toxics

MDPI

Article

Comparison of Artificial Neural Network and Multiple Linear Regression to Predict Cadmium Concentration in Rice: A Field Study in Guangxi, China

Junyang Zhao [1,†], Fuhai Zheng [2,†], Baoshan Yu [1], Guanchun Qin [1], Shunpiao Meng [1], Yuhang Qiu [1] and Bing He [1,*]

[1] Guangxi Key Laboratory of Agro-Environment and Agric-Products Safety, College of Agriculture, Guangxi University, Nanning 530004, China; zjy761174981@hotmail.com (J.Z.); a619650974@outlook.com (B.Y.); a2487960538@outlook.com (G.Q.); a2422450832@outlook.com (S.M.); a2470701350@hotmail.com (Y.Q.)

[2] Agricultural Resources and Environmental Research Institute, Guangxi Academy of Agricultural Sciences/Guangxi Key Laboratory of Arable Land Conservation, Nanning 530004, China; fuhaizheng110@126.com

* Correspondence: hebing@gxu.edu.cn

† These authors contributed equally to this manuscript.

Abstract

The translocation of cadmium (Cd) in the soil-rice system is complicated; therefore, most of the soil-plant models of Cd have not been extensively studied. Hence, we studied the back-propagation artificial neural network model (BP-ANN) and multiple regression model (MLR) to predict the cadmium (Cd) content in rice grain and soil through testing soil parameters. In this study, 486 pairs of rice grains and corresponding soil samples of 456 vectors were used for training + validation, and 30 vectors were collected from the southwestern karst area of Guangxi Province as a test data set. In this study, the Cd content in rice was successfully predicted by using the factors soil available cadmium (A_{Cd}), total soil cadmium (T_{Cd}), soil organic matter (SOM), and pH, which have a more significant impact on rice, as the main prediction variables. Root mean square error (RMSE), Relative Percent Difference (RPD), and correlation coefficient (R^2) were used to assess the models. The R^2, RPD, and RMSE values for R_{Cd} medium obtained by the MLR model with pH, T_{Cd}, and A_{Cd} as entered variables were 0.551, 2.398, and 0.049, respectively. The R^2 and RMSE values for R_{Cd} medium obtained by the BP-ANN model with pH, T_{Cd}, and A_{Cd} as entered variables were 0.6846, 2.778, and 0.104, respectively. Therefore, it was concluded that BP-ANN was useful in predicting R_{Cd} and had better performance than MLR.

Keywords: prediction model; cadmium; soil-rice system; BP neural network

check for updates

Academic Editor: Li Mu

Received: 1 July 2025
Revised: 23 July 2025
Accepted: 29 July 2025
Published: 30 July 2025

Citation: Zhao, J.; Zheng, F.; Yu, B.; Qin, G.; Meng, S.; Qiu, Y.; He, B. Comparison of Artificial Neural Network and Multiple Linear Regression to Predict Cadmium Concentration in Rice: A Field Study in Guangxi, China. *Toxics* **2025**, *13*, 645. https://doi.org/10.3390/toxics13080645

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cadmium (Cd) is one of the most toxic heavy metals for both plant growth and human health. Cd is highly mobile, non-degradable, tends to persist in soil solution, and is easily taken up by plants, especially in food crops, and also induces adverse effects such as slow and stunted growth, and thus reduces the yield [1]. The uptake of Cd from the consumption of food has been a primary source of human exposure. Further, Cd exposure may cause many chronic health disorders, including gastrointestinal cancer, nephrotoxicity, and pulmonary diseases [2]. In China, about 2.786×10^9 m^2 of agricultural soils were polluted with Cd due to human activities such as wastewater irrigation, application of agrochemicals and manure, fossil fuel combustion, atmospheric deposition, mining, and metal processing [3,4].

Rice consumption is a very prominent route of heavy metal exposure to humans in areas where rice is a staple food source. Therefore, a thorough understanding of the factors influencing the concentration of Cd absorbed by rice and the relationship models between soil properties and rice Cd content (R_{Cd}) has to be established to estimate the impact of rice consumption on human Cd burden. Soil pH was found to play the most critical role in determining the absorption, solubility, mobility, and eventual bio-availability of Cd in soil, while soil pH and R_{Cd} showed a negative correlation [5]. In addition, SOM can reduce the phyto-availability of Cd in soil through adsorption or forming stable complexes with humic substances [6]. While the changes in soil pH and dissolved organic carbon concentration could be critical for the increase in the availability of Cd in soils and Cd uptake [7]. In addition, R_{Cd} is also related to other soil properties, including clay content, cation exchange capacity (CEC), and other heavy metal concentrations, which also leads to a nonlinear relationship between R_{Cd} and soil properties [8]. Currently, many studies predict the Cd content in rice by constructing soil-rice coupling relationship models [9,10]. Compared to correlation analysis of a single variable, model analysis integrating multiple variables can predict the Cd content in rice more accurately, which is more conducive to assessing the harm of consuming rice from this region to the human body.

In the last few decades, several correlation models have been used to predict R_{Cd}, such as multiple linear regression models (MLR) and artificial neural network models (ANNs) [9–11]. MLR is a kind of model that uses two or more explanatory variables to explain the dependent variables, as well as to represent the correlations between a number of inputs and a response of interest [12]. In addition, it is used to provide a large amount of information from a limited number of experiments by considering one variable at a time [13]. Rőmkens [14] measured Cd levels in brown rice that were predicted well ($R^2 > 0.8$) based on Cd and Zinc in a 0.01 M $CaCl_2$ extract or a soil–plant transfer model using the reactive T_{Cd}, pH, and cation exchange capacity by MLR models. Brus [15] predicted R_{Cd} using the MLR model (R^2 = 66.1%) with log (HNO_3–Cd), pH, log (clay), and log (SOM) as predictors. Ding [16] used MLR analysis to predict Cd content of carrot (*Daucus carota* L.) with soil properties and found that T_{Cd}, pH, and organic carbon were the significant variables contributing to the Cd concentration in carrot. Chen et al. [17] used MLR models to predict R_{Cd} and found that soil pH and SOM were the major factors influencing metal translocation from soil to rice. However, MLR cannot describe the nonlinear complex relationship between rice Cd and related parameters. Secondly, the disadvantage of the regression technique is that the different portions of the database may have different relationships between the soil properties and predictors. Thirdly, the regression equations (e.g., linear, logarithmic, or exponential) and predictors need to be determined as a priority [18].

An artificial neural network (ANN) is a series of mathematical algorithms that endeavors to identify complex non-linear relationships between input and output datasets, which is patterned after the biological nervous system [19]. Unlike the regression technique, ANN does not require any pre-defined model concept and can reliably recognize patterns from noisy and complex data and hence estimate their non-linear relationships [18]. ANNs have been used to predict the uptake of heavy metals in plants and persistent organic pollutants. Hou et al. [20] used BP-ANN optimized by the genetic algorithm to predict R_{Cd} based on soil properties. Wu et al. [21] utilized the Bayes classification statistical method and established a risk forewarning model for rice grain Cd pollution. Khazaei et al. [22] had chosen four-layer back-propagation networks with two hidden layers for prediction of crop yield. Jin [23] developed an improved genetic algorithm and a back-propagation neural network model for surface water quality prediction in the Ashi River, China, and the model

had great performance both in prediction accuracy and reliability and effectively provided real-time early warning for emergency response.

MLR models do not meet the requirements of complex and nonlinear simulations and may fail under diverse conditions. Therefore, accurate and rapid detection of Cd concentration in rice grown in agricultural soil over large areas remains challenging. Moreover, most previous studies have primarily relied on soil-rice samples from specific regions, resulting in limited scope and scale, and hence exhibiting poor applicability to diverse regions [24]. The BP neural network is one of the most extensively used ANN models, consisting of a multi-layer network that uses a gradient descent-based algorithm for weight training [25]. Due to its strong fitting ability, the BP neural network is suitable for application of internal complex mechanisms such as the migration of heavy metals [20]. Therefore, in this study, we used the BP-ANN to predict R_{Cd} based on the influencing factors selected by correlation analysis. The specific objectives of this study include (1) using the basic soil property parameters (pH, SOM, A_{Cd}, and T_{Cd}) to accurately and rapidly predict the Cd content in rice and (2) comparing with the MLR model, verifying the feasibility of BP-ANN in predicting R_{Cd} in Northwest Guangxi.

2. Materials and Methods

2.1. Study Area

The study zone (106.57°–109.15° E, 23.68°–25.62° N) is located in the southwestern karst area of Guangxi Province, China (Figure 1). Over 80% of the excess in heavy metals in the karst area was caused by regional geological background and soil weathering [26]. In addition, mining and irrigation by polluted river water or groundwater are other important factors leading to excessive heavy metals in cultivated soil.

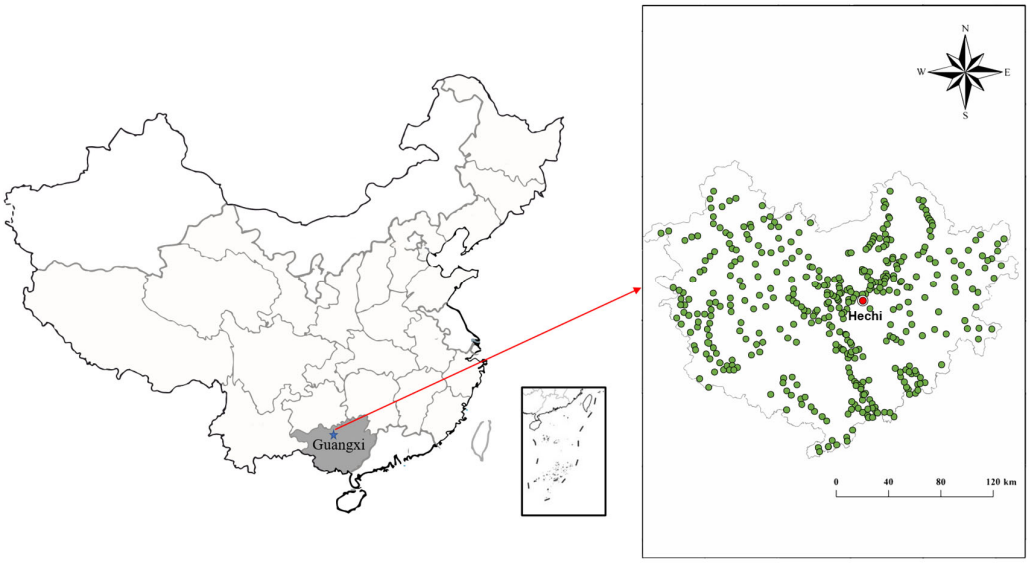

Figure 1. Sampling site and lithology sketch map of the study area.

2.2. Sampling and Pre-Treatment

During 2015 at harvest, a total of 486 paired samples of topsoil (0–20 cm in depth) and rice grain were collected at fixed points following a five-point mixing sampling method in the research area. Soil samples were air-dried and sieved through a 2 mm polyethylene sieve for measurement of soil pH. A portion of these samples were ground and sieved through a 0.149 mm (100 mesh) for T_{Cd}, A_{Cd}, and SOM analyses. Rice grains were washed with tap water and then three times with deionized water and oven-dried at 45 °C to a

constant weight. After removal of the hulls, oven-dried rice grains have been milled to <200 μm for measurement of Cd content.

2.3. Soil and Rice Sampling

The physical and chemical properties of soil were determined according to the method described in the Analysis Methods of Soil Agricultural Chemistry [27]. The pH was measured by the potentiometric method, and the water:soil ratio was 2.5:1.0; the SOM content was measured by the potassium dichromate external heating method; the T_{Cd} content in soil was determined by 2:2:1 HNO_3:$HClO_4$:HF (v:v:v) digestion, the A_{Cd} content was extracted by DTPA solution [28], and the Cd contents of the digestion solution and extract solution were determined by atomic absorption spectrophotometer (PinAAcle 900 T, PerkinElmer, Waltham, MA, USA). R_{Cd} was determined according to Lu et al. [29]. Brown rice was baked at 70 °C to constant weight, then pulverized and passed through a 100-mesh sieve. Nitric acid was used to carry out microwave digestion with a microwave digestion apparatus (MARSXpress, CEM, Matthews, NC, USA); the Cd content was measured using a graphite furnace atomic absorption spectrometer (PinAAcle900T, PerkinElmer, Waltham, MA, USA); and the quality control was carried out by using the GBW100348 [30] plant standard material of the National Institute of Metrology, China.

2.4. Statistical Analysis

The SPSS® 10.0 (SPSS Inc., Chicago, IL, USA) software was used for data analysis. Correlation analysis was performed by using Pearson's correlation coefficient test to determine the relations between different variables, while two-way analysis of variance (ANOVA) was conducted to identify the differences among groups, and relationships were considered significant at $p < 0.05$ and $p < 0.01$.

3. Development of Prediction Models

3.1. Data Pre-Processing Phase

To eliminate the impact of different platforms on the results of network training and improve the efficiency and quality of network training, this study normalized the rice Cd content and soil data to avoid significant differences between input and output data [31]. The sample data were normalized to [−1, 1] using the following formula:

$$X_{norm} = (a - b) \times (x - x_{min})/(x_{max} - x_{min}) + b \quad (1)$$

where x, x_{norm}, x_{max}, and x_{min} are the actual value, normalized value, maximum value, and minimum value of the sample, respectively, and a and b are the maximum and minimum values of the normalized interval, respectively.

3.2. Development of BP-ANN Model

In order to accurately and rapidly predict Cd content in rice by using basic soil property parameters. This study uses MATLAB (2015) to build a neural network model, with SOM, pH, A_{Cd}, and T_{Cd} as input parameters and R_{Cd} as output parameters (Figure 2). These indicators will be measured, and a database will be established during the national soil survey. Our model will also be submitted to the government, which will extract data from the database to use this model for predicting Cd content in rice and creating a predictive distribution map of rice quality and safety in Guangxi. The structure of the BP-ANN consists of an input layer, a hidden layer, and an output layer, which includes multiple neurons, and each layer is linked by connection weights. The number of neurons in the input and output layers depends on the input and output variables, respectively. One (or

more) hidden-neuron layer receives weight from each input neuron and adds them up to a weight value and then passes the results through a non-linear function [32].

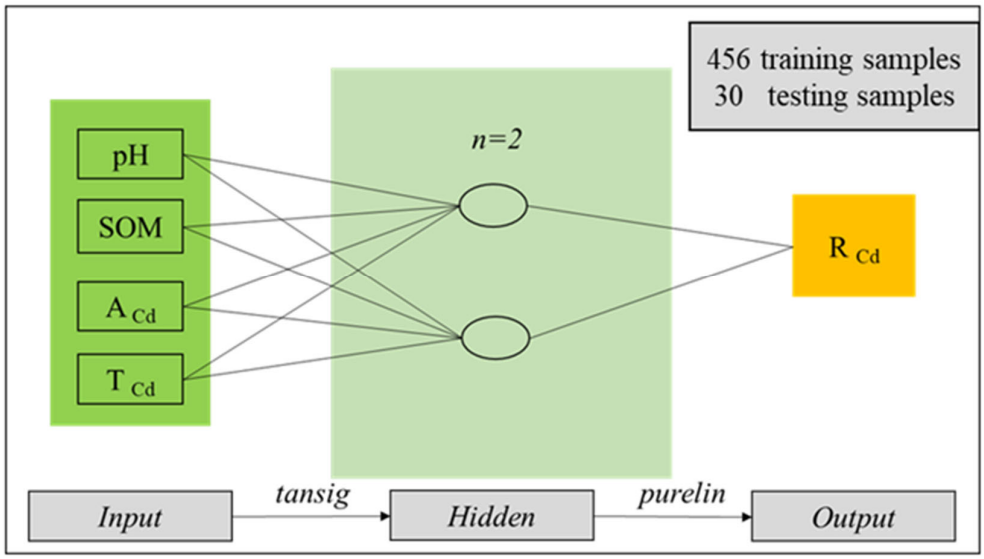

Figure 2. General flow-chart of BP-ANN model to predict Cd in rice.

Selecting the number of neurons in the hidden layers is a very crucial part of deciding the overall neural network architecture because the number of hidden layers and the number of neurons in each hidden layer may affect the training efficiency and the precision of prediction [18]. In most cases, BP-ANN with a single hidden layer is sufficient to provide an accurate approximation and useful for limiting the calculation [31]. The optimal number of neurons in the hidden layer was determined by the trial and error method to minimize the prediction error. The number of neurons was chosen in the hidden layer by the following empirical Formula (2):

$$\mathrm{n} = \sqrt{p+q} + a \quad (2)$$

where n is the number of hidden layer nodes, p is the number of input layer nodes, q is the number of output layer nodes, and a is an integer between 1 and 10. In this study, the numbers of neurons 2, 3, 4, 5, and 6 were selected for modeling.

In the ANN models, an essential step is the splitting of the available data into three subsets: training, validation, and test data sets [33]. In this paper, the required data are sampled into two subsets with a training + validation set size of 456 vectors and a test set size of 30 vectors (randomly selected data) that were used to predict R_{Cd}. The ratio of training set and validation set is 1:1 (231:225), 2:1 (306:150), and 3:1 (343:113), respectively. Meanwhile, the cross-validation technique is used during the training process to divide the training and validation data sets and prevent the training model from being overfitted. For the training of datasets, the trainlm (Levenberg-Marquardt backpropagation) function was used as a learning algorithm in the developed ANN model. Choose the tansig function for the connection between the input layer and the hidden layer, and opt for the purelin function for the connection between the hidden layer and the output layer.

3.3. Development of MLR Model

To assess and compare the performance of BP-ANN models, we used SPSS 21 to establish multiple linear regression models (MLR). MLR is a classic simple linear regression model from single to multiple predictors where the objective is to deduce a model that can exhibit the maximum deviations in the predictor data to evaluate their corresponding

regression coefficients [34]. MLR is often used as a reference to evaluate other non-linear models. The logarithm of T_{Cd}, A_{Cd}, and R_{Cd} based on 10 is taken, respectively, and then a regression analysis model is carried out. The following regression prediction model was obtained:

$$\log R_{Cd} = a + b\ pH + c\ SOM + d \log T_{Cd}$$

$$\log R_{Cd} = a + b\ pH + c\ SOM + d \log A_{Cd}$$

$$\log R_{Cd} = a + b\ pH + c \log T_{Cd} \quad (3)$$

3.4. Evaluation Criteria for Model Performance

In order to assess the performance and stability of the models, the difference between the predicted values and the experimental values was evaluated. The root mean square error (RMSE), coefficients of determination (R^2), and the ratio of standard deviation (RPD) values were as follows [35,36]:

$$SEP = \left[\frac{\sum_{i=1}^{n} \left(Y_i^p - \hat{Y}_i^p \right)^2}{n-1} \right]^{1/2} \quad (4)$$

SEP is the Standard Error of Prediction, Y_i^p, $\hat{Y}_i^p$ are the measured and predicted values of the predicted samples, and n is the number of modeling samples.

$$\mathrm{RPD} = \frac{S.D}{SEP} \quad (5)$$

S.D. is the measured standard deviation.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^{n} \left(Y_i - Y_i^J \right)^2} \quad (6)$$

The RMSE was used as a performance measure for the training performance using Equation (5). RMSE values close to zero also denote excellent performance on the part of the model, where Y_i and Y_i^J are the measured and predicted values of the modeling samples, n is the number of modeling samples, and k is the number of factors contained in the model.

4. Results and Discussion

4.1. Changes in Soil Properties and Cd Concentration

Data on soil properties and R_{Cd} are listed in Table 1. Soil pH ranged from 4.43 to 8.02. There are 242 samples having acidity values below 6.5, accounting for 50% of the total samples (Figure 3), and more differences in soil pH were due to the two main types of soil in the study area. One is red soil, with the characteristic of acidity to strong acidity, and the pH is usually in the range of 4–6. The other is limestone soil, usually 7.5–8.5 of alkaline pH. Further, SOM ranged from 4.07 g kg^{-1} to 75.43 g kg^{-1} with an average value of 38.44 g kg^{-1} (Figure 3). The content of organic matter was mainly in the range of 20–50 g kg^{-1}, which belongs to the normal content range. Samples with SOM below 20 g kg^{-1} account for 6.6% of the total samples, indicating that the surveyed soil was fertile.

Table 1. Soil properties and Cd in soil and rice in 2015.

Parameters	Mean	Minimum	Maximum
pH	6.14	4.56	8.00
SOM (g kg^{-1})	37.29	4.07	75.43
A_{Cd} (mg kg^{-1})	0.61	0.002	4.86
T_{Cd} (mg kg^{-1})	1.13	0.093	8.76
R_{Cd} (mg kg^{-1})	0.200	0.001	4.431

Note: A_{Cd} is the available Cd content in soil; T_{Cd} is the total Cd content in soil; R_{Cd} is the total Cd content in rice.

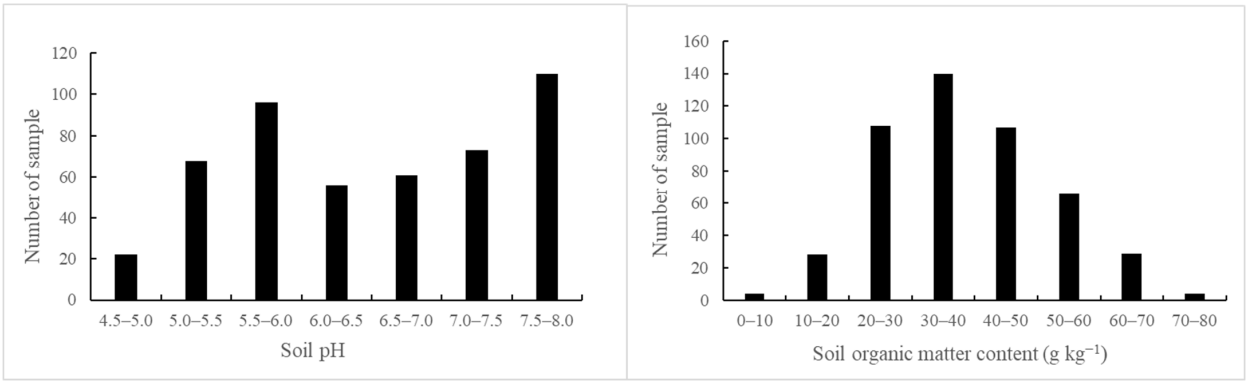

Figure 3. Variation in soil pH and soil organic matter content of different soil samples.

T_{Cd} ranged from 0.093 to 8.76 mg kg^{-1} (Figure 4), and about 23% of the samples had T_{Cd} lower than 0.3 mg kg^{-1}, which was lower than the risk screening values for the soil contamination of agricultural land. A_{Cd} ranged from 0.002 to 4.860 mg kg^{-1} with the mean value of 0.61 mg kg^{-1} (Table 1). A_{Cd} of the samples between 0 and 0.3 mg kg^{-1} was the most, accounting for 43% of the total sample size. High concentrations of A_{Cd} allow for permissive transfer from soil to rice grains and increased Cd accumulation in rice grains.

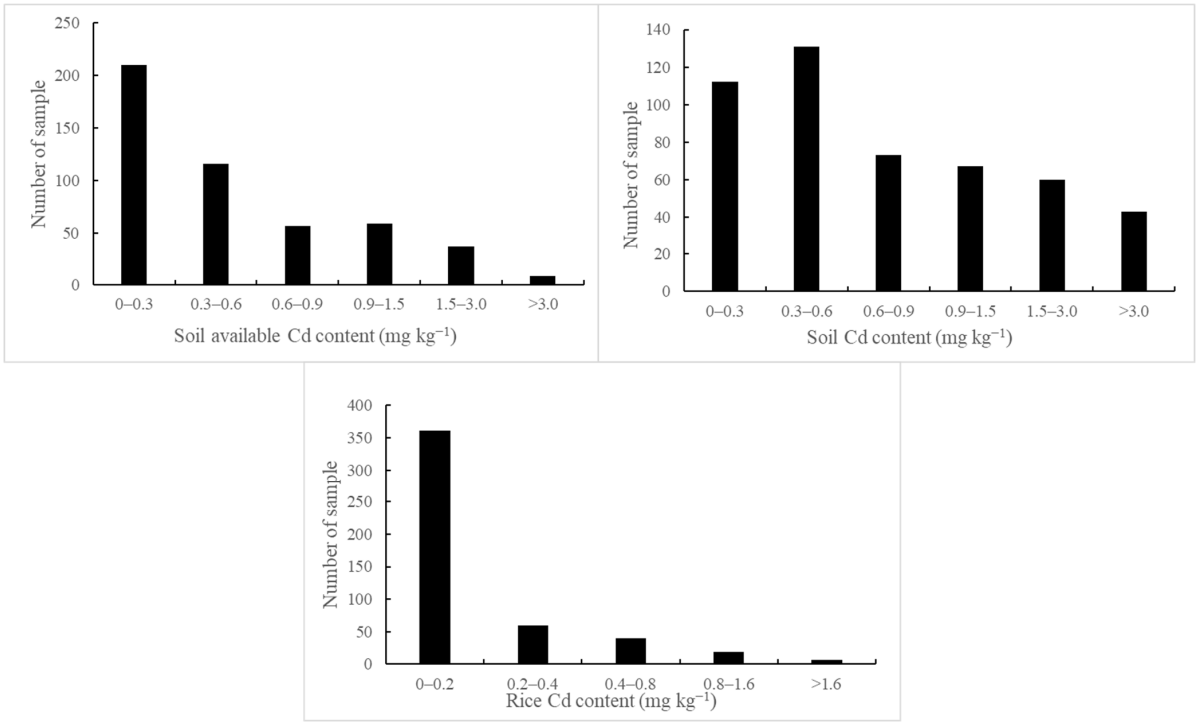

Figure 4. Variation in soil total and available Cd content and rice grain Cd content of different soil and plant samples.

R_{Cd} was mainly distributed between 0.001 and 4.43 mg kg^{-1}, with an average value of 0.2 mg/kg and a median value of 0.07 mg kg^{-1} (Table 1). About 74% of the rice samples had cadmium content between 0 and 2 mg kg^{-1} (Figure 4). According to the agricultural industry standard of the People's Republic of China [37], in this study area, rice samples have 26% more of the exceeding Cd than the standard cadmium.

4.2. Effect of Soil Properties on R_{Cd}

Pearson correlation analysis was performed to quantify the relationship between different parameters (Table 2). Correlation analysis showed that R_{Cd} was significantly positively correlated with T_{Cd} (0.124 **) and A_{Cd} (0.220 **) in soil and significantly negatively correlated with soil pH (−0.216 **), while negatively uncorrelated with SOM (−0.078). Chen considered SOM an important factor influencing the availability of heavy metals in soils [22], and we found that SOM was significantly correlated with A_{Cd} (Table 2).

Table 2. Correlation analysis of R_{Cd} and soil parameters.

Indications	pH	SOM	A_{Cd}	T_{Cd}	R_{Cd}
pH	1	0.123 **	0.112 *	0.232 **	0.216 **
SOM		1	0.256 **	0.286 **	−0.078
A_{Cd}			1	0.827 **	0.220 **
T_{Cd}				1	0.124 **
R_{Cd}					1

Note: ** Significant correlation at 0.01 level (bilateral); * Significant correlation at 0.05 level (bilateral).

Most studies showed that the increase in T_{Cd} and A_{Cd} led to the increase in R_{Cd} [38]. As shown in Table 1, T_{Cd} is an important factor in determining R_{Cd} (r = 0.124). In the previous study, A_{Cd} was positively correlated with T_{Cd}, SOM, and pH [39]. It is commonly accepted that the A_{Cd} could be a better indicator of bio-availability and toxicity than the T_{Cd} [40], and the bio-availability of Cd was strongly correlated with the ability to transfer from soil into plants [41].

A significant positive correlation between A_{Cd} and pH of the soils was also observed ($p < 0.05$). pH is an important factor affecting R_{Cd} by determining the bio-availability of Cd in soil. Increasing soil pH could immobilize heavy metals by increasing soil adsorption and enhancing the easily bio-available forms to immobile forms [42]. The decreased soil pH induced heavy metal desorption from soil constituents, increased mobility and bio-availability of Cd, and increased Cd uptake by rice [12]. In our research, the correlation between A_{Cd} and pH was not obvious. The possible reason is that our soil samples were mainly composed of acid red soil and alkaline lime soil. pH was different between soils, which may affect the correlation analysis of A_{Cd} and pH.

These findings indicate that soil properties other than SOM, such as pH and T_{Cd}, had a strong influence on the available Cd in soil and therefore the amount of Cd absorbed by rice grains. SOM may have different functions and generate noise to data. On the one hand, SOM can reduce the Cd availability in the soil through adsorption or forming stable complexes with humic substances [43]. On the other hand, SOM supplies organic chemicals, acting as chelates, to the soil solution and enhances Cd availability to rice [44]. The SOM appeared as an indirect variable affecting Cd uptake by rice; therefore, the SOM was also used as a predictor.

4.3. Development of BP-ANN Model for R_{Cd}

The best architecture (based on prediction accuracy on the training and test sets) was a fully connected three-layer, feed-forward network with one hidden layer. These architectures were used for all the networks, and their prediction accuracy was assessed

on the same set of validation cases. The current stage of this study aims to accurately and rapidly predict the Cd content in rice using BP-ANN model. We chose pH, SOM, A_{Cd}, and T_{Cd} as the four input parameters because the government has established databases for these parameters. We aim to utilize the existing government databases to establish a prediction model. For the purpose of rapid prediction, we test the four parameters in groups: (1) pH, SOM, T_{Cd}; (2) pH, SOM, A_{Cd}; (3) pH, A_{Cd}, T_{Cd}. For each subset of selected variables, BP-ANN models were developed (Table 3). There was no significant correlation between SOM and R_{Cd}, so there was no model test with parameters of SOM, A_{Cd}, and T_{Cd}. At the same time, the model with four input parameters (pH, SOM, A_{Cd}, T_{Cd}) is not established, because we have less data, the accuracy of the model may be reduced, and the increase in parameters will lead to the increase in the running time of the model, which cannot achieve the purpose of rapid prediction.

Table 3. Parameters of BP-ANN models.

Model	Input	Output
Model I	pH, SOM, T_{Cd}	R_{Cd}
Model II	pH, SOM, A_{Cd}	R_{Cd}
Model III	pH, T_{Cd}, A_{Cd}	R_{Cd}

Through comparative analysis, when the number of neurons in the hidden layer is 2 and the ratio of training set to correction set is 2:1, the prediction accuracy of Mode I is the best, and the RPD value is 2.165, which is significantly higher than the RPD value under other conditions (Table 4). The best structure of Mode II is the same as Mode I, but the RPD value is higher, which is 2.488. In Mode III, the best architecture was a fully connected three-layer, one hidden layer with 3 hidden nodes; the ratio of training set to correction set is 2:1.

Table 4. BP-ANN model prediction results.

Implied Layer of Neurons	Training Set and Calibration Set Ratio	RPD		
		Model I	Model II	Model III
2	1:1	0.992	1.488	2.368
	2:1	2.165	2.488	2.036
	3:1	1.892	2.409	2.015
3	1:1	1.172	1.504	2.225
	2:1	1.683	2.315	2.422
	3:1	1.464	1.328	2.282
4	1:1	1.037	1.389	2.221
	2:1	1.165	1.246	1.944
	3:1	1.492	2.386	2.172
5	1:1	0.846	1.703	2.278
	2:1	1.040	1.578	2.185
	3:1	1.132	1.153	2.047
6	1:1	1.024	0.981	2.296
	2:1	0.901	1.312	2.081
	3:1	0.910	1.749	2.157

Note: Mode I: pH, SOM, and T_{Cd} as input parameters; Mode II: pH, SOM, and A_{Cd} as input parameters; Mode III: pH, T_{Cd}, and A_{Cd} as input parameters.

4.4. MLR for Predicting Cd Content in Rice

A best-fit model using MLR to predict R_{Cd} was developed. Table 5 presented the RPD values for three models. After comparing the RPD, the MLR of Mode II had the best prediction accuracy. The following model, including independent variables, was adopted:

$$\log R_{Cd} = 0.851 - 0.226\ pH - 0.007\ SOM + 0.557\ \log A_{Cd}$$

Table 5. Regression equations for different models.

Model	Training and Calibration Set Ratio	MLR	RPD
Model I	1:1	$\log R_{Cd} = 0.517 - 0.23\ pH + 0.483\ \log T_{Cd}$	2.032
	2:1	$\log R_{Cd} = 0.973 - 0.263\ pH - 0.007\ SOM + 0.615\ \log T_{Cd}$	2.226
	3:1	$\log R_{Cd} = 1.0115 - 0.265\ pH - 0.008\ SOM + 0.596\ \log T_{Cd}$	2.224
Model II	1:1	$\log R_{Cd} = 0.385 - 0.204\ pH + 0.391\ \log A_{Cd}$	2.314
	2:1	$\log R_{Cd} = 0.851 - 0.226\ pH - 0.007\ SOM + 0.557\ \log A_{Cd}$	2.342
	3:1	$\log R_{Cd} = 0.814 - 0.221\ pH - 0.008\ SOM + 0.502\ \log A_{Cd}$	2.232
Model III	1:1	$\log R_{Cd} = 0.517 - 0.236\ pH + 0.48\ \log T_{Cd}$	2.032
	2:1	$\log R_{Cd} = 0.645 - 0.258\ pH + 0.533\ \log T_{Cd}$	2.172
	3:1	$\log R_{Cd} = 0.682 - 0.262\ pH + 0.509\ \log T_{Cd}$	2.164

Note: Mode I: pH, SOM, and T_{Cd} as input parameters; Mode II: pH, SOM, and A_{Cd} as input parameters; Mode III: pH, and T_{Cd} as input parameters.

It was found that RPD values were above 2.342, which means that rice Cd content could be attributed to soil pH, OM, and T_{Cd} using MLR models, but the accuracy was lower than the best ANN model II (RPD 2.488) and the best ANN model III (RPD 2.422).

5. Comparison of Different Models

To compare the performance of the MLR and BP-ANN predictions, the training and testing results of these methods were shown in Table 6 and Figure 5, respectively. Comparing the results of these methods and their models showed that the performance values of BP-ANN were better than MLR. The RMSE values of BP-ANN were lower than those of MLR in Mode I and Mode II, but not in Mode III (Table 6). For all models, according to the derived results of the ANN method, based on the testing data set, the RPD ranged from 2.670 to 2.853, while the corresponding range of 2.398–2.581 was obtained based on MLR (Table 6). The predicted outputs (R_{Cd}) of the testing datasets from both the well-trained BP-ANN model and the MLR model were compared with the actual values, as shown in Figure 5. The correlation will be better when the correlation coefficient is close to 1, namely, when the predicted value is closer to the observed value. It can be seen that the performance analysis of validation errors was similar to the performance analysis of test errors, indicating that the trained BP-ANN can generalize and the optimal network obtained in the model training process based on the training and validation data were therefore valid. A comparison of the performance analysis of test errors between BP-ANN and MLR shows the differences in prediction accuracy, which indicates that the BP-ANN model can effectively improve R_{Cd} prediction accuracy. It can be seen that the ANN

predictions have a good relation with the experimental results, which means that the ANN has been sufficiently trained.

Table 6. Forecast results for different models.

Model	Model Type	RPD	RMSE
Mode I	BP-ANN	2.670	0.119
	MLR	2.573	0.135
Mode II	BP-ANN	2.853	0.102
	MLR	2.581	0.123
Mode III	BP-ANN	2.778	0.104
	MLR	2.398	0.049

Note: Mode I: pH, SOM, T_{Cd}; Mode II: pH, SOM, A_{Cd}; Mode III: pH, ACd, T_{Cd}.

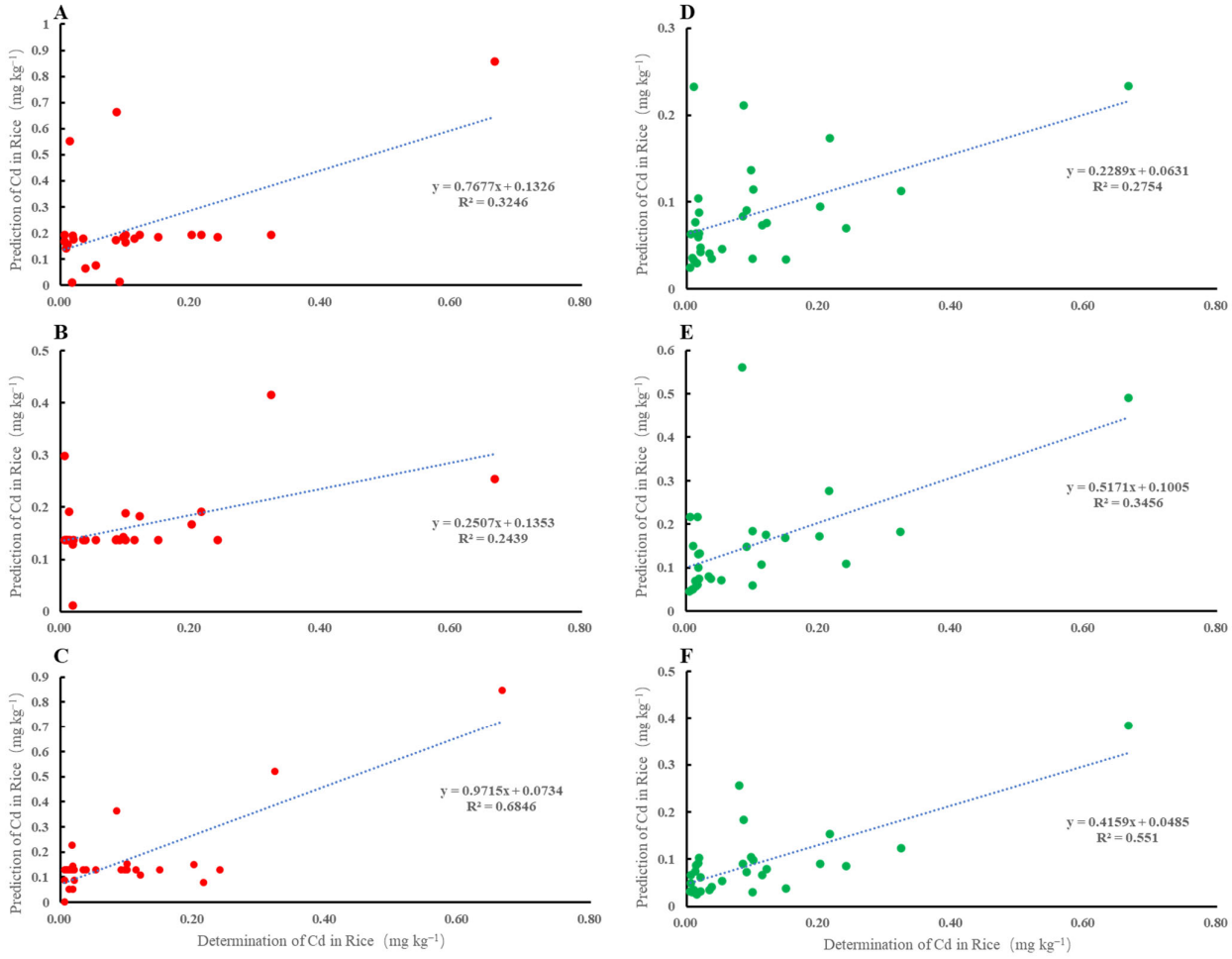

Figure 5. Comparison of the predicted values of the different BP-ANN and MLR models for the best input layer with the actual value ((**A**): Mode I, BP; (**B**): Mode II, BP; (**C**): Mode III, BP; (**D**): Mode I, MLR; (**E**): Mode II, MLR; (**F**): Mode III, MLR).

Table 6 showed that the BP-ANN model was superior to the MLR model in terms of predicting R_{Cd}. Although the RMSE decreased from 0.104 in the BP-ANN model to 0.049 in the MLR model, the R^2 increased from 0.551 in the MLR model to 0.6846 in the BP-ANN of Mode III (Figure 5C,F). Figure 5 shows the observed and predicted Cd values obtained through the BP-ANN model during the testing phase. The comparison results between Figure 5 indicated that the BP-ANN model outperformed the MLR model in the prediction of R_{Cd}. The findings can be explained as follows: The correlation between rice

Cd content and soil parameters tends to be nonlinear. The most significant problem in MLR is the assumption of a linear input-output relationship, but this assumption is unacceptable for complex systems. Conversely, a great advantage of ANNs is their ability to model non-linear relationships. There was no big difference between the goodness-of-fit measures for respective models developed by MLR (Figure 5D–F) methods.

Therefore, when compared with multiple regression analysis, the BP neural network can reveal the nonlinear relationship between Cd concentration in rice grain and soil properties better, which overcomes the shortcomings of simulation by the multiple regression model using complex factors. According to Maran [45], the BP-ANN model having the highest prediction performance among the models may be due to the tendency of ANNs to approximate the non-linearity of the system. Olawoyin [46] described results with the use of BP-ANN to model the relationship between soil input data and the content of soil carcinogenic PAHs, which performed better (R^2 = 0.99) than conventional models such as the MLR. Keshavarzi and Sarmadian [47] developed the performance of the MLR and ANN models for predicting soil parameters using easily measurable characteristics of clay and organic carbon. Results showed that ANN with seven neurons in the hidden layer had better performance in predicting soil CEC than MLR. The accuracy of rice cadmium prediction in contaminating soils using machine learning models was increased compared with linear regression. The comparison between ANN and MLR models has shown that ANN models have better precision with a higher coefficient of correlation than MLR models [48].

Overall, the optimal model for estimating the Cd concentration of rice grain is the BP neural network model because the BP neural network model can be used as a black-box model to predict an individual variable through a complex interaction factor and to process complex and fuzzy mapping relationships without knowing the relationship between the distribution form and variables. Therefore, when compared with multiple regression analyses, the BP neural network can reveal the nonlinear relationship between Cd concentration in rice grain and soil properties better, which overcomes the shortcomings of simulation by the multiple regression model using complex factors.

Our models did not incorporate some parameters that likely influence the Cd content of rice, such as irrigation patterns, CEC, clay content, rice varieties, Fe and S concentration, microbes in soil, and temporal or spatial associations in the ANNs. Among them, rice varieties exhibit significant variations in their propensity to accumulate heavy metals. Incorporating the genes primarily responsible for Cd absorption and translocation in these rice varieties into the model would further refine the relevant indicators influencing Cd migration within the soil-rice system. However, a limited dataset may result in weaker model performance. Additionally, this study is confined to soil-rice samples from specific regions, thus limiting its scope. Therefore, while enhancing the relevant indicators of the model, it is imperative to broaden the research scope and scale and augment the dataset substantially, which will enhance the accuracy of the predictive model.

6. Conclusions

The migration of Cd in the soil-rice system is very complicated. The black box characteristics of BP-ANN enable it to predict the content of Cd in rice grains through relatively stable physical and chemical properties in soil. In this study, the Cd content in rice was successfully predicted by using the factors that have a more significant impact on R_{Cd} (such as A_{Cd}, T_{Cd}, SOM, and pH) as the prediction variable. ANN models gave a higher R^2 and lower RMSE compared to the MLR and indicated that ANN was a more powerful tool than MLR. BP-ANN can be used as a fast and useful tool for predicting Cd concentration in soil, which can contribute to better assessments on the safe use of soil and the production of quality food. The research results can lay a theoretical foundation for the application

of neural network technology in the field of crop heavy metal content prediction and provide a more accurate rice prediction method. It provides a theoretical reference for the government to make and implement relevant policies on the classification and control of heavy metal pollution in rice fields.

Author Contributions: J.Z.: Conceptualization, Methodology, Writing—original draft. F.Z.: Conceptualization, Methodology. B.Y.: Methodology, Data curation. G.Q.: Data curation. S.M.: Formal analysis, Data curation. Y.Q.: Software, Visualization. B.H.: Writing—review and editing, Supervision, Funding acquisition, and Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the special project of the Department of Agriculture and Rural Affairs of Guangxi Zhuang Autonomous Region (20140466), 2014–2016.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

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

References

1. Zhao, J.; Qin, S.; Pan, P.; Chen, D.; Tang, S.; Chen, L.; Wang, X.; Gu, M.; Tang, F.; He, J.; et al. Microbial driving mechanism of soil conditioner on reducing cadmium uptake by rice and improving soil environment. *Agric. Ecosyst. Environ.* **2023**, *349*, 108452. [CrossRef]
2. Shahid, M.; Dumat, C.; Khalid, S.; Niazi, N.K.; Antunes, P.M.C. Cadmium Bioavailability, Uptake, Toxicity and Detoxification in Soil-Plant System. *Rev. Environ. Contam. Toxicol.* **2017**, *241*, 73–137. [PubMed]
3. Liu, F.; Liu, X.; Ding, C.; Wu, L. The dynamic simulation of rice growth parameters under cadmium stress with the assimilation of multi-period spectral indices and crop model. *Field Crops Res.* **2015**, *183*, 225–234. [CrossRef]
4. Hu, Y.; Cheng, H.; Tao, S. The Challenges and Solutions for Cadmium-contaminated Rice in China: A Critical Review. *Environ. Int.* **2016**, *92–93*, 515–532. [CrossRef]
5. Li, K.; Cao, C.; Ma, Y.; Su, D.; Li, J. Identification of cadmium bioaccumulation in rice (*Oryza sativa* L.) by the soil-plant transfer model and species sensitivity distribution. *Sci. Total Environ.* **2019**, *692*, 1022–1028. [CrossRef]
6. Zhou, T.; Wu, L.; Luo, Y.; Christie, P. Effects of organic matter fraction and compositional changes on distribution of cadmium and zinc in long-term polluted paddy soils. *Environ. Pollut.* **2018**, *232*, 514–522. [CrossRef]
7. Ye, X.; Hu, H.; Li, H.; Xiong, Q.; Gao, H. Combined nitrogen fertilizer and wheat straw increases the cadmium phytoextraction efficiency of Tagetes patula. *Ecotoxicol. Environ. Saf.* **2019**, *170*, 210–217. [CrossRef]
8. Wang, Y.; Zhang, Z.; Li, Y.; Liang, C.; Huang, H.; Wang, S. Available heavy metals concentrations in agricultural soils: Relationship with soil properties and total heavy metals concentrations in different industries. *J. Hazard. Mater.* **2024**, *471*, 134410. [CrossRef]
9. Hu, B.; Xue, J.; Zhou, Y.; Shao, S.; Fu, Z.; Li, Y.; Chen, S.; Qi, L.; Shi, Z. Modelling bioaccumulationof heavy metals in soil-crop ecosystems and identifying itscontrolling factors using machine learning. *Environ. Pollut.* **2020**, *262*, 114308. [CrossRef] [PubMed]
10. Mu, T.; Zhou, T.; Li, Z.; Hu, P.; Luo, Y.; Christie, P.; Wu, L. Prediction models for rice Cd accumulation in Chinese paddy fields and the im-plications in deducing soil thresholds based on food safety standards. *Environ. Pollut.* **2020**, *258*, 113879. [CrossRef]
11. Hou, Y.; Zhao, H.; Wu, K.; Li, K. Prediction of crop Cd content and zoning of safety planting based on BP neural network. *Resour. Sci.* **2018**, *40*, 2414–2424.
12. Liu, H.; Tarima, S.; Borders, A.S.; Getchell, T.V.; Getchell, M.L.; Stromberg, A.J. Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments. *BMC Bioinform.* **2005**, *6*, 106. [CrossRef]
13. Fawzy, M.; Nasr, M.; Nagy, H.; Helmi, S. Artificial intelligence and regression analysis for Cd(II) ion biosorption from aqueous solution by Gossypium barbadense waste. *Environ. Sci. Pollut. Res. Int.* **2018**, *25*, 5875–5888. [CrossRef] [PubMed]
14. Römkens, P.F.; Guo, H.Y.; Chu, C.L.; Liu, T.S.; Chiang, C.F.; Koopmans, G.F. Prediction of Cadmium uptake by brown rice and derivation of soil-plant transfer models to improve soil protection guidelines. *Environ. Pollut.* **2009**, *157*, 2435–2444. [CrossRef] [PubMed]

15. Brus, D.J.; Li, Z.; Song, J.; Koopmans, G.F.; Temminghoff, E.J.; Yin, X.; Yao, C.; Zhang, H.; Luo, Y.; Japenga, J. Predictions of spatially averaged cadmium contents in rice grains in the Fuyang Valley, P.R. China. *J. Environ. Qual.* **2009**, *38*, 1126–1136. [CrossRef] [PubMed]
16. Ding, C.; Zhang, T.; Wang, X.; Zhou, F.; Yang, Y.; Yin, Y. Effects of soil type and genotype on lead concentration in rootstalk vegetables and the selection of cultivars for food safety. *J. Environ. Manag.* **2013**, *122*, 8–14. [CrossRef]
17. Chen, H.; Yuan, X.; Li, T.; Hu, S.; Ji, J.; Wang, C. Characteristics of heavy metal transfer and their influencing factors in different soil-crop systems of the industrialization region, China. *Ecotoxicol. Environ. Saf.* **2016**, *126*, 193–201. [CrossRef]
18. Mojid, M.A.; Hossain, A.B.M.Z.; Ashraf, M.A. Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. *Environ. Pollut.* **2019**, *255 Pt 2*, 113355. [CrossRef]
19. Liu, W.; Guo, G.; Chen, F.; Chen, Y. Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm. *Atmos. Pollut. Res.* **2019**, *10*, 1482–1491. [CrossRef]
20. Hou, Y.X.; Zhao, H.F.; Zhang, Z.; Wu, K.N. A novel method for predicting cadmium concentration in rice grain using genetic algorithm and back-propagation neural network based on soil properties. *Environ. Sci. Pollut. Res. Int.* **2018**, *25*, 35682–35692. [CrossRef]
21. Wu, B.; Guo, S.; Zhang, L.; Li, F. Risk forewarning model for rice grain Cd pollution based on Bayes theory. *Sci. Total Environ.* **2018**, *618*, 1343–1349. [CrossRef] [PubMed]
22. Khazaei, J.; Naghavi, M.R.; Jahansouz, M.R.; Salimi-Khorshidi, G. Yield Estimation and Clustering of Chickpea Genotypes Using Soft Computing Techniques. *Agron. J.* **2008**, *100*, 1077–1087. [CrossRef]
23. Jin, T.; Cai, S.; Jiang, D.; Liu, J. A data-driven model for real-time water quality prediction and early warning by an integration method. *Environ. Sci. Pollut. Res. Int.* **2019**, *26*, 30374–30385. [CrossRef] [PubMed]
24. Chen, J.; Tang, L.; Xiang, M.; Zhang, C.; Ge, Y.; Chen, X. Model constructions and validations for regional cadmium coupling relationships in soilrice grain. *Chin. J. Ecol.* **2021**, *40*, 2341–2347.
25. Zhang, D.; Liu, J.; Jiang, C.; Liu, A.; Xia, B. Quantitative detection of formaldehyde and ammonia gas via metal oxide-modified graphene-based sensor array combining with neural network model. *Sens. Actuators B Chem.* **2017**, *240*, 55–65. [CrossRef]
26. China Geological Survey. *Report on Geochemical Survey of Cultivated Land in China*; China Geological Survey: Beijing, China, 2015. (In Chinese)
27. Bao, S. *Soil and Agricultural Chemistry Analysis*; China Agricultural Press: Beijing, China, 2000. (In Chinese)
28. *GB/T 23739-2009*; Soil Quality—Analysis of Available Lead and Cadmium Contents in Soils—Atomic Absorption Spectrometry. Ministry of Agriculture of the People's Republic of China: Beijing, China, 2009. (In Chinese)
29. Lu, H.; Qin, S.; Zhao, J.; Pan, P.; Wang, F.; Tang, S.; Chen, L.; Akhtar, K.; He, B. Silicon inhibits the upward transport of Cd in the first internode of different rice varieties in a Cd stressed farm land. *J. Hazard. Mater.* **2023**, *458*, 131860. [CrossRef]
30. *GBW100348*; Certified Reference Material for the Chemical Composition of Rice Flour. NCS Testing Technology Co., Ltd.: Beijing, China, 2015. (In Chinese)
31. Azadi, S.; Karimi-Jashni, A. Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province, Iran. *Waste Manag.* **2016**, *48*, 14–23. [CrossRef]
32. Blagojev, N.; Kukić, D.; Vasić, V.; Šćiban, M.; Prodanović, J.; Bera, O. A new approach for modelling and optimization of Cu(II) biosorption from aqueous solutions using sugar beet shreds in a fixed-bed column. *J. Hazard. Mater.* **2019**, *363*, 366–375. [CrossRef]
33. Zhang, B.; Zhao, B.; Zuo, P.; Huang, Z.; Zhang, J. Influencing factors and prediction of ambient Peroxyacetyl nitrate concentration in Beijing, China. *J. Environ. Sci.* **2019**, *77*, 189–197. [CrossRef] [PubMed]
34. Deo, R.C.; Sahin, M. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. *Renew. Sustain. Energy Rev.* **2017**, *72*, 828–848. [CrossRef]
35. Luo, X.; Chen, R.; Kabir, M.H.; Liu, F.; Tao, Z.; Liu, L.; Kong, W. Fast Detection of Heavy Metal Content in *Fritillaria thunbergii* by Laser-Induced Breakdown Spectroscopy with PSO-BP and SSA-BP Analysis. *Molecules* **2023**, *28*, 3360. [CrossRef]
36. An, B.; Wang, X.; Huang, X.; Kawuqiati, B. Hyperspectral Estimation of Heavy Metal Cadmium Content in Soil based on Continuous Wavelet Transform. *Earth Environ.* **2023**, *51*, 246–253.
37. *NY861-2004*; Limits of Eight Elements in Cereals, Legume, Tubes and Its Products. Ministry of Agriculture of the People's Republic of China: Beijing, China, 2004. (In Chinese)
38. Novotná, M.; Mikeš, O.; Komprdová, K. Development and comparison of regression models for the uptake of metals into various field crops. *Environ. Pollut.* **2015**, *207*, 357–364. [CrossRef] [PubMed]
39. Adams, M.L.; Zhao, F.; McGrath, S.P.; Nicholson, F.A.; Chambers, B.J. Predicting cadmium concentrations in wheat and barley grain using soil properties. *J. Environ. Qual.* **2004**, *33*, 532–541. [CrossRef]
40. Zhu, B.; Liao, Q.; Zhao, X.; Gu, X.; Gu, C. A multi-surface model to predict Cd phytoavailability to wheat (*Triticum aestivum* L.). *Sci. Total Environ.* **2018**, *630*, 1374–1380. [CrossRef]

41. Wu, L.; Zhou, J.; Zhou, T.; Li, Z.; Jiang, J.; Zhu, D.; Hou, J.; Wang, Z.; Luo, Y.; Christie, P. Estimating cadmium availability to the hyperaccumulator Sedum plumbizincicola in a wide range of soil types using a piecewise function. *Sci. Total Environ.* **2018**, *637–638*, 1342–1350. [CrossRef]
42. Wan, Y.; Huang, Q.; Wang, Q.; Yu, Y.; Su, D.; Qiao, Y.; Li, H. Accumulation and bioavailability of heavy metals in an acid soil and their uptake by paddy rice under continuous application of chicken and swine manure. *J. Hazard. Mater.* **2020**, *384*, 121293. [CrossRef]
43. Xu, W.; Li, Y.; He, J.; Ma, Q.; Zhang, X.; Chen, G.; Wang, H.; Zhang, H. Cd uptake in rice cultivars treated with organic acids and EDTA. *J. Environ. Sci.* **2010**, *22*, 441–447. [CrossRef] [PubMed]
44. Zhao, K.; Fu, W.; Ye, Z.; Zhang, C. Contamination and spatial variation of heavy metals in the soil-rice system in Nanxun County, Southeastern China. *Int. J. Environ. Res. Public Health* **2015**, *12*, 1577–1594. [CrossRef] [PubMed]
45. Prakash Maran, J.; Sivakumar, V.; Thirugnanasambandham, K.; Sridhar, R. Artificial neural network and response surface methodology modeling in mass transfer parameters predictions during osmotic dehydration of *Carcia papaya* L. *Int. J. Food Eng.* **2013**, *52*, 507–516.
46. Olawoyin, R. Application of backpropagation artificial neural network prediction model for the PAH bioremediation of polluted soil. *Chemosphere* **2016**, *161*, 145–150. [CrossRef] [PubMed]
47. Keshavarzi, A.; Sarmadian, F. Comparison of artificial neural network and multivariate regression methods in prediction of soil cation exchange capacity. *Int. J. Environ. Chem. Ecol. Geo. GeoEng* **2010**, *4*, 644–649.
48. Shi, J.; Xu, J.; Yao, Y.; Xu, B. Concept learning through deep reinforcement learning with memory-augmented neural networks. *Neural Netw.* **2019**, *110*, 47–54. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.