COMPARATIVE STUDY OF METAL(LOID) CONTENT IN VERTICAL AND HORIZONTAL LAYERS OF SOIL: AN APPROACH THROUGH MACHINE LEARNING ESPECIALLY TREE ALGORITHMS

Sudipta Biswas[1]*, Phani Bhusan Ghosh[1] and Soumendra Nath Talapatra[2]

[1]Department of Chemistry
Seacom Skills University
Kendradangal, Shantiniketan,
Birbhum – 731236, West Bengal, India

[2]Department of Bio-Science
Seacom Skills University
Kendradangal, Shantiniketan,
Birbhum – 731236, West Bengal, India

Abstract: The present study was attempted to predict the prediction accuracy of datasets related to Pb metal and As metalloid content in vertical and horizontal layers of soil through machine learning (ML) especially tree algorithms by using WEKA tool, version 3.8.5. Different tree algorithm viz. Random Forest (RF), Random Tree (RT) and Fast decision tree learner Tree (REPT) were studied separately based on 4 attributes such as Area, Seasons, Pb and As as well as effect viz. high (H) and low (L) to determine overall prediction accuracy as per 10-fold cross validation. It is concluded that ML algorithms performed accurately from the dataset and obtained rich information with statistical validation in both vertical and horizontal layers of soils, which obtained Pb and As content up to 6 cm depth as vertically and 12 m distance as horizontally during pre-monsoon and post-monsoon without seasonal variations. The future study in WEKA tool can easily be analysed with more dataset to predict classifier accuracy related to metal(loid) content speciation in soil.

Keywords: Machine learning, Tree algorithms, Prediction accuracy, WEKA tool, Predictive soil content, Pb and As

1. INTRODUCTION

Generally, the municipal solid waste (MSW) is generated as “garbage” or “trashes”, which is an unavoidable by product of human activity [1]. These wastes such as raw vegetables and cooked food wastes, garden wastes, papers, woods, plastics, construction and demolition wastes, glass, ceramics, electrical and electronic wastes, etc. are found in which few are biodegradables, but majority wastes are non-biodegradable [2].

The contamination of soil by heavy metal can cause adverse effects on human health, animals, and soil productivity [2,3,4]. Waste carries different metals, which persist onto soil for long time and then transferred to plants by different ways [5-12]. Recently, Choudhury et al. demonstrated the presence of toxic trace elements, such as As, Cr, Pb, Cu, Ni, Zn, Hg, etc., in the soil as well as the groundwater near municipal solid wastes dumping ground at Assam [13]. They observed following order in the abundance of the metals at all three depths (surface, 15 and 30 cm): Zn > Fe > Ni > Cu > Cr [13]. According to Azeez et al. [14], trace metal levels in soil caused by MSW deposition in an emerging city in Abeokuta, Nigeria where the highest concentrations of Cu, Cr, Mn, and Zn were observed at a depth of 0-40 cm while Pb, Fe, and Ni accumulations were observed at depths below 40 cm. It is well known fact that heavy metal species found in decomposed municipal solid waste [15,16].

Several studies revealed that big data mining or deep learning is interesting research in which the dataset provide valuable information through statistical interpretation. The big data mining is based on the study of ML models algorithms, which is predicted the performance accuracy of the dataset [17,18]. In earlier

study, availability of big data found associated with the soil environment [19,20]. These data can be used as alternate variables for the pollution sources and influencing factors can also be used as covariates to determine the prediction accuracy [21]. Moreover, a recent study by Biswas et al. [22] observed the better performances as per ML algorithms such as BN, NB, LgR, RF and CART followed by J48 and RT for soil, but the dataset did not use as per soil stratifications.

The objective of the present study was to predict the performance accuracy of datasets of vertical and horizonal soil contained metal(loids) through machine learning (ML) classification models in the WEKA (Waikato Environment for Knowledge Analysis) tool (version 3.8.5).

2. MATERIALS AND METHODS

As per earlier study by Biswas et al. [22] WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) was used for data mining developed by Frank et al. [23] to determine performance accuracy through ML algorithms. As mentioned earlier study [22] the mechanisms of the WEKA explorer [24]. In pre-processing step, both vertical and horizontal soil distribution datasets were analysed through unsupervised instance and 10-fold cross validation (CV).

The predictive accuracy of both dataset were compared as per higher and lower values of metal(loid) content in soil of MSW dumping ground through ML modelling algorithms especially different tree classifiers viz. decision tree (DT) J48, Logistic model tree (LMT), Random forest (RF), Random tree (RT), Fast decision tree learner (REPT) and Class implementing minimal cost-complexity pruning (CART) along with 4 attributes viz. area, seasons, Pb and As and effects (higher as H and lower as L content) studied from dataset to predict the overall performance accuracy from the dataset of our earlier study of Biswas et al. [25].

The performance accuracy of above-mentioned ML model classifications related to correctly and incorrectly classified instances, Kappa statistics (KS), mean absolute error (MAE) and root mean squared error (RMSE) were studied for 10-fold CV test as per earlier study by Talapatra et al. [37], Bhattacharya et al. [41] and Biswas et al. [22]. As per Bouckaert et al. [42], the results of six tree algorithms model summary were retrieved from WEKA tool [26]. The prediction accuracy of studied ML models as per 10-fold CV test was retrieved from summary results and the statistical parameters such as Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC) and Precision-recall curve (PRC), respectively were recorded separately for vertical and horizontal soil.

3. RESULTS AND DISCUSSION

In the pre-processing step, graphical representation of statistical data of different attributes [area, seasons, Pb and As and effects (higher as H and lower as L content)] for vertical and horizontal soil separately were obtained. It is not always possible to identify the metal(loid) content as per stratification and these problems can easily be elucidated by resorting to big data mining, which is the abstraction of implicit, previously unknown, and potentially useful information in data [24]. Generally, ML is used to extract valuable information from raw data of metal(loid) content in soil distributed vertically and horizontally [21]. The process is based on abstraction in which data were collected, with all their defects, and the underlying structure is represented [24].

The performance of model accuracy of studied ML algorithm classifications as per correctly and incorrectly classified instances, KS, MAE and RMSE were studied as per 10-fold CV test. In the case of algorithm model classification, the similar values were observed in all studied algorithms used for vertical and horizontal soil dataset (Table 1 and Table 2).

Table 1: Results on different classified instances and statistical values for different algorithm models for vertical soil

Classifier model	Correctly classified instances	Incorrectly classified instances	KS	MAE	RMSE
DT J48	100.0	0.0	1.0	0.0	0.0
LMT	100.0	0.0	1.0	0.12	0.12
RF	100.0	0.0	1.0	0.10	0.02
RT	100.0	0.0	1.0	0.0	0.0
REPT	100.0	0.0	1.0	0.0	0.0
SCT	100.0	0.0	1.0	0.0	0.0

DT J48 = Pruned and unpruned decision tree C4; LMT = Logistic model tree, RF = Random Forest; RT = Random tree; REPT = Fast decision tree learner; SC = Class implementing minimal cost-complexity pruning; KS = Kappa Statistics; MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Table 2: Results on different classified instances and statistical values for different algorithm models for horizontal soil

Classifier model	Correctly classified instances	Incorrectly classified instances	KS	MAE	RMSE
DT J48	100.0	0.0	1.0	0.0	0.0
LMT	100.0	0.0	1.0	0.12	0.12
RF	100.0	0.0	1.0	0.002	0.006
RT	100.0	0.0	1.0	0.0	0.0
REPT	100.0	0.0	1.0	0.0	0.0
SCT	100.0	0.0	1.0	0.0	0.0

DT J48 = Pruned and unpruned decision tree C4; LMT = Logistic model tree, RF = Random Forest; RT = Random tree; REPT = Fast decision tree learner; SCT = Class implementing minimal cost-complexity pruning tree; KS = Kappa Statistics; MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Table 3 evaluates the representation of the detailed accuracy of studied models for the studied dataset. In case of the accuracy of a class of values, MCC, ROC and PRC, the better performances (100%) were observed for all studied algorithms for both datasets. As per these values, the ROC curve, margin curve and cost curve were exhibited in Fig 1, 2 and 3.

Table 3: Statistical data for prediction accuracy of studied algorithms for vertical and horizontal soil

Classifier model	Effects	MCC	ROC area	PRC area
DT J48	H	100.0	100.0	100.0
	L	100.0	100.0	100.0
LMT	H	100.0	100.0	100.0
	L	100.0	100.0	100.0
RF	H	100.0	100.0	100.0
	L	100.0	100.0	100.0

RT	H	100.0	100.0	100.0
	L	100.0	100.0	100.0
REPT	H	100.0	100.0	100.0
	L	100.0	100.0	100.0
SCT	H	100.0	100.0	100.0
	L	100.0	100.0	100.0

DT J48 = Pruned and unpruned decision tree C4; LMT = Logistic model tree, RF = Random Forest; RT = Random tree; REPT = Fast decision tree learner; SCT = Class implementing minimal cost-complexity pruning tree; MCC = Matthew's correlation coefficient; ROC = Receiver operating characteristic; PRC = Precision-recall curve

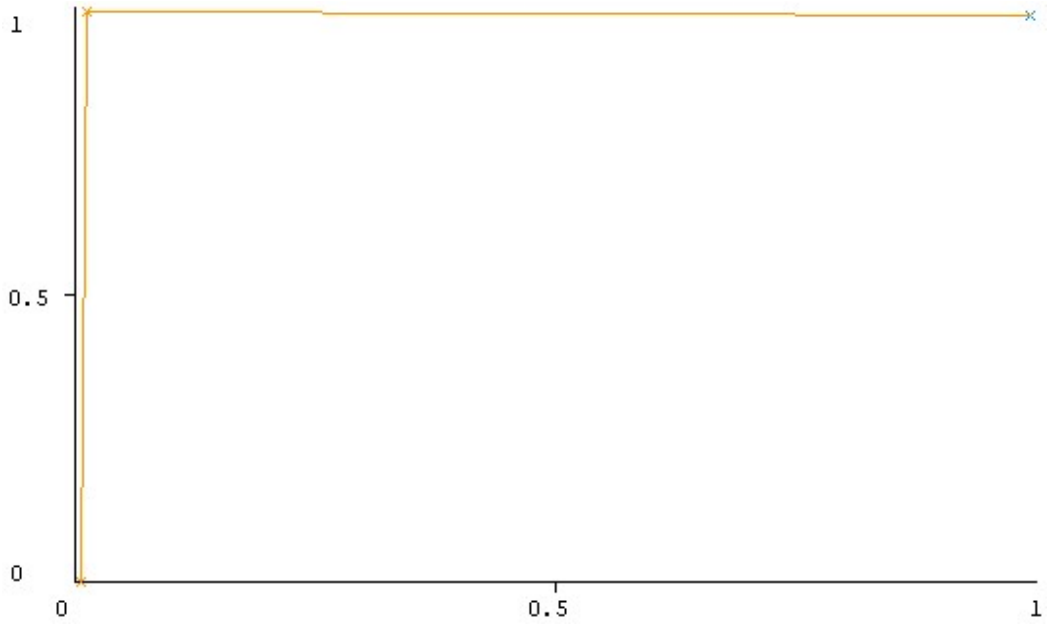

Figure 1: Area under ROC (=1) plot for all studied algorithms of vertical and horizontal soil

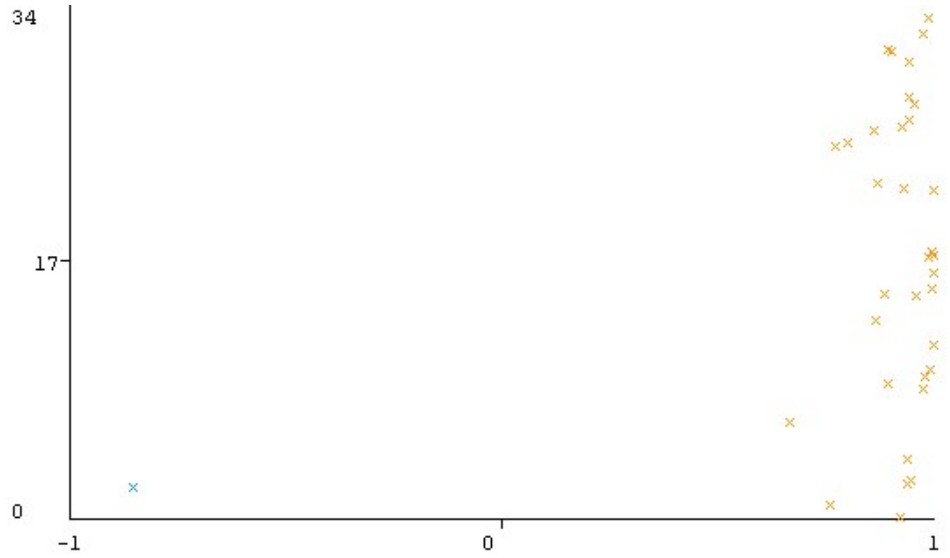

Figure 2: Margin curve plot for all studied algorithms of vertical and horizontal soil

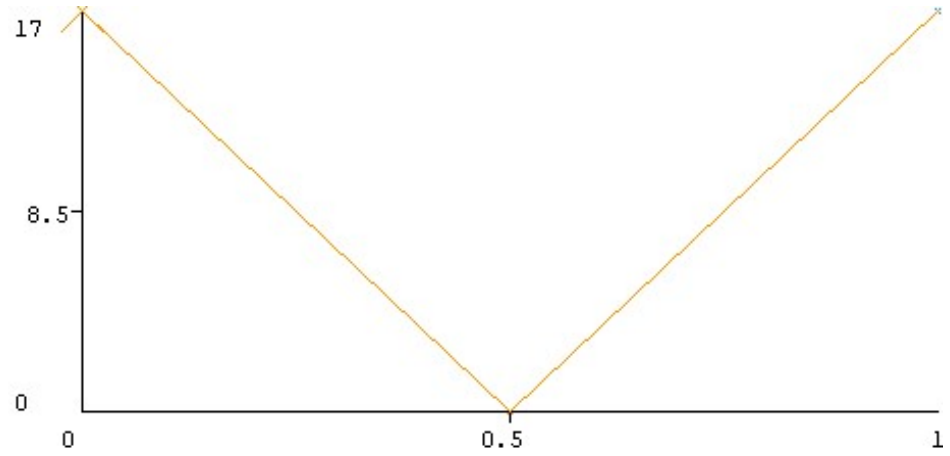

Figure 3: Cost curve plot for all studied algorithms of vertical and horizontal soil

In the present study, an ROC curve plot is based on true positive rate (TPR) vs. false positive rate (FPR) as per different classifiers thresholds and its performance with the value obtained of 1.0. Lowering the classifiers threshold classified maximum data as positive, thus increasing both TP and FP values. This curve is suitable for ML study [28]. While margin curve plot determines cumulative probability as per instance numbers and the performance of the 6-types classifier was observed better performance with a cumulative value of 34 [29]. Moreover, cost curve plot determines a cost function, which is used to know incorrect the model is in finding a relation between the input and output. It is known how badly the model is behaving/predicting the dataset [30]. In the present study, the classification prediction accuracy was 50%.

4. CONCLUSION

In the present predictive study, ML algorithms performed accurately from the dataset and obtained rich information with statistical validation in both vertical (up to 6 cm) and horizontal (up to 12 m distance) layers of soils, which obtained Pb and As content in soil during pre-monsoon and post-monsoon without

seasonal variations. In future, the predictive study in WEKA tool can easily be analysed with more dataset to predict classifier accuracy related to metal(loid) content speciation in soil.

Acknowledgement

Authors convey thanks to the developer of present tool, which used in the present study.

Funding source

This is a non-funded project.

Conflict of interest

Authors declare no conflict of interest.

REFERENCES

[1] S. M. Ali, A. Pervaiz, B. Afzal, N. Hamid, and A. Yasmin, Open dumping of municipal solid waste and its hazardous impacts on soil and vegetation diversity at waste dumping sites of Islamabad city. Journal of King Saud University - Science, vol. 26 no.1, **(2014)**, pp. 59-65.

[2] Kolkata Municipal Corporation. Assessment Report on Dhapa Disposal Site Kolkata, India. Prepared under the support of: U. S. Environmental Protection Agency Landfill Methane Outreach Program, **(2010)**.

[3] S. Esakku, K. Palanivelu, and K. Joseph, Assessment of heavy metals in a municipal solid waste dumpsite. In Proceedings of the Workshop on Sustainable Landfill Management, Chennai, India, 3–5 December; 35, **(2003)**, pp. 139-145.

[4] C.J. Smith, P. Hopmans, and F.J. Cook, Accumulation of Cr, Pb, Cu, Ni, Zn and Cd in soil following irrigation with treated urban effluent in Australia. Environ Pollut, vol. 94 no. 3, **(1996)**, pp. 317-323.

[5] H. Shaylor, M. McBride, E. Harrison, Sources and impacts of contaminants in soil. Cornell Waste Management Institute, **(2009)**. Available from: http://cwmi.css.cornell.edu.

[6] Y. Y. Long, D. S. Shen, H. T. Wang, W. J. Lu, and Y. Zhao, Heavy metal source analysis in municipal solid waste (MSW): Case study on Cu and Zn. Journal of Hazardous Materials, vol. 186, **(2011)**, 1082-1087.

[7] S. Kanmani, and R. Gandhimathi, Assessment of heavy metal contamination in soil due to leachate migration from an open dumping site. Applied Water Science, vol. 13, **(2013)**, pp.193-205.

[8] R. G. Van Ryan Kristopher, and R. Parilla, Analysis of heavy metals in Cebu city sanitary landfill, Philippines. Journal of Environmental Science and Management, vol. 17, **(2014)**, pp. 50-59.

[9] Y. N. Vodyanitskii, Biochemical processes in soil and groundwater contaminated by leachates from municipal landfills (mini review). Annals of Agrarian Science, vol. 14, **(2016)**, pp. 249-256.

[10] B. Gworek, W. Dmuchowski, E. Koda, M. Marecka, A. H. Baczewska, P. Bragoszewska, and P. Osin´ski, Impact of the municipal solid waste Łubna landfill on environmental pollution by heavy metals. Water, vol. 8, **(2016)**, p.470.

[11] S. R. Samadder, R. Prabhakar, D. Khan, D. Kishan, and M. S. Chauhan, Analysis of the contaminants released from municipal solid waste landfill site: A case study. Science of the Total Environment, vol. 580, **(2017)**, pp. 593-601.

[12] N. Vongdala, H. D. Tran, T. D. Xuan, R. Teschke, and T. D. Khanh, Heavy metal accumulation in water, soil, and plants of municipal solid waste landfill in Vientiane, Laos. International Journal of Environmental Research and Public Health, vol. 16, **(2019)**, p. 22.

[13] M. Choudhury, D. S. Jyethi, J. Dutta, S. P. Purkayastha, D. Deb, R. Das, G. Roy, T. Sen, and K. G. Bhattacharyya, Investigation of groundwater and soil quality near to a municipal waste disposal site in Silchar, Assam, India. Int J Energ Water Res, vol. 6 no. 1, (2022), pp. 37-47.

[14] J. O. Azeez, O. A. Hassan, and P. O. Egunjobi, Soil contamination at dumpsites: Implication of soil heavy metals distribution in municipal solid waste disposal system: A case study of Abeokuta, Southwestern Nigeria. Soil and Sediment Contamination: An International Journa, vol. 20, **(2011)**, pp. 370-386.

[15] S. Esakku, A. Selvam, K. Joseph, and K. Palanivelu, Assessment of heavy metal species in decomposed municipal solid waste. Chemical Speciation & Bioavailability, vol. 17 no. 3, **(2005)**, pp. 95-102.

[16] G. Demie, and H. Degefa, Heavy metal pollution of soil around solid waste dumping sites and its impact on adjacent community: the case of Shashemane open landfill, Ethiopia. Journal of Environment and Earth Science, vol. 5 no. 15, **(2015)**, pp. 169-179.

[17] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. Nature, vol. 521 no. 7553, **(2015)**, pp. 436-444.

[18] S. Mishra, A. Dash, and L. Jena, Use of deep learning for disease detection and diagnosis. In: Bio-inspired Neurocomputing. Studies in Computational Intelligence. (eds. Bhoi, A., Mallick, P., Liu, C.M. and Balas, V.), Springer, Singapore, **(2021)**, p. 903.

[18] X. He, L. Yang, A. Li, L. Zhang, F. Shen, Y. Cai, and C. Zhou, Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. CATENA, vol. 205, **(2021)**, p. 105442.

[19] H. Wu, A. Lin, X. Xing, D. Song, and Y. Li, Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method. Int J Appl Earth Obs Geoinf, vol. 103, **(2021)**, p. 102475.

[20] Y.-P. Lin, B.-Y. Cheng, H.-J. Chu, T.-K. Chang, and H.-L. Yu, Assessing how heavy metal pollution and human activity are related by using logistic regression and kriging methods Geoderma, vol. 163 no. 3-4, **(2011)**, pp. 275-282.

[21] W. Cao, and C. Zhang, Data prediction of soil heavy metal content by deep composite model. J Soils Sediments, vol. 21 no. 1, **(2020)**, pp. 487-498.

[22] S. Biswas, P. B. Ghosh, and S. N. Talapatra, The application of machine learning algorithms for prediction of performance accuracy for metal(loid) adsorption in soil and uptake

in weeds. Journal of Electronics Information Technology Science and Management. vol. 12 no. 11, **(2022)**, pp. 178-180.

[23] E. Frank, M. A. Hall, and I. H. Witten, The WEKA workbench, Online appendix for data mining: Practical machine learning tools and techniques. Morgan Kaufmann, 4th edition, **(2016)**.

[24] Witten, I. H., Frank, E., Hall, M. A. Data Mining: Practical machine learning tools and techniques. 3rd edn, Morgan Kaufmann, Burlington, MA, **(2011)**.

[25] S. Biswas, S. N. Talapatra, and P. B. Ghosh, Phytoremediation potential of elements by weed species around solid waste dumping ground, Berhampur, West Bengal, India. Pollution Research, vol. 40 no. 3, **(2021)**, pp. 344-352.

[26] S. N. Talapatra, R. Chaudhury, and S. Ghosh, CellProfiler and WEKA Tools: Image analysis for fish erythrocytes shape and machine learning model algorithm accuracy prediction of dataset. World Scientific News, vol. 154, **(2021)**, pp. 101-116.

[27] K. Bhattacharya, B. Mondal, and S. N. Talapatra, Multilayer perceptron network of machine learning for prediction accuracy after genetic biomonitoring of estuarine fish specimen. Journal of Electronics Information Technology Science and Management, vol. 12 no. 10, **(2022)**, pp. 42-50.

[28] M. Majnik, and Z. Bosnic, ROC analysis of classifiers in machine learning: A survey. Intelligent Data Analysis, vol. 17, **(2013)**, pp. 531-538.

[29] Y. Zhou, F. Yu, and T. Duong, Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning. PLoS ONE, vol. 9 no. 6, **(2014)**, p. e90405.

[30] M. Banoula, What is cost function in machine learning. Simplilearn tutorial. Updated **2022**, 13 September. Available from: https://www.simplilearn.com/tutorials/machine-learning-tutorial/cost-function-in-machine-learning