Article

Predicting the Zinc Content in Rice from Farmland Using Machine Learning Models: Insights from Universal Geochemical Parameters

Wenda Geng [1], Tingting Li [2], Xin Zhu [2], Lei Dou [2], Zijia Liu [3,4], Kun Qian [1], Guiqi Ye [1], Kun Lin [1], Bo Li [1], Xudong Ma [1], Qingye Hou [1,5], Tao Yu [5,6] and Zhongfang Yang [1,5,*]

[1] School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China; gengwd0911@163.com (W.G.); shiqiankun515@126.com (K.Q.); yegq77@163.com (G.Y.); 3001200073@cugb.edu.cn (K.L.); bobo_lee0425@126.com (B.L.); maxudongcugb@126.com (X.M.); qingyehou@cugb.edu.cn (Q.H.)

[2] Guangdong Institute of Geological Survey, Guangzhou 510080, China; tingtingliscsio@126.com (T.L.); ddystar@163.com (X.Z.); ggsdl@163.com (L.D.)

[3] Research Center of Geochemical Survey and Assessment on Land Quality, Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Sciences, Langfang 065000, China; zijialiu@outlook.com

[4] Key Laboratory of Geochemical Cycling of Carbon and Mercury in the Earth's Critical Zone, Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Sciences, Langfang 065000, China

[5] Key Laboratory of Ecological Geochemistry, Ministry of Natural Resources, Beijing 100037, China; yutao@cugb.edu.cn

[6] School of Science, China University of Geosciences, Beijing 100083, China

* Correspondence: yangzf@cugb.edu.cn

Abstract: Zinc (Zn) is an essential nutrient for the human body and is prone to deficiency. Supplementing Zn through zinc-enriched cereals is of great significance in addressing the widespread issue of zinc deficiency. However, there is no simple linear correlation between the soil zinc content and rice grain zinc content, which poses challenges for zoning zinc-enriched rice cultivation based on the soil Zn content. Therefore, accurately predicting the zinc content in rice grains is of great importance. To verify the robustness of the prediction model and expand its applicability, this study established a prediction model using 371 sets of previously collected and tested rice grain and root zone soil samples from the Pearl River Delta and Heyuan regions in Guangdong. The model was validated using the data from 65 sets of rice and root zone soil samples collected and analyzed in Zijin and Dongyuan counties, Heyuan, in 2023. The results show that zinc absorption by rice grains is controlled by multiple factors, primarily related to the soil S, P, CaO, Mn, TFe_2O_3, TOC, and SiO_2/Al_2O_3 ratio. Both the artificial neural network model and random forest model demonstrated a good predictive performance across large regions. However, in the Heyuan region, the random forest model outperformed the artificial neural network model, with an R^2 of 0.79 and an RMSE of 0.05 when the predicted data were compared against the measured BAF_{Zn} of the rice. This suggests that predicting the zinc content in rice grains based on the soil macro-elements (including oxides) and TOC is feasible, and, within certain regional boundaries, the prediction model is robust and widely applicable. This study provides valuable insights into the rational development of zinc-enriched rice in the Heyuan region and offers a useful reference for establishing prediction models of the beneficial element content in rice grains in areas with limited data.

Keywords: Zinc; soil; artificial neural network model; random forest model; Guangdong

1. Introduction

Zinc (Zn) is an essential trace element for the human body, playing a crucial role in physiological processes such as growth, reproduction, immunity, and endocrine functions [1]. Zinc deficiency can lead to weakened immunity and impaired digestive function, and, in children, it may cause stunted growth or, in severe cases, dwarfism. Studies have shown that approximately 40% of the global population faces health issues related to zinc deficiency [2]. Most people obtain zinc primarily from food, and insufficient zinc intake is considered one of the major causes of zinc deficiency [3]. Rice, a staple food for over 3.5 billion people worldwide, is closely linked to dietary zinc intake, especially in Asia [4]. However, in many regions, the zinc content in rice is very low [5,6], insufficient to meet human nutritional needs [7]. Therefore, the development of natural zinc-enriched rice from zinc-rich soils is of great importance for addressing the widespread issue of zinc deficiency.

Rice's absorption of trace elements is influenced not only by the total amount of these elements in the soil but also by factors such as their bioavailability, the forms in which the elements occur in the soil, soil physicochemical properties, and the interactions between elements [8]. Previous research has shown that a high soil zinc content does not necessarily correlate with a high zinc content in rice grains. In most cases, there is no significant correlation between the two [9], making it difficult to classify zinc-rich soils based solely on the soil zinc content. Therefore, understanding the characteristics of rice's zinc absorption and constructing predictive models for the zinc content in rice grains is crucial for the development of zinc-rich land resources.

In recent years, several studies have established models for predicting the absorption of heavy metals by rice [10,11], and machine learning methods have been widely used in the construction of these models, achieving good predictive results. Artificial neural networks (ANNs) and random forest (RF) are two of the most commonly applied machine learning methods and have shown excellent performance in predicting the heavy metal content in crops [12,13]. However, due to the high costs of field data collection and the intensive labor, equipment, and financial support required, as well as the complexity and variability of the soil environments in agricultural regions [14,15], predicting the heavy metal content in soil–rice systems using machine learning remains challenging in different research areas [9].

Wang et al. [9] selected soil pH, TOC (total organic carbon), Mn, and Zn as the input variables and used an artificial neural network model to predict the bioaccumulation factor of zinc (BAF_{Zn}) in rice in farmland in central and eastern Guangxi, achieving a strong predictive effect with an R^2 of 0.93. Ma et al. [16] used Fe_2O_3, Mn, and SiO_2/Al_2O_3 as the input variables for a maize prediction model and Fe_2O_3, SiO_2/Al_2O_3, and P for a rice prediction model. They applied a backpropagation neural network model to predict the BAF_{Zn} in maize and rice in Guang'an county, Sichuan Province, achieving good predictive results. Similarly, Liu et al. [17] used a neural network algorithm and selected soil Zn, SiO_2/Al_2O_3, Fe_2O_3, and K_2O as the input variables to construct a predictive model for the wheat grain zinc content in the Weining Plain of Ningxia, with a good model fitting. Several scholars' studies have shown that, despite the differences in input variables and machine learning methods across different research regions, predicting the crop zinc content based on the major soil elements (including oxides) and TOC is feasible. However, in these studies, both the training and validation datasets were collected from the same research region. Whether the training and validation datasets from different regions can be used in machine learning to predict the crop grain zinc content and whether the model's robustness and generalization ability can meet the requirements remains a key issue explored in this paper.

The zinc content in rice grains in Guangdong Province is higher than the national average, indicating a great potential for the development of zinc-enriched rice. Although

there is no linear correlation between the zinc content in rice grains and in the soil, there have been numerous cases where machine learning has been used to predict the elemental content of rice grains [9,11,14,18]. This study focuses on the Pearl River Delta and Heyuan areas in Guangdong Province, using 371 datasets collected from these regions for the model training and 65 datasets from Heyuan for the validation. The aim is to address the robustness and generalizability of the machine learning models for predicting the zinc content in rice grains and provide a scientific basis for delineating zinc-rich rice cultivation zones in the study area.

Against this background, the objectives of this study are as follows: (1) to investigate the impact of soil properties on the absorption of zinc in rice grains; (2) to explore whether machine learning models for predicting the zinc content in rice grains, based on the major soil elements and the TOC content, exhibit robustness and generalizability across a defined geographical area; (3) to provide useful references for the rational development of zinc-rich rice in Heyuan and for constructing models to predict the beneficial element content in rice grains when data availability is limited in the study area.

2. Materials and Methods

2.1. Study Area

The study area is located in the central and southern regions of Guangdong Province, including the Pearl River Delta and Heyuan city (Figure 1). Geographically, it spans from 111°22′ E to 115°36′ E and 21°28′ N to 24°47′ N, covering an area of 71,022 km^2. The area lies in the South Asian subtropical zone, characterized by a South Asian monsoon climate with abundant rainfall, ample heat, and a simultaneous rainy and hot season. The annual sunshine duration is 2000 h, and the seasonal distribution is relatively uniform. The average annual temperature ranges from 21.4 to 22.4 °C, and the annual precipitation is between 1600 and 2300 mm, with most rainfall occurring between April and September due to the monsoon influence. The study area has a total rice planting area of 1,030,440 hectares, mainly distributed across the cities of Zhaoqing, Jiangmen, Heyuan, and Huizhou, with two harvests per year and a total annual yield of 3.9191 million tons.

The Pearl River Delta is a complex delta formed by the deposition of sediments from the Xijiang, Beijiang, and Dongjiang rivers and their tributaries. Approximately one-fifth of the area consists of hills, terraces, and residual hills, with the overall landscape being a plain formed by Quaternary sediments. The Heyuan area, on the other hand, is a mountainous and hilly region, with the Dongjiang and Xinfengjiang rivers running through the area. The landscape is characterized by alternating mountain ranges and basins, with granite as the primary parent material for soil formation.

2.2. Sample Collection

According to the standards for the multi-target regional geochemical surveys (DZ/T 0258-2014) and land quality geochemical evaluations (DZ/T 0295-2016) [19,20], a total of 371 rice grain and root soil samples were collected from the study area (Figure 2). Of these, 306 samples were collected from the Pearl River Delta region during 2016 and 2017, and 65 samples were collected from the Heyuan region in 2023. The samples were taken at the rice maturity stage, with multiple points sampled within each plot, and the samples were mixed in equal amounts to form a composite sample. Simultaneously, the corresponding root soil samples were collected by uprooting the rice plants and gently shaking off the soil from the roots.
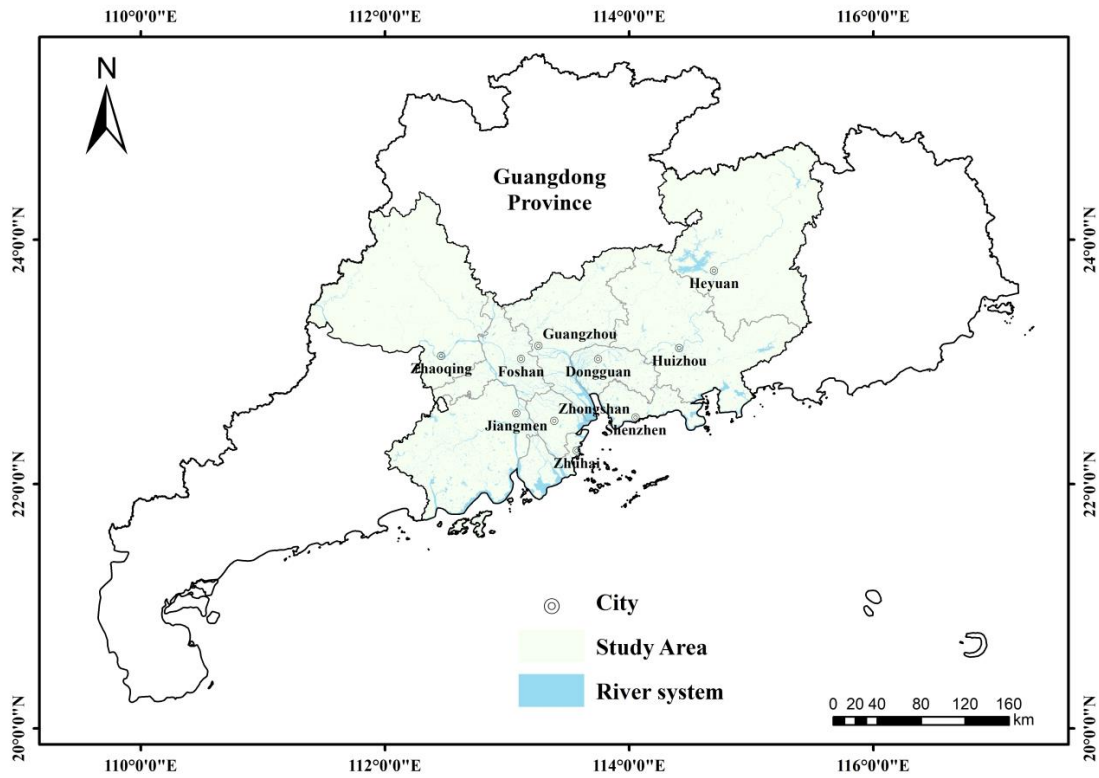
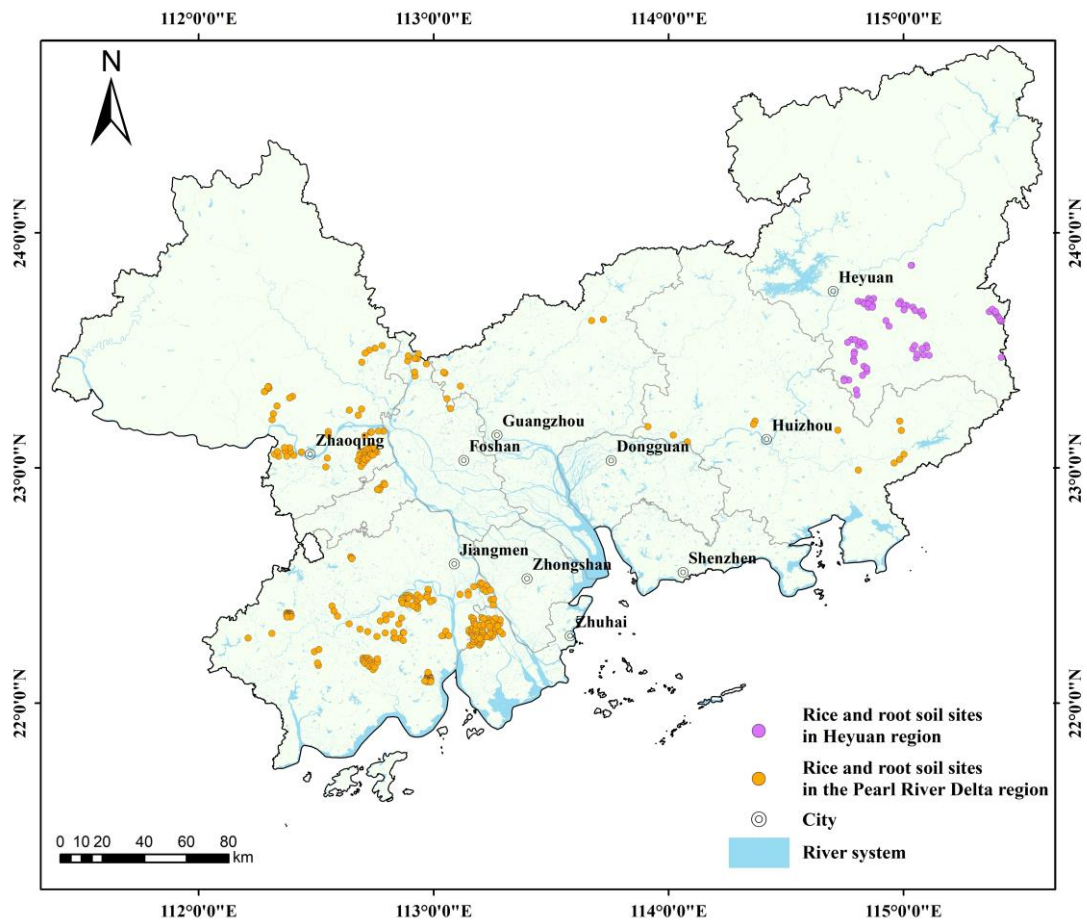**Figure 1.** Geographical location of study area.

Figure 2. Distribution of sampling sites in study area.

After collection, the rice grain and root soil samples were placed in a cool, ventilated area to air-dry. Once the rice grains were dried, they were threshed, weighed, and prepared for testing. During the drying process of the soil samples, the samples were periodically tapped with a rubber mallet to prevent clumping. After drying, the samples were sieved through a nylon sieve (with a 2 mm mesh size) using a shaking motion and stored in clean polyethylene bags for analysis.

2.3. Chemical Analysis

The rice grain and root soil samples were analyzed at the Anhui Geological Laboratory. The laboratory sample treatment and analytical process were as follows: After threshing the rice grains, they were rinsed with clean water, placed on a clean tray, and dried at temperatures below 65 °C, followed by dehulling. The dried and dehulled samples were then ground using a ceramic mortar and pestle until they passed through a 40-mesh (425 μm) to 60-mesh (250 μm) nylon sieve and were thoroughly mixed before testing. The root soil samples were ground in a ceramic mortar, sieved through a 200-mesh nylon sieve, and evenly mixed before testing.

The analysis followed the methods outlined in the DZ/T 0279-2016 "Regional Geochemical Sample Analysis Methods" and the GB5009.268-2016 "National Food Safety Standard for Multi-Element Determination in Foods" [21,22]. The specific analysis methods and detection limits are shown in Table 1. National primary standard materials and duplicate samples were used for quality control during the testing process. The accuracy of the methods was verified by calculating the logarithmic deviation (ΔlgC) or relative error (RE%) between the measured and standard values. The precision was evaluated by calculating the relative standard deviation (RSD%) between the measured and standard values [23]. All the detection limits, reporting rates, accuracy, and precision for the full element analysis methods met the monitoring limits specified, ensuring data reliability.

Table 1. Sample analysis method scheme and detection limit.

Samples	Item	Method	Detecting Limit
Soil	SiO_2	X-ray Fluorescent Spectroscopy (XRF)	0.05 *
	TFe_2O_3		0.02 *
	Al_2O_3		0.03 *
	CaO		0.02 *
	Mn		10
	P		10
	S		50
	Zn		4
	TOC	Volumetric method (VOL)	0.1 *
	pH	Ion-selective electrode (ISE)	0.08 **
Rice	Zn	Inductively coupled plasma mass spectrometry (ICP-MS)	0.05

The unit of measurement for "*" is %, "**" is dimensionless, and other elements are in mg/kg.

2.4. Models' Development

The bioaccumulation factor (BAF) is used to measure the accumulation of zinc (Zn) in rice grains, and, based on the soil properties, relevant predictive models, and corresponding theoretical framework, a predictive model for the BAF_{Zn} of rice grains is established [24].

The bioaccumulation factor of Zn is calculated using the following formula:

$$BAF_{Zn} = \frac{C_{Rice\ Zn}}{C_{Soil\ Zn}} \tag{1}$$

where BAF$_{Zn}$ is the bioaccumulation factor for Zn, C$_{Rice\ Zn}$ is the zinc content in the rice grains, and C$_{Soil\ Zn}$ is the zinc content in the corresponding root soil.

Before constructing the model, the input features are logarithmically transformed and normalized. This preprocessing step is aimed at making the data have similar scales and an approximately normal distribution, which accelerates the convergence of the machine learning algorithms and improves the prediction accuracy [12].

2.4.1. Artificial Neural Network (ANN)

Artificial neural networks (ANNs) are powerful data modeling tools consisting of an input layer, hidden layers, and an output layer [25]. The nodes in the input layer receive the input information, which is processed by activation functions and passed to the nodes in the hidden layers. The information from the hidden layers is then sent to the output layer. Each node is connected to the others through corresponding weights and thresholds [26]. If the output of any individual node is above the specified threshold, then that node is activated and the data are sent to the next layer of the network. Otherwise, no data are passed on to the next layer of the network [27]. Many factors can affect the stability of an ANN model, including the model used, the modeling process, and the simulated training data. The optimal combination of neurons could not be determined in advance; therefore, the optimal structure of the network was determined by testing multiple models using correlation coefficient and root mean square error (RMSE) guidelines [28]. Artificial neural networks are widely applied in various research fields due to their strong learning ability and good predictive performance [29–32]. In the field of ecology, Li et al. [14] used an artificial neural network model to predict the cadmium content in rice, and the prediction results were optimistic.

In this study, the elastic backpropagation algorithm (RPROP) was used to train the ANN model, where the maximum step size of the neural network training was set to 100,000, and the partial derivative of the error function was specified as a stopping criterion threshold of 0.01. As a variant of backpropagation, the RPROP eliminates the influence of derivative changes on the weight step size and only considers the sign of the derivative, which enhances the efficiency and stability of the algorithm [33]. The ANN prediction model in this study was built using the neuralnet package in R Studio.

2.4.2. Random Forest (RF)

Random forest (RF) is a widely used ensemble learning method that combines multiple decision trees to cluster data based on similar patterns [34]. It is an extension of decision trees. The algorithm divides the feature space using hierarchical rules and builds each decision tree using a random subset of the predictor variables [35]. Each decision tree is built using bootstrap sampling, where random subsets of the training data are selected with replacements. Additionally, at each split in a tree, a random subset of features is considered, introducing further randomness that helps reduce the correlation between individual trees and enhances the overall performance of the models [36]. The final prediction is obtained by averaging the regression results of all the decision trees in the ensemble [37,38].

Random forest reduces model variance by combining multiple decision trees, which effectively lowers the risk of overfitting. It also provides feature importance, which helps in understanding the model and the data [39]. This method is particularly suitable for handling high-dimensional data, making it applicable across various fields. Zhao et al. [40] used the random forest model to predict heavy metal pollution in soil, and Ma et al. [11] used the random forest model to predict the cadmium content in paddy rice, both achieving good predictive results. The RF prediction model in this study was built using the random forest package in R Studio.

2.5. Models' Evaluation

To assess the accuracy of the predictive models, the rice grain and root soil samples were divided into a training set and a test set at an 8:2 ratio. The model was trained using the training set and evaluated using the test set to verify its accuracy. The accuracy of the predictive models was validated using the regression coefficient of determination (R^2), normalized mean error (NME), mean relative error (MRE), and root mean square error (RMSE) [41]. The formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(o_i - e_i)^2}{\sum_{i=1}^{N}(o_i - \overline{o})^2} \tag{2}$$

$$NME = \frac{\overline{e} - \overline{o}}{\overline{o}} \tag{3}$$

$$MRE = \frac{\sum_{i=1}^{N}\frac{|e_i - o_i|}{o_i}}{N} \times 100\% \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(e_i - o_i)^2}{N}} \tag{5}$$

where e_i and o_i represent the predicted and observed values for each sample, respectively; $\overline{e}$ and $\overline{o}$ are the average predicted and observed values; and N is the total number of samples.

The model is considered more accurate when R^2 approaches 1. A negative NME indicates that the BAF is underestimated, while a positive value indicates overestimation. The smaller the MRE and RMSE, the more accurate the model is.

3. Results and Discussion

3.1. Zn Content Characteristics in the Soil–Rice Ecosystem

Table 2 summarizes the Zn content and bioaccumulation factor (BAF_{Zn}) characteristics of the rice grains and root soil in the study area, including both the Heyuan and Pearl River Delta regions. In the Heyuan region, the Zn content in the root soil ranged from 32.00 to 156.70 mg/kg, with an average content of 72.30 mg/kg. According to the GB15618-2018 Soil Environmental Quality Risk Control Standards for Agricultural Land [42], the screening value for Zn is set at 200 mg/kg, meaning that the Zn content in all the root soil samples from the Heyuan region was below the screening value. The Zn content in the rice grains from the Heyuan region ranged from 13.60 to 25.40 mg/kg, with an average content of 18.47 mg/kg. Referring to the Chinese Food Composition Table Standard Edition 6, the national average Zn content in rice grains is 15.40 mg/kg. The proportion of samples from the Heyuan region with a Zn content higher than the national average was 84.6%, indicating substantial potential for developing Zn-enriched rice. The bioaccumulation factor (BAF_{Zn}) of the soil–rice system in the Heyuan region ranged from 0.10 to 0.67, with an average value of 0.28.

In the Pearl River Delta region, the Zn content in the root soil ranged from 16.90 to 253.30 mg/kg, with an average of 79.78 mg/kg. Only one sample exceeded the screening value of 200 mg/kg. The Zn content in the rice grains from the Pearl River Delta region ranged from 11.33 to 30.37 mg/kg, with an average of 18.10 mg/kg, and 83.33% of the samples had a Zn content higher than the national average, indicating substantial potential for Zn development in rice. The BAFZn in the soil–rice system of the Pearl River Delta region ranged from 0.06 to 1.19, with an average value of 0.31.

Table 2. Zn content and BAF$_{Zn}$ of rice grains and root soil in study area.

Sample Distribution Area	Sample Size	Type	Min Value (mg/kg)	Max Value (mg/kg)	Mean Value (mg/kg)	Median Value (mg/kg)	Coefficient of Variation
Heyuan Region	65	Root Soil	32.00	156.70	72.30	68.40	0.32
		Rice Grains	13.60	25.40	18.47	18.40	0.15
		BAF$_{Zn}$	0.10	0.67	0.28	0.25	0.35
Pearl River Delta	306	Root Soil	16.90	253.30	79.78	71.32	0.49
		Rice Grains	11.33	30.37	18.10	17.73	0.16
		BAF$_{Zn}$	0.06	1.19	0.31	0.26	0.64

Compared to Heyuan, the Pearl River Delta region showed a broader range of Zn content in both rice grains and root soil, as well as a wider range of BAF$_{Zn}$. However, the average and median values of the Zn content and BAF$_{Zn}$ in both regions were quite similar. As shown in Figure 3, the distribution characteristics of the Zn content in both regions are similar. This suggests that, across the entire study area, there may be a similar relationship between the soil Zn content and rice grain Zn content.
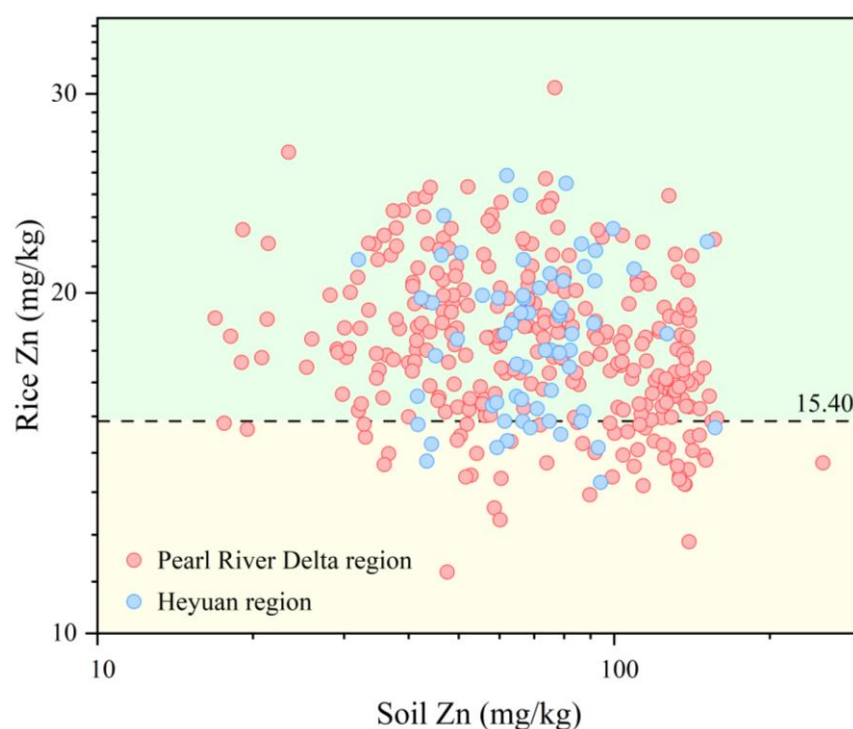

Figure 3. The relationship between soil Zn and rice Zn in different regions.

Figure 3 shows that there is no correlation between the soil Zn content and rice grain Zn content in the study area. Based on the Zn content in the soil, it is not possible to predict the Zn content in rice grains. Therefore, accurately predicting the Zn content in rice and identifying the Zn-enriched rice regions are crucial for the efficient and precise production of Zn-enriched rice.

3.2. Influence of Soil Properties on Zn Absorption by Rice Grains

The Zn content in rice grains is influenced by various factors. As shown in Figure 3, there is no significant correlation between the Zn content in the rice grains and the Zn

content in the root soil, indicating that the Zn content in the rice grains is not simply determined by the Zn content in the soil. Instead, it is closely related to the bioavailable form of Zn in the soil [43] and other physicochemical properties of the soil.

The relationship between the soil properties and the bioaccumulation factor of Zn (BAF_{Zn}) in the rice was evaluated using Pearson's correlation coefficient (Figure 4). The data were log-transformed to approximate a normal distribution. The results show that the BAF_{Zn} is negatively correlated with sulfur (S) ($r = -0.45$), phosphorus (P) ($r = -0.53$), calcium oxide (CaO) ($r = -0.75$), manganese (Mn) ($r = -0.72$), total iron (TFe_2O_3) ($r = -0.78$), and total organic carbon (TOC) ($r = -0.65$) and positively correlated with SiO_2/Al_2O_3 ($r = 0.69$). BAF_{Zn} has a strong correlation with CaO, Mn, and TFe_2O_3 and a moderate correlation with P, TOC, and SiO_2/Al_2O_3.
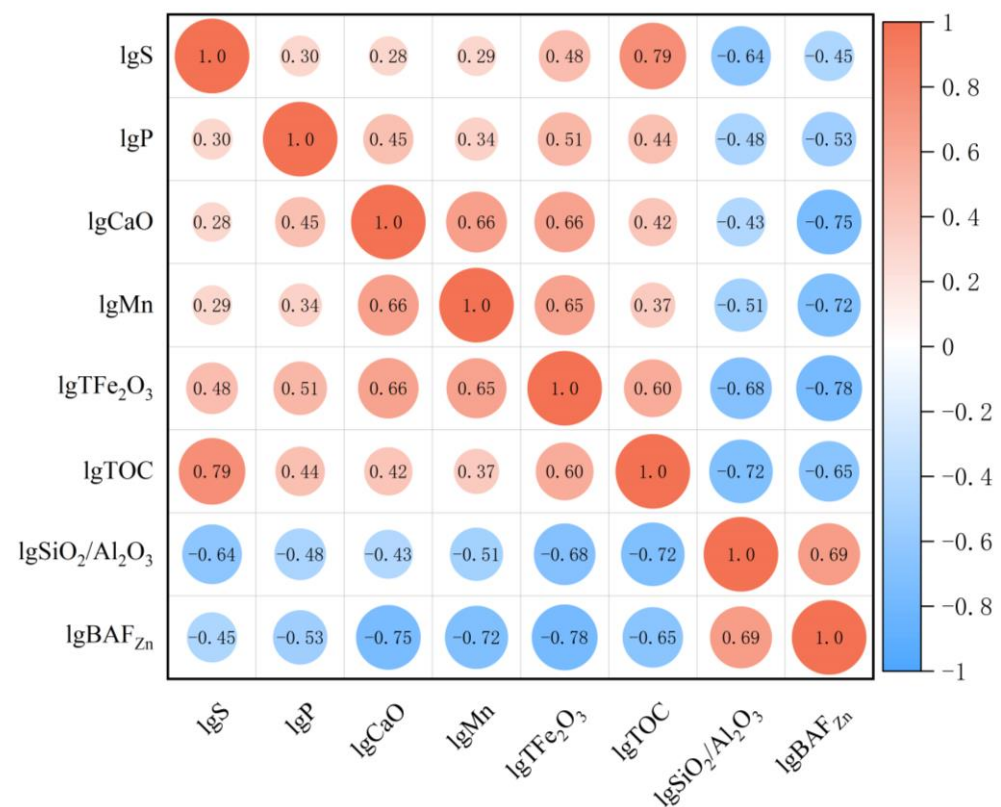
Figure 4. Correlation heatmap between soil properties and BAF_{Zn}.

Sulfur (S) can form insoluble zinc sulfide (ZnS), reducing the bioavailability of Zn in the soil [44,45]. In flooded paddy soils, the redox potential is low, and sulfur is often reduced to S^{2-}, which reacts with Zn^{2+} to form ZnS precipitates, reducing the mobility of Zn in the soil [46,47]. In addition to the above redox reactions, sulfur-containing organic compounds in the soil can form covalent bonds with heavy metals, playing a significant role in the activation and fixation of heavy metals.

The TOC content is used as an indicator of the soil organic matter content [48]. In general, the mobility and bioavailability of heavy metals increase as the organic matter content decreases [49]. Organic matter in the soil contains a variety of components, including abundant ligands or functional groups such as carboxyl (-COOH), hydroxyl (-OH), and phenolic hydroxyl groups (aromatic ring-OH), which can form complexes with metal ions and affect the migration and transformation of heavy metals [50].

Calcium oxide (CaO) is an effective buffering agent in soil. Increasing soil CaO raises soil pH [51], promoting the formation of the carbonate and hydroxide precipitates of heavy

metals, thereby reducing the bioavailability of heavy metals like Zn and Cd and inhibiting their absorption by plants.

Iron (Fe) and manganese (Mn) readily form iron and manganese oxide minerals in surface environments [52]. Zn and other heavy metals may be incorporated into the crystal lattice of these minerals [53], reducing their bioavailability. Additionally, iron and manganese hydroxide colloids have a significant effect on the adsorption of Zn^{2+} [54].

SiO_2/Al_2O_3 represents the degree of silicon depletion and aluminum enrichment in the soil. A higher SiO_2 content indicates a higher sand fraction in the soil and fewer clay minerals and organic matter, which reduces the soil's ability to adsorb Zn [55,56].

Phosphorus (P), as a nutrient element, shows a high negative correlation with the bioaccumulation factor of Zn in rice grains. Many studies suggest that P and Zn exhibit antagonistic effects [57,58]. Studies has found that high concentrations of P significantly reduced the Zn content and bioavailability in rice grains. Su et al. [59] observed that applying phosphorus fertilizer reduced the Zn content in wheat grains by 17% to 56%.

3.3. Development of Prediction Models

Based on the findings in Section 3.2, the soil properties, including S, P, CaO, Mn, TFe_2O_3, TOC, and SiO_2/Al_2O_3, were selected as the input variables and the BAF_{Zn} as the output variable for the model development. The dataset of 371 rice grain and root soil samples from the study area was divided into 80% for the training set and 20% for the validation set to build two models: the artificial neural network (ANN) and random forest (RF).

The ANN model structure is illustrated in Figure 5, with two hidden layers containing six and four neurons, respectively. The larger first layer and smaller second layer allow for the extraction of low-order features in the first layer, which are then used to extract higher order features in the second layer, potentially improving the ANN performance. The RF model had 800 trees (ntree = 800) and three features considered at each split (mtry = 3). The feature importance for the RF model is shown in Figure 6, where mtry plays a more significant role in optimizing the model compared to ntree.
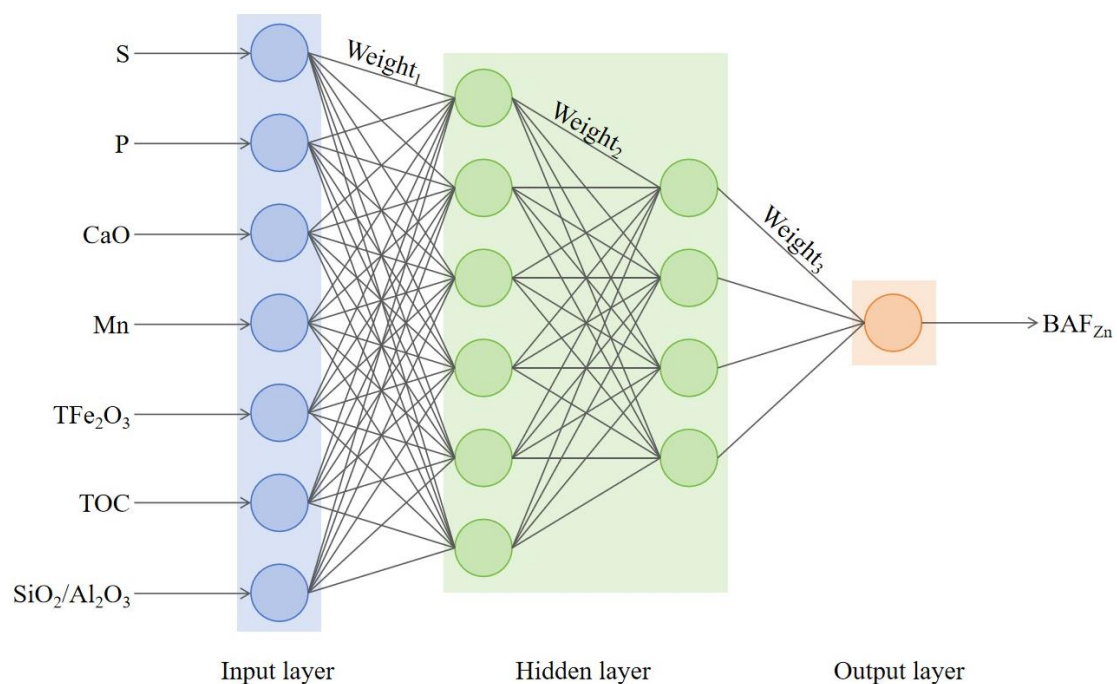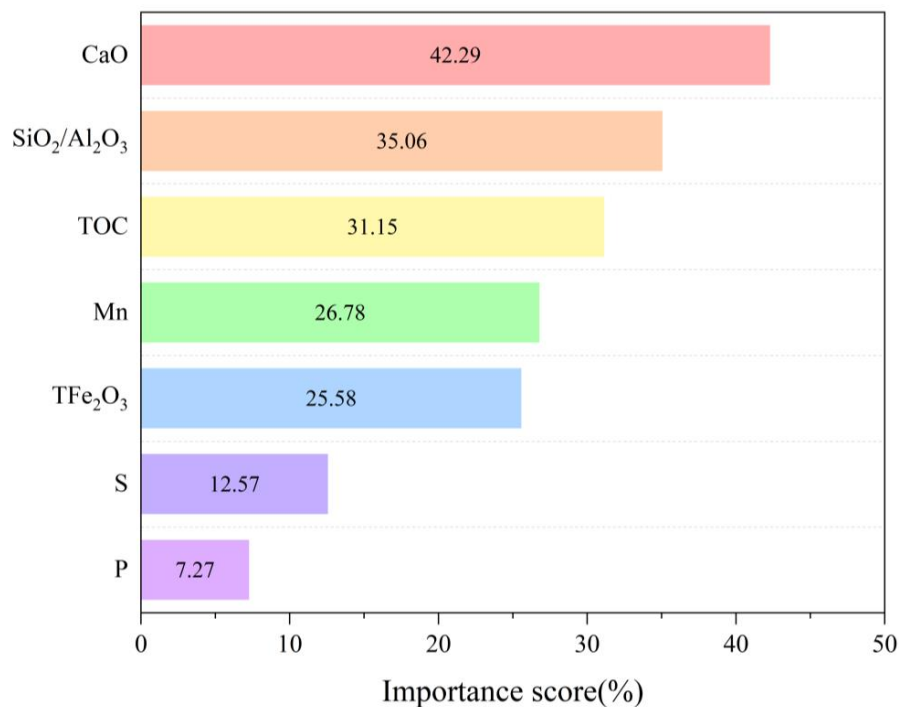

Figure 5. ANN model.

Figure 6. Importance score in random forest model.

The models' accuracies were assessed using R^2, the NME, the MRE, and the RMSE, and the results are summarized in Table 3. Both the ANN and RF models had an NME of 0.01, indicating that both models slightly overestimated the BAF_{Zn} values. The ANN model had MRE and RMSE values closer to zero than the RF model, although the difference was small. The predicted BAF_{Zn} values from both models were highly correlated with the measured values, and the predicted-to-observed ratio was nearly 1:1 (Figure 7). The R^2 value for the ANN model was slightly higher than for the RF model, indicating that both models are robust and provide reliable predictions.

Table 3. Accuracy of various models.

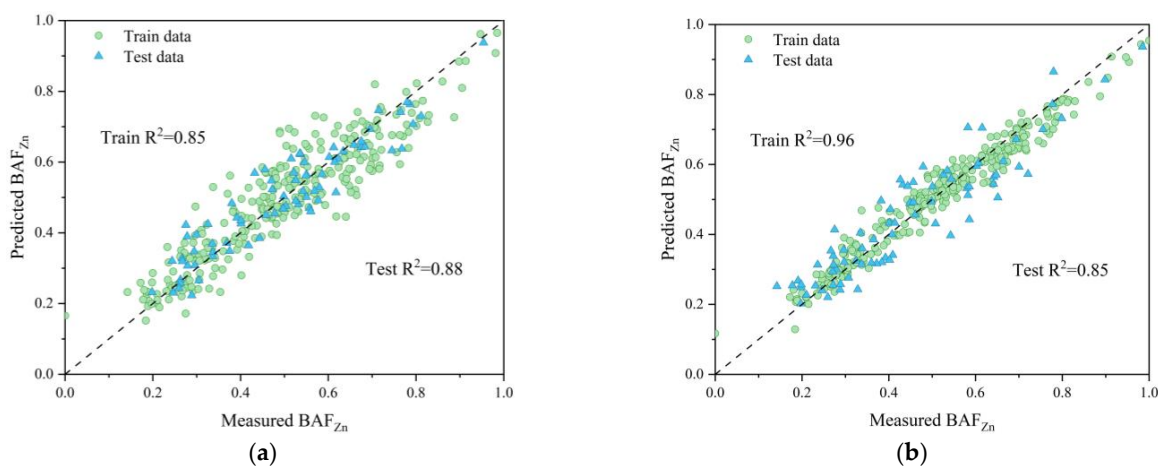| Model	Validation Sample Size	R^2	NME	MRE (%)	RMSE
Artificial Neural Network	75	0.88	0.01	11.02	0.06
Random Forest	77	0.85	0.01	15.09	0.07

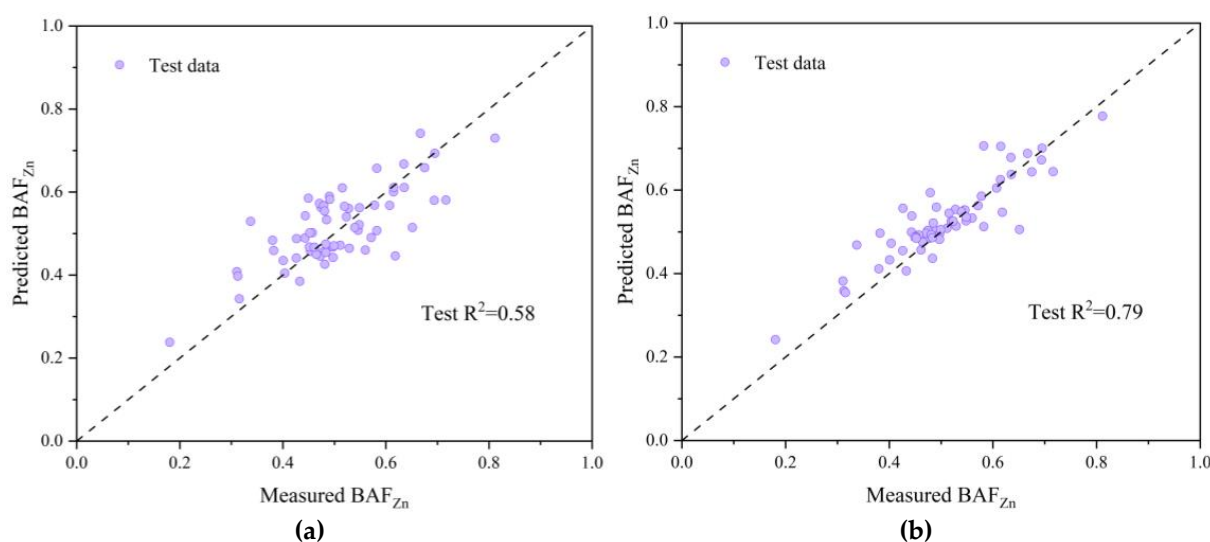**Figure 7.** Predicted BAF_{Zn} and measured BAF_{Zn} by (**a**) ANN and (**b**) RF.

3.4. Model Validation

To further validate whether the prediction models could be broadly applied to any small region within the study area, 65 rice grain and root soil samples from Heyuan, collected in 2023, were used as a validation set. The models' accuracies were again assessed using R^2, the NME, the MRE, and the RMSE, with the results summarized in Table 4.

Table 4. Accuracy of the model for Heyuan.

Model	Validation Sample Size	R^2	NME	MRE (%)	RMSE
Artificial Neural Network	65	0.58	0.02	11.93	0.07
Random Forest	65	0.79	0.04	8.28	0.05

Both the ANN and RF models had NME values of 0.02 and 0.04, respectively, indicating that both models overestimated the BAF_{Zn} values in Heyuan. The RF model performed better than the ANN model, with MRE and RMSE values closer to zero. The R^2 value for the RF model was higher, and the predicted values were closer to the 1:1 ratio (Figure 8). Overall, the RF model showed a better performance than the ANN model for Heyuan, demonstrating the broader applicability and robustness of the prediction models.
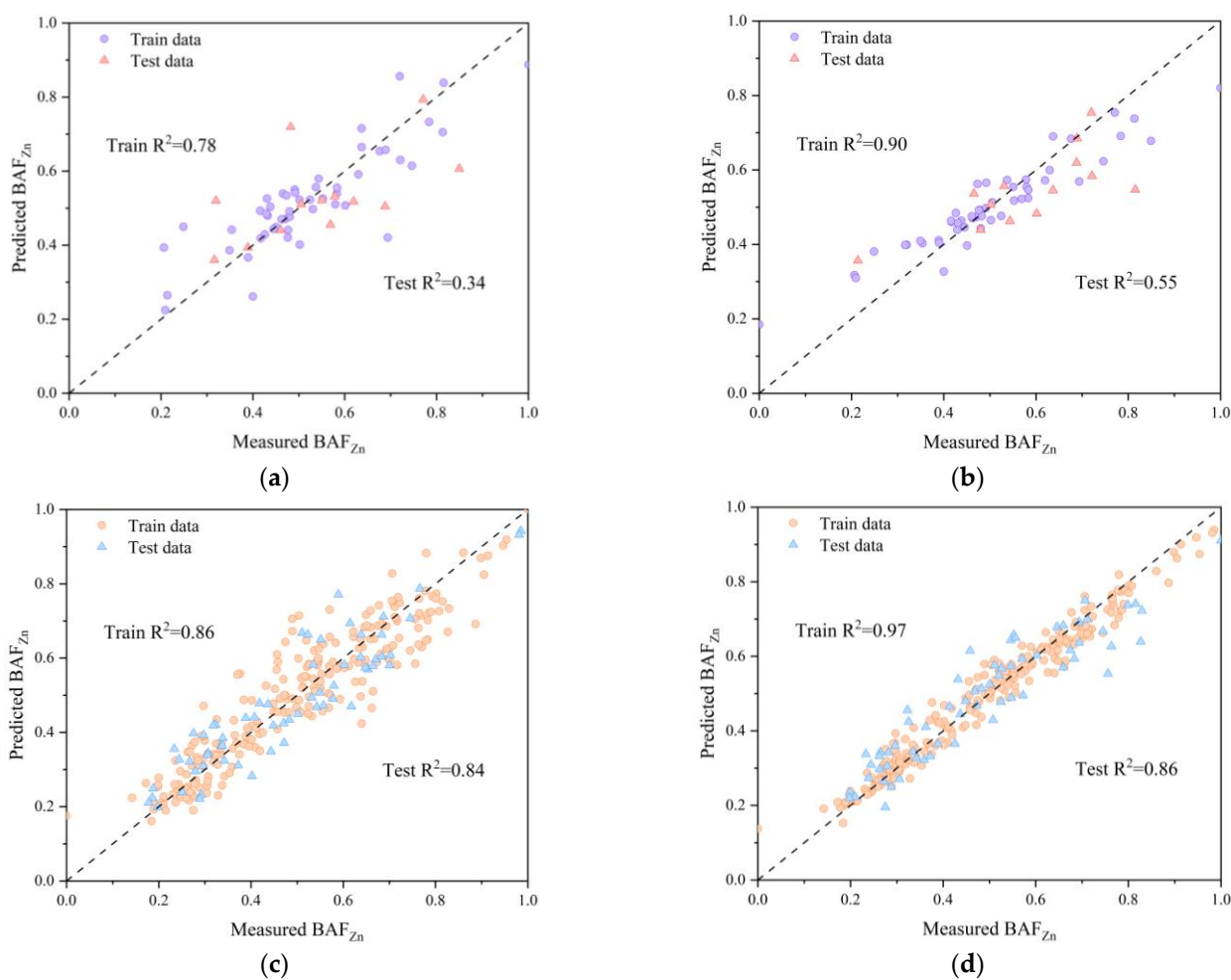

Figure 8. Predicted BAF_{Zn} and measured BAF_{Zn} for (**a**) ANN and (**b**) RF in Heyuan.

3.5. Model Robustness and Generalization

Previous studies often develop prediction models using the data from a single region for both the training and validation, which limits their applicability to other regions. This study aimed to assess the robustness and generalizability of the prediction models across regions. The models were developed using 65 samples from Heyuan and 306 samples from the Pearl River Delta. When applied to the Heyuan data, the models performed poorly, while the models for the Pearl River Delta performed well (Table 5, Figure 9).

Table 5. Predicting accuracy comparison for Heyuan and Pearl River Delta areas.

Region	Model	Validation Sample Size	R^2	NME	MRE (%)	RMSE
Heyuan	Artificial Neural Network	13	0.34	−0.03	18.47	0.13
	Random Forest	13	0.55	−0.07	16.30	0.11
Pearl River Delta	Artificial Neural Network	63	0.84	−0.003	15.09	0.07
	Random Forest	64	0.86	0.003	13.25	0.07

Figure 9. Predicted BAF_{Zn} and measured BAF_{Zn} for (**a**,**b**) Heyuan and (**c**,**d**) Pearl River Delta.

The sample-to-feature ratio (SFR) for Heyuan was 7.43, which is less than the ideal threshold of 10, leading to a weaker model performance. In contrast, the SFR for the Pearl River Delta was 34.71, ensuring better model stability. Combining the data from both regions improved the SFR and allowed the model to perform well across the study area, demonstrating its robustness and generalizability.

Finally, to further validate the models' generalization ability, the models were tested on 65 samples from Zhaoqing, collected in 2016. The results showed that the models performed

well, with the RF model slightly outperforming the ANN model (Table 6, Figure 10). This indicates that the prediction models have good generalizability within the study area, although further research is needed to refine the prediction methods for different regions.

Table 6. Accuracy of the model for Zhaoqing.

Model	Validation Sample Size	R^2	NME	MRE (%)	RMSE
Artificial Neural Network	65	0.80	0.02	13.18	0.08
Random Forest	65	0.90	-0.003	9.50	0.06
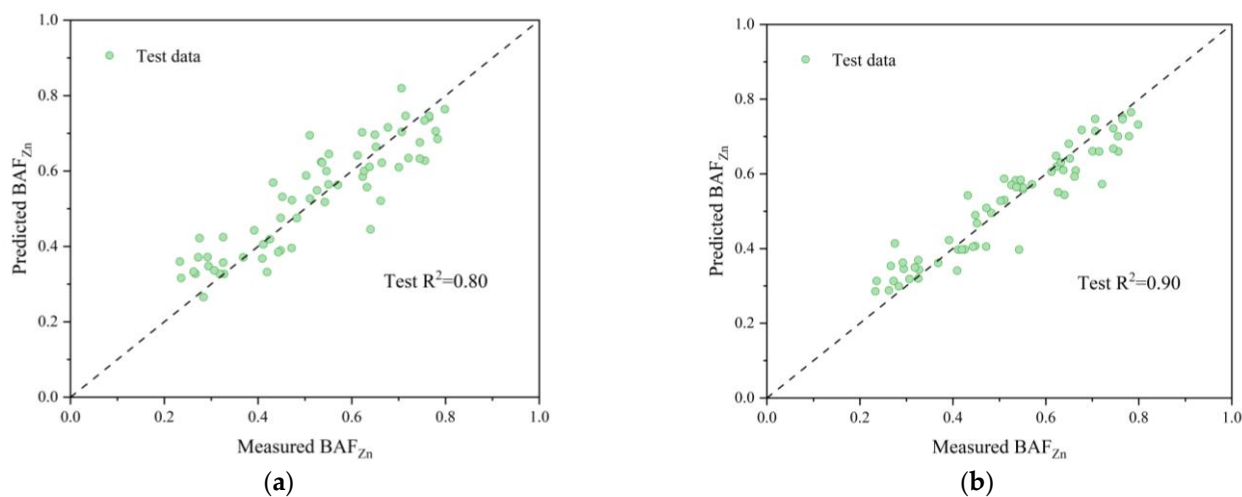
Figure 10. Predicted BAF_{Zn} and measured BAF_{Zn} for (**a**) ANN and (**b**) RF in Zhaoqing.

4. Summary and Conclusions

The rice grains in the study area are rich in zinc content, indicating a great potential for the development of zinc-enriched rice. However, the zinc content in the rice grains does not have a linear correlation with the zinc content in the soil. Research has found that the uptake of zinc by rice grains is controlled by multiple factors, mainly related to soil S, P, CaO, Mn, TFe_2O_3, TOC, and SiO_2/Al_2O_3. To accurately predict the zinc content in the rice grains in the study area, this study employed machine learning methods to establish a predictive model using 371 sets of data collected from rice grains and rhizosphere soil samples in the Pearl River Delta and Heyuan areas of Guangdong. The robustness and generalizability of the predictive model were further verified, and the model showed a good predictive performance across the entire study area.

This study provides valuable references for the rational development of zinc-enriched rice in the Heyuan area, as well as for establishing predictive models of rice grain elemental content in small areas with limited data. This paper focuses solely on the zinc content in rice grains in Guangdong, but similar properties are expected in other regions or for other elements. Therefore, it is anticipated that more research will be conducted in the future to precisely predict the content of beneficial or harmful elements in rice grains across more regions. At the same time, with the rapid development of machine learning in recent years, the interpretability and prediction accuracy of the predictive models will be further improved.

References

1. Sturikova, H.; Krystofova, O.; Huska, D.; Adam, V. Zinc, Zinc Nanoparticles and Plants. *J. Hazard. Mater.* **2018**, *349*, 101–110. [CrossRef]

2. Zhang, Y.; Huang, Y.; Yang, P.; Chen, J.; Zhu, L.; Zhao, Y.; Guo, J. Characteristics of zinc absorption and distribution after silking in the different maize inbred lines. *J. Hebei Agric. Univ.* **2018**, *41*, 14–20. [CrossRef]

3. Gupta, S.; Brazier, A.K.M.; Lowe, N.M. Zinc Deficiency in Low- and Middle-income Countries: Prevalence and Approaches for Mitigation. *J. Hum. Nutr. Diet.* **2020**, *33*, 624–643. [CrossRef]

4. Singh, A.P.; Johari, D. *Hymenophyllum javanicum* Spreng: An Addition to the Pteridophytic Flora of Central India. *Proc. Natl. Acad. Sci. India Sect. B Biol. Sci.* **2018**, *88*, 531–537. [CrossRef]

5. Prom-U-Thai, C.; Rashid, A.; Ram, H.; Zou, C.; Guilherme, L.R.G.; Corguinha, A.P.B.; Guo, S.; Kaur, C.; Naeem, A.; Yamuangmorn, S.; et al. Simultaneous Biofortification of Rice with Zinc, Iodine, Iron and Selenium Through Foliar Treatment of a Micronutrient Cocktail in Five Countries. *Front. Plant Sci.* **2020**, *11*, 589835. [CrossRef] [PubMed]

6. Utasee, S.; Jamjod, S.; Lordkaew, S.; Prom-U-Thai, C. Improve Anthocyanin and Zinc Concentration in Purple Rice by Nitrogen and Zinc Fertilizer Application. *Rice Sci.* **2022**, *29*, 435–450. [CrossRef]

7. Cakmak, I.; Kutman, U.B. Agronomic Biofortification of Cereals with Zinc: A Review. *Eur. J. Soil Sci.* **2018**, *69*, 172–180. [CrossRef]

8. Lin, Q.; Zhu, W.; Chen, Z.; Peng, X.; Zhao, S. Progress in Species and Bioavailability of Heavy Metals in Soil. *J. Guangdong Univ. Technol.* **2013**, *30*, 113–118.

9. Wang, Y.; Yu, T.; Yang, Z.; Bo, H.; Lin, Y.; Yang, Q.; Liu, X.; Zhang, Q.; Zhuo, X.; Wu, T. Zinc Concentration Prediction in Rice Grain Using Back-Propagation Neural Network Based on Soil Properties and Safe Utilization of Paddy Soil: A Large-Scale Field Study in Guangxi, China. *Sci. Total Environ.* **2021**, *798*, 149270. [CrossRef]

10. Gu, Q.; Yu, T.; Yang, Z.; Ji, J.; Hou, Q.; Wang, L.; Wei, X.; Zhang, Q. Prediction and Risk Assessment of Five Heavy Metals in Maize and Peanut: A Case Study of Guangxi, China. *Environ. Toxicol. Pharmacol.* **2019**, *70*, 103199. [CrossRef] [PubMed]

11. Ma, X.; Yu, T.; Guan, D.-X.; Li, C.; Li, B.; Liu, X.; Lin, K.; Li, X.; Wang, L.; Yang, Z. Prediction of Cadmium Contents in Rice Grains from Quaternary Sediment-Distributed Farmland Using Field Investigations and Machine Learning. *Sci. Total Environ.* **2023**, *898*, 165482. [CrossRef]

12. Hu, B.; Xue, J.; Zhou, Y.; Shao, S.; Fu, Z.; Li, Y.; Chen, S.; Qi, L.; Shi, Z. Modelling Bioaccumulation of Heavy Metals in Soil-Crop Ecosystems and Identifying Its Controlling Factors Using Machine Learning. *Environ. Pollut.* **2020**, *262*, 114308. [CrossRef]

13. Pyo, J.; Hong, S.M.; Kwon, Y.S.; Kim, M.S.; Cho, K.H. Estimation of Heavy Metals Using Deep Neural Network with Visible and Infrared Spectroscopy of Soil. *Sci. Total Environ.* **2020**, *741*, 140162. [CrossRef]

14. Li, C.; Zhang, C.; Yu, T.; Liu, X.; Yang, Y.; Hou, Q.; Yang, Z.; Ma, X.; Wang, L. Use of Artificial Neural Network to Evaluate Cadmium Contamination in Farmland Soils in a Karst Area with Naturally High Background Values. *Environ. Pollut.* **2022**, *304*, 119234. [CrossRef] [PubMed]

15. Zhao, B.; Zhu, W.; Hao, S.; Hua, M.; Liao, Q.; Jing, Y.; Liu, L.; Gu, X. Prediction Heavy Metals Accumulation Risk in Rice Using Machine Learning and Mapping Pollution Risk. *J. Hazard. Mater.* **2023**, *448*, 130879. [CrossRef] [PubMed]

16. Ma, X.; Yu, T.; Yang, Z.; Zhang, H.; Wu, Z.; Wang, J.; Li, M.; Lei, F. Geochemical characteristics of zinc in soil and prediction of zinc content in maize and rice grains in Linshui County, Sichuan Province. *Geol. China* **2022**, *49*, 324–335.

17. Liu, Z.; Zhang, X.; Dong, Y.; Qing, C.; Cheng, X.; Zhao, W.; Li, X.; Sang, L.; Hai, L. Geochemical characteristics of zinc in soil and prediction of Zn-rich wheat cultivating areas in Weining Plain, Northwest China. *Geol. China* **2024**, *51*, 1319–1330.

18. Guo, R.; Ren, R.; Wang, L.; Zhi, Q.; Yu, T.; Hou, Q.; Yang, Z. Using Machine Learning to Predict Selenium and Cadmium Contents in Rice Grains from Black Shale-Distributed Farmland Area. *Sci. Total Environ.* **2024**, *912*, 168802. [CrossRef] [PubMed]

19. *DZ/T 0258-2014*; Specification of Multi-Purpose Regional Geochemical Survey (1:250 000). MNR (Ministry of Land and Resources of the People's Republic of China): Beijing, China, 2014.

20. *DZ/T 0295-2016*; Specification of Land Quality Geochemical Assessment. MNR (Ministry of Land and Resources of the People's Republic of China): Beijing, China, 2016.

21. *GB5009.268-2016*; National Food Safety Standard for Multi-Element Determination in Foods. National Health and Family Planning Commission of the People's Republic of China: Beijing, China, 2016.

22. *DZ/T 0279-2016*; Regional Geochemical Sample Analysis Methods. MNR (Ministry of Land and Resources of the People's Republic of China): Beijing, China, 2016.

23. Li, M.; Xi, X.; Xiao, G.; Cheng, H.; Yang, Z.; Zhou, G.; Ye, J.; Li, Z. National Multi-Purpose Regional Geochemical Survey in China. *J. Geochem. Explor.* **2014**, *139*, 21–30. [CrossRef]

24. Yang, Y.; Li, C.; Yang, Z.; Yu, T.; Jiang, H.; Han, M.; Liu, X.; Wang, J.; Zhang, Q. Application of Cadmium Prediction Models for Rice and Maize in the Safe Utilization of Farmland Associated with Tin Mining in Hezhou, Guangxi, China. *Environ. Pollut.* **2021**, *285*, 117202. [CrossRef] [PubMed]

25. Li, P.; Hao, H.; Zhang, Z.; Mao, X.; Xu, J.; Lv, Y.; Chen, W.; Ge, D. A Field Study to Estimate Heavy Metal Concentrations in a Soil-Rice System: Application of Graph Neural Networks. *Sci. Total Environ.* **2022**, *832*, 155099. [CrossRef]

26. Yang, G.R.; Wang, X.-J. Artificial Neural Networks for Neuroscientists: A Primer. *Neuron* **2020**, *107*, 1048–1070. [CrossRef]

27. Georgevici, A.I.; Terblanche, M. Neural Networks and Deep Learning: A Brief Introduction. *Intensive Care Med.* **2019**, *45*, 712–714. [CrossRef] [PubMed]

28. Ebrahimi, M.; Sarikhani, M.R.; Safari Sinegani, A.A.; Ahmadi, A.; Keesstra, S. Estimating the Soil Respiration Under Different Land Uses Using Artificial Neural Network and Linear Regression Models. *Catena* **2019**, *174*, 371–382. [CrossRef]

29. Shao, W.; Guan, Q.; Tan, Z.; Luo, H.; Li, H.; Sun, Y.; Ma, Y. Application of BP—ANN Model in Evaluation of Soil Quality in the Arid Area, Northwest China. *Soil Tillage Res.* **2021**, *208*, 104907. [CrossRef]

30. Zhu, J.-J.; Yang, M.; Ren, Z.J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–17689. [CrossRef] [PubMed]

31. Zhao, B.; Song, J.; Xie, L.; Ma, H.; Li, H.; Ren, J.; Sun, W. Multiaxial Fatigue Life Prediction Method Based on the Back-Propagation Neural Network. *Int. J. Fatigue* **2023**, *166*, 107274. [CrossRef]

32. Wang, X.; Liu, L.; Zhang, W.; Ma, X. Prediction of Plant Uptake and Translocation of Engineered Metallic Nanoparticles by Machine Learning. *Environ. Sci. Technol.* **2021**, *55*, 7491–7500. [CrossRef]

33. Santra, A.K.; Chakraborty, N.; Sen, S. Prediction of Heat Transfer Due to Presence of Copper–Water Nanofluid Using Resilient-Propagation Neural Network. *Int. J. Therm. Sci.* **2009**, *48*, 1311–1318. [CrossRef]

34. Guio Blanco, C.M.; Brito Gomez, V.M.; Crespo, P.; Ließ, M. Spatial Prediction of Soil Water Retention in a Páramo Landscape: Methodological Insight into Machine Learning Using Random Forest. *Geoderma* **2018**, *316*, 100–114. [CrossRef]

35. Carranza, C.; Nolet, C.; Pezij, M.; van der Ploeg, M. Root Zone Soil Moisture Estimation with Random Forest. *J. Hydrol.* **2021**, *593*, 125840. [CrossRef]

36. Chen, W.; Li, Y.; Xue, W.; Shahabi, H.; Li, S.; Hong, H.; Wang, X.; Bian, H.; Zhang, S.; Pradhan, B.; et al. Modeling Flood Susceptibility Using Data-Driven Approaches of Naïve Bayes Tree, Alternating Decision Tree, and Random Forest Methods. *Sci. Total Environ.* **2020**, *701*, 134979. [CrossRef] [PubMed]

37. Chagas, C.D.S.; De Carvalho Junior, W.; Bhering, S.B.; Calderano Filho, B. Spatial Prediction of Soil Surface Texture in a Semiarid Region Using Random Forest and Multiple Linear Regressions. *Catena* **2016**, *139*, 232–240. [CrossRef]

38. Zhu, X.; Wang, X.; Ok, Y.S. The Application of Machine Learning Methods for Prediction of Metal Sorption onto Biochars. *J. Hazard. Mater.* **2019**, *378*, 120727. [CrossRef]

39. Bhagat, S.K.; Paramasivan, M.; Al-Mukhtar, M.; Tiyasha, T.; Pyrgaki, K.; Tung, T.M.; Yaseen, Z.M. Prediction of Lead (Pb) Adsorption on Attapulgite Clay Using the Feasibility of Data Intelligence Models. *Environ. Sci. Pollut. Res.* **2021**, *28*, 31670–31688. [CrossRef] [PubMed]

40. Zhao, W.; Ma, J.; Liu, Q.; Dou, L.; Qu, Y.; Shi, H.; Sun, Y.; Chen, H.; Tian, Y.; Wu, F. Accurate Prediction of Soil Heavy Metal Pollution Using an Improved Machine Learning Method: A Case Study in the Pearl River Delta, China. *Environ. Sci. Technol.* **2023**, *57*, 17751–17761. [CrossRef] [PubMed]

41. Pouladi, N.; Møller, A.B.; Tabatabai, S.; Greve, M.H. Mapping Soil Organic Matter Contents at Field Level with Cubist, Random Forest and Kriging. *Geoderma* **2019**, *342*, 85–92. [CrossRef]

42. *GB 15618-2018*; Soil Environment Quality—Risk Control Standard for Soil Contamination of Agriculture Land. Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2018.

43. Grant, C.A.; Clarke, J.M.; Duguid, S.; Chaney, R.L. Selection and Breeding of Plant Cultivars to Minimize Cadmium Accumulation. *Sci. Total Environ.* **2008**, *390*, 301–310. [CrossRef]

44. Bunquin, M.A.B.; Tandy, S.; Beebout, S.J.; Schulin, R. Influence of Soil Properties on Zinc Solubility Dynamics Under Different Redox Conditions in Non–Calcareous Soils. *Pedosphere* **2017**, *27*, 96–105. [CrossRef]

45. Khampuang, K. Effect of Sulfur Fertilization on Productivity and Grain Zinc Yield of Rice Grown Under Low and Adequate Soil Zinc Applications. *Rice Sci.* **2023**, *30*, 632–640. [CrossRef]

46. Du Laing, G.; Vanthuyne, D.R.J.; Vandecasteele, B.; Tack, F.M.G.; Verloo, M.G. Influence of Hydrological Regime on Pore Water Metal Concentrations in a Contaminated Sediment-Derived Soil. *Environ. Pollut.* **2007**, *147*, 615–625. [CrossRef] [PubMed]

47. De Livera, J.; McLaughlin, M.J.; Hettiarachchi, G.M.; Kirby, J.K.; Beak, D.G. Cadmium Solubility in Paddy Soils: Effects of Soil Oxidation, Metal Sulfides and Competitive Ions. *Sci. Total Environ.* **2011**, *409*, 1489–1497. [CrossRef] [PubMed]

48. Enya, O.; Heaney, N.; Iniama, G.; Lin, C. Effects of Heavy Metals on Organic Matter Decomposition in Inundated Soils: Microcosm Experiment and Field Examination. *Sci. Total Environ.* **2020**, *724*, 138223. [CrossRef]

49. Stefanowicz, A.M.; Kapusta, P.; Zubek, S.; Stanek, M.; Woch, M.W. Soil Organic Matter Prevails over Heavy Metal Pollution and Vegetation as a Factor Shaping Soil Microbial Communities at Historical Zn–Pb Mining Sites. *Chemosphere* **2020**, *240*, 124922. [CrossRef]

50. Li, R.; Tan, W.; Wang, G.; Zhao, X.; Dang, Q.; Yu, H.; Xi, B. Nitrogen Addition Promotes the Transformation of Heavy Metal Speciation from Bioavailable to Organic Bound by Increasing the Turnover Time of Organic Matter: An Analysis on Soil Aggregate Level. *Environ. Pollut.* **2019**, *255*, 113170. [CrossRef]

51. Ruehlmann, J.; Bönecke, E.; Meyer, S. Predicting the Lime Demand of Arable Soils from pH Value, Soil Texture and Soil Organic Matter Content. *Agronomy* **2021**, *11*, 785. [CrossRef]

52. Yan, D.; Li, J.; Hu, M.; Lang, Y. Characteristics and Genesis of Supergene Manganese Ores in Xialei, Guangxi. *Geol. Sci. Technol. Inf.* **2006**, *25*, 61–67.

53. Tang, J. Enrichment Characteristics and Poison Research to Crops of Heavy Metals in Fe-Mn Nodules Soil in Karst Area of Guangxi Central. Ph.D. Thesis, China University of Geosciences, Beijing, China, 2011.

54. Wang, K.; Xu, S.; Yang, Y.; Lin, Q. Study on Zn and Cd Colloid-affected Adsorption in Three Different Soils. *Soils* **2011**, *43*, 239–246. [CrossRef]

55. Li, Y.; Gao, Q.; Li, H.; Wang, X.; Wang, T.; Wang, Y. Study of Cu and Zn adsorption on the clay and its main components in surficial sediments. *J. North China Electr. Power Univ.* **2009**, *36*, 89–93.

56. Qiu, Z.; Ma, W.; Zhang, M. Formation Characteristics of Soils Developed from Metamorphic Rocks in the Hilly and Mountain Areas of Southwest Zhejiang Province. *Chin. J. Soil Sci.* **2020**, *51*, 1009–1015. [CrossRef]

57. Rose, T.; Kretzschmar, T.; Liu, L.; Lancaster, G.; Wissuwa, M. Phosphorus Deficiency Alters Nutrient Accumulation Patterns and Grain Nutritional Quality in Rice. *Agronomy* **2016**, *6*, 52. [CrossRef]

58. Zhang, W.; Liu, D.; Liu, Y.; Chen, X.; Zou, C. Overuse of Phosphorus Fertilizer Reduces the Grain and Flour Protein Contents and Zinc Bioavailability of Winter Wheat (*Triticum aestivum* L.). *J. Agric. Food Chem.* **2017**, *65*, 1473–1482. [CrossRef] [PubMed]

59. Su, D.; Wu, L.; Søren, K.R.; Zhou, L.; Pan, G.; Cheng, F. Influence of phosphorus on rice (*Oryza sativa* L.) grain zinc bioavailability and its relation to inositol phosphate profiles concentration. *Acta Agron. Sin.* **2020**, *46*, 228–237. [CrossRef]