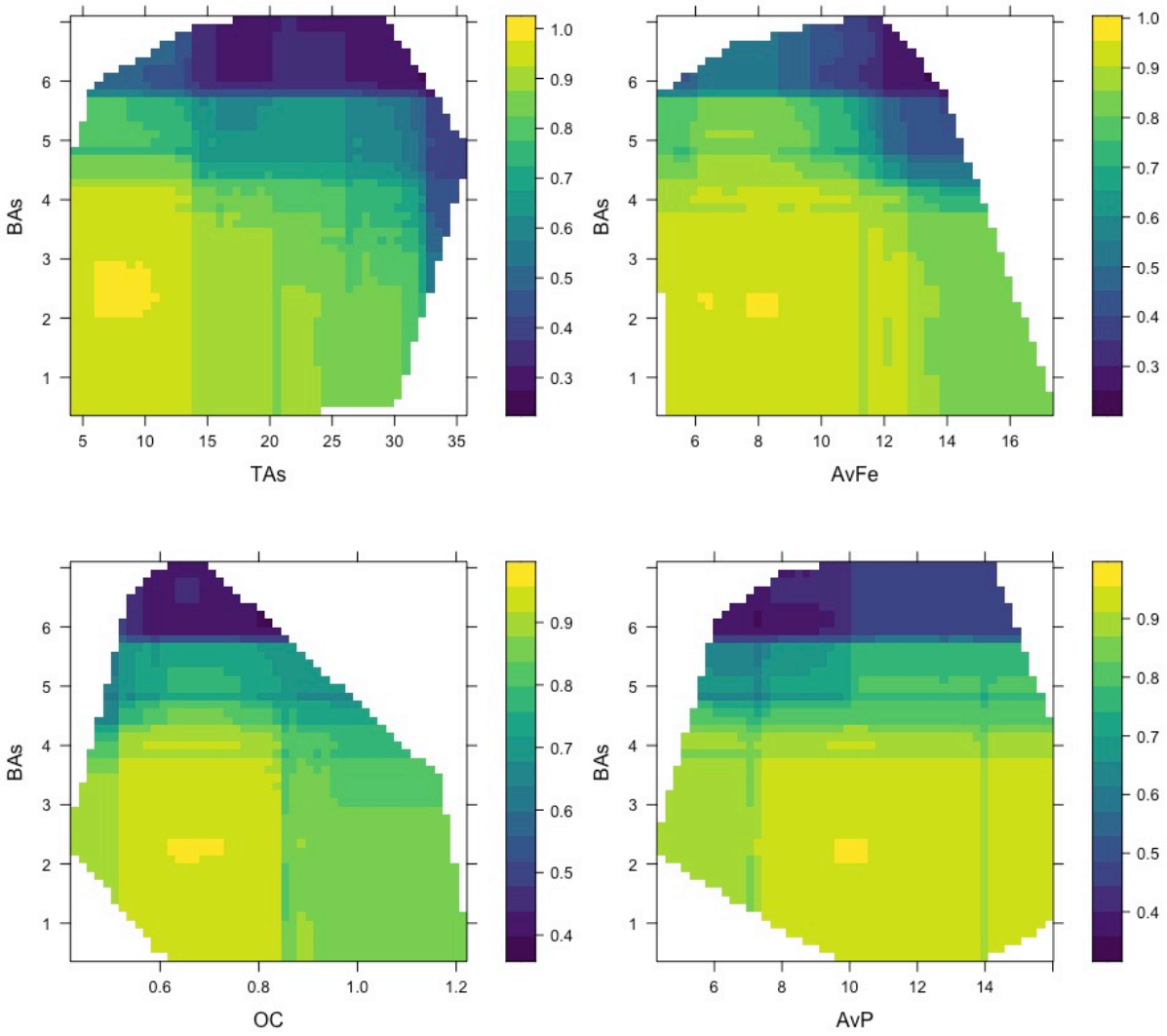


Figure 6.7. PDPs of two variables, BAs (mg kg^{-1}) with other important variables TAs AvFe, AvP (mg kg^{-1}) and OC (%) from RF model. Probability of <MTC is depicted in terms of colour intensities.

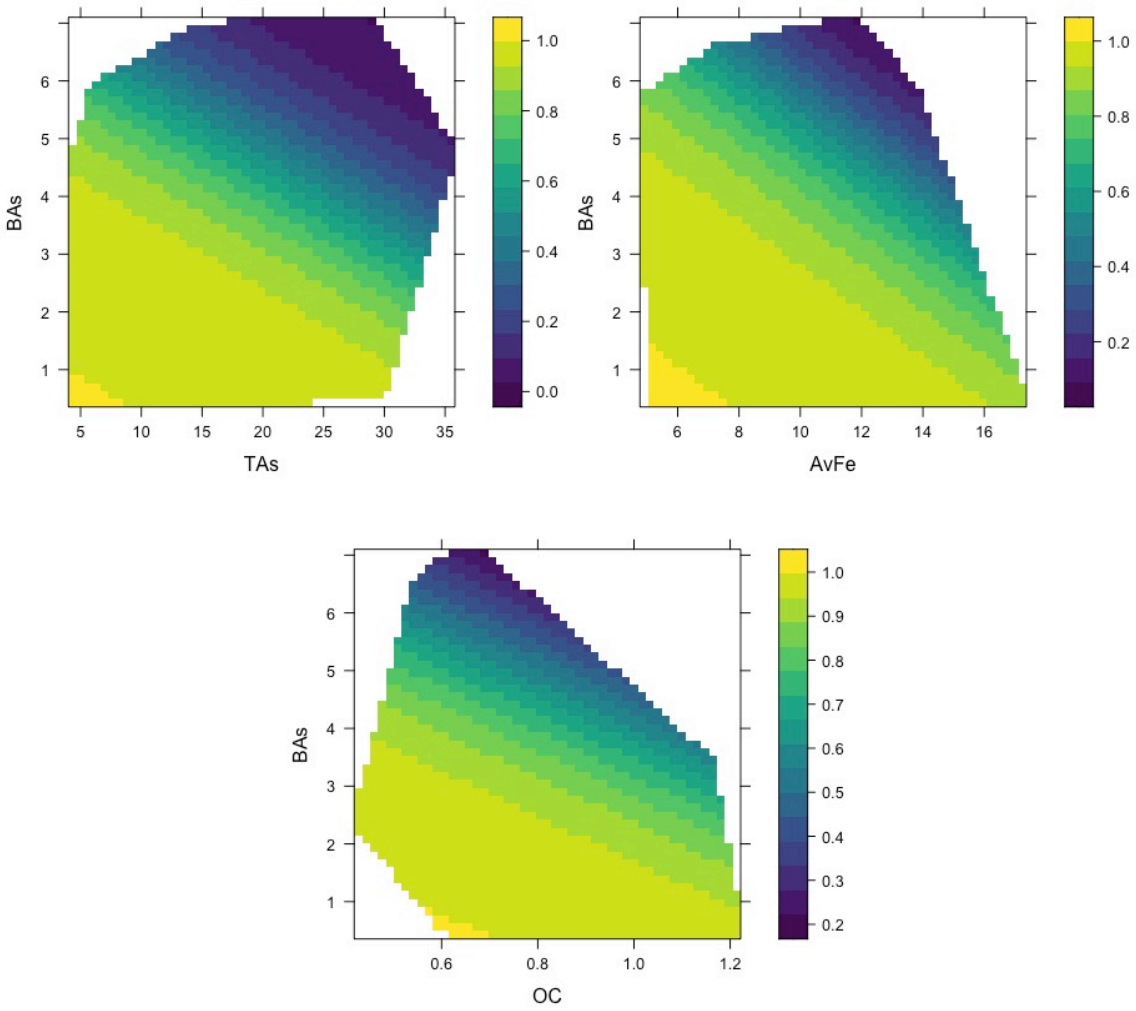


Figure 6.8. PDPs of two variables, AvAs (mg kg^{-1}) with other significant variables TAs, AvFe, (mg kg^{-1}) and OC (%) from LR model. Probability of <MTC is depicted in terms of colour intensities.

6.4 Discussion

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 have an edge over the RF model in terms of correctly predicting both the classes as previously recommended (Chicco et al., 2020). Although the log loss for GBM was minimum over the training set however over the test set it was highest. 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 as like the RF and LR. In general, the performance of LR is better

when their noise variables are less than or equal to the number of explanatory variables and random forest has a higher true and false positive rate as the number of explanatory variables surges in a dataset (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 dataset is simply insufficiently expressive (van der Ploeg et al., 2014) and hence a better performance of the LR model compared to RF and GBM.

To our knowledge this is the very first attempt of predicting the limit of soil available 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 the prediction is true. It represents the trade-off between the false positives and false negatives (Sarkar et al., 2022). Although from both the models the predicted limit for available As is very close to each other but considering the better performing LR model 5.70 mg kg-1 should be considered as the limit of soil available As. 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 maximum sensitivity-specificity on AUC. 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 of the effect of changes of 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 mg kg^{-1} from this study) and total As (14 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 surrounding rice plant roots, which reduces As uptake, and increases co-precipitation of Fe and As (Lee et

al., 2013). Metallic Fe and Fe-oxide have been observed to decrease As accumulation in rice by 51 and 47% (Matsumoto et al.,2015). BAs and OC relationship revealed the fact that 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 accumulator of As through formations of metal-humate complexes of varying stability (Sengupta et al., 2022; Kumar et al., 2021). 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 the same adsorption sites in the soil as well as on the Fe plaques mainly by ligand exchange which is a key characteristic in the rice field for bioavailability of As and uptake by plants (Peryea et al., 1995). This explains the relationship of BAs with AvP. Lee et al. (2016) proposed three key factors influencing As mobility in soil and uptake in rice: (1) enmity between As and P for adsorption sites, (2) antagonism between of inorganic P and As during transport in rice roots, and (3) role of P in As transfer from root to shoot. Thus, levels of AvFe, OC, and AvP (as shown in PDPs) in soil at which the BAs will be below the projected limit would aid in the development of acceptable management techniques in order to mitigate As build-up in rice. Both the LR and RF model did not identify pH as important predictor variable. This might be since water logging drives the pH of the paddy soils towards neutrality. The production of Carbon mono-oxide (CO) due to bacterial respiration along with its accumulation is the main reason for decrease in pH in alkaline soils. On the other hand, the reduction of Fe^{2+} in acidic soils are mainly responsible for the increase in soil pH (Kumari et al., 2021).

One of the limitations of predictability models is generalisability. Models trained on data from a specific location may not generalize well to other locations or contexts. The characteristics and patterns observed in one location may not be representative of other areas, leading to limited applicability of the model outside the training data's specific context. Field data collected at a specific time may not capture the temporal dynamics and changes that occur over time. If the models do not account for temporal variations, their predictive ability may be limited, especially if the relationships between variables change over time. Machine learning models are not perfect, and they are subject to uncertainty and error. The predictions made by these models should be interpreted with caution, and the associated uncertainty and error estimates should be considered. Field conditions and dynamics may

change over time, and the models may become less accurate or outdated if they are not regularly updated or recalibrated with new data.

6.5 Conclusion

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 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 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.

Chapter 7

Predicting the limit of arsenic concertation in irrigation water for cultivation of rice

7.1. Introduction

In addition to the human health impacts, contamination of irrigation water in South Asia also affects crop production. High levels of As in irrigation water accumulate in the grain of food crops such as rice. Rice is one of the most important crops in the region, and As contamination of irrigation water can lead to the build-up of As in grain surpassing the Codex recommendation (350 μg g^{-1} for husked rice and 200 μg g^{-1} for polished rice). Studies have shown that rice grown in fields with high levels of As in the water can have As levels in the grain that exceed safe limits for human consumption, which poses a risk to human health (Mandal et al., 2021; Mondal and Polya, 2008; Mondal et al., 2010). However, there has been little research on the soil and irrigation water concentrations that would cause these levels to be exceeded. Recently Mandal et al., (2023) reported total and bioavailable As of 14 mg kg^{-1} and 5.70 mg kg^{-1} as a guideline value for safe cultivation of rice in the As endemic regions of India considering the samples being collected from both the irrigated and rainfed system of rice cultivation. While it is known that the bioavailability of As from soil to rice is affected by a variety of factors (Kumari et al., 2021), As concentrations in soil and irrigation water are regarded as the most important factors influencing rice grain t-As concentrations (Mukherjee et al., 2017; Kumari et al., 2021). From a meta-analysis (Mandal et al., 2021) for the Asian paddy soils it was evidenced a stronger positive correlation between paddy soil and rice As concentrations compared to irrigation water and rice grain concentrations, the logistic regression (LR) model depicting a non-significant contribution of irrigation water As and the decision tree model predictions taking only the soil As as important predictor variable. In recent studies, authors identified the most substantial impact of soil As (Sengupta et al., 2021) and the "minimal if any" impact of irrigation water on rice As content (Van Geen et al., 2006). Regardless, several research have demonstrated that contaminated groundwater irrigation of paddy fields increases soil As concentrations (Huq et al., 2006; Panaullah et al., 2009; Dittmar et al., 2010), eventually leading to a rise in grain As (Rahman et al., 2007; Rahman et al., 2010). Given the complexity of the transmission of irrigation water As to rice grain via the soil, the non-significant influence of irrigation water on grain As concentrations were likely

unsurprising. For example, the build-up of As in soil from irrigation water is affected by a variety of parameters, including the temporal variation in As concentration during the crop-growth phase, the volume of irrigation water used, and the area of the field being watered (Chowdhury et al., 2018; Chowdhury et al., 2020). The link between irrigation water As and grain As may become more complex as a result of irrigation strategies that frequently include the use of both groundwater and precipitation. Although the FAO has set a limit of 100 µg L^{-1} for As in irrigation water, but this is for general agriculture rather than rice specifically (Food and Agriculture Organization FAO, 1992; Pescod, 1992).

As of now, no previous studies have predicted the threshold concentrations of As in irrigation water for rice soils. These thresholds would indicate the point at which the levels of As in rice grains would exceed the limits recommended by Codex. (JECFA, 2017). The threshold for irrigation water As was predicted considering the soil parameters (pH, organic carbon (OC), clay, available phosphorus (P), available (Fe), amorphous Fe and aluminium (Al) oxides) on the bioavailable As under different irrigation conditions for rice (rainfed and irrigated). Monolithic soil columns, allowing testing of soil which are close to actual field conditions (Lewis and Sjostrom, 2010) were used for this purpose. Further with the help of individual conditional expectation (ICE) and partial dependence plots (PDP) using the logistic regression (LR) and linear discriminant analysis (LDA), the limit for irrigation water has been predicted.

7.2. Materials and Methods

7.2.1 Incubation experiment with soil columns

Monolithic soil columns were collected from Gotera (S1), Dakshin Panchpota (S2), Ghetugachi (S3), Kalyani (S4), Jhumka (S5), Sujapur (S6), Beldanga (S7), Radhavallabhpur (S8), Baruipur (S9), Sonarpur (S10) comprising the As contaminated districts of Nadia (S1, S2, S3 and S4), Murshidabad (S6, S7, S8) and South-24 Parganas (S9, S10) of West Bengal India. Polyvinyl chloride (PVC) pipes of 4.75 cm diameter and 20 cm in height were used for collecting (0-15) cm of soil columns from the field. Each column was plugged at the bottom with a perforated PVC cap and Whatman 42 filter paper to ensure regular drainage of only water throughout the experiment duration. The dose of As in irrigation water has 7 levels (0, 100, 200, 300, 400, 500, 600 µ L^{-1}), with 2 types of irrigation conditions (rainfed and irrigated) and each replicated thrice resulting in 42 soil columns from each field. Forty-two soil columns were collected from a specific crop field, altogether resulting in 420 soil columns. The characteristics of the soils of

the sites are depicted in Table 7.1. As contaminated water (0, 100, 200, 300, 400, 500, 600 µ L^{-1}) was applied (IrriAs) to the set of 210 columns as per the irrigation practices followed for cultivation of rice under rainfed conditions and remaining set of 210 columns as per the irrigated condition (Table 7.2. The levels of As dose in irrigation water were decided based on the mean value of 235.49 ± 215.48 (mean ± SD) As concentration in irrigation water in the Asian countries as reported by Mandal et al., (2021). The soil columns were kept at laboratory temperature (25-30°C) for 12 weeks. Rice grain and irrigation water samples were also collected from As contaminated sites of Maldah (n=29) and Nadia (n=44) districts of West Bengal, India. The soil parameters (pH, OC, available P (AvP), available Fe (AvFe), amorphous Fe-oxide (AmFe), amorphous Al-oxide (AmAl), bioavailable As (BAs) and total As (TAs)), rice grain and irrigation water samples has been mentioned in Chapter 3. The grain As content was converted to categorical variables (<MTC and >MTC) as per the methods outlined in Mandal et al. (2021, 2023).

Table 7.1. Characteristics of the soil used in the column study (Mean±SE, n=4)

Site Name	pH	OC (%)	Clay (%)	Available P (mg kg^{-1})	Available Fe (mg kg^{-1})	Amorphous Fe-Oxide (g kg^{-1})	Amorphous Al-Oxide (g kg^{-1})	Total As (mg kg^{-1})	Bioavailable As (mg kg^{-1})
Gotera (S1)	7.48±0.07	0.39±0.08	16.60±0.08	12.76±0.06	7.12±0.13	9.34±0.05	4.70±0.04	8.87±0.10	4.98±0.03
Dakshin Panchpota (S2)	7.91±0.11	0.81±0.05	48.40±0.09	18.21±0.05	6.81±0.10	7.11±0.04	3.85±0.06	24.77±0.12	6.06±0.06
Ghetugachi (S3)	7.45±0.07	0.54±0.08	28.61±0.10	20.34±0.07	5.22±0.08	5.05±0.05	3.35±0.06	21.56±0.11	5.21±0.08
Kalyani (S4)	7.41±0.06	0.52±0.07	18.70±0.10	10.37±0.09	3.62±0.07	4.12±0.09	2.87±0.05	7.69±0.13	1.78±0.09
Jhumka (S5)	6.74±0.08	0.38±0.05	39.60±0.09	9.49±0.05	4.48±0.06	5.72±0.04	4.12±0.06	7.93±0.07	5.26±0.02
Sujapur (S6)	6.76±0.13	0.56±0.06	57.30±0.07	13.58±0.07	7.39±0.09	7.72±0.07	4.57±0.03	24.8±0.09	5.68±0.03
Beldanga (S7)	7.17±0.08	0.47±0.06	34.40±0.10	18.48±0.05	6.47±0.05	5.35±0.06	2.64±0.07	6.60±0.08	3.88±0.07
Radhavallabhpur (S8)	7.07±0.06	0.61±0.08	42.10±0.07	12.19±0.07	2.64±0.02	3.82±0.04	2.67±0.06	7.95±0.10	4.83±0.09
Baruipur (S9)	6.43±0.07	0.64±0.09	14.10±0.10	10.46±0.08	9.60±0.02	10.35±0.04	5.84±0.05	7.57±0.09	1.46±0.03
Sonarpur (S10)	7.20±0.04	0.39±0.05	11.70±0.05	18.58±0.09	9.02±0.06	10.31±0.08	6.07±0.06	5.57±0.07	2.35±0.04

Table 7.2. Schedule of application of As contaminated water being applied to the soil columns under rainfed and irrigated condition over 12 weeks at laboratory temperature (25-30°C).

Type of Rice	**Stages (week)**	***Volume of water used in farmer's field (L)**	**Volume water/ sq cm. of soil (L)**	**Volume of water for each column (mL)**
Rainfed rice	Field Preparation (1-2)	60,000	0.0045	80
	Transplanting (2)	1,00,000	0.0075	140
	Vegetative phase (3-4)	1,00,000	0.0075	140
	Reproductive phase (5-8)	1,00,000	0.0075	140
	Ripening & Harvesting (9-12)	1,20,000	0.0089	165
Irrigated rice	Field Preparation (1-2)	1,40,000	0.0104	196
	Transplanting (2)	40,000	0.0029	54
	Vegetative phase (2-4)	1,00,000	0.0075	140
		1,00,000	0.0075	140
	Reproductive phase (5-8)	80,000	0.0060	112
		80,000	0.0060	112
		80,000	0.0060	112
		80,000	0.0060	112
	Ripening & Harvesting (9-12)	1,00,000	0.0075	140
		1,00,000	0.0075	140
		1,40,000	0.0104	196

*Roy Chowdhury et al., 2020 (rainfed rice) and Roy Chowdhury et al., 2018 (irrigated rice)

7.3. Statistical Analysis

The boxplots and the violin plots along with the non-parametric Wilcox test were performed using the packages *"ggplot2" (version 3.3.5)* and "*ggpubr*" *(version 0.4.0).* The null hypothesis was (H_0): The median differences (bioavailable As) between the paired observations (rainfed and irrigated) was zero. The alternative hypothesis is (H_1): The median difference is not zero. The Kruskal-Wallis test was done to compare the effect of As doses of irrigation water on the soil bioavailable As under irrigated and rainfed condition. The null hypothesis was (H_0): The median value of bioavailable As across different doses of As is equal and the alternative hypothesis was (H_1): at least one median value of bioavailable As for a dose of As is different from others. The pairwise comparison between the doses of As was done by Wilcox test. Further Kruskal-Wallis test was done to compare the difference in bioavailable As across the 10 soil types. The null hypothesis was (H_0): The median value of bioavailable As across different soil types is equal and the alternative hypothesis was (H_1): at least one median value of bioavailable As for a soil type is different from others. The pairwise comparison between the soil types was done by Wilcox test. The regression with GLM was performed using the

"stats" (4.0.3) package with BAs as the dependent variable and all other soil parameters (IrriAs, Tas, pH, OC, clay, AvP, AvFe, AmFe and AmAl) as independent variables.

7.3.1. Training of the models

For predicting irrigation water As concentration alongside the impact of other soil parameters with the LR and LDA the whole data set (n=420) was used. The BAs content was converted to categorical variables as class "A" representing above 5.70 mg kg^{-1} and class "B" representing below 5.70 mg kg^{-1} as per the limit of soil bioavailable As predicted by Mandal et al. (2023) for rice. The category of BAs (A and B) were considered as the dependent variable whereas, dose of As-contaminated water IrriAs, Tas, pH, OC, clay, AvP, AvFe, AmFe and AmAl as the predictor variables. Multicollinearity can affect logistic regression models. Multicollinearity occurs when predictor variables are highly correlated, which raises the variance of parameter estimations and can lead to incorrect inferences about the relationship between the dependent and independent variables. The presence of multicollinearity in the models can also affect the assumptions for PDP. To assess this, the variance inflation factor (VIF) was used to test the severity of multicollinearity for each variable. The VIF measures the extent of multicollinearity among the predictor variables in a regression analysis. According to Franke (2010), a VIF greater than 10 indicates high levels of multicollinearity. In this study, the VIF values were found to be 1.01 for IrriAs, 3.98 for pH, 3.40 for OC, 3.66 for clay, 6.01 for AvFe, 2.84 for AvP, 4.05 for Tas, 12.53 for AmFe and 12.77 for AmAl. The AmFe and AmAl was having a VIF >10 and hence were removed as the predictor variables for training the models. The data was transformed to ensure normality before the LDA model was trained. The whole data set was randomized and split into two, 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. Basically, the training set was used to generate multiple splits of the training and validation sets to reduce over fitting of the model. The "*caret*" (*version 6.0–91*) and "*caretEnsemble*" (*version 2.0.1*) package were used to train the models with 10-fold cross-validation repeated 5 times using *R-Studio* (*version 1.4.1103*).

7.3.2. Model performance parameters

The evaluation of a model's performance involves the analysis of a single confusion matrix, which consists of four categories: true positive (TP), true negative (TN), false positive (FP), and

false negative (FN). It is important to assess the relationships between these categories, rather than evaluating them individually, to accurately evaluate the model's performance. The model performance parameters include accuracy, sensitivity or recall, specificity or true negative rate, and precision or positive predictive value. The F1 score and the Mathews correlation coefficient (MCC) were also calculated.

The Area Under the Curve (AUC) is calculated by plotting the receiver operating characteristic (ROC) curve, which plots sensitivity against specificity at different classification thresholds. The "*ROCR" (version 1.0-11)* and the "pROC" *(version 1.17.0.1)* were used to plot the ROC curves over the training and testing phase.

7.3.3. ICE and PDP

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. 2017). 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 LR was prepared using the *"pdp"* (*version 0.7.0*) package.

7.4. Results

7.4.1 Effect of As dose on Bioavailable As and its relationship with the soil parameters

The Figure 7.1 represents the effect of As dose on the different soil columns over the rainfed and irrigated system of irrigation. From the non-parametric Kruskal-Wallis test it was observed that the overall effect was statistically non-significant ($p > 0.05$) rainfed and significant ($p < 0.05$) for irrigated type of irrigation. However, paired-wise comparison (Wilcox test) of dose of As it was observed that there was a significant difference between the As0 and As600 over the irrigated condition of irrigation. The boxplot in Figure 7.2 represents the comparison between the soil types in terms of BAs in the post incubated soil samples under two different irrigation systems. It was observed that in both the irrigation types an overall statistically significant ($p <= 0.01$) variation of BAs with respect to the soil types was observed. The paired-wise comparison between the soil types revealed that except S1 and S8 in both irrigation types all the soil types were statistically significant with each other ($p <= 0.05$, $p <= 0.01$, $p <= 0.001$, $p <= 0.0001$). From the violin plots in Figure 7.3 it was observed that the effect on BAs from

type of irrigation was statistically non-significant ($p > 0.05$) irrespective of soil type and dose of irrigation. The violin plots provide a comprehensive visualization of the distributional characteristics of data (non-normal distribution in this case), including central tendency, spread, shape, and the presence of outliers. It was observed that the soil properties played a significant role on the bioavailability of As in soil. From the regression analysis BAs = -2.77 + 0.001IrriAs + 0.03Tas – 0.33AvFe + 0.05AvP + 0.60pH -2.46OC + 0.06Clay + 0.35AmFe + 0.11AmAl ($R^2 = 0.82$, Adj-$R^2 = 0.81$, AIC = 890.74) it was observed that all the soil parameters significantly ($p < 0.001$) affected the BAs except AmAl. From the spearman's correlation studies in Figure 7.4 it was observed that there was a significant positive correlation ($p <= 0.001$) of BAs was with IrriAs (0.22), Tas (0.55), AvP (0.27), pH (0.20) and clay (0.76) and a significant negative correlation with AvFe (-0.18). Apart from correlation data the Figure 7.4 also provides the information about the distribution of data through histograms for each variable along with the shape of the distribution. As the data was not normally distributed spearman's correlation was undertaken in this case.

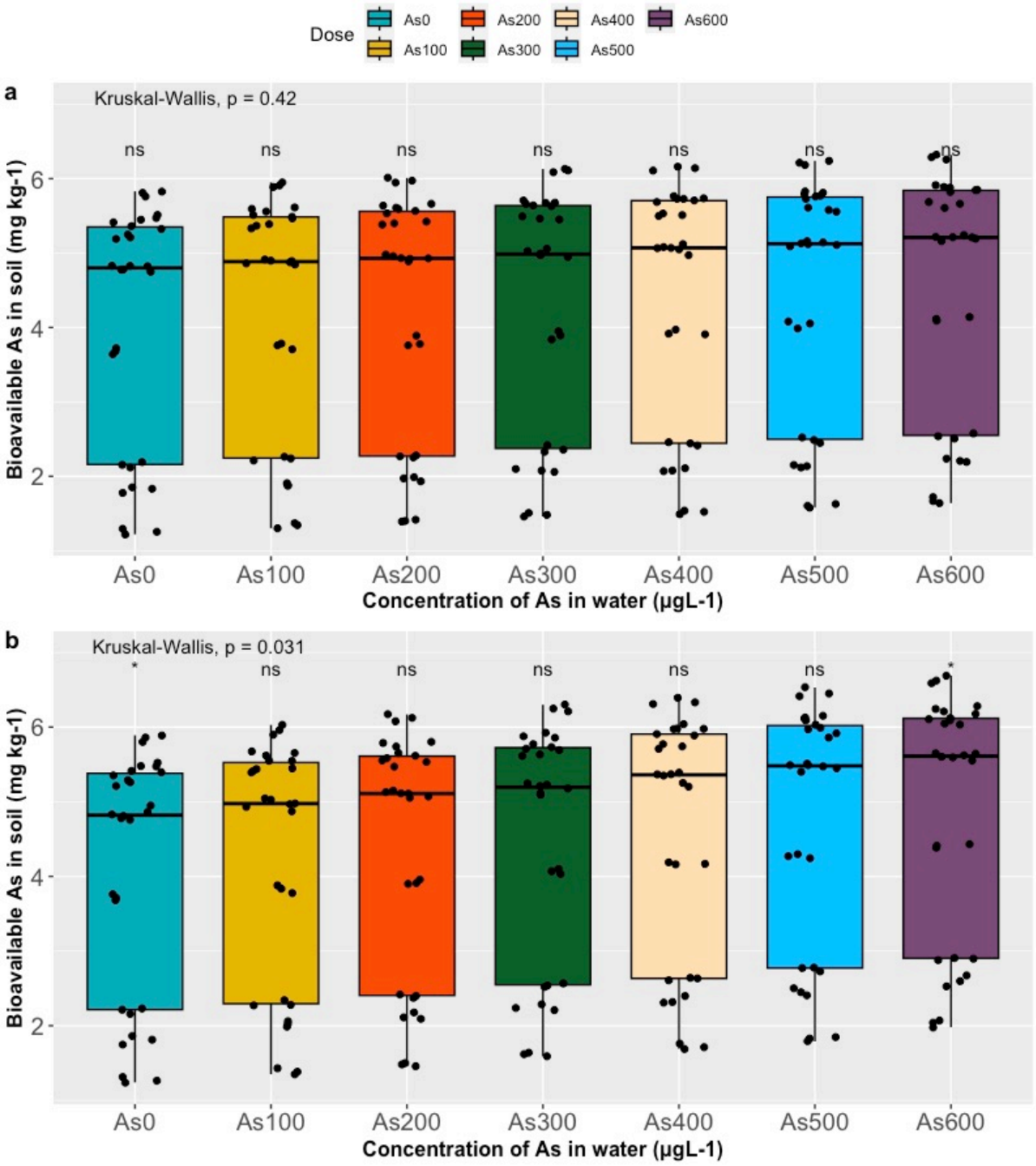


Figure 7.1. Boxplots representing the variation of bioavailable As in post-incubated soil samples with respect to dose of As irrespective of soil types under rainfed (a) and irrigated (b) conditions. (ns: p > 0.05 *: p <= 0.05).

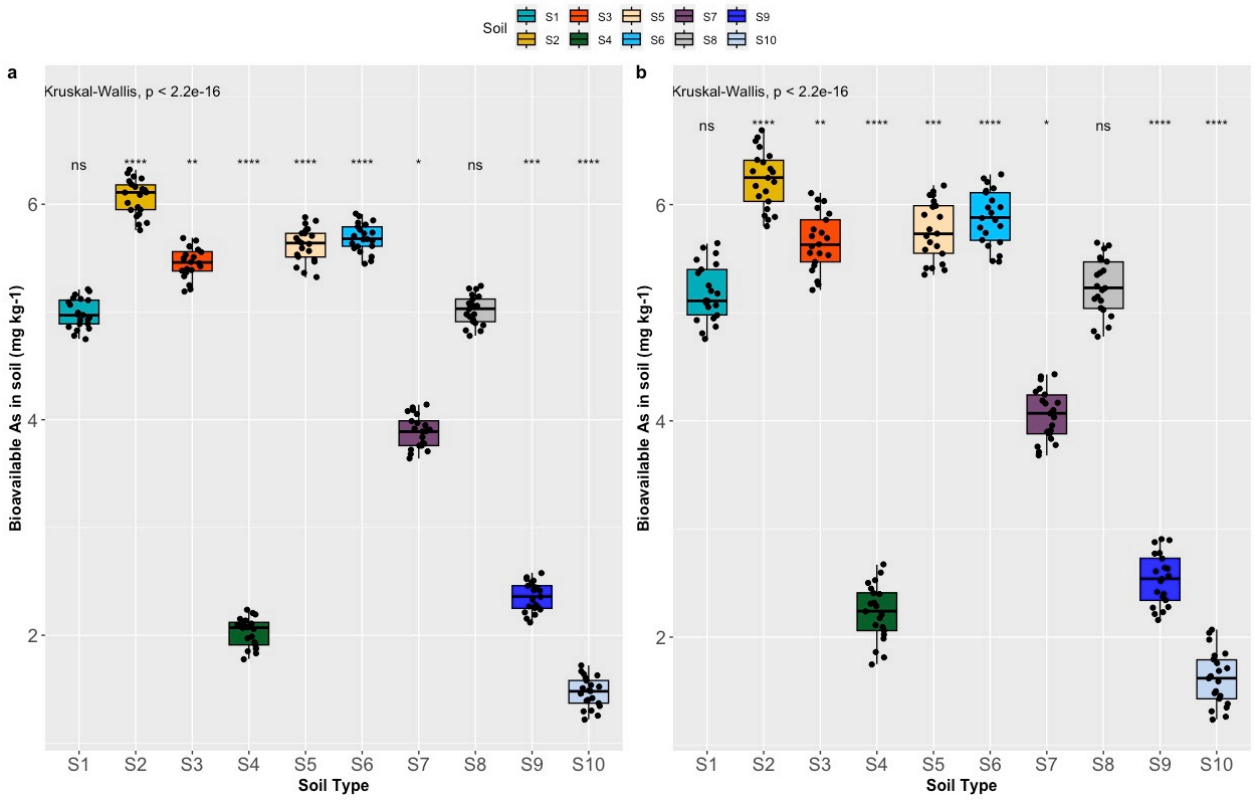


Figure 7.2. Boxplots representing the variation of bioavailable As in post-incubated soil samples with respect to different soil types irrespective of dose of As over rainfed (a) and irrigated (b) conditions. (ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001)

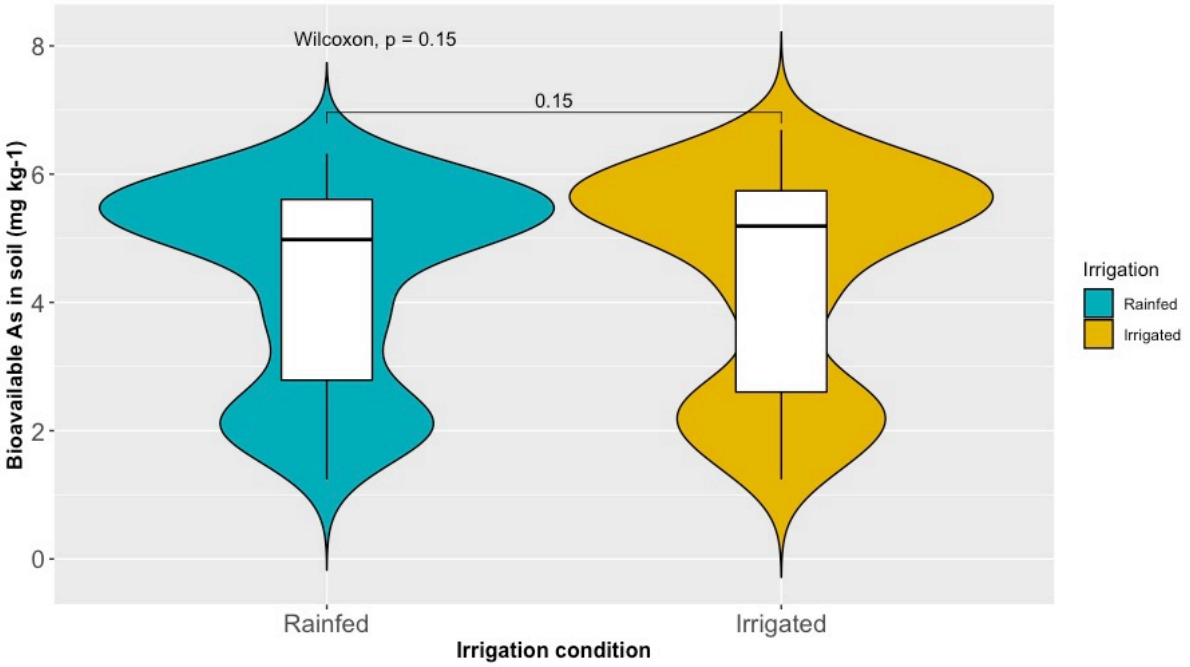


Figure 7.3. Violin plots representing the comparison between irrigation condition irrespective of soil types and As dose.

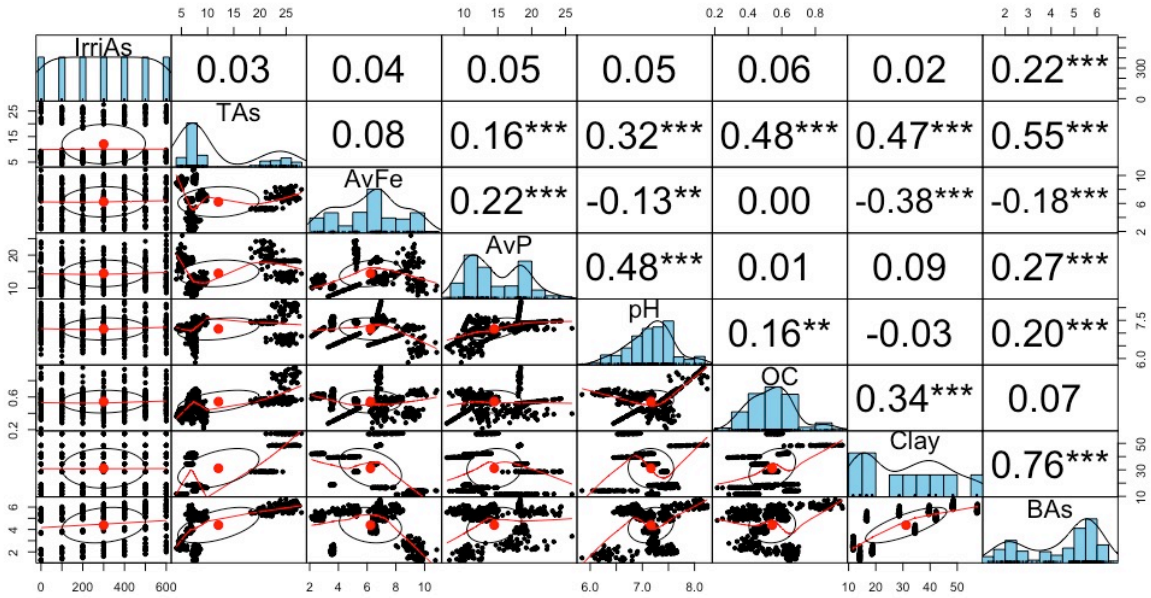


Figure 7.4. Spearman's correlation matrix between the soil properties and doses of irrigation water (n=420). (*: p <= 0.05, **: p <= 0.01, ***: p <= 0.001)

7.4.2. Training and selection of models

The final LR model after repeated cross-validation was selected at AUC of 0.9579 of ROC curve having a sensitivity of 0.735 and specificity of 0.9330 at the accuracy (average) of 0.8886. The LR model equation predicted Probability (A | B) = 69.61 – 0.009IrriAs – 0.21Tas + 0.20AvP – 9.25pH + 19.79OC + 0.42AvFe – 0.30Clay (AIC=141.61). The intercept, IrriAs, Tas, AvP, pH, OC,

and clay coefficients were statistically significant at ($p < 0.01$) and AvFe coefficient was significant at ($p < 0.05$). The log odds ratio is a measure that represents the change in the log odds of the outcome (A or B) for a unit change in the value of a predictor variable. The log odds ratio provides information about the strength and direction of the relationship between the predictor variable and the outcome. For example, in our case, the coefficient of IrriAs = –0.009, can be exponentiated (as per the Eq.2) $e^{-0.009} = 0.99$, which suggests that the odds of class A changes by a factor of 0.99 for an increase of concentration of As in irrigation water. Equivalently it can also be interpreted as a decrease by a factor of 0.01 in odds of class A with the increase of concentration of As in irrigation water. So, with the decrease in concentration of As in irrigation water, the log odds for class B will increase accordingly.

The LDA model was selected at AUC of 0.9663 having a sensitivity and specificity of 0.9410 and 0.8586 respectively. The prior probabilities for class A and B as per the LDA model were 0.225 and 0.774 respectively. A prior probability is a probability that an observation will fall into a group before someone collects the data. The group means for class A were 0.92 , 0.08, 0.02, 0.06, 0.02, 0.002 and 0.17 for IrriAs, Tas, AvFe, AvP, pH, OC and clay respectively. For class B 0.82, 0.07, 0.05, 0.11, 0.06, 0.004, and 0.21 were group means for the IrriAs, Tas, AvFe, AvP, pH, OC, and clay respectively. The coefficients of linear discriminants were 0.61 for IrriAs, -13.78 for Tas, 1.15 for AvFe, -2.85 for AvP, 37.68 for pH, -141.78 for OC and 0.37 for clay. Figure 7.5 represents the comparison between the LR and LDA models over the training phase in terms of ROC, sensitivity, and specificity.

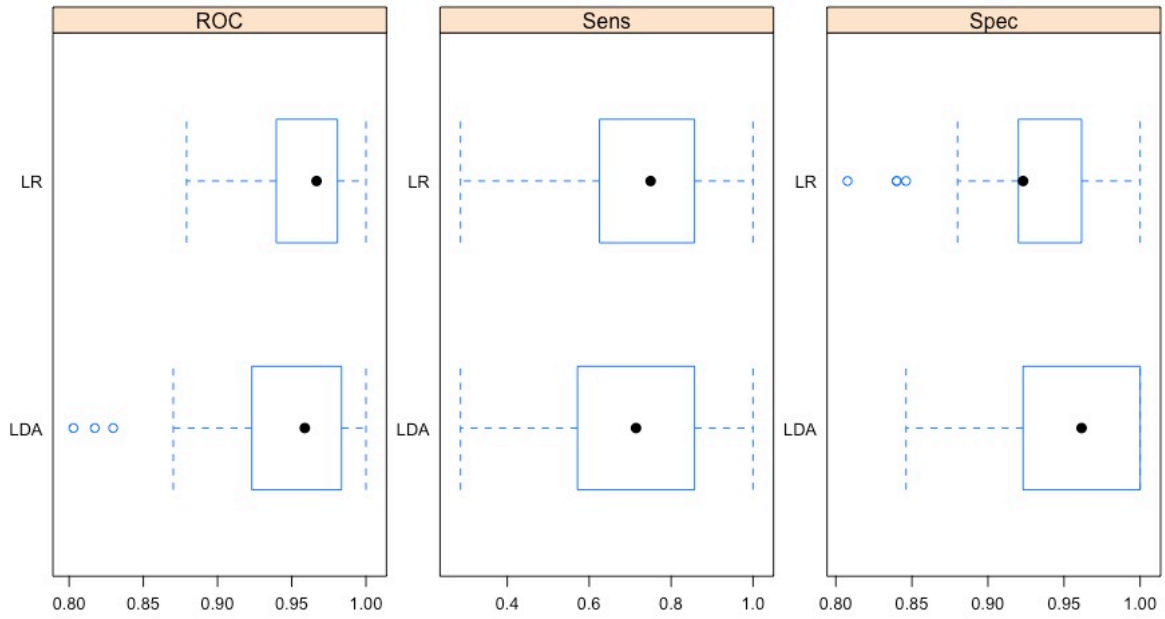


Figure 7.5. Comparison between the logistic regression (LR) and linear discriminant analysis (LDA) in terms of area under the receiver operating characteristic (ROC) curve, sensitivity (Sens) and specificity (Spec) during the training phase.

7.4.3. Confusion matrix and performance of the LR and LDA model

The performance of the LR and LDA models over the training and testing phase can be observed in Table 7.3. From the confusion matrix it was observed that over the training set the model prediction accuracy was similar for both LR and LDA (0.8946). However, the LR model has higher kappa, sensitivity, F1 Score and MCC compared to the LDA over the training set. The AUC of LR (0.968) > LDA(0.954) over the training phase (Figure 4). Although the LDA have an edge over LR in terms of specificity and precision, the log loss was more in LDA (0.1891) compared to LR (0.2907). During the testing phase of the models LR > LDA in terms of all model matrices like accuracy, kappa, specificity, precision, F1 score and MCC except sensitivity the LR have an edge over LDA. Besides the AUC was higher in LR (0.950) compared to LDA (0.940) during the testing phase (Figure 4). Additionally, the log loss for LR was lower than LDA both during the training and testing phase, indicating better predictability.

Table 7.3. Confusion matrix of LR and LDA and model parameters over training and testing phase.

Training set (n=332)			**Testing set (n=88)**	
Logistic Regression (LR)				
	Actual Class		*Actual Class*	
Predicted Class	**A**	**B**	**A**	**B**
A	57 (TP)	17 (FP)	14 (TP)	2 (FP)
B	18 (FN)	240 (TN)	7 (FN)	65 (TN)
Accuracy (%)	0.8946		0.8977	
95% CI	(0.8564, 0.9255)		(0.8147, 0.9522)	
Kappa	0.6971		0.6935	
Sensitivity/Recall	0.7600		0.6667	
Specificity	0.9339		0.9701	
Precision	0.7702		0.8750	
Log Loss	0.1891		0.2348	
F1 Score	0.7651		0.7567	
MCC	0.7017		0.7037	
Linear Discriminant Analysis (LDA)				
	Actual Class		*Actual Class*	
Predicted Class	**A**	**B**	**A**	**B**
A	52 (TP)	12 (FP)	15 (TP)	5 (FP)
B	23 (FN)	245 (TN)	6 (FN)	62 (TN)
Accuracy (%)	0.8946		0.875	
95% CI	(0.8564, 0.9255)		(0.7873, 0.9359)	
Kappa	0.6821		0.6503	
Sensitivity/Recall	0.6933		0.7143	
Specificity	0.9533		0.9254	
Precision	0.8125		0.7500	
Log Loss	0.2907		0.3594	
F1 Score	0.7482		0.7317	
MCC	0.6854		0.6506	

7.4.4. Predicting the limit of irrigation water As from LR model and comparison with field data

The final class probability from the better-performing LR model was calculated considering the log odds of all the coefficients of the significant variables as per the LR model and the final class was assigned based on the cut-off probability. A cut-off probability refers to a threshold value in a logistic regression model that determines the boundary between two classifications (class A and B in our case). If the predicted probability of an event is above the cut-off, the observation is classified A; if it is below the cut-off, it is classified as B. From better performing LR model the cut-off probability was at 0.836 over the testing phase at highest sensitivity (0.952) and specificity (0.851) (Figure 7.6). From the ICE (a) and the PDP (b) of the LR model at Figure 7.7 it can be observed that at cut-off probability of 0.836 the limit of irrigation water

As to classify soil As as class B (BAs < 5.70 mg kg^{-1}) was 190 μ L^{-1}. The boxplot in Figure 7.8 shows the comparison between category of grain As (< MTC and >MTC) with respect to As concentration in irrigation water from two As-contaminated sites (Nadia and Maldah) and the limit predicted by the LR model. The black points below the red line (representing 190 μ L^{-1}) represent the instances at which the rice grain As was > MTC.

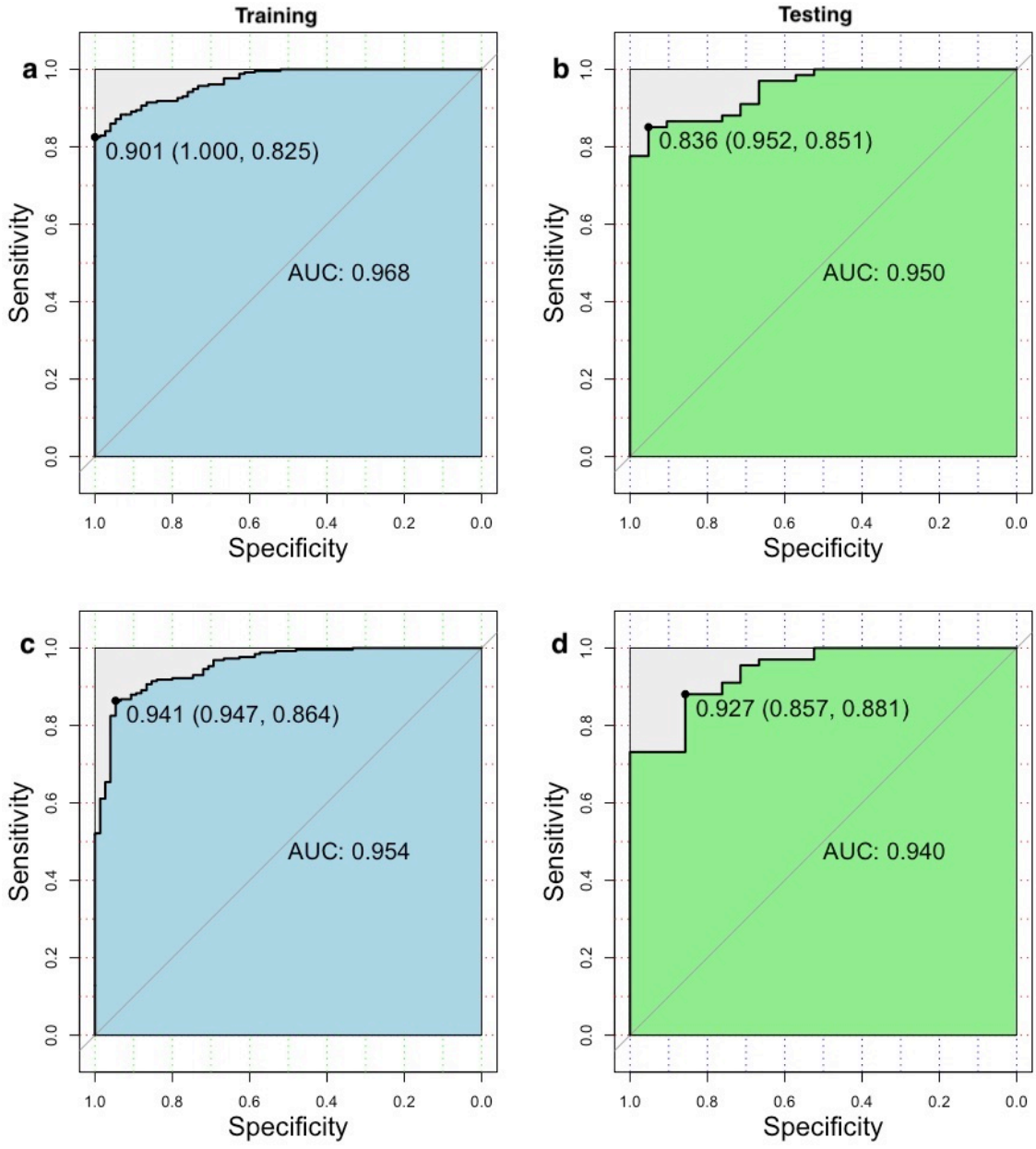


Figure 7.6. Sensitivity vs. specificity plot and cut-off probability (at maximum sensitivity and specificity) for LR (a) and (b) and LDA (c) and (d) models over training phase and testing phase.

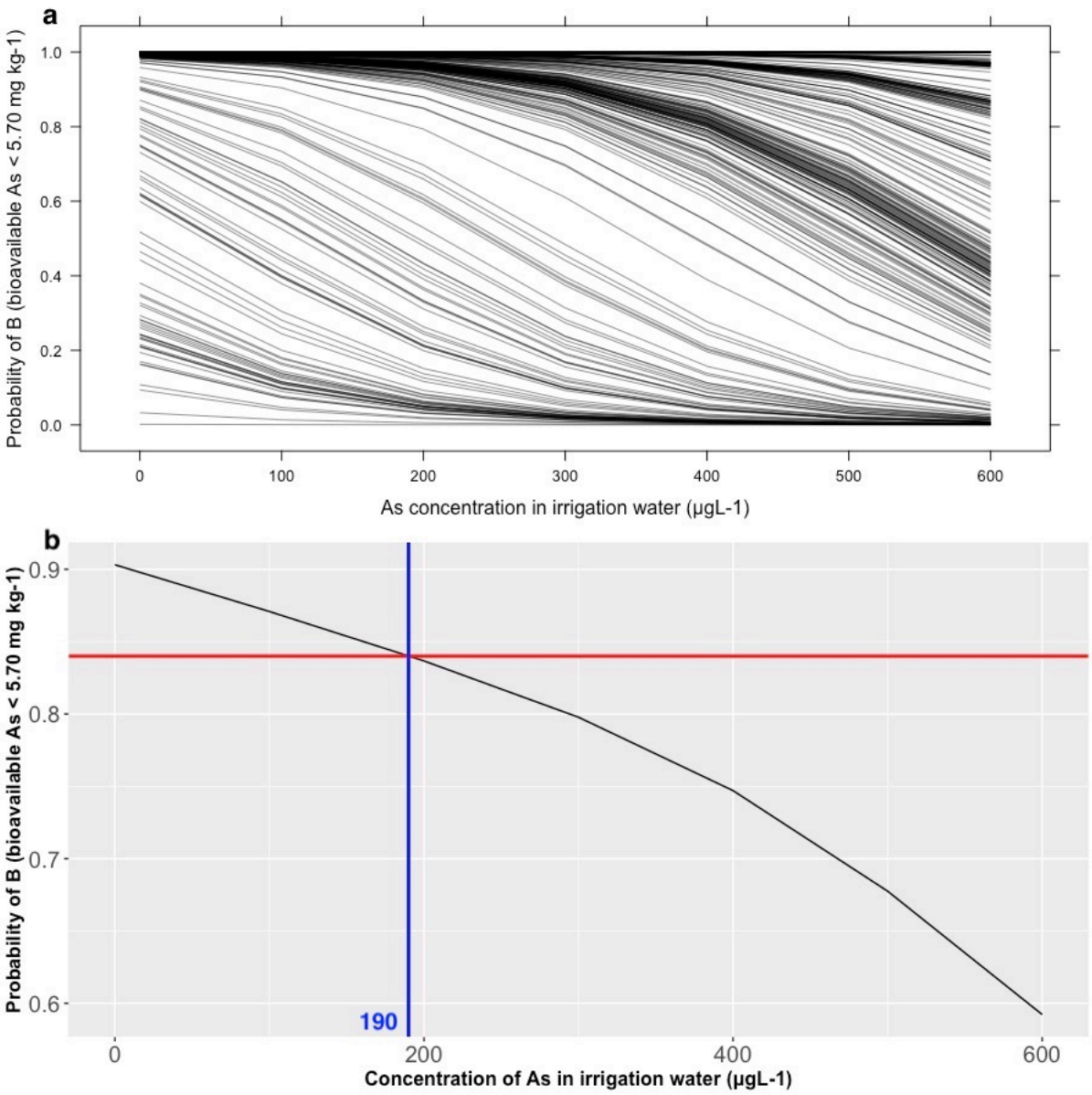


Figure 7.7. ICE (a) and PDP (b) of irrigation water As (µg L^{-1}) from logistic regression with respect to probability of B (bioavailable As < 5.70 mg kg^{-1}) representing the threshold limit of irrigation water As at cut-off probability.

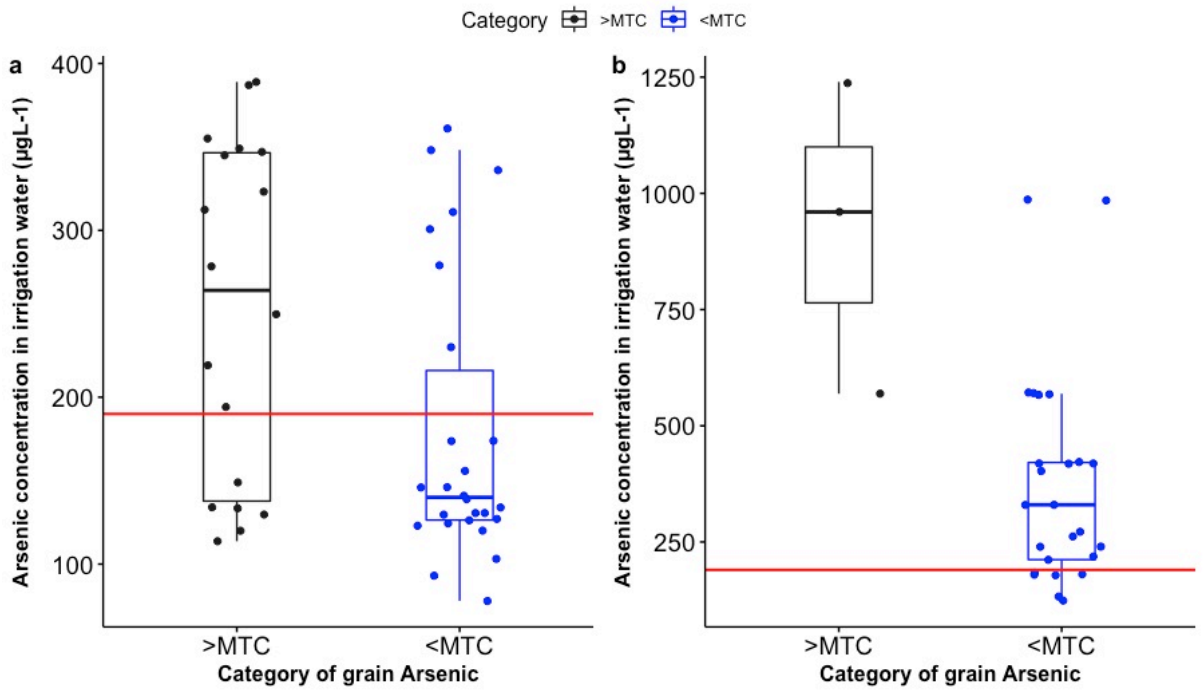


Figure 7.8. Boxplots of irrigation water As concentration (µg L^{-1}) with respect to the category of grain As concentration (<MTC and >MTC) of a) Nadia (and b) Maldah district, West Bengal, India. The horizontal red line indicates the limit of irrigation water As (190 µg L^{-1}) predicted by the predicted by logistic regression.

7.4.5. Bioavailable As and its relationship with the soil parameters from PDP of LR model

The PDP in Figure 7.9 from LR model shows the probability of soil As category (B) with respect to IrriAs along with other variables for LR (pH, OC, clay, AvFe, AvP and Tas). It can be observed that at lower pH (6-6.5) values the probability of class B was maximum (1.0-0.9). However, at higher pH and also at higher concentrations of As in irrigation water the probability of class B decreases. The probability of class B was maximum at a higher OC (0.6) at lower concentration of As in irrigation water. The probability decreased at lower levels of OC and higher levels of As in irrigation water. The effect of clay on BAs considering the IrriAs followed the same trend of pH as with high clay (more than 30%) content the bioavailability of As increased even at lower concentrations of As in irrigation water. At the higher level of AvFe the probability of class B was stable (0.90-0.85) even at higher concentrations of As in irrigation water signifying the adsorptive capacity of Fe. The same trend was followed by AvP as it competes with As for the adsorptive sites of the soil. At lower levels of TAs the probability of class B was high (1.0-0.8) up to 300 µ L^{-1} of As in irrigation water.

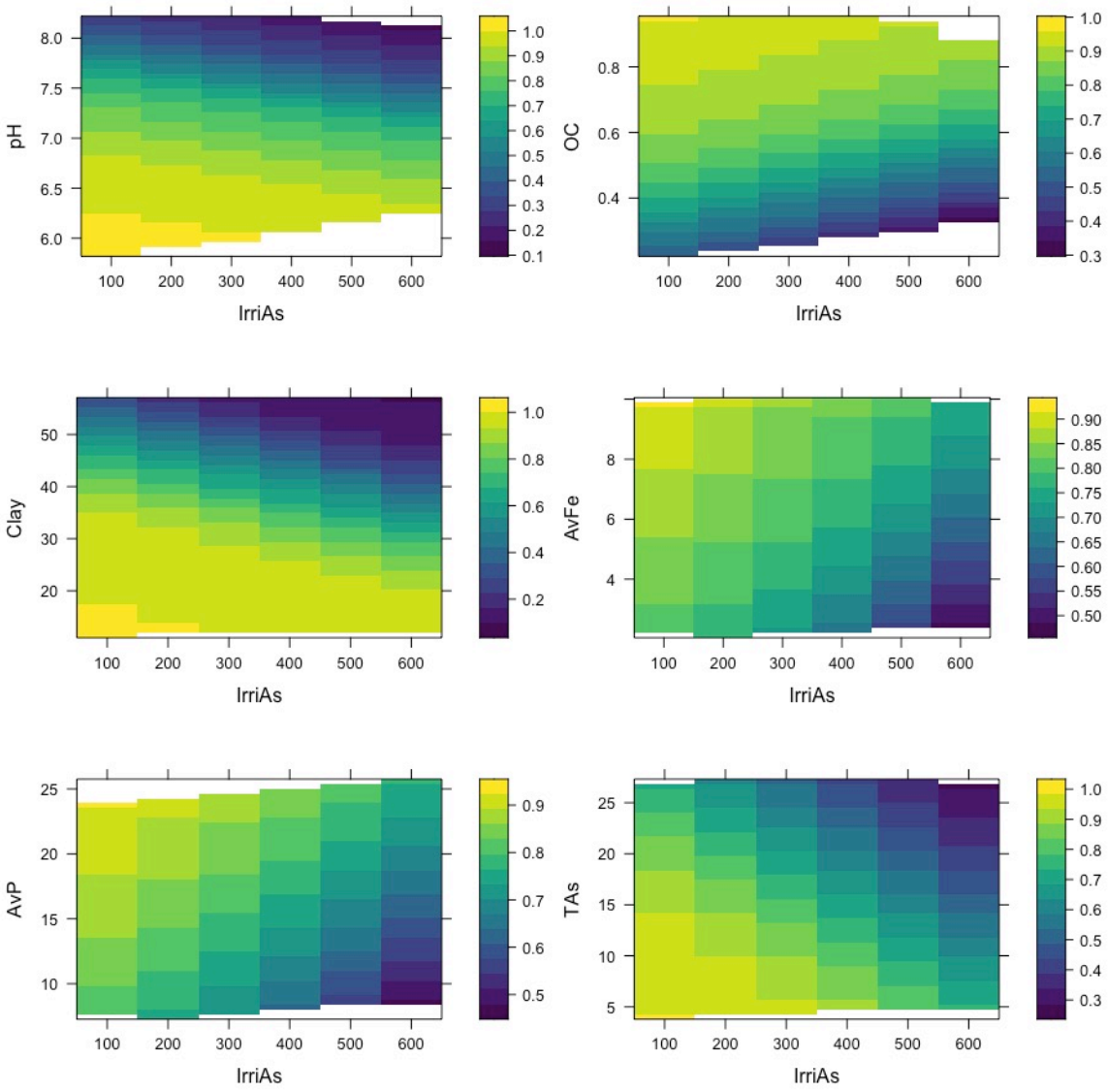


Figure 7.9. PDPs of two variables, IrriAs (μg L^{-1}) with other variables pH, OC (%), Clay (%) Tas, AvFe, AvP (mg kg^{-1}) and AmFe, AmAl (g kg^{-1}) from logistic regression. Probability of class B is depicted in terms of colour intensities.

7.5. Discussion

7.5.1. As dose, Bioavailable As and soil properties

The regression equation, correlation studies and more specifically the PDP from the LR model gives us an insight regarding the behaviour of BAs in relation to As from irrigation water (IrriAs) along with the significant soil parameters. The significant positive relationship of BAs from regression study with IrriAs aligns with the previous findings as reported by (Kumari et al., 2021; Mandal et al., 2021). However, from PDP it was observed that this relationship is more complex rather than linear due to effects of different soil parameters. The transfer of As in its

natural environment is governed by several factors, especially by pH, soil organic matter, clays, Fe (available as well as its oxides) and P (available) content (Kumari et al., 2021 and Raj et al., 2021). The availability of As in soil is dependent on soil pH, and generally, an increase in soil pH leads to an increase in the availability of As particularly above pH 7.5 (Sanyal et al., 2015 and Raj et al., 2021). At high pH levels, the soil becomes more alkaline, which can lead to the solubilization and release of As from mineral phases and make it more available for uptake by plants and other organisms. On the other hand, at low pH levels, the soil is more acidic, and As may be less available due to its binding to Fe and Al compounds in the soil (Kumari et al., 2021). Organic matter can bind to As, mainly by humic and fulvic acids as reported by (Sengupta et al., 2022). Additionally, dissolved organic matter can reduce the sorption of As to soil particles through redox reactions and by forming soluble complexes with As (Mandal et al., 2019a, 2019b). This has resulted in decrease in probability of class B at lower OC content irrespective of concentration of IrriAs and negative relationship of (regression) with BAs. The amount of As that can be absorbed by soil particles depends on the presence of clay. Clay can contribute to the sorption of As, increasing the amount of As that is retained in the soil and reducing its availability (Raj et al., 2021). However, under certain conditions, such as high levels of soil organic matter, clay can also increase the bioavailability of As by releasing it into the soil solution. The relationship between clay and As in soil is complex and depends on several factors, including the type and amount of clay present, the soil pH, and the presence of other minerals and organic matter (Raj et al., 2021). A significant negative relationship of As with Fe signifies that it can act as a barrier to the bioavailability of As, as iron oxides and hydroxides can adsorb As, making it less available for plant uptake (Chowdhury et al., 2018 and Sengupta et al., 2021). Fe can also form complex compounds with As, which can enhance or reduce As bioavailability, depending on the speciation of the compounds (Kumari et al., 2021). Additionally, iron when present in excess amount can also mobilize As in soil, making it more available for uptake by plants (Mandal et al., 2023). Although a positive significant relationship between BAs and AvP has been observed however, between phosphorus (P) and As in soil is complex and can vary depending on soil type, pH, redox conditions, and presence of other minerals (Sanyal et al., 2015). An antagonistic relationship between P and As has been previously reported by Raj et al., (2021). On the other hand, a positive relationship between AvP and AvAs has also been reported in paddy soils by Sengupta et al. (2021) and Jiang et al. (2014). Similarly, the positive significant relationship of soil Tas with BAs have been previously

reported by Mandal et al. (2023) from As contaminated paddy fields. The relationship between Tas and BAs in soil is also influenced soil pH, organic matter content, and the presence of other minerals that may bind or stabilize Tas in the soil (Kumari et al., 2021 and Raj et al., 2021). The Am-Fe oxides act as sorbent for both the arsenite and arsenate ions. The Am-Fe oxides involve As to form inner sphere surface complex with the displacement of OH^- and H_2O as a ligand substitution transfer of electron from As (Aide et al., 2016).

7.5.2. Model performance and prediction of irrigation water limit

From the accuracy, kappa, recall, precision, F1 score and MCC of the models over the test set, it can be concluded that the performance of the LR model was better as compared LDA. The accuracy and F1 score have long been popular measures for binary classification tasks, but they can be misleading in unbalanced datasets. The MCC provides a more comprehensive evaluation of binary classifications than accuracy and F1 score (Chicco et al. 2021 and 2020). In this study, the LR model outperformed the LDA model in terms of accurately predicting both classes according to the MCC results. The MCC score is based on the idea that the classifier must make accurate predictions for both positive and negative cases, regardless of their ratio in the dataset as previously reported by Mandal et al., (2023). A higher AUC value in a ROC curve indicates that the model is more effective at distinguishing between the positive (A) and negative (B) classes. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values and the AUC measures the overall performance of the model by computing the area under the curve (Mandal et al.,2021). A higher AUC means the model has a higher TPR for a given FPR or that it has a lower FPR for a given TPR. Hence the LR model performed better compared to LDA. The log loss measures how closely the predicted probabilities align with the actual values that is higher the log-loss number, the more the predicted probability deviates from the actual value (Vovk, 2015). In conclusion, after considering all model parameters, the LR model performed better than the LDA model.

From the LR model the predicted limit for irrigation water was 190 μ L^{-1} Previously an attempt to derive the limit of irrigation water was undertaken by Kumari et al., (2021) considering only soil As in two soil types but no particular limit was prescribed. To our knowledge, this is the first attempt at predicting the limit of irrigation water As using the PDP with respect to the cut-off probability from the model. The ICE plot shows the individual conditional relationship between a specific predictor variable and the predicted response,

while holding all other predictors constant at their median values. It helps to understand how the response changes with the changes in a single predictor (Goldstein et al. 2017). The PDP plot is an aggregate representation of the relationship between all the predictor variables and the predicted response. It shows the average prediction over all observations for each possible value of a single predictor variable, while holding all other predictors constant at their median value. It provides information about the global relationship between the predictor variable and the response. In summary, ICE plot shows the relationship between a single predictor and the response at the individual level, while PDP plot shows the relationship at the aggregate level (Mandal et al., 2023). It is not advisable to use either an overly strict or an overly lenient threshold limit. This is because India and Southeast Asia do not have enough fertile land to meet the needs of their populations and using a permissive limit would not adequately protect human health from the risks of As.

The comparison of the predicted limit with the field data revealed that at certain instances the grain As concentration was > MTC. This is due to the fact as the models were trained on the data from an incubation study although monolithic soil columns were used for the purpose, still the extraneous variables (temperature, relative humidity, rainfall, surface run-off, cropping sequence, cropping system) that persist in real field conditions were not considered. The challenge in modelling was to enhance its generalizability so that it could be applied more broadly and not just limited to a specific set of data from a certain geographic region. This warrants further investigation collecting irrigation water, soil, rice grain across different cultivation systems (irrigated and rainfed), cropping systems, and also considering the soil parameters like pH, OC, available Fe, P, clay and total As to predict the limit of irrigation water. This will increase the model's generalizability so that its applicability is not limited.

7.6. Conclusion

The study investigated the effect of As dose on monolithic soil columns in two different irrigation systems (rainfed and irrigated). Results from the non-parametric test showed that the overall effect of As on the soil was statistically non-significant in both irrigation systems. However, a significant difference was observed between the As0 and As600 doses in the irrigated condition. The comparison of soil types in terms of BAs in post-incubated soil samples showed a statistically significant variation with respect to the soil types in both irrigation systems. Regression analysis showed that all soil parameters significantly affected BAs except

for AmAl. The spearman's correlation studies revealed a significant positive correlation between BAs and IrriAs, Tas, AvP, pH and clay, and a significant negative correlation with AvFe. The LR and LDA models were developed to predict the probability of BAs on the soil parameters and irrigation water As concentration. The LR performed better in terms of model matrices like accuracy, kappa, recall, precision, F1 score and MCC over the LDA. The limit of irrigation water As to classify soil As was found to be 190 μ L^{-1}using the PDP and ICE plots at a cut-off probability of 0.836 at the highest sensitivity and specificity. Further the LR model developed in this study provides a comprehensive understanding of the relationship between soil As levels and the predictor variables. The study highlights the importance of considering multiple soil parameters such as pH, OC, AvFe, AvP, Tas and clay in determining safe levels of As in irrigation water, and the need for further research to validate the findings in real-world conditions.

Chapter 8

Evaluation of biochar as an amendment for mitigation of arsenic contamination in rice through meta-analysis

8.1. Introduction

Management solutions that are both efficient and feasible (under local conditions) for the remediation of As-contaminated soil is of importance to reduce human health risks from soil-crop-food transfer. Both phytoremediation and bioremediation of the As contaminated sites have been undertaken (Laha et al., 2021 & 2022; Upadhyaya et al., 2018 and Mondal et al., 2021). Solutions include amendments used for the remediation of As-contaminated soil such as the use of inorganic elements like phosphorus (Hossain et al, 2009), silicon (Ma et al., 2008), iron (Ultra et al., 2009) and selenium (Wan et al., 2018); and complexation of As by the application of organic amendments such as sugarcane bagasse (Mandal et al., 2019a and 2019b) and vermicompost (Sengupta et al., 2021; 2022). One of the limitations of organic amendments, however, is that they need to be applied in large quantities.

Biochar is an effective amendment in reducing the ecotoxicity of soils that are contaminated with heavy metals because it can effectively bind metal(loid)s in water and immobilise them in soil (Guo et al., 2020; Ahmad et al., 2014; Tan et al., 2015; O'Connor et al., 2018). Biochar is prepared by carbonising organic materials through pyrolysis at high temperatures (between 300 and 1000 °C) with little or no oxygen (Lehman and Joseph, 2015). The surface functional groups on biochar, such as hydroxyls, carbonyls and carboxyls, serve as binding sites for metal(loid)s (Tan et al., 2015). Due to the presence of negatively charged surface functional groups, biochar can electrostatically bind heavy metal cations and adsorb them. Additionally, the electron-rich aromatic biochar surface may electrostatically draw electron-deficient metal cations to itself through donor-acceptor interactions (Vithanage et al., 2017). The encouraging results of many studies regarding the efficacy of biochar in binding contaminants have stimulated much interest in using it as a soil amendment for environmental rehabilitation (Guo et al., 2020). At present, most biochar research publications looked at its use from a technical or economic perspective in relation to soil quality and the remediation of surface-, ground-, and waste-water. An integrated understanding of the mechanisms of remediation of As-contaminated soils (specifically in the rice rhizosphere) through pristine and

modified biochar to improve the functional properties of biochar could result in future larger-scale applications.

In fact, crop residues are one of the most common feedstocks for biochar production. The main crops that generate crop residues in India are rice, wheat, sugarcane, maize, and millets. Rice crop alone contributes 34% to the crop residues. According to the Indian Ministry of New and Renewable Energy (MNRE), India generates on an average 500 Million tons (Mt) of crop residue per year. The same report shows that a majority of this crop residue is in fact used as fodder, fuel for other domestic and industrial purposes (Bhuvaneshwari et al.,2019). However, there is still a surplus of 140 Mt out of which 92 Mt is burned each year. The burning of crop residues is a major environmental problem in India. It contributes to air pollution, including the formation of smog and PM2.5, which can have serious health consequences. It also releases greenhouse gases into the atmosphere, contributing to climate change (Bhuvaneshwari et al., 2019). The management of crop residues in India is a complex issue. There is no single solution that will work for all areas. However, by using a combination of the methods (composting, biogas production, biochar production), it is possible to reduce the environmental impacts of crop residue burning and improve the sustainability of agriculture in India. Overall, the use of crop residues for biochar production is a promising option for managing crop residues in India and improving the sustainability of agriculture (Yrjälä et al., 2022). Biochar has the potential to solve the problem of crop residue management as well as the management of As-contaminated soils when used as an amendment which is required to be explored.

The aim of this systematic review is to examine recent research on the use of biochar for the removal and/or immobilisation of As in paddy soils. A meta-analysis using random effect model (REM) and forest plots have been undertaken to evaluate the prospect of using biochar as an amendment for As-contaminated paddy soils. The efficacy of different biochars (pristine and modified) in reducing As content in rice and soil is compared and some insight into the mechanisms involved in As immobilisation in the rice rhizosphere is provided. Finally, the future scope and direction of biochar research specifically for As-contaminated paddy soils is underlined.

8.2. Materials and Methods

We systematically reviewed published articles reporting the use of biochar to remediate As in rice. We used Boolean operators (e.g., “OR” and “AND”) to develop search terms from the keywords (“arsenic,” “soil,” “biochar,” “rice”; “arsenic,” “soil,” “biochar,” “paddy”). Searching ISI Web of Science and PubMed with these terms, we identified relevant research papers published from 2006, since the term "biochar" was first formally used in 2006 (Lehman et al., 2006). The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flowchart is presented in Figure 8.1. Studies were only included in the meta-analysis if (1) the research was carried out using rice as the test crop; (2) information regarding the characteristics of the biochar has been reported; (3) As concentration in the soil, rice grain has been reported; (4) details of the analytical method(s) and quality assurance procedures used for the study was provided. Based on the inclusion criteria 23 studies that were selected from the published literature. Selected studies were used to gather data on the sample size and mean values of As concentrations in rice and soil, tiller number, plant height, leaf, grain, shoot and root biomass. We quantitatively combined the data from the individual research papers to assess the relative influence of pristine and modified biochar on As immobilisation. The data were converted to the same units for comparisons. The final finding was expressed as a weighted mean difference between each group of individual studies and the marginal level of each parameter under study at a confidence level of 95%. As a result, the sample size determined how much weight was to be given to each study. The REM was used for this purpose. A REM model is a statistical method used in meta-analysis to combine data from multiple studies that have a common research question. It is a type of meta-analysis model that accounts for both within-study variation and between-study variation in estimating the overall effect size. In a REM model, it is assumed that the true effect size varies across studies, and the observed effect sizes are drawn from a normal distribution with a mean equal to the true effect size and a variance that includes both within-study and between-study variation (Romano et al., 2006). This means that each study has its own true effect size, but these true effect sizes are not identical across studies, and the variation in true effect sizes is assumed to follow a normal distribution (Jeffery et al., 2016). Observed study estimates vary not only due to random sampling error but also due to inherent differences in the way studies have been designed and conducted (Langan, 2022). To summarise, the data from individual studies in the meta-analysis, and to give a visual indication of the degree of heterogeneities, a forest

plot was developed. The absence of difference between the study group and the marginal level, also known as the no effect or zero effect line (mean difference was zero at this point), was shown by a vertical line in the plot's centre. The subsequent squares represent the mean difference values for each study and the size of the squares indicates the effect of the estimate and the weight of the studies. Each horizontal segment's succeeding endpoints showed 95% confidence intervals (CI) that were symmetrical about the mean. When all the diverse studies were combined and averaged, the diamond represents the point estimate and confidence intervals (Langan 2022). The data analysis was performed in R-Studio (*version 1.3.1093 2.3.1*) using the '*metafor*' package (*version 3.8-1*) (Viechtbauer et al., 2010).

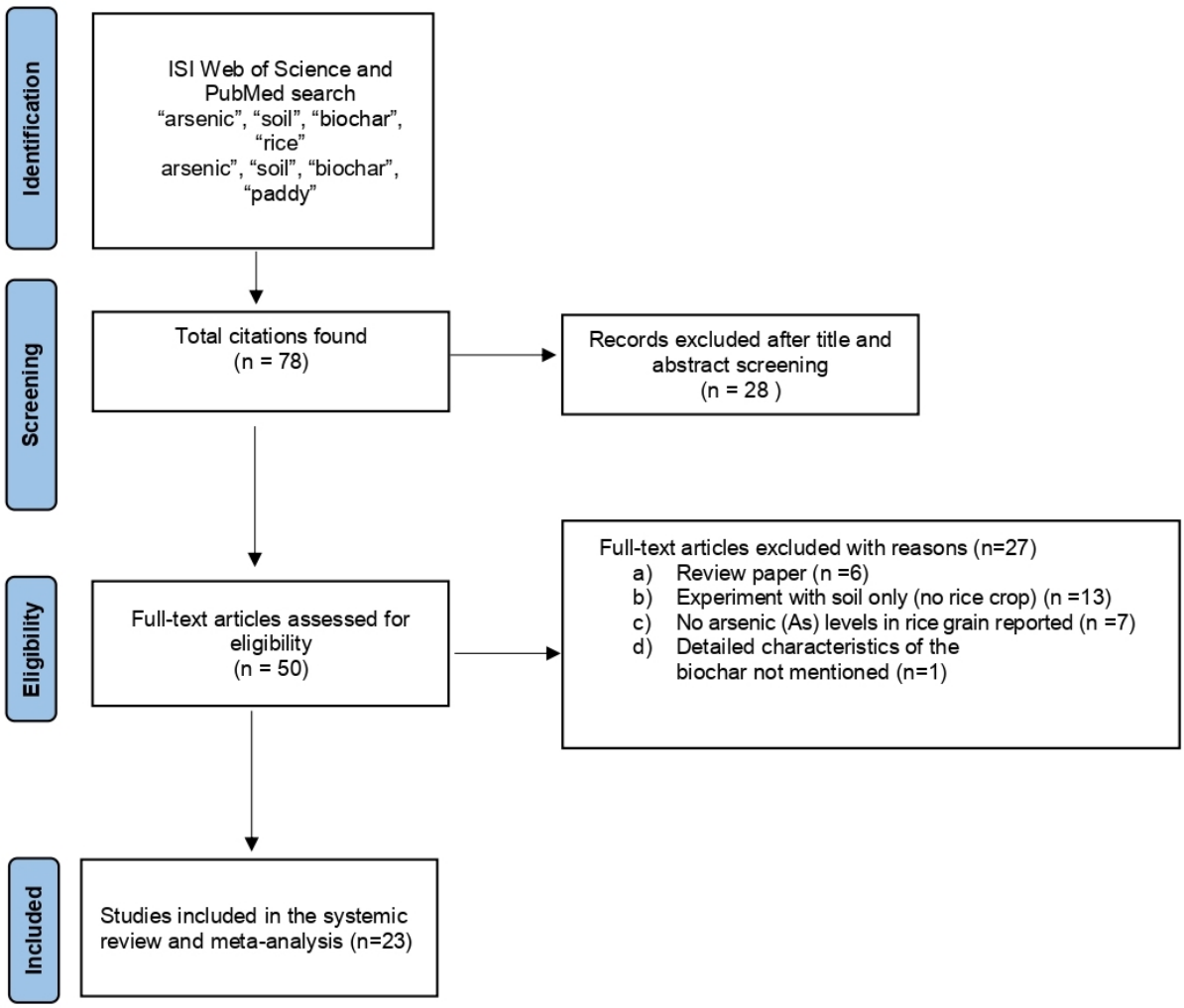


Figure 8.1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the selection of studies eligible for a meta-analysis.

8.3. Results

8.3.1 Details of the studies and properties of the biochar

The details of the biochars (pristine and modified) used for the remediation of the As-contaminated paddy soil are summarised in Table 8.1. Of the 23 studies selected, 20 studies were undertaken in China, two in Australia and one in Thailand. All the studies were pot experiments except for two (Pan et al., 2019 and Zheng et al. 2015) which were field-based experiments. The raw materials used for the preparation of biochar mostly included rice plant parts (straw, husk and hull) (nine studies) followed by maize/corn plant parts (straw and cobs) (seven studies). Other feedstocks for biochar included palm plant parts (shell and fibres) (three studies), wheat straw (two studies), eggshells (two studies), sewage sludge (two studies), bamboo (one study), oriental pane (one study), bean stalks (one study) and cordgrass (one study). The materials used for biochar modification included iron (Fe) (six studies); goethite (two studies); zerovalent Fe (one study); nano zerovalent Fe (one study); hydroxyapatite and zeolite (one study); silicon (Si) with nano-montmorillonite (one study); a combination of Fe, manganese (Mn) and cerium (Ce) (one study); a combination of Fe and Mn (one study); a combination of Fe, Mn and lanthanum (La) (one study); a combination of Fe, Mn and lanthanum (La) and manganese oxide (one study); and mixing of two pristine biochars (one study). The minimum temperature used for the preparation of biochar across all the studies was 450 °C while the maximum temperature was 800 °C with a mean of 591.30 $\pm$ 91.26 °C. The pH of the biochars used ranged from 3.17 to 11.30 with a mean of 8.20 $\pm$ 2.09. The carbon (C) content of the biochars ranged from 7.8 to 296.0 g kg^{-1} with mean value of 56.50 $\pm$ 43.0 g kg^{-1}. The cation exchange capacity (CEC) ranged from 11.5 to 73.0 cmol kg^{-1} and mean value was 31.04 $\pm$ 15.97 cmol kg^{-1}. The mean specific surface area (SSA) was 97.41 $\pm$ 87.64 m^2 g^{-1} and ranged from 2.76 to 276.24 m^2 g^{-1}. The mean pore volume was 0.1506 $\pm$ 0.00784 mL g^{-1} and ranged from 0.0144 to 0.2026 mL g^{-1}. The mean pore diameter was and 5.68 $\pm$ 2.76 mm and ranged from 4.02 - 10.6 mm respectively.

Table 8.1. Details of the biochar used for remediation of arsenic contaminated paddy soil.

Sl.No	Reference	Country	Raw material	Temperature (°C)	Biochar modification	**Biochar name	pH	C (g kg^{-1})	CEC (cmol kg^{-1})	***SSA (m^2 g^{-1})	Pore volume (mL g^{-1})	Pore diameter (nm)
1	Gu et al. (2018)	China	rice straw	500	Hydroxyapetite, Zeolite	Biochar	6.13	57.63	28.93	49.57	---	---
						HZB	7.27	21.24	73.0	---	---	---
2	Herath et al. (2020)	China	rice husk	700	Silicon, Nano-montmorillonite	RHBC	9.98	55.03	---	187.7	0.1968	4.317
						Si-RHBC	10.41	53.65	---	182.3	0.1968	4.317
						NM-RHBC	9.94	39.86	---	189.6	0.1972	4.159
3	Irshad et al.(2020)	China	wheat straw	600	Goethite	Biochar	7.45	---	---	44.97	---	---
						GB	7.99	---	---	276.24	---	---
4	Irshad et al.(2022)	China	wheat straw	600	Goethite	Biochar	7.45	---	---	44.97	---	---
						GB	7.99	---	---	276.24	---	---
5	Islam et al. (2021a)	China	egg shell, corn cobs	450	Mixing (1:1)	EB	9.58	13.1	---	6.5	---	---
						CB	9.83	81.6	---	75.2	---	---
						ECB	9.66	49.35	---	40.9	---	---
6	Islam et al. (2021b)	China	egg shell, corn cobs	450	Iron	FCB	8.71	62.6	---	205.8	---	---
						FEB	8.65	7.8	---	25.3	---	---
						FCEB	8.67	35.3	---	115.5	---	---
7	Jin et al. (2020)	China	rice husk, codgrass	600	---	RBC	11.3	---	---	---	---	---
						SBC	10.5	---	---	---	---	---
8	Khan et al. (2013)	China	sewage sludge	550	---	SSBC	7.22	28.0		5.50	0.0144	10.5
						SSBC5%	4.86	14.40	---	---	---	---
						SSBC10%	5.39	20.20	---	---	---	---
9	Khan et al. (2014)	China	sewage sludge	600	---	SSBC	7.18	27.8		5.57	0.015	10.6
						SSBC5%	5.66	30.5	---	---	---	---
						SSBC10%	5.83	51.5	---	---	---	---
10	Kumarathilaka et al.(2021a)	Australia	rice hulls	600	Iron	RBC	9.81	---	---	201.39	0.2026	4.024
						Fe-RBC	5.33	---	---	142.6	0.1650	4.6285
11	Kumarathilaka et al.(2021b)	Australia	rice hulls	600	Iron	RBC	9.81	---	---	201.39	0.2026	4.024
						Fe-RBC	5.33	---	---	142.6	0.1650	4.6285
12	Leksungnoen et al. (2019)	Thailand	rice husk	750	ash, acid wash	RHB	7.3	37.92	26	---	---	---
						RHA	9.3	31.03	29	---	---	---
						AWB	3.2	31.03	---	---	---	---
13	Lian et al.(2020)	China	corn straw	600	Iron, Manganese,	BC	8.91	84.92	38.4	60.16	---	---
						FMCBC1	9.41	63.25	43.28	27.16	---	---

					Cerium	FMCBC2	9.64	56.27	43.86	36.21	---	---
						FMCBC3	9.72	43.48	44.52	46.87	---	---
14	Lin et al. (2019)	China	corn straw	600	Iron, Manganese	BC	8.93	75.5	---	61	---	---
						FMBC1	9.60	67.3	---	208	---	---
						FMBC2	3.17	53.8	---	7.53	---	---
15	Lin et al. (2020)	China	corn stem	600	Iron, Manganese, Lanthanum	BC	8.93	75.5	---	61	---	---
						FMLBC1	6.96	62.3	---	2.76	---	---
						FMLBC2	6.83	54.8	---	12.2	---	---
						FMLBC3	6.9	47.1	---	30.6	---	---
16	Liu et al. (2017)	China	palm shell	500	Nano zerovalent Iron	---	---	296	---	244	---	---
17	Lv et al. (2021)	China	rice straw, corn straw, bamboo	600	---	RSBC	11	47.1	---	175.5	---	---
						CSBC	10.5	60.3	---	3.6	---	---
						BABC	9.6	87.3	---	9.2	---	---
18	Pan et al. (2019)*	China	palm fibres	800	Iron	Fe-BC	6.0	60	---	---	---	---
19	Qiao et al. (2018)	China	palm fibres	700	Zerovalent Iron different levels	Biochar	9.3	86.7	11.5	241.6	---	---
20	Wen et al. (2021)	China	Oriental plane	650	Iron	RawBC	9.25	69.34	21.59	110.7	---	---
						FeBC	4.41	59.91	16.70	74.5	---	---
21	Yin et al. (2017)	China	rice straw	450	Iron	Biochar	10.7	---	15.1	---	---	---
						Fe-Biochar	4.87	---	14.2	---	---	---
22	Yu et al. (2017)	China	corn straw	600	Manganese Oxide	BC	10.4	85.3	---	60.9	---	---
						MBC	10.8	73	---	3.18	---	---
23	Zheng et al. (2015)*	China	bean stalk, rice straw	500	---	BBC	9.2	44.5	27.5	---	---	---
						RBC	10.5	27.4	32.1	---	---	---
Mean±SD				591.30±91.26	--------	--------	8.20±2.09	56.50±43.0	31.04±15.97	97.41±87.64	0.1506±0.0784	5.68±2.76
Range (Minimum-Maximum)				450-800	--------	--------	3.17-11.3	7.8-296.0	11.5-73.0	2.76-276.24	0.0144-0.2026	4.02-10.6

*Indicates field experiment

**Biochar name: HZB: hydroxiapatite zeolite biochar, RHBC: rice husk biochar, Si-RHBC: silicon-RHBC, NM-RHBC: nano-montmorillonite-RHBC, GB: goethite biochar, EB: egg shell biochar, CB: corn cob biochar, FCB: iron-CB, ECB: iron-CB, SSBC: sewage sludge biochar, RHA: rice husk acid wash, FMBC: iron-manganese-cerium biochar, FMLBC: iron-manganese-lanthanum biochar, BBC: bean stalk biochar, RBC: rice straw biochar

***SSA: Specific Surface Area

8.3.2 Effect of biochar on As content in rice grain

The continuous REM revealed a significant ($p < 0.001$) weighted mean value of −539.21 (95% CI: -663.15 to -415.27) (Figure 8.2). The impact of the negative value signified that the estimated reduction of As concentration in rice grain was statistically significant with respect to the control. From the mean different effect sizes of the different studies, it can be observed that the biochars prepared from sewage sludge (pristine and modified) and maize straw, irrespective of the dose, significantly ($p < 0.001$) reduced the grain As content as the subsequent confidence intervals did not overlap the zero-effect line. Pristine biochar prepared from bean stalks and rice straw (Zheng et al., 2015) was not effective in reducing the grain As content as observed in field experiments. Pristine biochar refers to the original or untreated form of biochar, which is a carbon-rich material derived from biomass such as wood, crop residues, or organic waste through a process called pyrolysis. Modified biochar refers to biochar that has undergone specific treatments or modifications to enhance its properties or tailor it for specific applications. These modifications can involve physical, chemical, or biological processes that alter the structure, surface chemistry, or properties of the biochar (Vithanage et al., 2017). The sewage sludge and maize straw feedstock were more effective in reducing the grain As content. The inconsistency index of 99.98% indicated substantial and significant heterogeneity ($p <0.001$).

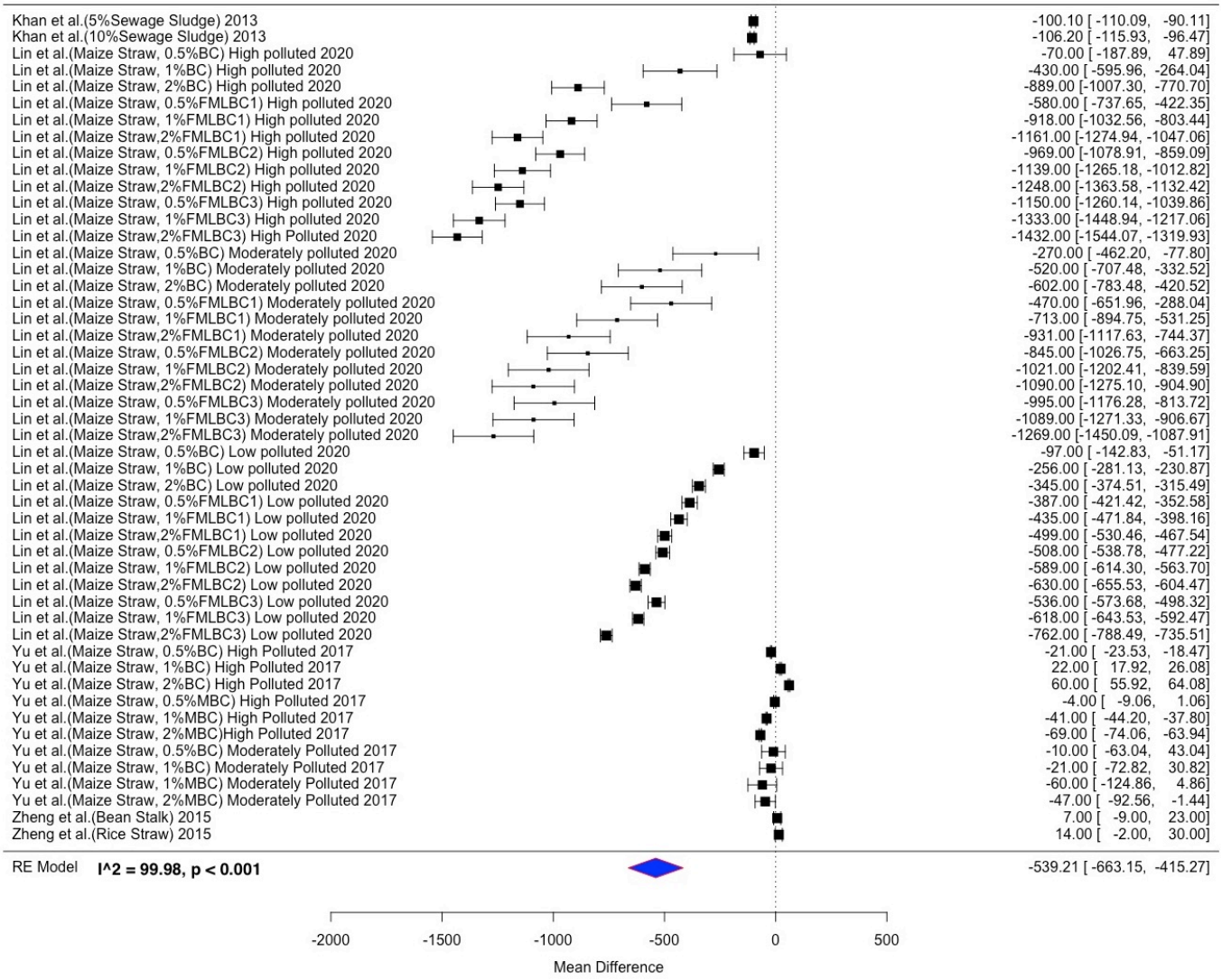


Figure 8.2. Forest plot showing the effect of biochar on the weighted mean difference of arsenic concentration in rice grain (μg kg^{-1}) between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics. (BC: biochar; MBC: manganese oxide biochar composites; BC:Fe:Mn:La at different weight ratios: (FMLBC1) 25:4:1:1, (FMLBC2) 25:4:1:3, and (FMLBC3) 25:4:1:5)

8.3.3 Effect of biochar on plant parameters

From Figure 8.3a it can be observed that overall application of biochar resulted in a significant ($p < 0.001$) increase in plant height with a weighted mean value was 7.51 cm (95% CI 3.39 to 11.63), though having 98.91% heterogeneity ($p < 0.001$). The sewage sludge, wheat straw (pristine and modified) and rice hull (modified) biochar significantly increased plant height. The overall effect of biochar on tiller number had a weighted mean value of -1.91 (95% CI -7.14 to 3.31) and was not statistically significant ($p > 0.001$) but indicated having heterogeneity

of 98.96% (p <0.001) (Figure 8.3b). Only the sewage sludge biochar significantly increased the tiller number irrespective of the doses.

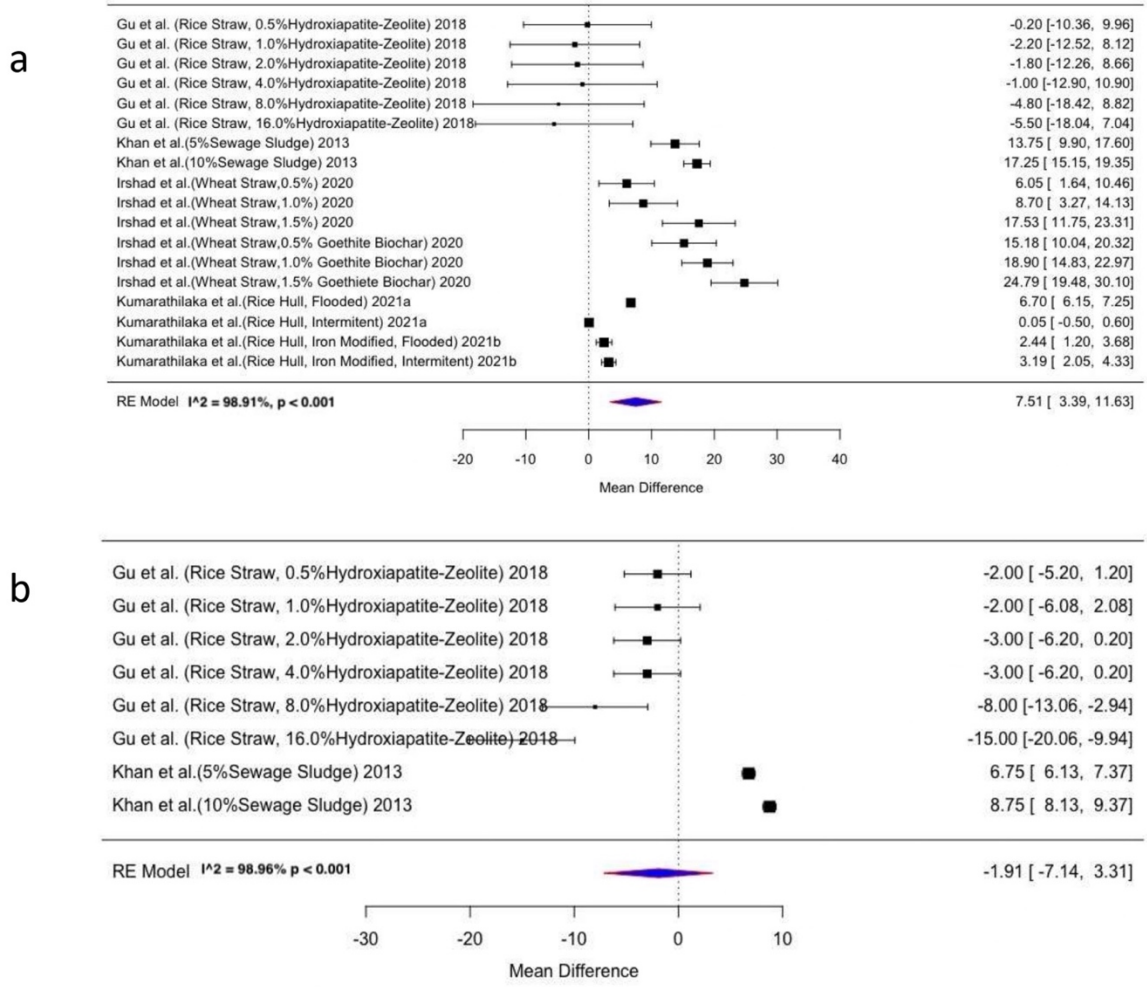


Figure 8.3. Forest plot showing the effect of biochar on the weighted mean difference of (a) plant height (cm) and (b) tiller number of rice between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics.

Figure 8.4 (a and b) revealed that biochar significantly (p < 0.001) increased the root (weighted mean value: 0.78, 95% CI: 0.59 to 0.98) and shoot (weighted mean value: 2.15, 95% CI: 1.30 to 2.99) biomass. Similar to previous meta-analysis, significant heterogeneity (p < 0.001) was observed at 94.69% and 97.56% respectively. The wheat straw (pristine and modified), maize straw (modified), and oil palm fibre (modified) biochars significantly increased both the root and shoot biomass of rice plants.

The effect of biochar on the leaf and grain biomass is presented in Figure 8.5 (a and b). Biochar application significantly (p < 0.001) increased the leaf biomass (weighted mean value: 1.54, 95% CI: 0.95 to 2.12) and grain biomass (weighted mean value: 3.59, 95% CI: 2.48 to

4.71). Again, significant heterogeneity of 97.04% and 99.31% ($p < 0.001$) respectively, were noted. The maize straw (pristine and modified) was effective in increasing the leaf biomass whereas the sewage sludge (pristine), maize straw (pristine and modified), and oil palm fibre (modified) biochar increased the grain biomass significantly.

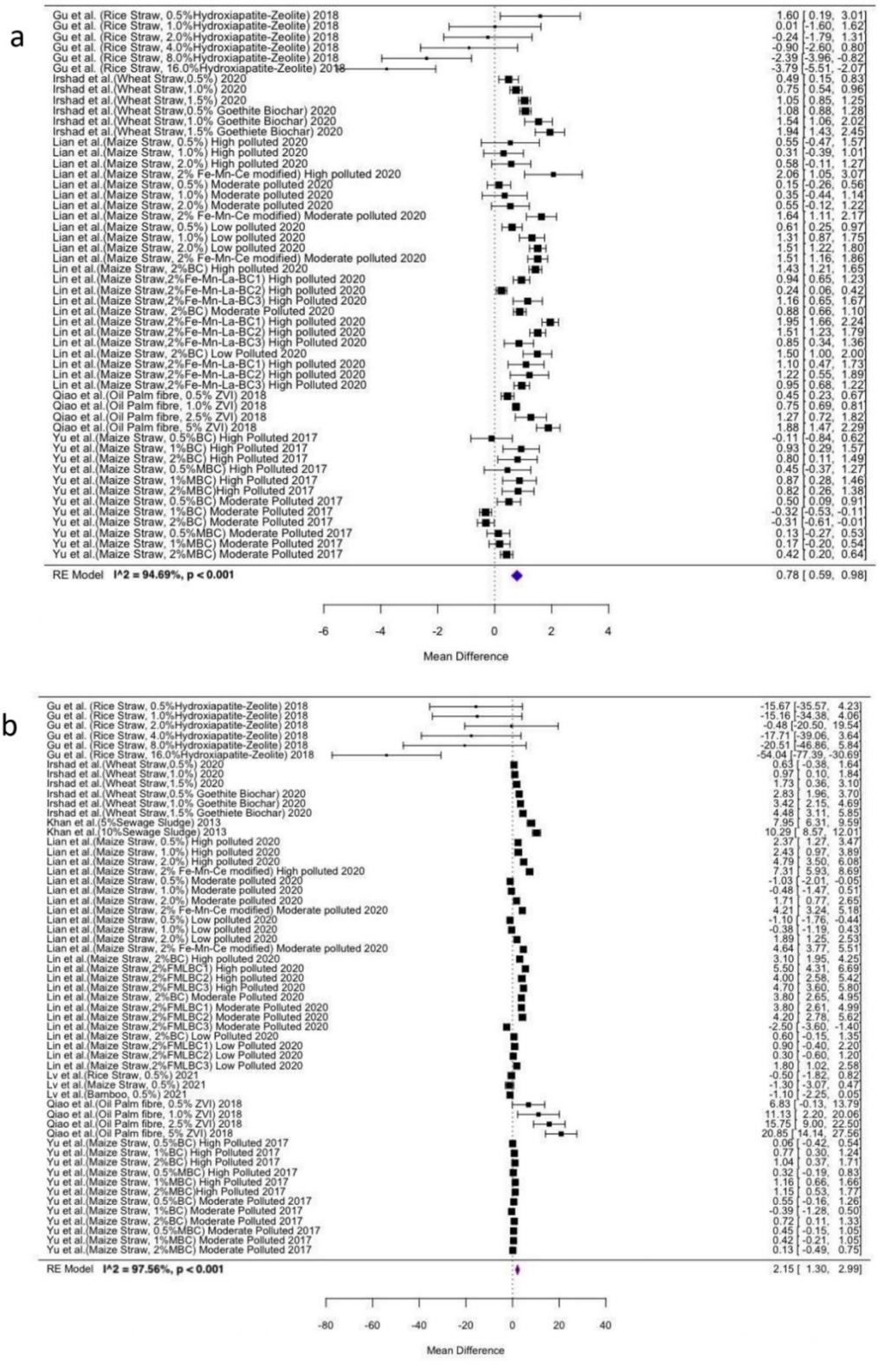


Figure 8.4. Forest plot showing the effect of biochar on the weighted mean difference of rice (a) root biomass (g) and (b) shoot biomass (g) between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics.

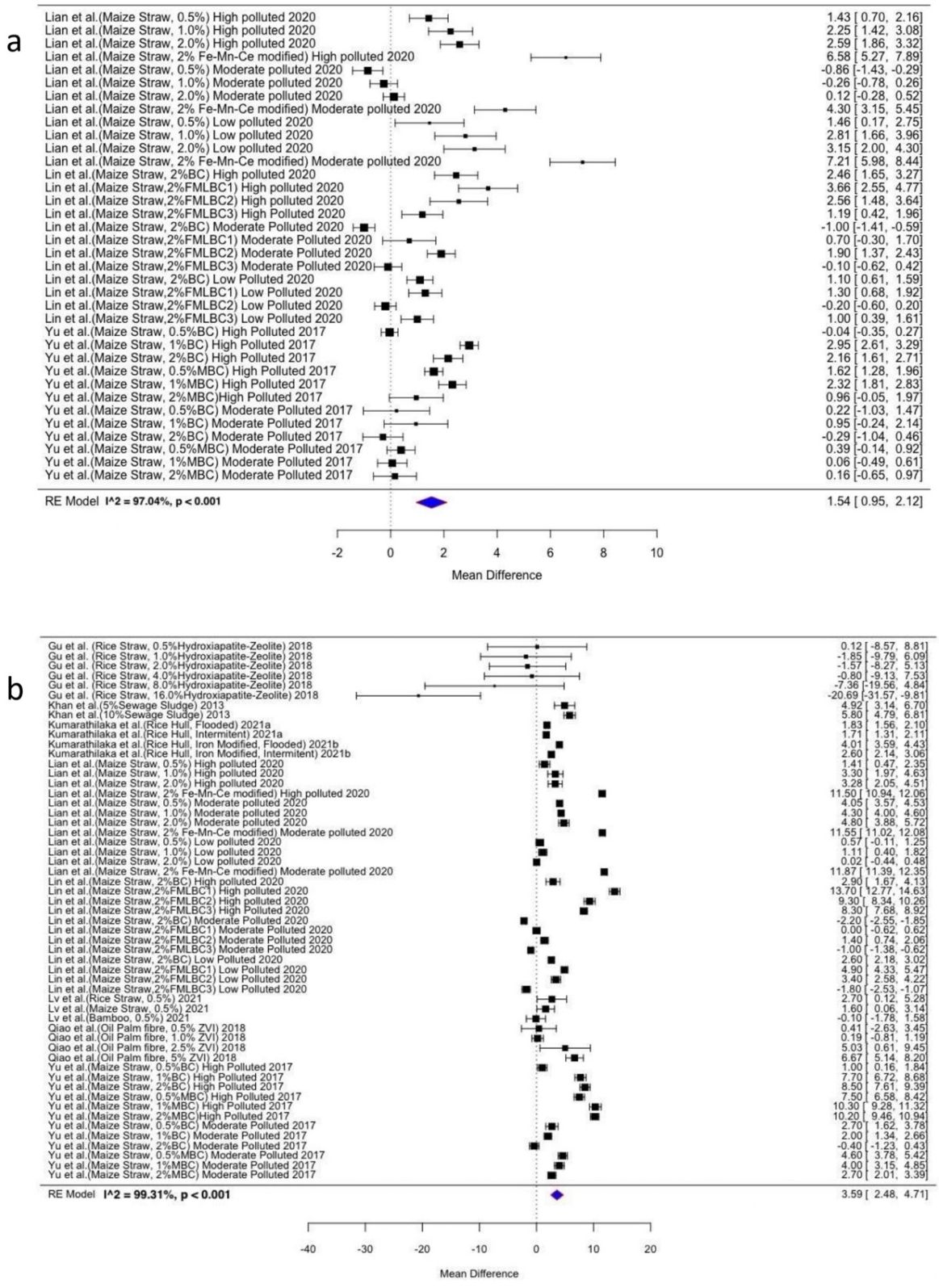


Figure 8.5. Forest plot showing the effect of biochar on the weighted mean difference of rice (a) leaf biomass (g) and (b) grain biomass (g) between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics.

8.3.4 Effect of biochar on fractions of soil As

Figure 8.6 (a and b) represents the effect of biochar on the exchangeable and Al-bound As fractions in the rice rhizosphere. The application of biochar significantly ($p < 0.001$) reduced exchangeable As (weighted mean value: -0.04, 95% CI: -0.06 to -0.02) and Al-bound As (weighted mean value: -3.17, 95% CI: -4.15 to -2.18). Both analysis had significant heterogeneity of 99.96% and 99.73% respectively. Maize straw (pristine and modified), and rice straw (pristine and modified) significantly reduced the exchangeable As and the same trend was observed for Al-bound As fraction.

The effect of biochar on the Fe and Ca bound As fractions in the rice rhizosphere is presented in Figure 8.7 (a and b). Both Fe-bound (weighted mean value: 1.40, 95% CI: 0.74 to 2.05) and Ca-bound (weighted mean value: 1.69, 95% CI: 1.13 to 2.26) As increased significantly ($p < 0.001$). Significant heterogeneity of 95.24% and 99.21% respectively was observed. The rice straw (pristine and modified) and maize straw (pristine and modified) biochar increased the Ca-bound fraction of As and only maize straw (pristine and modified) biochar increased the Fe-bound As fraction resulting in reducing the exchangeable or available fraction of As in soil.

Only two of the studies (Pan et al., 2019 and Zheng et al., 2015) under consideration in the meta-analysis were undertaken in field conditions. Pan et al., (2019) reported that Fe-modified palm fibre biochar was effective in reducing the availability of As in soil. A conflicting report of As mobilisation in rice soils due to the application of pristine rice straw and bean stalk biochar was reported by Zheng et al. (2015).

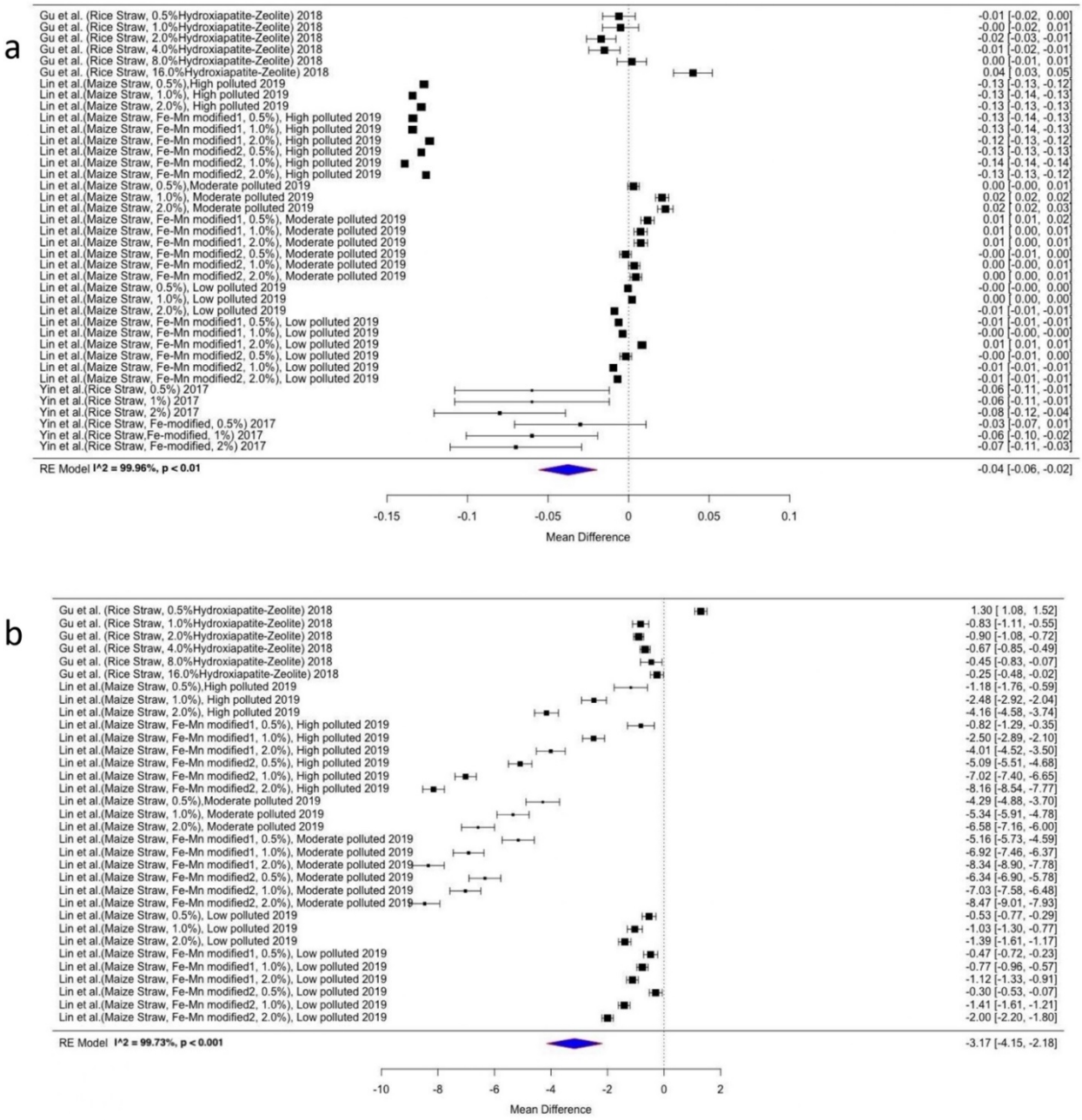


Figure 8.6. Forest plot showing the effect of biochar on the weighted mean difference of soil (a) exchangeable As (mg kg^{-1}) and (b) Al-bound As (mg kg^{-1}) in rice rhizosphere between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics.

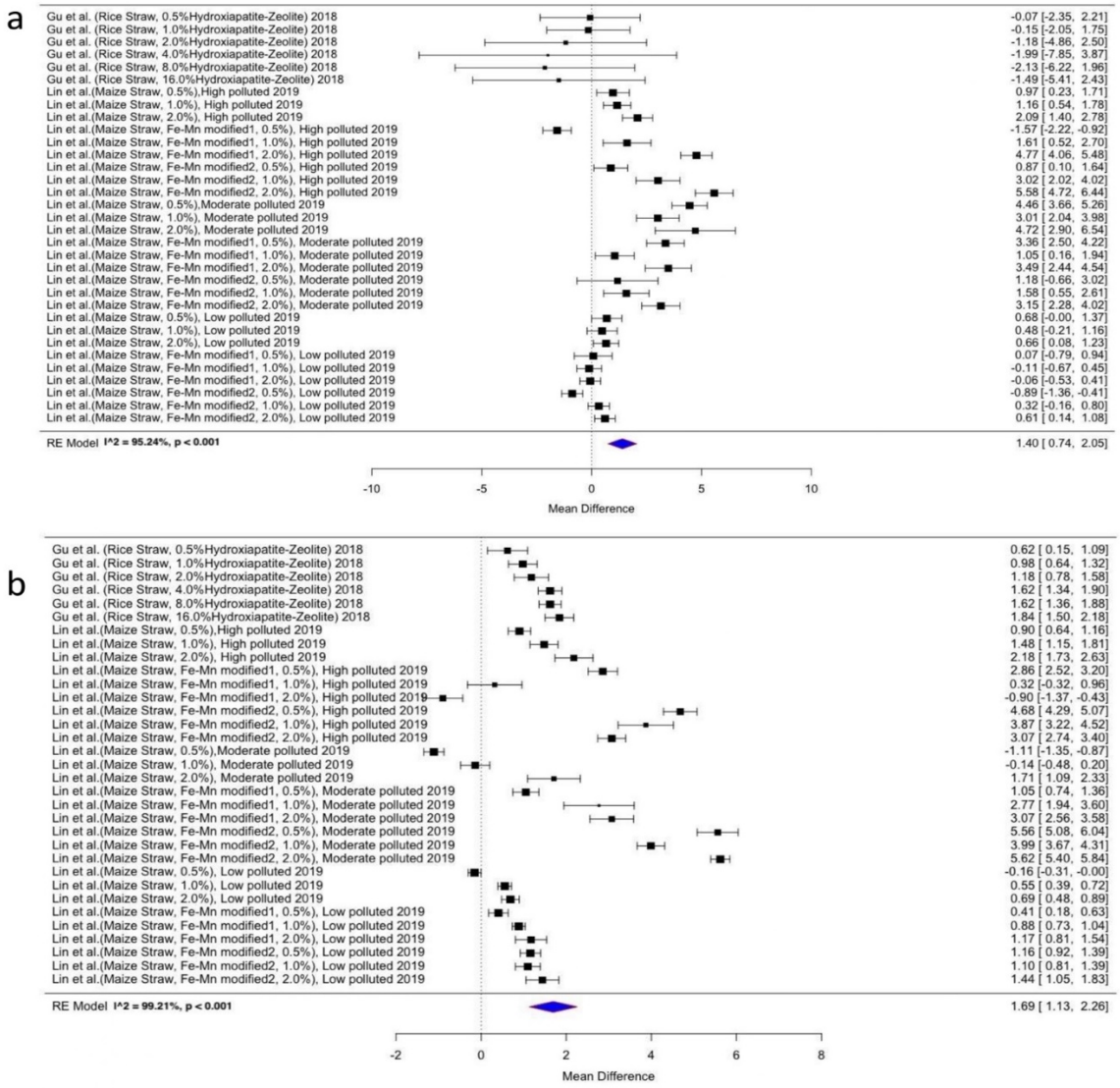


Figure 8.7. Forest plot showing the effect of biochar on the weighted mean difference of soil (a) Fe-bound As (mg kg^{-1}) and (b) Ca-bound As (mg kg^{-1}) in rice rhizosphere between the different studies with their respective confidence intervals and weight in the meta-analysis together with the heterogeneity statistics.

8.4. Discussion

8.4.1 Effect of biochar properties

A high pyrolysis temperature has been used in all the studies considered in this review. The yield and characteristics of biochar were determined based on thermochemical methods, operating conditions, and feedstock. It is widely known that low-temperature-generated biochars from slow pyrolysis have low hydrophobicity and aromaticity but significant surface acidity and polarity. Major biomass decomposition occurs between 200 °C and 500 °C through a series of phases that include partial hemicellulose decomposition, complete hemicellulose

decomposition, full cellulose, and partial lignin decomposition (Rutherford et al., 2012). The elemental composition of biochar varies with pyrolytic temperature and is dependent on feedstock type. Carbon concentration increases as production temperature rises, but the concentrations of nitrogen (N), sulphur (S), hydrogen (H), and oxygen (O), which are components of the gaseous products during pyrolysis, fall. Biochars made from biosolids, and animal manures are often high in N, phosphorus (P), potassium (K), and S (Ahmad et al., 2014; Jin et al., 2014). The pyrolytic temperature and feedstock content influence physical qualities such as pore structure, surface area, and adsorption properties (Vithanage et al., 2017). Volatile chemicals in the biomass tend to be eliminated from the medium as the pyrolytic temperature rises. This increases surface area and ash content while reducing surface functional groups and exchange sites. As the pyrolytic temperature rises, aliphatic C species are transformed into aromatic rings, generating a graphene-like arrangement that enhances the pore volume, pore distribution and surface area of the biochar (Ahmad et al., 2014). Biochars with a high concentration of C in condensed aromatic rings have few functional groups. Surface functional groups are important in biochar adsorption capacity, and the amount and type of functional groups vary depending on the feedstock and pyrolytic temperature (Kim et al., 2013). High-temperature pyrolysis carbonisation is appropriate for forestry and agricultural wastes with higher levels of lignin, cellulose, and hemicellulose (Labanya et al., 2022). Use of a high pyrolytic temperature (>450 °C) for preparation of biochar has been observed in all the studies considered in this review. In a meta-analysis, Arabi et al. (2021) reported that low pyrolysis temperature biochars (≤ 450 °C) did not affect As mobility in the soil, but high pyrolysis temperature biochars (> 450 °C) considerably mobilised the soil As. Biochars pyrolysed at high temperatures are more successful than those generated at low temperatures for As immobilisation, which could be attributed to the high aromaticity and porous structure, as well as the presence of mineral-phases (e.g., $CaPO_4$, $CaCO_3$) (Amen et al., 2020). Heating the biomass to temperatures ranging from 350 °C to 650 °C breaks down and reorganises the chemical links, resulting in the formation of new functional groups (e.g., carboxyl, lactone, lactol, quinine, chromene, anhydride, phenol, ether, pyrone, pyridine, pyridone, and pyrrole) (Mia et al. 2017). High-temperature biochar (600-700 °C) has a very hydrophobic character with well-organised C layers (Uchimiya et al. 2011). However, due to dehydration and deoxygenation of the biomass, it has lower amounts of H- and O-containing functional groups (Ahmad et al. 2014; Uchimiya et al. 2011). The surface groups can function

both as electron acceptors or donors. This results in the formation of simultaneous zones with features varying from acidic to basic and hydrophilic to hydrophobic (Amonette and Joseph 2009). Biochar produced at lower temperatures (300-400 °C) has a more diverse organic character due to the presence of aliphatic and cellulose type arrangements (Novak et al. 2009). As the pyrolysis temperature rises, the structure of biochar appears to have more orderly C layers (e.g., graphene-like structure) and less content of surface functional groups (Ahmad et al. 2014).

The presence of acidic groups such as carboxylic, phenolic, and cationic groups on the surface of biochar produced at a shorter pyrolysis duration result in biochar with a relatively low pH (Shaaban et al., 2014). According to Beiyuan et al. (2017), biochar generated at low temperature has a greater O/C ratio than biochar produced at high temperature. These findings suggested that As stabilisation may be significantly aided by O-containing functional groups, such as carboxyl and hydroxyl groups (Shaaban et al., 2014). In contrast, a little higher pH in the higher temperature pyrolysis likely led to increase As mobilisation (Beiyuan et al., 2017; Zhao et al., 2018). However, in all the studies considered in this review, the biochars used were prepared at high temperature resulting in high pH of the biochar in most of the cases. The pH of the medium (i.e., soil solution) can affect the charge characteristics of the biochar surface as well as As speciation, but not the pH of biochar. For example, depending on the solution pH, various functional groups such as amine, alcohols, carboxylic, on the surface of biochar tend to be protonated, hence altering the surface charge of biochar (Vithanage et al., 2017). Contrarily, the pH of the solution affects the speciation of As into its many neutral and anionic forms such as H_3AsO_4, H_2AsO_4, $HAsO_4^{2-}$, and AsO_4^{3-}. At pH 3-6, the $H_2AsO^-_4$ species can dominate; however, at pH 8 and above, $HAsO_4^{2-}$ and AsO_4^{3-} species become dominant (Kumari et al., 2021; Raj et. al., 2020). Consequently, multiple species of As can be adsorbed on the surface of biochar at different pH values, making it difficult to determine the predominant species of As. The dominant As species on the surface, as well as how the As-surface complexation takes place, varies with changing solution pH.

Owing to the existence of many exchange sites on biochar surfaces, the application of biochar often increases soil CEC (Mohamed et al., 2015; Moreno-Jim enez et al., 2016; Zhang et al., 2017), which promotes heavy metal immobilisation in soil. CEC has no influence on As mobilisation as it is a metalloid and can easily generate anions, and its non-metal characteristics predominate. Furthermore, as per the meta-analysis by Arabi et al., (2021)

demonstrated that biochar immobilises Cr and Ni in soil but was not effective in immobilising As in soil.

Numerous investigations have revealed that the surface area of biochars rises with pyrolysis temperature. Increased pyrolysis temperature was thought to improve lignin and cellulose degradation in feedstocks and remove H- and O-containing functional groups, resulting in an increase in biochar surface area (Phuong et al., 2016). In the studies considered in this review, biochars made from rice biomass (Herath et al.,2020; Kumarathilaka et al., 2021a and 2021b; Lv et al., 2021) at high temperature had a higher surface area. As per the findings of Jeong et al. (2012), softwood biochar had a lesser surface area compared to hardwood biochar with 159 m^2 g^{-1} for softwood biochar as against 242 m^2 g^{-1} for hardwood biochar.

According to Steiner et al. (2016), the porous features and surface area are responsible not only for nutrients, organic compounds, and water retention, but also offer a favoured refuge for bacteria and other microorganisms. The surface properties of biochars can vary greatly depending on the biomass used and methods of production, such as pyrolysis temperature or duration, according to Verheijen et al. (2010). Large macropores (> 10nm diameter) are observed in biochars derived from wood due to the presence of large cells whereas cellulosic straw biochars have a pore size range of 1-10 nm due to the presence of thinner walls and channels. The biochars prepared from rice husk (Herath et al., 2021), and rice hulls (Kumarathilaka et. al., 2021a and 2021b) feedstock had a pore diameter of the range 1-10 nm. The As molecules are attracted to the surface of the biochar by the electrostatic forces between the negatively charged As ions and the positively charged functional groups on the biochar surface. The higher the surface area, the greater the capacity of the biochar to adsorb As. The pore size distribution of biochar refers to the size and distribution of the pores on the biochar surface. The pores provide a place for the As molecules to be held, and the size of the pores determines how well the As molecules can be held (Kumarathilaka et. al., 2021a and 2021b).

8.4.2. Effect of biochar on plant parameters

The surge in readily accessible nutrients and the enhancement of the rhizosphere environment are both responsible for the effect on plant parameters resulting in the increase of plant height, root, shoot, leaf and grain biomass of rice. The application of biochar improved

the soil's characteristics over time (such as pH, C content, available P, and available K), however, these improvements are somewhat dependent on the materials used to make the biochar. Increased soil carbon stocks, nutrient retention, soil fertility, and crop output can all be achieved with biochar (Chan et al., 2007; Lehmann et al., 2003; Novak et al., 2009; Steiner et al., 2007). The pyrolysis process and kind of feedstock have a significant impact on the properties of biochar. Numerous biomass sources, including crop by-products, shrubs, green trash, and even animal manures, can be used to make biochar. Biochar application alters soil pH (Kimetu et al., 2008; Novak et al., 2009; Rondon et al., 2007), and it may also have an indirect impact on how plants receive nutrients (Atkinson et al., 2010). Therefore, it seemed unlikely that pH variations would have significantly impacted nutritional availability. Different soil and biochar mixtures may have various pH buffering capabilities (Mukherjee et al., 2011). Numerous studies have demonstrated that adding biochar to soil can greatly increase its nutritional content (Liang et al., 2014 and Luo et al., 2014). This is due in part to the direct supply of nutrients, like P and K as well as a decrease in runoff and leaching (Enders et al., 2012). Previous findings demonstrated that nutrient concentrations in brown rice were raised by utilising soil amendments and improved the soil CEC and organic matter content due to retained and reduced leaching of nutrients (Ippolito et al. 2016). Improvements in soil physical properties brought about by adding pristine or modified biochar can mostly be attributed to the rise in root weight, grain weight, and biomass. This is because adding biochar to the soil reduces the bulk density of the topsoil, which also lowers the endurance to root growth into the soil profile. As a result, the growth of roots would undoubtedly boost the bioavailability of nutrients for rice in soils, further resulting in an increase in grain weight and biomass. On the other hand, when nutrients were not the main limiting factor, enhanced soil chemical characteristics, such as pH and cation exchange capacity (CEC), might also lessen the toxicity of As present in rice helping in the growth of the rice plants. However, the growth of roots, rice grains, and biomass did not increase along with the increase in biochar doses, indicating that adding too much biochar might not increase biomass and that the biochar level should be kept at optimum (Yu et al., 2017).

8.4.3. Mechanism of As immobilisation/mobilisation in biochar amended soil

Following the application of biochar in paddy soil, many crucial processes lead to the mechanisms of As immobilisation. The type of biochar used and the modifications made

determine the mechanisms of As immobilisation/mobilisation. The presence of numerous functional groups like alcoholic, phenolic, and carboxylic groups associated with the biochar can play the role of electron donors regulating the reduction of As(V) to As(III) as has been seen in soils treated with biochar (Choppala et al., 2016). Anionic forms of As predominate. Therefore, the functional groups that are carboxylic and phenolic on the surface of biochar particles might not have a strong attraction for As (Irshad et al., 2020). Modification of biochar may be a solution to this problem. Rice straw biochar modified with hydroxyapatite and zeolite increased the amount of Ca in soil which sequesters As from the exchangeable fraction into insoluble Ca-bound As (Gu et al., 2018). The presence of a large surface area and the number of pores in biochars modified with zeolite further fixed more As from the exchangeable phase. Additionally, modified biochar having more oxygen-containing functional groups (Chen et al. 2015) may covalently link to As (Hu et al. 2015) and fix the active As rendering it unavailable for crop uptake.

Modification of wheat straw biochar with goethite resulted in the restricted mobility of As in paddy soil due to the presence of Fe functional groups. Applications of Fe and Mn oxide residues have been shown in numerous prior studies to minimise As uptake in rice by controlling the mobility and bioavailability of As in the soil through dissolution and mineralisation (Jindo et al., 2016). The application of goethite-modified biochar may enhance the amount of iron oxide in paddy soil, which is essential for controlling the uptake of As by rice plants (Yu et al., 2017). Another potential explanation for the enhanced production of Fe-plaques is the raised pH of the soil following the application of goethite-modified biochar. After adding modified biochar to the soil, there was an increase in the rhizospheric Fe^{2+} concentration, which is thought to be the cause of the increased Fe-plaque formation. The findings demonstrated that amendment improved As sequestration by increasing Fe plaque formation. According to previous findings, adding Fe compounds to contaminated paddy soil raised the Fe content of the Fe-plaque on the surface of rice roots, which in turn played a significant effect in regulating the bioavailability and uptake of As (Yu et al., 2017). Similarly, the modification of rice straw biochar with Fe resulted in reducing the As bioavailability by the formation of Fe-plaque at the rice rhizosphere (Wen et al., 2021).

To control As movement in rice soils as well as the accretion of As in rice grains, the supplement of Si-rich amendments, such as chemical Si-minerals, Si-fertilizers, straw biomasses, rice husk and related chars has been reported previously (Bogdan and Schenk,

2008; Seyfferth and Fendorf, 2012). As(III) and Fe(II) are mostly released from hematite (-Fe_2O_3) and goethite (-FeOOH) in paddy soil, resulting in dissolved As(III) and Fe(II) being the dominating species in soil-porewater in anoxic situations. This implies that using these Si-impregnated biochar composites by Herath et al., (2020) can reduce the release of high As levels in the rice rhizosphere by limiting the dissolution of Fe minerals in the soil. In order to interact electrostatically with the positively charged Fe(II) ions in the rice rhizosphere, the O-SiO- groups linked to the surface of biochar are able to produce a high negative charge on the surface of the material. The dissolved O in soil-porewater and silicate groups present in biochar both quickly oxidise Fe(II) ions bonded on the biochar surface to Fe(III) at near-neutral pH levels. Fe(III) rapidly forms a Si-based ferrihydrite complex on the surface of the biochar by complexing with hydroxyl ions found in the soil-porewater. Eventually, As(III) and As(V) form a complex after being adsorbed on this ferrihydrite layer. Depending on the pH of the medium, As(III) can bind to ferrihydrite through inner- or outer-sphere complexation. Nearly neutral pH allows for the thermodynamically stable development of inner-sphere complexation, and As(III) inner-sphere complexes formed on ferrihydrite through bidentate interactions (Zhao et al., 2011).

The application of maize straw biochar modified with manganese oxide decreased As(III) and As(V) mobility and availability in rice both moderately and heavily polluted rice soils (Yu et al., 2017). Thus, the application of biochar enhanced Mn content in soils as compared to the control. Fe/Mn plaque is reported to have a greater affinity for arsenate than arsenite. Reduced mobility and bioavailability are the result of As(III) and As(V) in soil due to the formation of complexes with a variety of oxides, including Fe, Mn, and Al oxides.

The combination of zero-valent iron (ZVI) with oil palm fibre biochar resulted in a significant decrease in As bioavailability in rice soils (Qiao et al., 2018). The O_2 in the paddy soils can quickly oxidise the highly reactive reducing agent ZVI, converting it into amorphous iron oxyhydroxides. Fe^{2+} is produced through the reactions of ZVI and H_2O, which can then be further oxidised by biotic (iron oxidising bacteria, FeOBs) or abiotic processes. Amorphous iron oxyhydroxides, such as ferrihydrite, are then formed (am-FeOOH). In addition to offering a large number of new As surface adsorption sites, the newly produced iron minerals from the aforementioned ZVI retain As, increasing As immobilisation and decreasing its bioavailability.

Several earlier investigations employing pot experiments showed that adding biochar to the soil throughout rice cultivation increased the amount of dissolved As in soil pore water

and accelerated the conversion of As fractions, such as As combined into Fe oxides (amorphous and crystalline) reducing the bioavailability. Lv et al., (2021) reported that rice straw, corn stalks and bamboo-derived biochars exhibited a partial reduction (12-16%) in As accumulation in rice grains. The authors by microcosm-based anaerobic incubation studies, revealed that As levels in soil solution treated with biochar significantly rose, by 2.8–6.6 times, with the increase in biochar doses (0.5–5%, w/w), especially at higher concentrations (3–5%, w/w). Fe and As were shown to be significantly positively correlated during rice culture ($r2 = 0.73$, $p < 0.001$), suggesting that the microbially mediated reductive breakdown of Fe (oxyhydr-) oxides may be the primary cause of the release of As during rice cultivation (Somenahally et al., 2011; Wu et al., 2020). Arsenic concentrations in pore water were higher in the biochar-treated pots compared to the control during the growth period (Wu et al., 2020). The increased activity of *Geobacter* and *Desulfosporosinus* (As(V)-/Fe(III)-reducing bacteria) triggered by an excessive biochar dose was predominantly responsible for the discharge of As under anaerobic settings (Chen et al., 2016 and Wang et al., 2017). This suggests that an elevated biochar application rate could boost the As availability in polluted paddy soils resulting in As build-up in rice plants (Wu et al. 2020; Yu et al. 2017). Biochar dose could be a major element in regulating As kinesis, and biochar at high-dose could increase As toxicity in contaminated locations by increasing bioavailability in paddy soil.

Due to the relatively permanent and insoluble nature of the soil carbon pool, the use of biochar as a soil amendment lowers the dissolved organic C in pore water obtained from field and pot experiments (Jones et al., 2012; Karami et al., 2011; Beesely et al., 2014). Whereas Bessely et al., (2014) showed that the application of biochar prepared from orchard prune residue increased the organic C content in soil by 50% and resulted in the mobilisation of As. Because phosphate ions are connected to the mobility and bioavailability of As species, interactions between As and phosphate ions associated with biochar, are also of concern which resulted in the increased dissolution of As and hence its bioavailability (Bolan et al., 2013). With increasing application rates of biochar, soil CEC and pH show a significant rise, which may lead to conflicting conclusions. High electrical conductivity (EC) may cause more cations to be present, which may cause As sorption or precipitation, whereas high pH may lessen the positive charge on the soil-biochar system and reduce anionic As sorption (Liang et al., 2006). However, none of the research has specifically examined how the addition of biochar affects As sorption by raising pH and CEC.

8.5. Conclusion

All the experiments were conducted in pots except two which took place under field conditions. The soils used for the pot experiments were mine-impacted soils rather than the soils that were geogenically contaminated with As, like the soils of south and southeast Asia (contamination due to the use of contaminated irrigation water). Our meta-analysis showed that the application of biochar (pristine and modified) in As-contaminated rice soil not only effectively reduced the As accumulation in rice grain, but also resulted in the increase of grain shoots and grain biomass (yield). The surface area, pore volume, functional groups, and organic/inorganic makeup of biochar surfaces, as well as their adsorption capacity, differed substantially (e.g., contents of C, O, inorganic elements, ash, mobile matter, etc.). These crucial characteristics are influenced by several factors, including the type of biochar used as the feedstock, the machinery used in its production, the pyrolysis temperature and duration, heating rate and post- and pre-treatments. The maize straw and sewage sludge pristine biochars were most effective in reducing the As content in rice grain in terms of the type of feedstock used to produce biochar. The modified biochars were most effective in the immobilisation of As in soil and hence reducing As bioavailability. The rice straw biochar modified with hydroxyapatite, zeolite and Fe, and maize straw biochar modified with Fe and Mn effectively increased the Fe- and Ca-bound As fractions in the soil which resulted in the reduced bioavailability of As. Following the addition of modified biochars to paddy soils, As may accumulate less in rice tissues due to a variety of processes and mechanisms, such as the chemi- and physi-sorption of As species onto biochars, sequestration of As on Fe plaque, and decreased As uptake by rice roots due to the competitive uptake with silicate ions. Further, the addition of biochar leads to a decrease in the number of Fe(III)-reducing bacteria, which, in turn, reduces the mobility and bioavailability of inorganic As species in the rice rhizosphere. A low dose of biochar 0.5-2 % (w/w) was effective in reducing the As content in rice grain even in highly polluted soils. However, reports of As mobilisation due to the application of biochar at higher doses (5 % w/w) was also observed. One of the benefits of using biochar as an amendment is that it is required in much smaller quantities compared to organic amendments such as vermicompost, farmyard manure, which need to be applied in large quantities.

There are several research areas, which should be considered in future studies using biochar-based sorbents for the remediation of As in rice soils:

- Competition is anticipated for As sorption on the surface of biochar since different anions and cations may be present concurrently with As in real systems. Hence conducting experiments (pot or incubation) should be undertaken in the presence of competing ions such as phosphates, sulphates, silicates, and environmental contaminants, such as, metals, pesticides, and per- and poly-fluoro alkyl substances.
- In rice fields, the pH and Eh of the rhizospheric soil fluctuate drastically during the crop cultivation period due to flood irrigation after a dry spell. Further research is required to assess the efficacy of biochars at varying pH and redox (Eh) conditions.
- Typically, mostly pot experiments employing As-contaminated soils (mine-impacted) have been carried out. Instead, soils that are naturally/geogenically contaminated with As, mirroring the soils of south and southeast Asia, must be utilised for pot studies.
- Further investigation is required to assess the efficacy of biochars in reducing As in rice grain in field conditions (where water sources for irrigation are contaminated with As) as, in the real systems, various factors will come into play. The amount of irrigation water used the level of contamination, the use of fertilizers, and the cropping sequence are likely to affect the efficacy of biochar in As immobilisation/mobilisation which needs to be studied.
- Research on novel thiol functionalised biochars, biochar/nano-zero-valent Fe composites, and nano-particle-impregnated (e.g., mackinawite) biochars is necessary. Although little research has concentrated on these features, modification with amide functional groups on biochar's surface may significantly aid in sequestering As.
- Cost benefit analysis of the modified biochar production and its application in field conditions is required to evaluate the viability of its use in practical terms, as pristine biochars are relatively easier to manufacture and are also effective in the immobilisation of As in soil.
- From a pragmatic standpoint, methods for low-tech production of biochar at the farmer level are required to ensure the adoption of pristine and modified biochar for the remediation of As-contaminated paddy soil.

Chapter 9 – Summary and Conclusion

Arsenic is a naturally occurring toxic trace element that is of environmental and public health concern. Irrigation water contaminated with As acts as a potent source of contamination to humans through water-soil-crop transfer, especially in areas of India, Bangladesh, Nepal, Taiwan, Vietnam and Thailand. Rice accumulates a higher concentration of As in grains than other cereals, such as wheat and barley. However, whilst drinking water has a permissible/safe guideline value that has been defined internationally, appropriate guideline values for irrigation water and soil are lacking.

Initially through meta-analysis, decision tree (DT) and logistic regression (LR) based machine learning modelling, the relationship between As concentrations in rice grain, soil and irrigation water was evaluated. Soil total As (rather than irrigation water As) was a stronger predictor of As in rice grain. Both the decision tree and, to a lesser extent, the logistic regression models successfully predicted the concentrations of soil above which As in grain would exceed the Codex recommendation of 350 μg kg^{-1} for husked rice and 200 μg kg^{-1} for polished rice. From the logistic regression the limit for soil total As was 11.75 mg kg^{-1} and from the better performing decision tree model, the proposed guideline value for soil total As was 14 mg kg^{-1}. The prediction efficacy of both the DT and LR models was validated using purposely collected field data. On the basis of the model performance metrics, it was observed that the decision tree has an edge over the logistic regression and hence soil total As of 14 mg kg^{-1} will be an appropriate guideline value. Further the concentration of bioavailable As was predicted in the paddy soil with the help of random forest (RF), gradient boosting machine (GBM) and LR models using the collected field samples (n=233) considering other soil parameters: total As, pH, organic carbon, available iron and phosphorus. From the better performing LR model, bioavailable As (BAs), total As (TAs), available iron (AvFe) and organic carbon (OC) were significant variables for grain As. From the partial dependence plots (PDP) and individual conditional expectation (ICE) plots of the LR model 5.70 mg kg^{-1} was found to be the limit for BAs in soil. 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.

An attempt was undertaken to predict the limit for irrigation water by an incubation study with (n=420) monolithic soil columns collected from 10 As-contaminated sites. Six levels of As contaminated water was applied to the soil columns considering two types of irrigation (rainfed and irrigated) and incubated for 4 months. The LR and linear discriminant analysis (LDA) was used to predict the limit of irrigation water considering the dose of As, soil pH, organic carbon, clay, available iron and phosphorus and total As. The LR model performed better compared to LDA in terms of model performance matrices and predicted 190 μ L^{-1} as the guideline value for irrigation water. The predicted value satisfactorily classified the rice grain As when compared with the field samples. Further the LR model developed in this study provides a comprehensive understanding of the relationship between soil As levels and the predictor variables. The study highlights the importance of considering multiple soil parameters such as pH, OC, AvFe, AvP, TAs and clay in determining safe levels of As in irrigation water, and the need for further research to validate the findings in real-world conditions.

Lastly due to high toxicity and widespread pollution by As, developing an appropriate and effective method for remediation of As-contaminated soil, specifically for rice cultivation, is crucial. Among the various amendments that have been investigated, biochar is a promising immobilising agent for metals and metalloids in water and soil. A meta-analysis was conducted to assess the efficacy of biochar in the mitigation of As-contaminated rice soils. Both the rice crop and the soil were included in this study as it is now well known that soil-crop-food transfer is a potent source of As pollution particularly in south and southeast Asian countries. Altogether 23 studies met the selection criteria and were considered for meta-analysis. Most of the studies were conducted in pots with only two undertaken in field conditions. It was observed that rice (straw, husk, hulls) and maize (straw, cobs, stems) were predominantly used as the feedstock for biochar preparation, apart from eggshells, oil palm fibre and bean stalks. In all the studies the pyrolysis temperature was high (> 450 °C). From our meta-analysis, it was observed that both maize-based biochar (pristine and modified) and sewage sludge (pristine) were effective in reducing rice grain As content. Further application of biochar resulted in an increase in the root, shoot, leaf, and grain biomass (yield) over controls. Modified biochar was most effective in immobilizing As to the Fe and Ca-bound fractions in soil, compared to pristine biochar, and thus reduced the bioavailability of As. However, at higher doses of biochar application, As mobilization was observed in some instances. Biochar can be an effective tool

in the sequestration of As in soil, but further research is required under realistic field conditions.

The findings from this research are critical for the public health response in As-contaminated regions. By minimizing As intake from food consumption, we can reduce the risk of serious health problems and improve the quality of life for millions of people. The guideline values will also help to ensure the safety of rice cultivation and reduce exposure to harmful levels of As. By setting regulatory limits for contamination, government agencies can help prevent the occurrence of health problems and reduce the risk of exposure to As. Furthermore, regulatory limits provide a framework for monitoring and enforcing compliance with these limits, which will help ensure that the public is protected from exposure to unsafe levels of As contamination.

In conclusion, As contamination is a significant public health issue that requires attention. My research has provided essential data to establish guideline values for As in soil and irrigation water, which will help policymakers and farmers to implement strategies to reduce As exposure. It is my hope that these findings will be widely disseminated and acted upon to protect the health and well-being of people living in As contaminated regions.

Chapter 10 - References

Abedin, M. J., & Meharg, A. A. (2002). Relative toxicity of arsenite and arsenate on germination and early seedling growth of rice (Oryza sativa L.). *Plant and soil*, *243*(1), 57–66.

Ahmad, M., Rajapaksha, A. U., Lim, J. E., Zhang, M., Bolan, N., & Mohan, D. (2014). Biochar as a sorbent for contaminant management in soil and water: a review. *Chemosphere, 99*, 19–33.

Ahmed, Z. U., Panaullah, G. M., Gauch, H., McCouch, S. R., Tyagi, W., Kabir, M. S., & Duxbury, J. M. (2011). Genotype and environment effects on rice (Oryza sativa L.) grain arsenic concentration in Bangladesh. *Plant and Soil*, *338*(1), 367–382.

Akosa, J.S. (2017). Predictive accuracy: a misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum 2017 Conference. Cary, North Carolina: SAS Institute Inc., p. 942–2017.

Amen, R., Bashir, H., Bibi, I., Shaheen, S. M., Niazi, N. K., Shahid, M., ... & Rinklebe, J. (2020). A critical review on arsenic removal from water using biochar-based sorbents: the significance of modification and redox reactions. *Chemical Engineering Journal*, *396*, 125195.

Amonette, J. (2009). Characteristics of Biochar: Microchemical Properties in Biochar for Environmental Management: Science and Technology ed J Lehmann and S Joseph (London: Earthscan).

Arabi, Z., Rinklebe, J., El-Naggar, A., Hou, D., Sarmah, A. K., & Moreno-Jiménez, E. (2021). (Im) mobilization of arsenic, chromium, and nickel in soils via biochar: A meta-analysis. *Environmental Pollution*, *286*, 117199.

Atkinson, C. J., Fitzgerald, J. D., & Hipps, N. A. (2010). Potential mechanisms for achieving agricultural benefits from biochar application to temperate soils: a review. *Plant and Soil*, *337*(1), 1-18.

Ayers, R.S., Westcot, D.W. (1985). Water quality for agriculture (Vol.29, p. 96): Food and Agricultural Organization, United Nations: Rome, Italy; ISBN 9251022631.

Bakhat, H. F., Zia, Z., Fahad, S., Abbas, S., Hammad, H. M., Shahzad, A. N., … Shahid, M. (2017). Arsenic uptake, accumulation and toxicity in rice plants: Possible remedies for its detoxification: A review. *Environmental Science and Pollution Research*, *24*(10), 9142–9158.

Beesley, L., Inneh, O. S., Norton, G. J., Moreno-Jimenez, E., Pardo, T., Clemente, R., & Dawson, J. J. (2014). Assessing the influence of compost and biochar amendments on the mobility and toxicity of metals and arsenic in a naturally contaminated mine soil. *Environmental Pollution*, *186*, 195-202.

Beiyuan, J., Awad, Y. M., Beckers, F., Tsang, D. C., Ok, Y. S., & Rinklebe, J. (2017). Mobility and phytoavailability of As and Pb in a contaminated soil using pine sawdust biochar under systematic change of redox conditions. *Chemosphere*, *178*, 110-118.

Bhattacharya, P., Samal, A. C., Majumdar, J., & Santra, S. C. (2010a). Accumulation of arsenic and its distribution in rice plant (Oryza sativa L.) in Gangetic West Bengal, India. *Paddy and Water Environment*, *8*(1), 63–70.

Bhattacharya, P., Samal, A. C., Majumdar, J., & Santra, S. C. (2010b). Arsenic contamination in rice, wheat, pulses, and vegetables: A study in an arsenic affected area of West Bengal, India. *Water, Air, and Soil Pollution*, *213*(1–4), 3–13.

Bhattacharyya, A. K. G. Æ. P. (2004). Arsenate sorption by reduced and reoxidised rice soils under the influence of organic matter amendments, 1010–1016.

Bhuvaneshwari, S., Hettiarachchi, H., & Meegoda, J. N. (2019). Crop residue burning in India: policy challenges and potential solutions. *International Journal of Environmental Research and Public Health*, *16*(5), 832.

Biswas, A., Biswas, S., & Santra, S. C. (2014). Arsenic in irrigated water, soil, and rice: perspective of the cropping seasons. *Paddy and water environment*, *12*, 407-412.

Biswas, A., Biswas, S., Das, A., & Roychowdhury, T. (2018). Spatial variability and competing dynamics of arsenic, selenium, iron and bioavailable phosphate from ground water and soil to paddy plant parts. Groundwater for Sustainable Development, 7, 328-335.

Bogdan, K., & Schenk, M. K. (2008). Arsenic in rice (Oryza sativa L.) related to dynamics of arsenic and silicic acid in paddy soils. *Environmental Science & Technology*, *42*(21), 7885-7890.

Bogdan, K., & Schenk, M. K. (2009). Evaluation of soil characteristics potentially affecting arsenic concentration in paddy rice (Oryza sativa L.). *Environmental Pollution*, *157*(10), 2617–2621.

Bolan, N., Mahimairaja, S., Kunhikrishnan, A., & Choppala, G. (2013). Phosphorus–arsenic interactions in variable-charge soils in relation to arsenic mobility and bioavailability. *Science of the Total Environment*, *463*, 1154-1162.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L.(1984). Classification and Regression Trees (1st ed.). Routledge.

Catarecha, P., Segura, M. D., Franco-Zorrilla, J. M., García-Ponce, B., Lanza, M., Solano, R., … Leyva, A. (2007). A mutant of the Arabidopsis phosphate transporter PHT1;1 displays enhanced arsenic accumulation. *Plant Cell*, *19*(3), 1123–1133.

Chakraborti, D., Mukherjee, S. C., Pati, S., Sengupta, M. K., Rahman, M. M., Chowdhury, U. K., … Basu, G. K. (2003). Arsenic groundwater contamination in Middle Ganga Plain, Bihar, India: A future danger? *Environmental Health Perspectives*, *111*(9), 1194–1201.

Chakraborti, D., Rahman, M. M., Mukherjee, A., Alauddin, M., Hassan, M., Dutta, R. N., ... & Hossain, M. M. (2015). Groundwater arsenic contamination in Bangladesh—21 Years of research. *Journal of Trace elements in Medicine and Biology*, *31*, 237-248.

Chan, K. Y., Van Zwieten, L., Meszaros, I., Downie, A., & Joseph, S. (2007). Agronomic values of greenwaste biochar as a soil amendment. *Soil Research*, *45*(8), 629-634.

Chen, H. L., Lee, C. C., Huang, W. J., Huang, H. T., Wu, Y. C., Hsu, Y. C., & Kao, Y. T. (2016). Arsenic speciation in rice and risk assessment of inorganic arsenic in Taiwan population. *Environmental Science and Pollution Research*, *23*(5), 4481–4488.

Chen, T., Zhou, Z., Xu, S., Wang, H., & Lu, W. (2015). Adsorption behavior comparison of trivalent and hexavalent chromium on biochar derived from municipal sludge. *Bioresource technology*, *190*, 388-394.

Chen, Y., Saud, S., Li, X., Chen, Y., Zhang, L., Fahad, S., … Sadiq, A. (2014). Silicon application increases drought tolerance of Kentucky bluegrass by improving plant water relations and morphophysiological functions. *Scientific World Journal*.

Chen, Z., Wang, Y., Xia, D., Jiang, X., Fu, D., Shen, L., ... & Li, Q. B. (2016). Enhanced bioreduction of iron and arsenic in sediment by biochar amendment influencing microbial community composition and dissolved organic matter content and composition. *Journal of Hazardous Materials*, *311*, 20-29.

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

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.

Choppala, G., Bolan, N., Kunhikrishnan, A., & Bush, R. (2016). Differential effect of biochar upon reduction-induced mobility and bioavailability of arsenate and chromate. *Chemosphere*, 144, 374-381.

Chowdhury, N. R., Das, A., Joardar, M., De, A., Mridha, D., Das, R., ... & Roychowdhury, T. (2020). Flow of arsenic between rice grain and water: Its interaction, accumulation and distribution in different fractions of cooked rice. *Science of the Total Environment*, *731*, 138937.

Chowdhury, N. R., Das, A., Mukherjee, M., Swain, S., Joardar, M., De, A., … Roychowdhury, T. (2020). Monsoonal paddy cultivation with phase-wise arsenic distribution in exposed and control sites of West Bengal , alongside its assimilation in rice grain. *Journal of Hazardous Materials*, *400*,123206.

Chowdhury, N. R., Das, R., Joardar, M., Ghosh, S., Bhowmick, S., & Roychowdhury, T. (2018). Arsenic accumulation in paddy plants at different phases of pre-monsoon cultivation. *Chemosphere*, *210*, 987-997.

Chowdhury, N. R., Ghosh, S., Joardar, M., & Kar, D. (2018). Impact of arsenic contaminated groundwater used during domestic scale post harvesting of paddy crop in West Bengal : Arsenic partitioning in raw and parboiled whole grain. *Chemosphere*, *211*, 173–184.

Cohen, S. M., Arnold, L. L., Beck, B. D., Lewis, A. S., & Eldan, M. (2013). Evaluation of the carcinogenicity of inorganic arsenic. *Critical reviews in toxicology, 43*(9), 711-752.

Dahal, B. M., Fuerhacker, M., Mentler, A., Karki, K. B., Shrestha, R. R., & Blum, W. E. H. (2008). Arsenic contamination of soils and agricultural plants through irrigation water in Nepal. *Environmental Pollution*, *155*(1), 157–163.

Daniel, W. W. (1990). Applied nonparametric statistics pws. In *KENT Publishing Company, Boston, MA, 2001 PROCEEDINGS AAZV, AAWV, ARAV, NAZWV JOINT CONFERENCE* (Vol. 121, pp. 262-275).

Dittmar, J., Voegelin, A., Roberts, L. C., Hug, S. J., Saha, G. C., Ali, M. A., ... & Kretzschmar, R. (2010). Arsenic accumulation in a paddy field in Bangladesh: seasonal dynamics and trends over a three-year monitoring period. *Environmental science & technology*, *44*(8), 2925-2931.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*, 352-359

Duan, G., Liu, W., Chen, X., Hu, Y., & Zhu, Y. (2013). Association of arsenic with nutrient elements in rice plants. *Metallomics*, 5(7), 784–792.

Dubey, A., & Tarar, S.(2018). Evaluation of approximate rank-order clustering using Matthews correlation coefficient. *International Journal of Engineering and Advanced Technology*, *8*(2), 106–13.

Duxbury, J.M., & Panaullah, G. (2007). Remediation of arsenic for agriculture sustainability, food security and health in Bangladesh. FAO Water working paper, Food and Agriculture Organization, Rome.

Egbinola, C. N., & Amanambu, A. C. (2014). Groundwater contamination in Ibadan, South-West Nigeria. SpringerPlus, *3*(1), 2–7.

Enders, A., Hanley, K., Whitman, T., Joseph, S., & Lehmann, J. (2012). Characterization of biochars to evaluate recalcitrance and agronomic performance. *Bioresource Technology*, *114*, 644-653.

FAO. (2014). Water used to grow rice. In Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture (Chapter 9). Rome: Food and Agriculture Organization of the United Nations. https://www.fao.org/3/y5682e/y5682e09.htm

FAO. (2016). Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture. Retrieved from https://www.fao.org/3/i7959e/i7959e.pdf

Finland. Ministry of the Environment (2007). Government Decree on the Assessment of Soil Contamination and Remediation Needs.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179-188.

Food and Agriculture Organization (FAO). (1992). Wastewater treatment and use in agriculture - FAO irrigation and drainage paper 47. FAO, Rome, Italy.

Franke, G. R. (2010). Multicollinearity. Wiley international encyclopedia of marketing (pp. 197–198). Chichester, UK: Wiley-Blackwell.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189-1232.

Garba, Z. ., Hamza, S. ., & Galadima, A. (2010). Arsenic level speciation in fresh water from Karaye , local government area , Kano State. *International Journal of Chemistry*, *20*(22), 113–117.

Garba, Z. N., Gimba, C. E., & Galadima, A. (2012). Arsenic Contamination of Domestic Water from Northern Nigeria. *International Journal of Science and Technology*, *2*(1), 55–60.

Geng et al. (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. MSc thesis, Graduate School of Public Health, University of Pittsburgh.

Ghosh, A. K., Sarkar, D., Nayak, D. C., & Bhattacharyya, P. (2004). Assessment of a sequential extraction procedure for fractionation of soil arsenic in contaminated soils. *Archives of Agronomy and Soil Science*, *50*(6), 583–591.

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.

Golui, D., Guha Mazumder, D. N., Sanyal, S. K., Datta, S. P., Ray, P., Patra, P. K., … 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.

GRiSP (Global Rice Science Partnership). 2013. Rice almanac (4th ed). Los Baños (Philippines): International Rice Research Institute.

Gu, J. F., Zhou, H., Yang, W. T., Peng, P. Q., Zhang, P., Zeng, M., & Liao, B. H. (2018). Effects of an additive (hydroxyapatite–biochar–zeolite) on the chemical speciation of Cd and As in paddy soils and their accumulation and translocation in rice plants. *Environmental Science and Pollution Research*, *25*(9), 8608-8619.

Guo, M. (2020). The 3R principles for applying biochar to improve soil health. *Soil Systems*, 4:9.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E.R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*(6), 822–30.

Hao, J., & Priestley, Jennifer L. (2016). A Comparison of Machine Learning Techniques and Logistic Regression Method for the Prediction of Past-Due Amount. Grey Literature from PhD Candidates.

Herath, I., Zhao, F. J., Bundschuh, J., Wang, P., Wang, J., Ok, Y. S., ... & Vithanage, M. (2020). Microbe mediated immobilization of arsenic in the rice rhizosphere after incorporation of silica impregnated biochar composites. *Journal of hazardous materials*, *398*, 123096.

Higgins, J. P., & Green, S. (Eds.). (2011). Cochrane handbook for systematic reviews of interventions. Wiley.

Hilber, I., Bastos, A.C., Loureiro, S., Soja, G., Marsz, A., Cornelissen, G., & Bucheli, T.D. (2017). The different faces of biochar: contamination risk versus remediation tool. *Journal of Environmental Engineering & Landscape Management*., *25*, 86–104.

Hossain, M. B., Jahiruddin, M., Loeppert, R. H., Panaullah, G. M., Islam, M. R., & Duxbury, J. M. (2009). The effects of iron plaque and phosphorus on yield and arsenic accumulation in rice. *Plant and Soil*, *317*(1), 167-176.

Hsu, W. M., Hsi, H. C., Huang, Y. T., Liao, C. Sen, & Hseu, Z. Y. (2012). Partitioning of arsenic in soil-crop systems irrigated using groundwater: A case study of rice paddy soils in southwestern Taiwan. *Chemosphere*, *86*(6), 606–613.

Hu, X., Ding, Z., Zimmerman, A. R., Wang, S., & Gao, B. (2015). Batch and column sorption of arsenic onto iron-impregnated biochar synthesized through hydrolysis. *Water research*, *68*, 206-216.

Huq, S. I., Joardar, J. C., Parvin, S., Correll, R., & Naidu, R. (2006). Arsenic contamination in food-chain: transfer of arsenic into food materials through groundwater irrigation. *Journal of Health, Population, and Nutrition*, *24*(3), 305.

Hussain, M. M., Bibi, I., Niazi, N. K., Shahid, M., Iqbal, J., Shakoor, M. B., ... & 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.

Ippolito, J. A., Ducey, T. F., Cantrell, K. B., Novak, J. M., & Lentz, R. D. (2016). Designer, acidic biochar influences calcareous soil characteristics. *Chemosphere*, *142*, 184-191.

Irshad, M. K., Noman, A., Alhaithloul, H. A., Adeel, M., Rui, Y., Shah, T., ... & Shang, J. (2020). Goethite-modified biochar ameliorates the growth of rice (Oryza sativa L.) plants by

suppressing Cd and As-induced oxidative stress in Cd and As co-contaminated paddy soil. *Science of The Total Environment*, *717*, 137086.

Irshad, M. K., Noman, A., Wang, Y., Yin, Y., Chen, C., & Shang, J. (2022). Goethite modified biochar simultaneously mitigates the arsenic and cadmium accumulation in paddy rice (Oryza sativa) L. *Environmental Research*, *206*, 112238.

Islam, M. S., Magid, A. S. I. A., Chen, Y., Weng, L., Arafat, M. Y., Khan, Z. H., ... & Li, Y. (2021). Arsenic and cadmium load in rice tissues cultivated in calcium enriched biochar amended paddy soil. *Chemosphere*, *283*, 131102.

Islam, M. S., Magid, A. S. I. A., Chen, Y., Weng, L., Ma, J., Arafat, M. Y., ... & Li, Y. (2021). Effect of calcium and iron-enriched biochar on arsenic and cadmium accumulation from soil to rice paddy tissues. *Science of The Total Environment*, *785*, 147163.

Islam, S., Rahman, M. M., Islam, M. R., & Naidu, R. (2017). Effect of irrigation and genotypes towards reduction in arsenic load in rice. *Science of the Total Environment*, *609*, 311–318.

Jackson, M.L. (1973). Soil Chemical Analysis. Prentice Hall India Pvt.Ltd., New Delhi, p. 498.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer. (Accessed on January, 2021).

JECFA,. (2017). Report of the Eleventh Session of the Codex Committee on Contaminants in Foods. Joint FAO/WHO Food Standards Programme Codex Alimentarius Commission: Rio de Janeiro, Brazil, pp.8.

Jeong, C. Y., Dodla, S. K., & Wang, J. J. (2016). Fundamental and molecular composition characteristics of biochars produced from sugarcane and rice crop residues and by-products. *Chemosphere*, *142*, 4-13.

Jian, F. M., Yamaji, N., Mitani, N., Xu, X. Y., Su, Y. H., McGrath, S. P., & Zhao, F. J. (2008). Transporters of arsenite in rice and their role in arsenic accumulation in rice grain. Proceedings of the National Academy of Sciences of the United States of America, 105(29), 9931–9935.

Jin, H., Capareda, S., Chang, Z., Gao, J., Xu, Y., Zhang, J. (2014). Biochar pyrolytically produced from municipal solid wastes for aqueous As(V) removal: adsorption property and its improvement with KOH activation, *Bioresource Technology*, *169*, 622-629.

Jin, W., Wang, Z., Sun, Y., Wang, Y., Bi, C., Zhou, L., & Zheng, X. (2020). Impacts of biochar and silicate fertilizer on arsenic accumulation in rice (Oryza sativa L.). *Ecotoxicology and Environmental Safety*, *189*, 109928.

Jindo, K., Sonoki, T., Matsumoto, K., Canellas, L., Roig, A., & Sanchez-Monedero, M. A. (2016). Influence of biochar addition on the humic substances of composting manures. *Waste Management*, *49*, 545-552.

Jones, D. L., Rousk, J., Edwards-Jones, G., DeLuca, T. H., & Murphy, D. V. (2012). Biochar-mediated changes in soil quality and plant growth in a three year field trial. *Soil biology and Biochemistry*, *45*, 113-124.

Jouanneau, S., Durand, M., Courcoux, P., Blusseau,T., & Thouand, G. (2011). Improvement of the Identification of Four Heavy Metals in Environmental Samples by Using Predictive Decision Tree Models Coupled with a Set of Five Bioluminescent Bacteria. *Environmental Science & Technology*, *45*, 2925–2931.

Karami, N., Clemente, R., Moreno-Jiménez, E., Lepp, N. W., & Beesley, L. (2011). Efficiency of green waste compost and biochar soil amendments for reducing lead and copper mobility and uptake to ryegrass. *Journal of hazardous materials*, *191*(1-3), 41-48.

Khan, S., Chao, C., Waqas, M., Arp, H. P. H., & Zhu, Y. G. (2013). Sewage sludge biochar influence upon rice (Oryza sativa L) yield, metal bioaccumulation and greenhouse gas emissions from acidic paddy soil. *Environmental Science & Technology*, *47*(15), 8624-8632.

Khan, S., Reid, B. J., Li, G., & Zhu, Y. G. (2014). Application of biochar to soil reduces cancer risk via rice consumption: a case study in Miaoqian village, Longyan, China. *Environment International*, *68*, 154-161.

Khosravi-Darani, K.; Rehman, Y.; Katsoyiannis, I.A.; Kokkinos, E.; Zouboulis, A.I.(2022). Arsenic Exposure via Contaminated Water and Food Sources. *Water*, *14*, 1884. https://doi.org/10.3390/w14121884

Kim, W.-K., Shim, T., Kim, Y.-S., Hyun, S., Ryu, C., Park, Y.-K., Jung, J. (2013). Characterization of cadmium removal from aqueous solution by biochar produced from a giant Miscanthus at different pyrolytic temperatures, *Bioresource Technology*, *138*, 266-270.

Kimetu, J. M., Lehmann, J., Ngoze, S. O., Mugendi, D. N., Kinyangi, J. M., Riha, S., ... & Pell, A. N. (2008). Reversibility of soil productivity decline with organic matter of differing quality along a degradation gradient. *Ecosystems*, *11*(5), 726-739.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review: Vol. 1: No. 3, Article 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*(3), 1–10.

Kumarathilaka, P., Bundschuh, J., Seneweera, S., & Ok, Y. S. (2021a). An integrated approach of rice hull biochar-alternative water management as a promising tool to decrease inorganic arsenic levels and to sustain essential element contents in rice. *Journal of hazardous materials*, *405*, 124188.

Kumarathilaka, P., Bundschuh, J., Seneweera, S., Marchuk, A., & Ok, Y. S. (2021b). Iron modification to silicon-rich biochar and alternative water management to decrease arsenic accumulation in rice (Oryza sativa L.). *Environmental Pollution*, *286*, 117661.

Kumari, P. B., Singh, Y. K., Mandal, J., Shambhavi, S., Sadhu, S. K., Kumar, R., … 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.

Labanya, R., Srivastava, P. C., Pachauri, S. P., Shukla, A. K., Shrivastava, M., & Srivastava, P. (2022). Effect of three plant biomasses and two pyrolysis temperatures on structural characteristics of biochar. *Chemistry and Ecology*, *38*(5), 430-450.

Laha, A., Bhattacharyya, S., Sengupta, S., Bhattacryya K., & GuhaRoy, S. (2021). Investigation of arsenic-resistant, arsenite-oxidizing bacteria for plant growth promoting traits isolated from arsenic contaminated soils. *Archives in Microbiology, 203*, 4677–4692.

Laha, A., Sengupta, S., Bhattacharya, P., Mandal, J., Bhattacharyya, S., & Bhattacharyya, K. (2022). Recent advances in the bioremediation of arsenic-contaminated soils: a mini review. *World Journal of Microbiology and Biotechnology*, *38*(11), 1-15.

Langan, D. (2022). Assessing heterogeneity in random-effects meta-analysis. *Meta-Research: Methods and Protocols*, 67-89.

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..

Lehmann, J., & Joseph, S. (Eds.). (2015). *Biochar for environmental management: science, technology and implementation*. Routledge.

Lehmann, J., Gaunt, J., & Rondon, M. (2006). Bio-char sequestration in terrestrial ecosystems–a review. *Mitigation and adaptation strategies for global change*, *11*, 403-427.

Lehmann, J., Pereira da Silva, J., Steiner, C., Nehls, T., Zech, W., & Glaser, B. (2003). Nutrient availability and leaching in an archaeological Anthrosol and a Ferralsol of the Central Amazon basin: fertilizer, manure and charcoal amendments. *Plant and Soil*, *249*(2), 343-357.

Leksungnoen, P., Wisawapipat, W., Ketrot, D., Aramrak, S., Nookabkaew, S., Rangkadilok, N., & Satayavivad, J. (2019). Biochar and ash derived from silicon-rich rice husk decrease inorganic arsenic species in rice grain. *Science of the Total Environment*, *684*, 360-370.

Lewis, J., & Sjöstrom, J. (2010). Optimizing the experimental design of soil columns in saturated and unsaturated transport experiments. *Journal of contaminant hydrology*, *115*(1-4), 1-13.

Lian, F., Liu, X., Gao, M., Li, H., Qiu, W., & Song, Z. (2020). Effects of Fe-Mn-Ce oxide–modified biochar on As accumulation, morphology, and quality of rice (Oryza sativa L.). *Environmental Science and Pollution Research*, *27*(15), 18196-18207.

Liang, B., Lehmann, J., Solomon, D., Kinyangi, J., Grossman, J., O'Neill, B. J. O. J. F. J. J. E. G., ... & Neves, E. G. (2006). Black carbon increases cation exchange capacity in soils. *Soil Science Society of America journal*, *70*(5), 1719-1730.

Liang, F., LI, G. T., LIN, Q. M., & ZHAO, X. R. (2014). Crop yield and soil properties in the first 3 years after biochar application to a calcareous soil. *Journal of Integrative Agriculture*, *13*(3), 525-532.

Lin, L., Gao, M., Song, Z., & Mu, H. (2020). Mitigating arsenic accumulation in rice (Oryza sativa L.) using Fe-Mn-La-impregnated biochar composites in arsenic-contaminated paddy soil. *Environmental Science and Pollution Research*, *27*(33), 41446-41457.

Lin, L., Li, Z., Liu, X., Qiu, W., & Song, Z. (2019). Effects of Fe-Mn modified biochar composite treatment on the properties of As-polluted paddy soil. *Environmental Pollution*, *244*, 600-607.

Lindsay, W.L., & Norvell, W.A. (1978). Development of DTPA soil test for zinc, iron, manganese and copper. *Soil Science Society of America Journal*, *42*, 421-428.

Liu, S., Lu, Y., Yang, C., Liu, C., Ma, L., & Dang, Z. (2017). Effects of modified biochar on rhizosphere microecology of rice (Oryza sativa L.) grown in As-contaminated soil. *Environmental Science and Pollution Research*, *24*(30), 23815-23824.

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.

Lv, D., Wang, Z., Sun, Y., Jin, W., Wang, Y., Zhou, L., & Zheng, X. (2021). The effects of low-dose biochar amendments on arsenic accumulation in rice (Oryza sativa L.). *Environmental Science and Pollution Research*, *28*(11), 13495-13503.

Ma, J.F., Yamaji, N., Mitani, N., Xu, X.Y., Su, Y.H, McGrath, S. P. and Zhao, F.J. (2008). Transporters of arsenite in rice and their role in arsenic accumulation in rice grain. Proceedings of the National Academy of Science, U.S.A 105, 9931–9935.

Majumdar, K., Sanyal, S. K. (2003). pH dependent Arsenic sorption in an Alfisol and am Entisol of West Bengal. *Agropedology*, *13*, 25–29.

Mandal, B. K., Chowdhury, T. R., Samanta, G., Basu, G. K., Chowdhury, P. P., Chanda, C. R., … Chakraborti, D. (1996). Arsenic in groundwater in seven districts of West Bengal, India - The biggest arsenic calamity in the world. *Current Science*, *70*(11), 976–986.

Mandal, J., Golui, D., & Datta, S. P. (2019a). Assessing equilibria of organo-arsenic complexes and predicting uptake of arsenic by wheat grain from organic matter amended soils. *Chemosphere*, *234*, 419-426.

Mandal, J., Golui, D., Raj, A., & Ganguly, P. (2019b). Risk assessment of arsenic in wheat and maize grown in organic matter amended soils of Indo-Gangetic plain of Bihar, India. *Soil and Sediment Contamination: An International Journal*, *28*(8), 757-772.

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*, 547.

Mandal, J., Jain, V., Sengupta, S., Rahman, M. A., Bhattacharyya, K., Rahman, M. M., ... & Mondal, D. (2023). Determination of bioavailable arsenic threshold and validation of modeled permissible total arsenic in paddy soil using machine learning (Vol. 52, No. 2, pp. 315-327).

Meharg, A. A. (2004). Arsenic in rice - Understanding a new disaster for South-East Asia. *Trends in Plant Science*, *9*(9), 415–417.

Meharg, A. A., & Hartley-Whitaker, J. (2002). Arsenic uptake and metabolism in arsenic resistant and nonresistant plant species. *New Phytologist*, *154*(1), 29–43.

Meharg, A. A., & Rahman, M. (2003). Arsenic contamination of Bangladesh paddy field soils: Implications for rice contribution to arsenic consumption. *Environmental Science and Technology*, *37*(2), 229–234.

Meharg, A. A., Williams, P. N., Adomako, E., Lawgali, Y. Y., Deacon, C., Villada, A., … Yanai, J. (2009). Geographical variation in total and inorganic arsenic content of polished (white) rice. *Environmental Science and Technology*, *43*(5), 1612–1617.

Mia, S., Singh, B., & Dijkstra, F. A. (2017). Aged biochar affects gross nitrogen mineralization and recovery: a 15N study in two contrasting soils. *Gcb Bioenergy*, 9(7), 1196-1206.

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253–267.

Ministry of the Environment, Finland (2007). Government Decree on the Assessment of Soil Contamination and Remediation Needs (214/2007 (Accessed on 08 01, 2021).

Mohamed, I., Zhang, G. S., Li, Z. G., Liu, Y., Chen, F., & Dai, K. (2015). Ecological restoration of an acidic Cd contaminated soil using bamboo biochar application. *Ecological Engineering*, *84*, 67-76.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *The BMJ*, 339(3), b2535. https://doi.org/10.1136/bmj.b2535

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*(11), 2987-2998.

Mondal, D., Banerjee, M., Kundu, M., Banerjee, N., Bhattacharya, U., Giri, A. K., … 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*(6), 463–477.

Mondal, D., Periche, R., Tineo, B., Bermejo, L. A., Rahman, M. M., Siddique, A. B., … Cruz, G. J. F. (2020). Arsenic in Peruvian rice cultivated in the major rice growing region of Tumbes river basin. *Chemosphere*, 241, 125070

Mondal, D., Rahman, M.M., Suman, S.,Sharma, P., Siddique, A.B., Rahman, M.A., Bari, A.S.M.F., Kumar, R., Bose,N., Singh, S.K.,Ghosh,A.,Polya D.A.(2021). Arsenic exposure

from food exceeds that from drinking water in endemic area of Bihar, India. *Science of the Total Environment*,*754*,142082.

Mondal, S., Pramanik, K., Ghosh, S. K., Pal, P., Mondal, T., Soren, T., & Maiti, T. K. (2021). Unraveling the role of plant growth-promoting rhizobacteria in the alleviation of arsenic phytotoxicity: A review. *Microbiological Research*, *250*, 126809.

Moreno-Jiménez, E., Esteban, E., & Peñalosa, J. M. (2012). The fate of arsenic in soil-plant systems. *Reviews of environmental contamination and toxicology*, 1-37.

Muchhal, U. S., Pardo, J. M., & Raghothama, K. G. (1996). Phosphate transporters from the higher plant Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, *93*(19), 10519-10523.

Mukherjee, A., Kundu, M., Basu, B., Sinha, B., Chatterjee, M., Bairagya, M. D., ... & Sarkar, S. (2017). Arsenic load in rice ecosystem and its mitigation through deficit irrigation. *Journal of Environmental Management*, *197*, 89-95.

Mukherjee, A., Sarkar, S., Chakraborty, M., Duttagupta, S., Bhattacharya, A., Saha, D., Bhattacharya, P., Mitra, A., & Gupta, S. (2021). Occurrence, predictors and hazards of elevated groundwater arsenic across India through field observations and regional-scale AI-based modeling. *Science of the Total Environment*, *759*, 143511.

Mukherjee, A., Zimmerman, A. R., & Harris, W. (2011). Surface chemistry variations among a series of laboratory-produced biochars. *Geoderma*, 163(3-4), 247-255.

Mukhopadhyay, D., Mani, P., & Sanyal, S. (2002). Effect of phosphorus, arsenic and farmyard manure on arsenic availability in some soils of West Bengal. *Journal of the Indian Society of Soil Science*, *50*(1), 56–61.

Mwale, T., Rahman, M. 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), 1–11.

Nookabkaew, S., Rangkadilok, N., Mahidol, C., Promsuk, G., & Satayavivad, J. (2013). Determination of arsenic species in rice from Thailand and other Asian countries using simple extraction and HPLC-ICP-MS analysis. *Journal of Agricultural and Food Chemistry*, *61*(28), 6991–6998.

Novak, J. M., Busscher, W. J., Laird, D. L., Ahmedna, M., Watts, D. W., & Niandou, M. A. (2009). Impact of biochar amendment on fertility of a southeastern coastal plain soil. *Soil Science*, *174*(2), 105-112.

Novak, J., Sigua, G., Watts, D., Cantrell, K., Shumaker, P., Szogi, A., ... & Spokas, K. (2016). Biochars impact on water infiltration and water quality through a compacted subsoil layer. *Chemosphere*, *142*, 160-167.

O'Connor, D., Peng, T., Zhang, J., Tsang, D. C. W., Alessi, D. S., Shen, Z.,et al. (2018). Biochar application for the remediation of heavy metal polluted land: a review of in situ field trials. *Science of the Total Environment, 619*, 815–826.

Olsen, S.R., Cole, C.V., Watanabe, F.S., & Dean, L.A. (1954). Estimation of Available Phosphorus in Soils by Extraction with Sodium Bicarbonate. Gov. Printing Office Washington DC, USDA Circular.939, pp. 1-19.

Pal, A., Chowdhury, U. K., Mondal, D., Das, B., Nayak, B., Ghosh, A., … Chakraborti, D. (2009). Arsenic burden from cooked rice in the populations of arsenic affected and nonaffected areas and Kolkata City in West-Bengal, India. *Environmental Science & Technology*, *43*(9), 3349–3355.

Pan, D., Liu, C., Yu, H., & Li, F. (2019). A paddy field study of arsenic and cadmium pollution control by using iron-modified biochar and silica sol together. *Environmental Science and Pollution Research*, *26*(24), 24979-24987.

Panaullah, G. M., Alam, T., Hossain, M. B., Loeppert, R. H., Lauren, J. G., Meisner, C. A., ... & Duxbury, J. M. (2009). Arsenic toxicity to rice (Oryza sativa L.) in Bangladesh. *Plant and soil*, *317*, 31-39.

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.

Pescod, M.B. (1992). Wastewater treatment and use in agriculture - FAO irrigation and drainage paper 47. FAO, Rome, Italy.

Pillai, I., Fumera, G., & Roli, F. (2017). Designing multi-label classifiers that maximize F measures: state of the art. *Pattern Recognition*, *61*, 394–404.

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. *Science*, 368(6493), 845-850.

Podgorski, J.,Wu, R., Chakravorty, B., & Polya, D.A.(2020).Groundwater Arsenic Distribution in India by Machine Learning Geospatial Modeling. *International Journal of Environmental Research and Public Health*, *17*(19), 7119.

Purkait, B., & Mukherjee, A. (2008). Geostatistical analysis of arsenic concentration in the groundwater of Malda district of West Bengal, India. *Frontiers of Earth, 2*(3), 292–301.

Qiao, J. T., Liu, T. X., Wang, X. Q., Li, F. B., Lv, Y. H., Cui, J. H., ... & Liu, C. P. (2018). Simultaneous alleviation of cadmium and arsenic accumulation in rice by applying zero-valent iron and biochar to contaminated paddy soils. *Chemosphere, 195*, 260-271.

Rahaman, S., & Sinha, A. C. (2013). Water regimes: An approach of mitigation arsenic in summer rice (Oryza sativa L.) under different topo sequences on arsenic-contaminated soils of Bengal delta. *Paddy and Water Environment*, *11*(1–4), 397–410.

Rahaman, S., Sinha, A. C., Pati, R., & Mukhopadhyay, D. (2013). Arsenic contamination: a potential hazard to the affected areas of West Bengal, India. *Environmental Geochemistry and Health, 35*, 119-132.

Rahman, A., Persson, L. Å., Nermell, B., Arifeen, S. E., Ekström, E. C., Smith, A. H., & Vahter, M. (2010). Arsenic exposure and risk of spontaneous abortion, stillbirth, and infant mortality. *Epidemiology*, 797-804.

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.

Rahman, M. M., Chen, Z., & Naidu, R. (2009a). Extraction of arsenic species in soils using microwave-assisted extraction detected by ion chromatography coupled to inductively coupled plasma mass spectrometry. *Environmental Geochemistry and Health*, *31*(1), 93-102.

Rahman, M. M., Owens, G., & Naidu, R. (2009b). Arsenic levels in rice grain and assessment of daily dietary intake of arsenic from rice in arsenic-contaminated regions of Bangladesh—implications to groundwater irrigation. *Environmental Geochemistry and Health*, *31*(1), 179-187.

Rahman, M. S., Miah, M. A. M., Khaled, H. M., Islam, A., & Panaullah, G. M. (2010). Arsenic concentrations in groundwater, soils, and irrigated rice in Southwestern Bangladesh. *Communications in Soil Science and Plant Analysis*, *41*(16), 1889–1895.

Rahman, M., Islam, M., Hassan, M., Islam, S., & Zaman, S. (2014a). Impact of Water Management on the Arsenic Content of Rice Grain and Cultivated Soil in an Arsenic Contaminated Area of Bangladesh. *Journal of Environmental Science and Natural Resources*, *7*(2), 43–46.

Rahman, M. A., Rahman, M. M., & Naidu, R. (2014). Arsenic in rice: Sources and human health risk. In *Wheat and rice in disease prevention and health* (pp. 365-375). Academic Press.

Raj, A., Mandal, J., Golui, D., Sihi, D., Dari, B., Kumari, P. B., … 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*(6), 1–11.

Rajapaksha, A.U., Chen, S.S., Tsang, D.C.W., Zhang, M., Vithanage, M., Mandal, S., Gao, B., Bolan, N.S., & Ok, Y.S. (2016). Engineered/designer biochar for contaminant removal/immobilization from soil and water: potential and implication of biochar modification. *Chemosphere*, *148*, 276–291.

Reid, M. C., Asta, M. P., Falk, L., Maguffin, S. C., Cong Pham, V. H., Le, H. A., … Le Vo, P. (2021). Associations between inorganic arsenic in rice and groundwater arsenic in the Mekong Delta. *Chemosphere*, 265.

Ripley, B.D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Rondon, M. A., Lehmann, J., Ramírez, J., & Hurtado, M. (2007). Biological nitrogen fixation by common beans (Phaseolus vulgaris L.) increases with bio-char additions. *Biology and Fertility of Soils*, *43*(6), 699-708.

Roychowdhury, T., Uchino, T., & Tokunaga, H. (2008). Effect of arsenic on soil, plant and foodstuffs by using irrigated groundwater and pond water from Nadia district, West Bengal. *International Journal of Environment and Pollution*, *33*(2-3), 218-234.

Roychowdhury, T. (2008b). Impact of sedimentary arsenic through irrigated groundwater on soil, plant, crops and human continuum from Bengal delta: Special reference to raw and cooked rice. *Food and Chemical Toxicology*, *46*(8), 2856–2864.

Rutherford, D. W., Wershaw, R. L., Rostad, C. E., & Kelly, C. N. (2012). Effect of formation conditions on biochars: Compositional and structural properties of cellulose, lignin, and pine biochars. *Biomass and Bioenergy*, *46*, 693-701.

Sadiq, M. (1997). Arsenic chemistry in soils: An overview of thermodynamic predictions and field observations. *Water, Air, & Soil Pollution*, *93*(1–4), 117–136.

Sanyal, S. K., Gupta, S. K., Kukal, S. S., & Jeevan Rao, K. (2015). Soil degradation, pollution and amelioration. *State of Indian Agriculture-Soil. National Academy of Agricultural Sciences, New Delhi*, 234-266.

Sanyal,S.K.(2014). Arsenic contamination in ground water: an environmental issue. *Journal of Crop and Weed*, *10*(1),1-12.

Sarkar, S., Basu, B., Kundu, C. K., & Patra, P. K. (2012). Deficit irrigation: An option to mitigate arsenic load of rice grain in West Bengal, India. *Agriculture, Ecosystems and Environment*, *146*(1), 147–152.

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. *Environmental Science and Tecnology Enginering*, *2*, 689−702.

Saud, S., Yajun, C., Fahad, S., Hussain, S., Na, L., Xin, L., & Alhussien, S. A. A. F. E. (2016). Silicate application increases the photosynthesis and its associated metabolic activities in Kentucky bluegrass under drought stress and post-drought recovery. *Environmental Science and Pollution Research*, *23*(17), 17647–17655.

Schmidt, C. W. (2015). In Search of "Just Right": the challenge of Regulating Arsenic in rice. *Environmental Health Perspectives*, *123* (1), A16–A19.

Schoof, R. A., Yost, L. J., Crecelius, E., Irgolic, K., Goessler, W., Guo, H. R., & Greene, H. (1998). Dietary arsenic intake in Taiwanese districts with elevated arsenic in drinking water. Human and Ecological Risk Assessment (HERA), *4*(1), 117–135.

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.

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, Ecosystems & Environment*, *319*, 107516.

Seyfferth, A. L., & Fendorf, S. (2012). Silicate mineral impacts on the uptake and storage of arsenic and plant nutrients in rice (Oryza sativa L.). *Environmental Science & Technology*, *46*(24), 13176-13183.

Shaaban, A., Se, S. M., Dimin, M. F., Juoi, J. M., Husin, M. H. M., & Mitan, N. M. M. (2014). Influence of heating temperature and holding time on biochars derived from rubber wood sawdust via slow pyrolysis. *Journal of Analytical and Applied Pyrolysis*, *107*, 31-39.

Sharma, S., Kaur, I., & Nagpal, A. K. (2017). Assessment of arsenic content in soil, rice grains and groundwater and associated health risks in human population from Ropar wetland, India, and its vicinity. *Environmental Science and Pollution Research*, *24*(23), 18836–18848.

Shin, H., Shin, H. S., Dewbre, G. R., & Harrison, M. J. (2004). Phosphate transport in Arabidopsis: Pht1;1 and Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate environments. *Plant Journal*, *39*(4), 629–642.

Siegel, Sidney (1956). Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill. pp. 75–83. ISBN 9780070573482.

Singh, S. K., & Ghosh, A. K. (2011). Entry of Arsenic into food material -A case study. *World Applied Sciences Journal*, *13*(2), 385–390.

Sinha, B., & Bhattacharyya, K. (2011). Retention and release isotherm of arsenic in arsenic-humic/fulvic equilibrium study. *Biology and Fertility of Soils*, *47*(7), 815–822.

Sinha, B., & Bhattacharyya, K. (2014). Arsenic toxicity in rice with special reference to speciation in Indian grain and its implication on human health. *Journal of the Science of Food and Agriculture*, *95*(7), 1435–1444.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Proceedings of Advances in Artificial Intelligence (AI 2006), Lecture Notes in Computer Science, vol. 4304. Heidelberg: Springer, p. 1015–21.

Somenahally, A. C., Hollister, E. B., Yan, W., Gentry, T. J., & Loeppert, R. H. (2011). Water management impacts on arsenic speciation and iron-reducing bacteria in contrasting rice-rhizosphere compartments. *Environmental science & technology*, *45*(19), 8328-8335.

Srivastava, P. K., Singh, M., Gupta, M., Singh, N., Kharwar, R. N., Tripathi, R. D., & Nautiyal, C. S. (2015). Mapping of arsenic pollution with reference to paddy cultivation in the middle Indo-Gangetic Plains. *Environmental Monitoring and Assessment*, *187*(4).

Steiner, C. (2016). Considerations in biochar characterization. Agricultural and environmental applications of biochar: advances and barriers, *63*, 87-100.

Steiner, C., Teixeira, W. G., Lehmann, J., Nehls, T., de Macêdo, J. L. V., Blum, W. E., & Zech, W. (2007). Long term effects of manure, charcoal and mineral fertilization on crop

production and fertility on a highly weathered Central Amazonian upland soil. *Plant and soil*, *291*(1), 275-290.

Syu, C., Huang, C., Jiang, P., Lee, C., & Lee, D. (2015). Arsenic accumulation and speciation in rice grains influenced by arsenic phytotoxicity and rice genotypes grown in arsenic-elevated paddy soils. *Journal of Hazardous Materials*, *286*, 179–186.

Takahashi, Y., Minamikawa, R., Hattori, K. H., Kurishima, K., Kihou, N., & Yuita, K. (2004). Arsenic behavior in paddy fields during the cycle of flooded and non-flooded periods. *Environmental Science & Technology*, *38*(4), 1038–1044.

Talukder, A. S. M. H. M., Meisner, C. A., Sarkar, M. A. R., & Islam, M. S. (2011). Effect of water management, tillage options and phosphorus status on arsenic uptake in rice. *Ecotoxicology & Environmental Safety*, *74*(4), 834–839.

Tan, X., Liu, Y., Zeng, G., Wang, X., Hu, X., Gu, Y., et al. (2015). Application of biochar for the removal of pollutants from aqueous solutions. *Chemosphere*, *125*, 70–85.

Tan, Z., Yang, Q., & Zheng, Y. (2020). Machine Learning Models of Groundwater Arsenic Spatial Distribution in Bangladesh: Influence of Holocene Sediment Depositional History. *Environment Science & Technology*, *54*, 9454−9463.

Torres-Escribano, S., Leal, M., Velez, D.& Montoro, R.(2008). Total and inorganic arsenic concentrations in rice sold in Spain, effect of cooking, and risk assessments. *Environment Science & Technology*, *42*, 3867-3872.

Toth G., Hermann, T., Silva M.R.Da. & Montanarella L. (2016). Heavy metals in agricultural soils of the European Union with implications to food safety. *Environment International*, *88*, 299-309.

Turpeinen, R., Pantsar-Kallio, M., Häggblom, M., & Kairesalo, T. (1999). Influence of microbes on the mobilization, toxicity and biomethylation of arsenic in soil. *Science of the Total Environment*, *236*(1–3), 173–180.

Uchimiya, M., Chang, S., & Klasson, K. T. (2011). Screening biochars for heavy metal retention in soil: role of oxygen functional groups. *Journal of Hazardous Materials*, *190*(1-3), 432-441.

Ultra Jr, V. U., Nakayama, A., Tanaka, S., Kang, Y., Sakurai, K., & Iwasaki, K. (2009). Potential for the alleviation of arsenic toxicity in paddy rice using amorphous iron-(hydro) oxide amendments. *Soil Science and Plant Nutrition*, *55*(1), 160-169.

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

Van Geen, A., Zheng, Y., Cheng, Z., He, Y., Dhar, R. K., Garnier, J. M., ... & Ahmed, K. M. (2006). Impact of irrigating rice paddies with groundwater containing arsenic in Bangladesh. *Science of the Total Environment*, *367*(2-3), 769-777.

Verheijen, F., Jeffery, S., Bastos, A. C., Van der Velde, M., & Diafas, I. (2010). Biochar application to soils. *A critical scientific review of effects on soil properties, processes, and functions. EUR*, *24099*(162), 2183-2207.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.

Vithanage, M., Herath, I., Almaroai, Y. A., Rajapaksha, A. U., Huang, L., Sung, J. K., et al. (2017). Effects of carbon nanotube and biochar on bioavailability of Pb, Cu and Sb in multi-metal contaminated soil. *Environment Geochemistry and Health*, *39*, 1409–1420.

Vovk, V. (2015). The fundamental nature of the log loss function. *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday*, 307-318. https://doi.org/10.1007/978-3-319-23534-9_20.

Walkley, A., & Black, C.A. (1934). An examination of wet acid method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science*, *37*, 29-38.

Wan, Y., Camara, A. Y., Huang, Q., Yu, Y., Wang, Q., & Li, H. (2018). Arsenic uptake and accumulation in rice (Oryza sativa L.) with selenite fertilization and water management. *Ecotoxicology and environmental safety*, *156*, 67-74.

Wang, L., Chu, F., & Xie, W.(2007). Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans *Computational Biology and Bioinformatics*, *4*(1), 40–53.

Wang, Y., Wu, S., Yan, D., Li,F., Chengcheng, W., Min, C., & Wenyu, S. (2020). Determining and mapping the spatial mismatch between soil and rice cadmium (Cd) pollution based on a decision tree model. *Environmental Pollution*, *265*, 115029.

Wen, E., Yang, X., Chen, H., Shaheen, S. M., Sarkar, B., Xu, S., ... & Wang, H. (2021). Iron-modified biochar and water management regime-induced changes in plant growth, enzyme activities, and phytoavailability of arsenic, cadmium and lead in a paddy soil. *Journal of Hazardous Materials*, *407*, 124344.

Williams, P. N., Islam, M. R., Adomako, E. E., Raab, A., Hossain, S. A., Zhu, Y. G., … Meharg, A. A. (2006). Increase in rice grain arsenic for regions of Bangladesh irrigating paddies with elevated arsenic in groundwaters. *Environmental Science and Technology*, *40*(16), 4903–4908.

Williams, P. N., Price, A. H., Raab, A., Hossain, S. A., Feldmann, J., & Meharg, A. A. (2005). Variation in arsenic speciation and concentration in paddy rice related to dietary exposure. *Environmental Science and Technology*, *39*(15), 5531–5540.

Williams, Paul N, Villada, A., Deacon, C., Raab, A., Figuerola, J., Green, A. J., … Meharg, A. A. (2007). Greatly enhanced arsenic shoot assimilation in rice leads to elevated grain levels compared to wheat and barley. *Environmental Science & Technology*, *41*(19), 6854–6859.

Woolson, E. A., & Axley, J. H. (1971). Correlation between available soil arsenic , estimated by six methods, and response of corn (Zea mays L.), *Soil Science Society of America Journal*, *35*, 101–105.

Worth, A.P., & Cronin, M.T.D. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure*, *662*, 97-111.

Wu, J., Li, Z., Huang, D., Liu, X., Tang, C., Parikh, S. J., & Xu, J. (2020). A novel calcium-based magnetic biochar is effective in stabilization of arsenic and cadmium co-contamination in aerobic soils. *Journal of Hazardous Materials*, *387*, 122010.

Yaashikaa, P. R., Kumar, P. S., Varjani, S., & Saravanan, A. (2020). A critical review on the biochar production techniques, characterization, stability and applications for circular bioeconomy. *Biotechnology Reports*, *28*, e00570.sk

Yin, D., Wang, X., Peng, B., Tan, C., & Ma, L. Q. (2017). Effect of biochar and Fe-biochar on Cd and As mobility and transfer in soil-rice system. *Chemosphere*, *186*, 928-937.

Yrjälä, K., Ramakrishnan, M., & Salo, E. (2022). Agricultural waste streams as resource in circular economy for biochar production towards carbon neutrality. *Current Opinion in Environmental Science & Health*, 100339.

Yu, L. U. O., JIAO, Y. J., ZHAO, X. R., LI, G. T., ZHAO, L. X., & MENG, H. B. (2014). Improvement to maize growth caused by biochars derived from six feedstocks prepared at three different temperatures. *Journal of Integrative Agriculture*, *13*(3), 533-540.

Yu, Z., Qiu, W., Wang, F., Lei, M., Wang, D., & Song, Z. (2017). Effects of manganese oxide-modified biochar composites on arsenic speciation and accumulation in an indica rice (Oryza sativa L.) cultivar. *Chemosphere*, *168*, 341-349.

Zhang, R. H., Li, Z. G., Liu, X. D., Wang, B. C., Zhou, G. L., Huang, X. X., ... & Brooks, M. (2017). Immobilization and bioavailability of heavy metals in greenhouse soils amended with rice straw-derived biochar. *Ecological Engineering*, *98*, 183-188.

Zhao, B., O'Connor, D., Zhang, J., Peng, T., Shen, Z., Tsang, D. C., & Hou, D. (2018). Effect of pyrolysis temperature, heating rate, and residence time on rapeseed stem derived biochar. *Journal of Cleaner Production*, *174*, 977-987.

Zhao, F. J., McGrath, S. P., & Meharg, A. A. (2010). Arsenic as a food chain contaminant: mechanisms of plant uptake and metabolism and mitigation strategies. *Annual review of plant biology*, *61*, 535-559.

Zhao, Z., Jia, Y., Xu, L., & Zhao, S. (2011). Adsorption and heterogeneous oxidation of As (III) on ferrihydrite. *Water research*, *45*(19), 6496-6504.

Zheng, R., Chen, Z., Cai, C., Tie, B., Liu, X., Reid, B. J., ... & Baltrėnaitė, E. (2015). Mitigating heavy metal accumulation into rice (Oryza sativa L.) using biochar amendment—a field experiment in Hunan, China. *Environmental Science and Pollution Research*, *22*(14), 11097-11108.

Appendices

Appendix A: Publications

A.1.Research Papers

1. **Mandal, J**., Sengupta, S., Sarkar, S., Mukherjee, A., Wood, M.D., Hutchinson, S.M. and Mondal, D. (2021). Meta-Analysis Enables Prediction of the Maximum Permissible Arsenic Concentration in Asian Paddy Soil. *Frontiers in Environmental Science,* 9:760125.
2. **Mandal J**., Bakare, W.A., Rahman, M.M., Rahman, M.A., Siddique, A.B., Oku, E., Wood, M.D., Hutchinson, S.M. and Mondal, D. (2022). Varietal differences influence arsenic and lead contamination of rice grown in mining impacted agricultural fields of Zamfara State, Nigeria, *Chemosphere*, 135339.
3. **Mandal, J**., Jain, V., Sengupta, S., Rahman, M.A., Bhattacharyya, K., Rahman, M.M., Golui, D., Wood, M.D. and Mondal, D. (2023). Determination of bioavailable arsenic threshold and validation of modelled permissible total arsenic in paddy soil using machine learning. *Journal of Environmental Quality*. https://doi.org/10.1002/jeq2.20452
4. Moulick, D., Ghosh, D., **Mandal, J**., Bhowmick, S., Mondal, D., Choudhury, S., Santra, S.C., Vithanage, M. and Biswas, J.K. (2023). A cumulative assessment of plant growth stages and selenium supplementation on arsenic and micronutrients accumulation in rice grains. *Journal of Cleaner Production*, p.135764. https://doi.org/10.1016/j.jclepro.2022.135764
5. Sengupta, S., Bhattacharyya, K., **Mandal, J**. and 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*, p.115531. https://doi.org/10.1016/j.jenvman.2022.115531

A.2. Book Chapters

1. **Mandal, J**., Golui, D., Ray, P. and Bhattacharyya, P., 2022. Heavy Metal Pollution in Soil and Remediation Strategies. In *Soil Management For Sustainable Agriculture* (pp. 505-529). Apple Academic Press. https://doi.org/10.1201/9781003184881
2. Sengupta, S., Roychowdhury, T., Phonglosa, A. and **Mandal, J**., 2022. Arsenic Contamination in Rice and the Possible Mitigation Options. In *Global Arsenic Hazard:*

Ecotoxicology and Remediation (pp. 35-48). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16360-9_3

3. Chowdhury, N.R., Das, A., Joardar, M., Mridha, D., De, A., Majumder, S., **Mandal, J**., Majumdar, A. and Roychowdhury, T., 2022. Distribution of Arsenic in Rice Grain from West Bengal, India: Its Relevance to Geographical Origin, Variety, Cultivars and Cultivation Season. In *Global Arsenic Hazard: Ecotoxicology and Remediation* (pp. 509-531). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16360-9_23

A.3. Others

- Participated and presented a research paper (online) at 9th Annual Convention and National Webinar on "Managing Agro-Chemicals for Crop and Environmental Health" organised by Society of Fertilizers and Environment and received the **Best Paper Presentation Award.**
- Successfully **chaired** virtual conference session on 'Managing environmental pollution for sustainable development' with my supervisor Prof. Mike Wood at International Postgraduate Research Conference (IPGRC) from 4th -6th April 2022.
- Delivered an **invited talk** on "Prospects of using Machine Learning Algorithms in Natural Resource Management" at the National Webinar on Sustainable Intervention towards Resource Conservation and Natural Farming organised by Assam Agricultural University, Jorhat, Assam.
- Media coverage of our research work by **India Water Portal** and **Aaj Tak and Bartaman newspaper** (Bengali language). The Links are as follows:

https://bangla.aajtak.in/specials/story/arsenic-concentration-asian-paddy-soil-irrigation-water-contaminated-too-abk-336076-2022-01-22

https://www.indiawaterportal.org/articles/soil-arsenic-above-14-mgkg-can-cause-rice-be-unsafe-consumption

https://bartamanpatrika.com/home?cid=13&id=459796

Appendix B – Training, Conference/Seminars attended and Supervision Records

B.1. Training/Courses/Workshop Undertaken

Date	Training Undertaken	Aim of the training
03/10/2020	PGR Welcome and Induction	Designed to navigate research project and provides essential information to programme of study.
03/10/2020	Researcher Integrity and Research Ethics	Ensure compliance with university and legal requirements for research ethics. Guide and support researchers through the research ethics application submission process
04/10/2020	Introductory Research Skills	Statistical Software
16/10/2020	PGR Return to Campus and Remote Working Safety Induction	Safety measures to be undertaken
16/03/2021	Statistical Learning (STATSX0001) Stanford Online, Stanford University. Course Duration: 3months	An Introduction to Statistical Learning, with Applications in R
12/04/2021	Peer Reviewer Course Researcher Academy, Elsevier	Techniques for peer reviewing a scientific article
07/03/2022	Research Impact Workshop	To embed real-world impact from the outset of research projects to ensure maximum benefit to public and stakeholders

B.2. Conferences/Seminars

Date	Conference	Presentation
15th December 2020	Workshop on Nature and nurture in Arsenic induced toxicity in Bihar	Oral presentation – 10 minutes
10th June 2021	8th UK & Ireland Occupational and Environmental Exposure Science Meeting 2021	Oral presentation – 5 minutes
29th June 2021	Salford Postgraduate Annual Research Conference (SPARC), 2021	Oral Presentation-7 minutes
17th November 2021	85th Annual Convention of the Indian Society of Soil Science	Oral Presentation-15 minutes
25th February 2022	9th Annual Convention and National Webinar on "Managing Agro-Chemicals for Crop and Environmental Health" organised by Society of Fertilizers and Environment	Oral Presentation-10 minutes

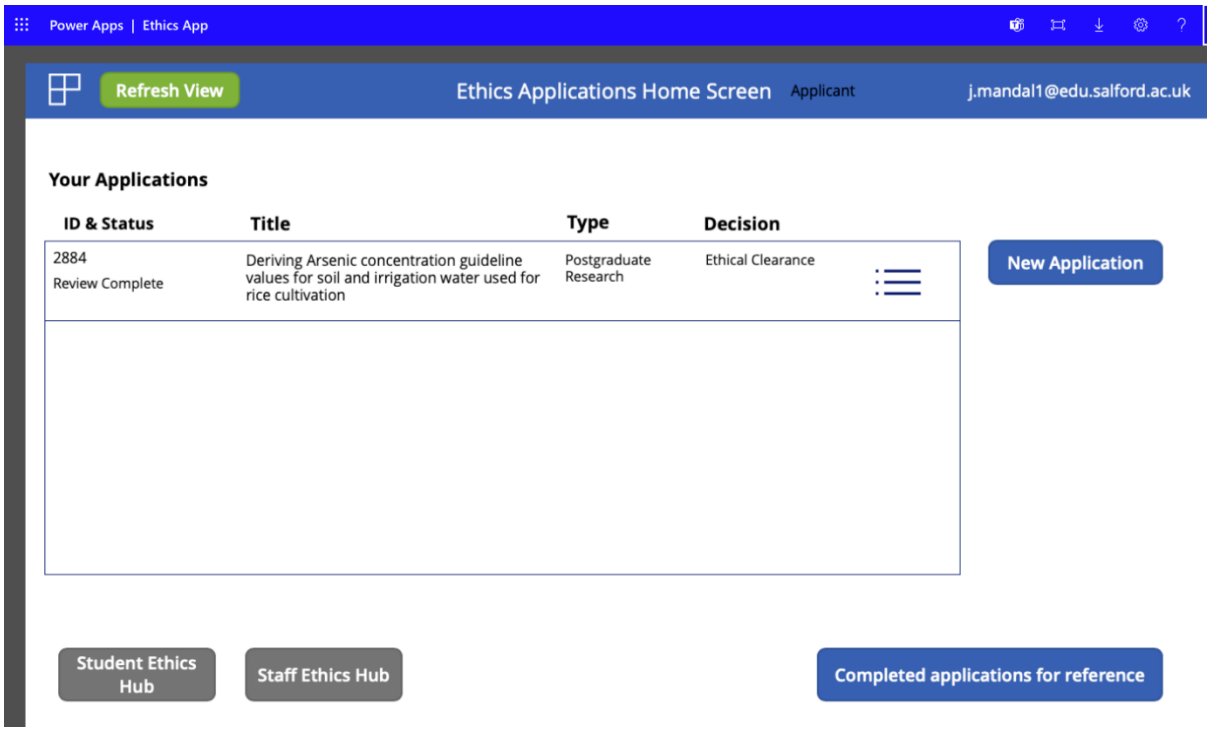
5th April 2022	International Postgraduate Research Conference (IPGRC), University of Salford	Oral Presentation-15 minutes
30th July 2022	Salford Postgraduate Annual Research Conference (SPARC), 2022	Oral Presentation-7 minutes

B.3. Supervision Meeting Records 1st, 2nd and 3rd year

Meeting No.	Date	Medium
1	13/10/2020	MS-Teams
2	16/11/2020	
3	17/12/2020	
4	28/01/2021	
5	11/02/2021	
6	24/03/2021	
7	29/04/2021	
8	21/05/2021	
9	14/06/2021	
10	12/07/2021	
11	12/08/2021	
12	07/08/2021	
13	07/09/2021	
14	25/10/2021	
15	24/11/2021	
16	17/12/2021	
17	19/01/2022	
18	15/02/2022	In-person
19	23/03/2022	MS-Teams
20	25/04/2022	In-person
21	23/05/2022	MS-Teams
22	13/06/2022	MS-Teams
23	08/07/2022	MS-Teams
24	15/08/2022	MS-Teams

25	16/09/2022	MS-Teams
26	03/10/2022	MS-Teams
27	07/11/2022	MS-Teams
28	06/12/2022	MS-Teams
29	25/01/2023	MS-Teams
30	08/02/2023	MS-Teams
31	06/03/2023	MS-Teams

Appendix C: Ethical Clearance



Appendix D- R-Codes

D.1.R-Codes for training of the models (Logistic Regression and Decision tree) with meta data and testing with field data

```
#For Meat-Training data
# Read data file
str(Meta_Model)
Meta_Model$Category <- as.factor(Meta_Model$Category)
xtabs(~Category,data = Meta_Model)

# Partition data - train (80%) & test (20%)
set.seed(123)
ind <- sample(2, nrow(Meta_Model), replace = T, prob = c(0.8, 0.2))
train <- Meta_Model[ind==1,]
test <- Meta_Model[ind==2,]


#Logistic Regression
library(caret)
# define training control
train_control <- trainControl(method = "repeatedcv", number = 10,repeats =
5)


# train the model on training set
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
model <- train(Category ~SoilAs,
               data = train,
               trControl = train_control,
               method = "glm",
               family=binomial())
model
model$finalModel
summary(model)

# Prediction with test data in terms of probability
PredLR_test_prob <- predict(model,test, type = 'prob')
PredLR_test_prob
PredictLR_test_prob <- ifelse(PredLR_test_prob>0.5, 1, 2)
PredictLR_test_prob

# Misclassification error - test data and Accuracy
PredLR_test <- predict(model,test, type = 'raw')
confusionMatrix(PredLR_test, test$Category)


# Prediction with train data in terms of probability
PredLR_train_prob <- predict(model,train, type = 'prob')
PredLR_train_prob

PredictLR_train_prob <- ifelse(PredLR_train_prob>0.5, 0, 1)
PredictLR_train_prob

# Misclassification error - train data and Accuracy
PredLR_test <- predict(model,train, type = 'raw')
confusionMatrix(PredLR_test, train$Category)
```

```
# ROC Curve
par(pty="s")
ROC_train<-roc(train$Category,PredictLR_train_prob [,
2],plot=TRUE,legacy.axes=TRUE,percent=TRUE,
               col="#AD4433",lwd=4,print.auc=TRUE)

par(pty="s")
ROC_test<-roc(test$Category,PredictLR_test_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)

#Plotting of Probability vs Soil As
X1<-0.1429
b<--1.6822
X1_range <- seq(from=min(train$SoilAs), to=max(train$SoilAs), by=0.01)
a_logits <- b + X1*X1_range
a_probs <- exp(a_logits)/(1 + exp(a_logits))
a_probs

plot(X1_range, a_probs,
     ylim=c(0,1),
     type="l",
     lwd=3,
     lty=2,
     col="black",
     xlab="Soil Arsenic mg/kg", ylab="Probability", main="Probability of
super important outcome")
abline(h=.5, lty=2)  # add a horizontal line at p=.5


library(ggplot2); library(tidyr)
# first you have to get the information into a long dataframe, which is
what ggplot likes
plot.data <- data.frame(a=a_probs,X1=X1_range)
plot.data <- gather(plot.data,key = group, value=prob,a)
head(plot.data)
A<-ggplot(plot.data, aes(x=X1, y=prob)) +
  geom_line(lwd=1) + ylim(0,1.0)+ xlim(0,100)+
  labs(x="Soil Arsenic (mg/kg)",y= "Pr(≤MTC |
>MTC)=-1.6822+0.1429SoilAs***",cex=2.5)+geom_hline(yintercept =
0.5,colour="red",lwd=1)+
  geom_vline(xintercept = 11.75,colour="blue",show.legend = TRUE,lwd=1)
FinalA<-A+theme(axis.title.y = element_text(color="black", size=12,
face="bold"),
                axis.title.x = element_text(color="black", size=12,
face="bold"))
FinalA
B<-ggplot(plot.data, aes(x=X1, y=prob)) +
  geom_line(lwd=1) + ylim(0,1.0)+xlim(11.55,11.85)+
  labs(x="Soil Arsenic (mg/kg)", y="Pr(≤MTC |
>MTC)=-1.619+0.1429SoilAs***",cex=5)+geom_hline(yintercept =
0.5,colour="red",lwd=1)+
  geom_vline(xintercept = 11.75,colour="blue",show.legend = TRUE,lwd=1)
FinalB<-B+theme(axis.title.y = element_text(color="black", size=12,
face="bold"),
                axis.title.x = element_text(color="black", size=12,
face="bold"))
FinalB
```

```
ggarrange(FinalA,FinalB,nrow =1,align = "hv",labels = "AUTO" )



#Decision Tree
library(caret)
# Set up caret to perform 10-fold cross validation repeated 5 times
caret.control <- trainControl(method = "repeatedcv",
                              number = 10,
                              repeats = 5)
set.seed(1234)
DT<- train(Category~SoilAs,data=train,
           method = "rpart",
           trControl = caret.control,
           tuneLength = 10)
DT
DT_best <- DT$finalModel
DT_best
DT_best$cptable
rpart.plot(DT_best,extra=104)
rpart.plot(DT_best,extra = 6)
summary(DT_best)

# Prediction with test data in terms of probability
PredDT_test_prob <- predict(DT_best,test, type = 'prob')
PredDT_test_prob

# Misclassification error - test data and Accuracy
PredDT_test <- predict(DT_best,test, type = 'class')
confusionMatrix(PredDT_test,test$Category)

# Prediction with train data in terms of probability
PredDT_train_prob <- predict(DT_best ,train , type = 'prob')
PredDT_train_prob

# Misclassification error - train data and Accuracy
PredDT_train <- predict(DT_best,train, type = 'class')
confusionMatrix(PredDT_train,train$Category)


# ROC Curve
par(pty="s")
ROC_train<-roc(train$Category,PredDT_train_prob [,
2],plot=TRUE,percent=TRUE,
               col="#AD4433",lwd=4,print.auc=TRUE)
par(pty="s")
ROC_test<-roc(test$Category,PredDT_test_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col="#4C00FF",lwd=4,print.auc=TRUE)


#Testing with the Test Sets
# For test set
# Read data file

str(TestSet)
TestSet$Category <- as.factor(TestSet$Category)
xtabs(~Category,data = TestSet)
```

```
###Testing with the Test Sets with Logistic Regression

#Testing with the Test Set1
PredLR_test_prob <- predict(model,TestSet, type = 'prob')
PredLR_test_prob
PredictLR_test1_prob <- ifelse(PredLR_test_prob>0.5, 1, 2)
PredictLR_test1_prob

# Misclassification error - test data and Accuracy
PredLR_test1 <- predict(model,TestSet, type = 'raw')
confusionMatrix(PredLR_test1, TestSet$Category)

# ROC Curve
par(pty="s")
ROC_test<-roc(TestSet$Category,PredictLR_test1_prob[,
1],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)


#Testing with the Test Set2
PredLR_test2_prob <- predict(model,TestSet, type = 'prob')
PredLR_test_prob
PredictLR_test2_prob <- ifelse(PredLR_test_prob>0.5, 1, 2)
PredictLR_test_prob

# Misclassification error - test data and Accuracy
PredLR_test2 <- predict(model,TestSet, type = 'raw')
confusionMatrix(PredLR_test2, TestSet$Category)

# ROC Curve
par(pty="s")
ROC_test<-roc(TestSet$Category,PredictLR_test2_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)

#Testing with the Test Set3
PredLR_test3_prob <- predict(model,TestSet, type = 'prob')
PredLR_test_prob
PredictLR_test3_prob <- ifelse(PredLR_test_prob>0.5, 1, 2)
PredictLR_test_prob

# Misclassification error - test data and Accuracy
PredLR_test3 <- predict(model,TestSet, type = 'raw')
confusionMatrix(PredLR_test3, TestSet$Category)

# ROC Curve
par(pty="s")
ROC_test<-roc(TestSet$Category,PredictLR_test3_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)


###Testing with the Test Sets with Decision Tree

#Testing with the Test Set1
PredDT_test1_prob <- predict(DT_best,TestSet, type = 'prob')
```

```
PredDT_test1_prob

# Misclassification error - test data and Accuracy
PredDT_test1 <- predict(DT_best,TestSet, type = 'class')
confusionMatrix(PredDT_test1,TestSet$Category)

par(pty="s")
ROC_test<-roc(TestSet$Category,PredDT_test1_prob[,
1],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col="#4C00FF",lwd=4,print.auc=TRUE)


#Testing with the Test Set2
PredDT_test2_prob <- predict(DT_best,TestSet, type = 'prob')
PredDT_test2_prob

# Misclassification error - test data and Accuracy
PredDT_test2 <- predict(DT_best,TestSet, type = 'class')
confusionMatrix(PredDT_test2,TestSet$Category)

par(pty="s")
ROC_test<-roc(TestSet$Category,PredDT_test2_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col="#4C00FF",lwd=4,print.auc=TRUE)


#Testing with the Test Set3
PredDT_test3_prob <- predict(DT_best,TestSet, type = 'prob')
PredDT_test3_prob

# Misclassification error - test data and Accuracy
PredDT_test3 <- predict(DT_best,TestSet, type = 'class')
confusionMatrix(PredDT_test2,TestSet$Category)

par(pty="s")
ROC_test<-roc(TestSet$Category,PredDT_test3_prob[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col="#4C00FF",lwd=4,print.auc=TRUE)


# Boxplots of the Test sets on the basis of <MTC and >MTC
A<-ggboxplot(TestSet, x = "Category",
             y = "SoilAs",
             merge = TRUE,
             ylab = "Total As in soil (mg kg-1)", add = "jitter",width =
0.3,
             color="Category", palette =c("#1E1E1E", "Blue"))+
  geom_hline(yintercept = 14,colour="red",lwd=0.75)+
  geom_hline(yintercept = 11.75,colour="green",lwd=0.75)
A

B<-ggboxplot(TestSet, x = "Category",
             y = "SoilAs",
             merge = TRUE,
             ylab = "Total As in soil (mg kg-1)", add = "jitter",width =
0.3,
             color="Category", palette =c("#1E1E1E", "Blue"))+
  geom_hline(yintercept = 14,colour="red",lwd=0.75)+
```

```
  geom_hline(yintercept = 11.75,colour="green",lwd=0.75)
B

C<-ggboxplot(TestSet, x = "Category",
             y = "SoilAs",
             merge = TRUE,
             ylab = "Total As in soil (mg kg-1)", add = "jitter",width =
0.3,
             color="Category", palette =c("#1E1E1E", "Blue"))+
  geom_hline(yintercept = 14,colour="red",lwd=0.75)+
  geom_hline(yintercept = 11.75,colour="green",lwd=0.75)
C

ggarrange(A,B,C,nrow =1,align = "hv",labels = "AUTO",common.legend = TRUE)

# Spearman correlation between SoilAs and Grain As of the testing sets
ggscatter(TestSet, x = "SoilAs", y = "GrainAs",
          add = "reg.line",
          color = "TestSet", palette = c("red","blue","black"),
          shape = 16,
          fullrange = FALSE,
          labels = "top",
          xlab = "Arsenic concentration in soil (mg kg-1)",
          ylab="Arsenic concentration in rice grain (µg kg-1) ",
)+ theme_grey()+
  stat_cor((aes(color = TestSet)),method = "spearman", label.x=1.7)


summary(TestSet)
sd(TestSet$SoilAs)
sd(TestSet$GrainAs)
```

D.2. R-Codes for training and testing of the models (Logistic Regression, Gradient Boost Machine and Random Forest)

```
# Read data file
str(RF)
RF$Category <- as.factor(RF$Category)
xtabs(~Category,data = RF)

# Partition data - train (80%) & test (20%)
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
ind <- sample(2, nrow(RF), replace = T, prob = c(0.8, 0.2))
train <- RF[ind==1,]
test <- RF[ind==2,]

library(randomForest)
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
RF<-randomForest(Category~.,data=train,ntree=300,mtry=1,
                 importance=TRUE,proximity=TRUE)
print(RF)
attributes(RF)


# Prediction with test data RF
#Confusion Matrix  - test data
PredRF_test<- predict(RF,test, type="response")
confusionMatrix(PredRF_test,test$Category)


# Prediction with train data
#Confusion Matrix train data
PredRF_train<- predict(RF,train, type="response")
confusionMatrix(PredRF_train,train$Category)

# Error rate of RF
plot(RF)
train <- as.data.frame(train)
t<-tuneRF(train[,-1],train[,1],stepFactor=0.5,plot=TRUE,
        ntreeTry=300,trace=TRUE,improve=0.05)

varImpPlot(RF)
varImp(RF)

# Get importance values as a data frame
imp = as.data.frame(importance(RF))
imp = cbind(vars=rownames(imp), imp)
imp = imp[order(imp$MeanDecreaseAccuracy),]
imp$vars = factor(imp$vars, levels=unique(imp$vars))

barplot(imp$MeanDecreaseGini, names.arg=imp$vars)

imp %>%
  pivot_longer(cols=matches("Mean")) %>%
  ggplot(aes(value, vars)) +
```

```
  geom_col() +
  geom_text(aes(label=round(value), x=0.5*value), size=5, colour="white")
+
  facet_grid(. ~ name, scales="free_x") +
  scale_x_continuous(expand=expansion(c(0,0.04))) +
  theme_bw() +
  theme(panel.grid.minor=element_blank(),
        panel.grid.major=element_blank(),
        axis.title=element_blank())



#Accuracy and cutoff on train
PredRF_train<- predict(RF,train, type="prob")
predRF<-prediction(PredRF_train[,2],train$Category)
evalRF<-performance(predRF,"acc")
plot(evalRF)

#Identifying the best values
max<-which.max(slot(evalRF,"y.values")[[1]])
acc<-slot(evalRF,"y.values")[[1]][max]
cut<-slot(evalRF,"x.values")[[1]][max]
print(c(Accuracy=acc,Cutoff=cut))

acc.perf = performance(predRF, measure = "acc")
plot(acc.perf,lwd=3,col="blue")
abline(v=0.57)

# ROC Curve
PredRF_train<- predict(RF,train,type = "prob")
par(pty="s")
ROC_train<-roc(train$Category,PredRF_train[,
2],plot=TRUE,legacy.axes=TRUE,percent=TRUE,
               col="#AD4433",lwd=4,print.auc=TRUE)


PredRF_test<- predict(RF,test, type="prob")
par(pty="s")
ROC_test<-roc(test$Category,PredRF_test[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)

partialPlot(RF,train,AvAs,"<MTC",ylab="Probability of <MTC",
            lwd=3,col="blue",abline(h=0))

partialPlot(RF,train,AvAs,">MTC",ylab="Probability of >MTC",
            lwd=3,col="blue")



fitControl1<-trainControl(method = "boot",
                          search = "random")

fitControl2 <- trainControl(method = "repeatedcv",
                            number = 10,repeats = 5)
```

```
#Repeated cross validation performed better than bootstrap on test data

#Random Forest
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
RF<-train(Category~.,data=train,
           method='rf', trControl=fitControl2,tuneLength=10,ntree=1000)
print(RF)
RF$bestTune
varImp(RF)
plot(varImp(RF,scale = T))
plot(RF)

# Prediction with test data RF
#Confusion Matrix  - test data
PredRF_test<- predict(RF,test, type="raw")
confusionMatrix(PredRF_test,test$Category)

# Prediction with train data
#Confusion Matrix train data
PredRF_train<- predict(RF,train, type="raw")
confusionMatrix(PredRF_train,train$Category)


# ROC Curve
PredRF_train<- predict(RF,train,type = "prob")
par(pty="s")
ROC_train<-roc(train$Category,PredRF_train[,
2],plot=TRUE,legacy.axes=TRUE,percent=TRUE,
               col="#AD4433",lwd=4,print.auc=TRUE)


PredRF_test<- predict(RF,test, type="prob")
par(pty="s")
ROC_test<-roc(test$Category,PredRF_test[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)

#Accuracy and cutoff on train
predRF<-prediction(PredRF_train[,2],train$Category)
evalRF<-performance(predRF,"acc")
plot(evalRF)

#Identifying the best values
max<-which.max(slot(evalRF,"y.values")[[1]])
acc<-slot(evalRF,"y.values")[[1]][max]
cut<-slot(evalRF,"x.values")[[1]][max]
print(c(Accuracy=acc,Cutoff=cut))

acc.perf = performance(predRF, measure = "acc")
plot(acc.perf,lwd=3,col="blue")
abline(h=1.0,v=0.62)

#Accuracy and cutoff on test
predRF<-prediction(PredRF_test[,2],test$Category)
```

```
evalRF<-performance(predRF,"acc")
plot(evalRF)

#Identifying the best values
max<-which.max(slot(evalRF,"y.values")[[1]])
acc<-slot(evalRF,"y.values")[[1]][max]
cut<-slot(evalRF,"x.values")[[1]][max]
print(c(Accuracy=acc,Cutoff=cut))

acc.perf = performance(predRF, measure = "acc")
plot(acc.perf,lwd=3,col="blue",)
abline(h=0.90,v=0.43)
```

```
# PDP partial dependence plots for AvAs
library("pdp")
p1<- partial(RF,pred.var = c("TAs","AvAs"),prob = TRUE, plot = TRUE, chull
= TRUE)
p1

p2 <- partial(RF,pred.var = c("AvFe","AvAs"),prob = TRUE, plot = TRUE,
chull = TRUE)
p2

p3 <- partial(RF,pred.var = c("OC","AvAs"),prob = TRUE, plot = TRUE, chull
= TRUE)
p3

p4 <- partial(RF,pred.var = c("AvP","AvAs"),prob = TRUE, plot = TRUE,
chull = TRUE)
p4

grid.arrange(p1,p2,p3,p4, nrow = 2,ncol=2)

library(magrittr)
library(fpp2)
# ggplot2-based PDP partial dependence plots
p5 <- RF %>%  # the %>% operator is read as "and then"
  partial(pred.var = "AvAs",prob = TRUE,which.class = ">MTC") %>%
  autoplot(smooth=FALSE,ylab = expression(Probability>MTC),xlab =
expression(AvailableAs)) +
  theme_grey()
p5

p6 <- RF %>%  # the %>% operator is read as "and then"
  partial(pred.var = "AvAs",prob = TRUE,which.class = "<MTC") %>%
  autoplot(smooth=FALSE,ylab = expression(Probability<MTC),xlab =
expression(AvailableAs)) +
  theme_grey()
p6


p7<- RF %>%  # the %>% operator is read as "and then"
```

```
  partial(pred.var = "AvAs",prob = TRUE,which.class = "<MTC") %>%
  autoplot(smooth=FALSE,ylab = expression(Probability<MTC),xlab =
expression(AvailableAs)) +
  theme_grey()+ ylim(0.25,0.8)+ xlim(4,7)+
  geom_hline(yintercept =0.62,color="red",lwd=1)+
  geom_vline(xintercept =5.72,color="green",lwd=1)
p7



library(reshape2)
#ICE Plots
AVAsice <- partial(RF, pred.var = "AvAs",prob = TRUE, ice = TRUE,center.at
= 0)
I1 <- plotPartial(AVAsice, alpha = 0.5,pdp.col = "yellow",
                  pdp.lwd = 3, pdp.lty = 1, ylab="Probability of <MTC",
                  xlab="Available As in soil (mg kg-1)")
I1

grid.arrange(p5,I1,ncol=2,nrow=1)
AVFeice <- partial(RF,prob = TRUE, pred.var = "AvFe", ice = TRUE,center =
TRUE)
I2 <- plotPartial(AVFeice, alpha = 0.5,pdp.col = "yellow",
                  pdp.lwd = 2, pdp.lty = 1, ylab="GrainAs")
I2

OCice <- partial(RF,prob = TRUE, pred.var = "OC", ice = TRUE)
I3 <- plotPartial(OCice, alpha =0.5 ,pdp.col = "yellow",
                  pdp.lwd = 2, pdp.lty = 1, ylab="GrainAs")
I3

pHice <- partial(RF,prob = TRUE, pred.var = "pH", ice = TRUE)
I4 <- plotPartial(pHice, alpha = 0.5,pdp.col = "yellow",
                  pdp.lwd = 2, pdp.lty = 1, ylab="GrainAs")
I4

AvPice <- partial(RF, prob = TRUE, pred.var = "AvP", ice = TRUE)
I5 <- plotPartial(AvPice, alpha = 0.5,pdp.col = "yellow",
                  pdp.lwd = 2, pdp.lty = 1, ylab="GrainAs")
I5

grid.arrange(I1,I2,I3,I4,I5, nrow = 2,ncol=2,layout_matrix= rbind(c(1,2),
3))


# Logistic Regression
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
LR<- train(Category ~.,
           data = train,
           trControl=fitControl1,
           method = "glm",
           family=binomial())
print(LR)
LR$finalModel
```

```
summary(LR)

# Logistic Regression with OC, AvFe and AvAs
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
LR<- train(Category ~OC+AvFe+AvAs+TAs,
           data = train,
           trControl=fitControl2,
           method = "glm",
           family=binomial())
LR
LR$finalModel
summary(LR)

#Accuracy and cutoff on train
predLR<-prediction(PredLR_train[,2],train$Category)
evalLR<-performance(predLR,"acc")
plot(evalLR)

#Identifying the best values
max<-which.max(slot(evalLR,"y.values")[[1]])
acc<-slot(evalLR,"y.values")[[1]][max]
cut<-slot(evalLR,"x.values")[[1]][max]
print(c(Accuracy=acc,Cutoff=cut))

acc.perf = performance(predLR, measure = "acc")
plot(acc.perf,lwd=3,col="blue")
abline(v=0.51,h=0.92)


# Prediction with test data Logistic Regression
#Confusion Matrix  - test data
PredLR_test<- predict(LR,test,type="raw")
confusionMatrix(PredLR_test,test$Category)


# Prediction with train data
#Confusion Matrix  - test data
PredLR_train<- predict(LR,train,type="raw")
confusionMatrix(PredLR_train,train$Category)

# ROC Curve
PredLR_train<- predict(LR,train,type = "prob")
par(pty="s")
ROC_train<-roc(train$Category,PredLR_train[,
2],plot=TRUE,legacy.axes=TRUE,percent=TRUE,
               col="#AD4433",lwd=4,print.auc=TRUE)


PredLR_test<- predict(LR,test, type="prob")
par(pty="s")
ROC_test<-roc(test$Category,PredLR_test[,
2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
              col= "#4C00FF",lwd=4,print.auc=TRUE)
```

```
P1 <- LR %>%  # the %>% operator is read as "and then"
  partial(pred.var = "AvAs",prob = TRUE) %>%
  autoplot(smooth=FALSE,ylab = expression(Probability<MTC),xlab =
expression(AvailableAs)) +
  theme_grey()
P1

P2<- LR %>%  # the %>% operator is read as "and then"
  partial(pred.var = "AvAs",prob = TRUE) %>%
  autoplot(smooth=FALSE,ylab = expression(Probability<MTC),xlab =
expression(AvailableAs)) +
  theme_grey()+ ylim(0.4,0.8)+ xlim(5,6)+
  geom_hline(yintercept =0.51,color="red",lwd=1)+
  geom_vline(xintercept =5.70,color="green",lwd=1)
P2

#ICE Plots
AVAsice <- partial(LR, pred.var = "AvAs",prob = TRUE, ice = TRUE,center.at
= 0)
I1 <- plotPartial(AVAsice, alpha = 0.5,pdp.col = "yellow",
                  pdp.lwd = 3, pdp.lty = 1, ylab="Probability of <MTC",
                  xlab="Available As in soil (mg kg-1)")
I1

P3<- partial(LR,pred.var = c("TAs","AvAs"),prob = TRUE, plot = TRUE, chull
= TRUE)
P3

P4 <- partial(LR,pred.var = c("AvFe","AvAs"),prob = TRUE, plot = TRUE,
chull = TRUE)
P4

P5 <- partial(LR,pred.var = c("OC","AvAs"),prob = TRUE, plot = TRUE, chull
= TRUE)
P5

grid.arrange(P3,P4,P5, nrow = 2,ncol=2,layout_matrix= rbind(c(1,2), 3))

grid.arrange(P3,P4,P5, nrow = 2,ncol=2,layout_matrix=rbind(c(1,1, 2,2),
c(NA, 3, 3,NA)))


#Decision Tree
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
DT<- train(Category~.,data=train,
           method = "rpart",
           trControl=train_control,
           tuneLength = 10)
DT
DT_best <- DT$finalModel
DT_best
DT_best$cptable
rpart.plot(DT_best,extra=104)
```

```
rpart.plot(DT_best,extra = 6)
summary(DT_best)
##Decision Tree Not working



#Correlation Plot
library(psych)
pairs.panels(RF[-2],
             gap = 0,
             method = "spearman",
             cor = TRUE,sclae= TRUE,stars = TRUE,
             hist.col = "skyblue",cex.cor = 1,
             alpha = 0.05,
             pch=20)

M<-lm(GrainAs~pH+OC+AvAs+AvFe+AvP+TAs,data=RF)
summary(M)
vif(M)
```

D.3. R-Codes for training and testing of the models (Logistic Regression and Linear Discriminant Analysis) for prediction of irrigation water As

```
 Read data file
# Converting the categorical variable as factor
str(Data_All)
Data_All$Category <- as.factor(Data_All$Category)
xtabs(~Category,data = Data_All)

# Partition data - train (80%) & test (20%)-----------------------------------------------------
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
ind <- sample(2, nrow(Data_All), replace = T, prob = c(0.8, 0.2))
train <- Data_All[ind==1,]
test <- Data_All[ind==2,]


library(caret)
library(caretEnsemble)
library(pdp)
library(ROCR)
library(pROC)


#########################################################################################################

# Logistic Regression-------------------------------------------------------------------
fitControl2 <- trainControl(method = "repeatedcv",
                            number = 10,repeats = 5,
                            savePredictions = TRUE,
                            classProbs = TRUE,
                            summaryFunction = twoClassSummary)

set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
LR<- train(Category ~.,data = train,
                                trControl=fitControl2,
                                method = "glm",
                                family=binomial(),
                                metric="ROC")

LR
LR1$finalModel
summary(LR)
confusionMatrix(LR)
anova(LR,test="Chisq")




# Plot Average ROC ----------------------------------------------------------------------
for_lift <- data.frame(Category = LR$pred$obs, glm = LR$pred$B)
lift_obj <- lift(Category ~ glm, data = for_lift, class = "B")

ggplot(lift_obj$data) +
  geom_line(aes(1 - Sp, Sn, color = liftModelVar)) +
  scale_color_discrete(guide = guide_legend(title = "method"))



# Prediction with test data Logistic Regression-----------------------------------------
#Confusion Matrix  - test data
PredLR_test<- predict(LR,test,type="raw")
confusionMatrix(PredLR_test,test$Category)



# Prediction with train data-------------------------------------------------------------
#Confusion Matrix  - train data
PredLR_train<- predict(LR,train,type="raw")
confusionMatrix(PredLR_train,train$Category)
```

```
# ROC Curve + CutOff-------------------------------------------------------------------------
PredLR_train<- predict(LR,train,type = "prob")
par(pty="s")
ROC_train_LR<-roc(train$Category,PredLR_train[,2],plot=TRUE,legacy.axes=TRUE,percent=TRUE,
                  col="#AD4433",lwd=4,print.auc=TRUE,ci.thresholds=TRUE)

ROC_train_LR<-roc(train$Category,PredLR_train[,2])
coords(ROC_train_LR, "best", ret="threshold", transpose = FALSE)



PredLR_test<- predict(LR,test, type="prob")
par(pty="s")
ROC_test_LR<-roc(test$Category,PredLR_test[,2],plot=TRUE,percent=TRUE,legacy.axes=TRUE,
                 col= "#4C00FF",lwd=4,print.auc=TRUE)

coords(ROC_test_LR, "best", ret="threshold", transpose = FALSE)


#Arranging the ROC plots, Run the codes at one go-----------------------------------------
par(mfrow = c(2, 2))

PredLR_train<- predict(LR,train,type = "prob")
ROC_train_LR<-roc(train$Category,PredLR_train[,2])
plot(ROC_train_LR,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("green","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightblue",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

PredLR_test<- predict(LR,test, type="prob")
ROC_test_LR<-roc(test$Category,PredLR_test[,2])
plot(ROC_test_LR,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("blue","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightgreen",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

#Calculating Log Loss for test set-------------------------------------------------------
mResults = predict(LR, test, na.action = na.pass, type = "prob")
mResults$obs = test$Category
head(mResults)
mnLogLoss(mResults, lev = levels(mResults$obs))

#Calculating Log Loss for train set------------------------------------------------------------
mResults = predict(LR,train, na.action = na.pass, type = "prob")
mResults$obs = train$Category
head(mResults)
mnLogLoss(mResults, lev = levels(mResults$obs))
```

```
# PDP of IrriAs with significant soil parameters from the LR Model------------------------------

p1<- partial(LR,pred.var = c("IrriAs","pH"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p1

p2<- partial(LR,pred.var = c("IrriAs","OC"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p2

p3<- partial(LR,pred.var = c("IrriAs","Clay"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p3

p4<- partial(LR,pred.var = c("IrriAs","AvFe"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p4

p5<- partial(LR,pred.var = c("IrriAs","AvP"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p5

p6<- partial(LR,pred.var = c("IrriAs","TAs"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p6

# Arranging the PDPs-----------------------------------------------------------------------------
grid.arrange(p1,p2,p3,p4,p5,p6, nrow = 3,ncol=2)

grid.arrange(P1,P2,P3, nrow = 2,ncol=2,layout_matrix=rbind(c(1,1, 2,2), c(NA, 3, 3,NA)))


# PDP for IrriAs with respect to class B with LR Model at test cut-Off---------------------------------------
P1<-partial(LR, pred.var = "IrriAs", prob = TRUE, plot = TRUE,which.class = "B",
            chull = TRUE, progress = "text",contour = TRUE,plot.engine = "ggplot2")+
  ylab("Probability of B (bioavailable As < 5.70 mg kg-1)")+
  xlab("Concentration of As in irrigation water (µgL-1)")+
  geom_hline(yintercept =0.84,color="red",lwd=1)+
  geom_vline(xintercept =190,color="blue",lwd=1)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=12,face="bold"))

P1

# PDP for IrriAs with respect to class A with LR Model at train cut-Off---------------------------------------
P2<-partial(LR, pred.var = "IrriAs", prob = TRUE, plot = TRUE,which.class = "A",
            chull = TRUE, progress = "text",contour = TRUE,plot.engine = "ggplot2")+
  ylab("Probability of A (bioavailable As > 5.70 mg kg-1)")+
  xlab("Concentration of As in irrigation water (µgL-1)")+
  geom_hline(yintercept =0.16,color="red",lwd=1)+
  geom_vline(xintercept =190,color="blue",lwd=1)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=12,face="bold"))

P2

P1<-partial(LR, pred.var = "IrriAs", prob = TRUE, plot = TRUE,which.class = "B",
            chull = TRUE, progress = "text",contour = TRUE,plot.engine = "ggplot2")
P1

#ICE Plot of IrriAs from LR Model---------------------------------------------------------------------
IrriAsice <- partial(LR, pred.var = "IrriAs",prob = TRUE, ice = TRUE,center.at =
                       0,which.class="B")
I1 <- plotPartial(IrriAsice, alpha = 0.5,plot.pdp = FALSE,
                  pdp.lty = 1,
                  ylab="Probability of B (bioavailable As < 5.70 mg kg-1)",
                  xlab="As concentration in irrigation water (µgL-1)")
I1

IrriAsice <- partial(LR, pred.var = "IrriAs",prob = TRUE, ice = TRUE,center.at =
                       0,which.class="A")
I2 <- plotPartial(IrriAsice, alpha = 0.5,plot.pdp = FALSE,
                  pdp.lty = 1,
                  ylab="Probability of A (bioavailable As > 5.70 mg kg-1)",
                  xlab="As concentration in irrigation water (µgL-1)")
I2


grid.arrange(I1,P1,ncol=1)
grid.arrange(I2,P2,ncol=1)
```

```
# Linear Discriminant Analysis----------------------------------------------------------------------
fitControl2 <- trainControl(method = "repeatedcv",
                            number = 10,repeats = 5,
                            savePredictions = TRUE,
                            classProbs = TRUE,
                            summaryFunction = twoClassSummary)
set.seed(1234)
set.seed(1000)
set.seed(5678)
set.seed(1090)
set.seed(7186)
LDA<- train(Category ~.,data = train,
            trControl=fitControl2,
            method = "lda",
            family=binomial(),
            metric="ROC")

LDA$finalModel
summary(LDA)
confusionMatrix(LDA)

# Prediction with test data Logistic Regression-------------------------------------
#Confusion Matrix  - test data
PredLDA_test<- predict(LDA,test,type="raw")
confusionMatrix(PredLDA_test,test$Category)


# Prediction with train data-------------------------------------------------------
#Confusion Matrix  - train data
PredLDA_train<- predict(LDA,train,type="raw")
confusionMatrix(PredLDA_train,train$Category)

#Arranging the ROC plots, Run the codes at one go-------------------------------------------
par(mfrow = c(2, 2))

PredLDA_train<- predict(LDA,train,type = "prob")
ROC_train_LDA<-roc(train$Category,PredLDA_train[,2])
plot(ROC_train_LDA,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("green","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightblue",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

PredLDA_test<- predict(LDA,test, type="prob")
ROC_test_LDA<-roc(test$Category,PredLDA_test[,2])
plot(ROC_test_LDA,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("blue","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightgreen",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

#Calculating Log Loss for test set-----------------------------------------------------
mResults = predict(LDA, test, na.action = na.pass, type = "prob")
mResults$obs = test$Category
head(mResults)
mnLogLoss(mResults, lev = levels(mResults$obs))

#Calculating Log Loss for train set-----------------------------------------------------------
mResults = predict(LDA,train, na.action = na.pass, type = "prob")
mResults$obs = train$Category
head(mResults)
mnLogLoss(mResults, lev = levels(mResults$obs))
```

```
# PDP of IrriAs with significant soil parameters from the LDA Model------------------------------

p7<- partial(LDA,pred.var = c("IrriAs","pH"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p7

p8<- partial(LDA,pred.var = c("IrriAs","OC"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p8

p9<- partial(LDA,pred.var = c("IrriAs","Clay"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p9

p10<- partial(LDA,pred.var = c("IrriAs","AvFe"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p10

p11<- partial(LDA,pred.var = c("IrriAs","AvP"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p11

p12<- partial(LDA,pred.var = c("IrriAs","TAs"),which.class="B",prob = TRUE, plot = TRUE, chull = TRUE)
p12

grid.arrange(p7,p8,p9,p10,p11,p12, nrow = 3,ncol=2)


# PDP for IrriAs with respect to class B with LDA Model at test cut-Off--------------------------------------
P3<-partial(LDA, pred.var = "IrriAs", prob = TRUE, plot = TRUE,which.class = "B",
            chull = TRUE, progress = "text",contour = TRUE,plot.engine = "ggplot2")+
  ylab("Probability of B (bioavailable As < 5.70 mg kg-1)")+
  xlab("Concentration of As in irrigation water (µgL-1)")+
  geom_hline(yintercept =0.84,color="red",lwd=1)+
  geom_vline(xintercept =260,color="blue",lwd=1)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=12,face="bold"))

P3

# PDP for IrriAs with respect to class A with LDA Model at train cut-Off--------------------------------------
P4<-partial(LDA, pred.var = "IrriAs", prob = TRUE, plot = TRUE,which.class = "A",
            chull = TRUE, progress = "text",contour = TRUE,plot.engine = "ggplot2")+
  ylab("Probability of A (bioavailable As > 5.70 mg kg-1)")+
  xlab("Concentration of As in irrigation water (µgL-1)")+
  geom_hline(yintercept =0.16,color="red",lwd=1)+
  geom_vline(xintercept =190,color="blue",lwd=1)+
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=12,face="bold"))

P4



#ICE Plot of IrriAs from LDA Model-------------------------------------------------------------------
IrriAsice <- partial(LDA, pred.var = "IrriAs",prob = TRUE, ice = TRUE,center.at =
                       0,which.class="B")
I3 <- plotPartial(IrriAsice, alpha = 0.5,plot.pdp = FALSE,
                  pdp.lty = 1,
                  ylab="Probability of B (bioavailable As < 5.70 mg kg-1)",
                  xlab="As concentration in irrigation water (µgL-1)")
I3

IrriAsice <- partial(LDA, pred.var = "IrriAs",prob = TRUE, ice = TRUE,center.at =
                       0,which.class="A")
I4 <- plotPartial(IrriAsice, alpha = 0.5,plot.pdp = FALSE,
                  pdp.lty = 1,
                  ylab="Probability of A (bioavailable As > 5.70 mg kg-1)",
                  xlab="As concentration in irrigation water (µgL-1)")
I4


grid.arrange(I3,P3,ncol=1)
grid.arrange(I4,P4,ncol=1)
```

```
#Arranging plots of LDA and LR-------------------------------------------------------

par(mfrow = c(2, 2))

PredLR_train<- predict(LR,train,type = "prob")
ROC_train_LR<-roc(train$Category,PredLR_train[,2])
plot(ROC_train_LR,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("green","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightblue",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

PredLR_test<- predict(LR,test, type="prob")
ROC_test_LR<-roc(test$Category,PredLR_test[,2])
plot(ROC_test_LR,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("blue","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightgreen",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

PredLDA_train<- predict(LDA,train,type = "prob")
ROC_train_LDA<-roc(train$Category,PredLDA_train[,2])
plot(ROC_train_LDA,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("green","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightblue",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)

PredLDA_test<- predict(LDA,test, type="prob")
ROC_test_LDA<-roc(test$Category,PredLDA_test[,2])
plot(ROC_test_LDA,
     print.auc=T,
     print.auc.cex=1.5,
     auc.polygon=T,
     grid=c(0.1,0.1),
     grid.col=c("blue","red"),
     max.auc.polygon=T,
     auc.polygon.col="lightgreen",
     print.thres=T,
     print.thres.cex=1.5,
     cex.lab=1.7, cex.sub=1.2)
##############################################################################################
```