Article Title: Delineating Groundwater Potential Zones in Tropical Hard-Rock Aquifers Using Machine and Deep Learning, Spatial Cross-Validation, and Spatially Explicit Uncertainty Analysis

Journal Name: Earth Science Informatics

Author Names: Suraj Dule[1], Arabinda Sharma[*1]

Affiliation and E-mail of Corresponding Author:

[1] *School of Geography, Gangadhar Meher University, Sambalpur, Odisha, India*

[*]*Email: arbind_78@rediffmail.com*

Online Resource 1: Supplementary Tables

Table S1 Hyperparameter search spaces, distributions, and final optimized values selected via the Optuna framework.

Model	Hyperparameter	Search Range / Distribution	Final Value*	Description
Random Forest (RF)	n_estimators	300 – 1200 (Int)	1184	Number of trees in the forest.
	max_depth	5 – 16 (Int)	15	Maximum depth of the tree.
	max_features	0.2 – 0.8 (Float)	0.38	Fraction of features to consider at each split.
	min_samples_split	20 – 40 (Int)	22	Min. samples required to split an internal node.
	min_samples_leaf	1 – 15 (Int)	2	Min. samples required to be at a leaf node.
	bootstrap	[True]	True	Whether bootstrap samples are used.
XGBoost (Standard & SF)	n_estimators	400 – 1200 (Int)	1082	Number of gradient boosted trees.

	max_depth	4 – 8 (Int)	7	Maximum tree depth for base learners.
	learning_rate (η)	0.01 – 0.3 (Log-Uniform)	0.018	Step size shrinkage used in update.
	subsample	0.5 – 1.0 (Uniform)	0.68	Subsample ratio of the training instances.
	colsample_bytree	0.3 – 1.0 (Uniform)	0.74	Subsample ratio of columns per tree.
	gamma	1.5 – 3.5 (Uniform)	1.92	Min. loss reduction required to make a partition.
	reg_alpha (L1)	0.1 – 10.0 (Log-Uniform)	0.24	L1 regularization term on weights.
	reg_lambda (L2)	0.1 – 10.0 (Log-Uniform)	1.85	L2 regularization term on weights.
CatBoost	iterations	400 – 1500 (Int)	1350	The maximum number of trees built.
	depth	4 – 12 (Int)	8	Depth of the tree.
	learning_rate	0.01 – 0.2 (Log-Uniform)	0.042	Learning rate.
	l2_leaf_reg	0.001 – 10.0 (Log-Uniform)	3.5	Coefficient at the L2 regularization term.
	rsm	0.3 – 1.0 (Uniform)	0.88	Random subspace method (colsample).
Support Vector Machine (SVM)	C	0.1 – 20.0 (Log-Uniform)	8.4	Regularization parameter.
	gamma	0.0001 – 0.1 (Log-Uniform)	0.03	Kernel coefficient for 'rbf'.

epsilon (ϵ)	0.1 – 1.0 (Log-Uniform)	0.15	Epsilon in the ϵ-insensitive loss function.

*Final values correspond to the configuration yielding the highest spatial cross-validation R^2 score.

Table S2 Fixed architectural configuration and training hyperparameters for the Multi-Layer Perceptron Deep Learning (MLP-DL) model.

Component	Configuration	Description
Backbone Architecture	[256, 256, 128, 64]	Layer-wise hidden dimensions of the residual MLP.
Optimization	Adam (lr=5e-4)	Adaptive Moment Estimation optimizer.
Batch Size	32	Number of samples per gradient update step.
Regularization	Dropout (p=0.1)	Probability of element-wise zeroing for regularization.
Embedding Strategy	$N_{unique}/2$	Factor to determine embedding size for categorical features.

Table S3 Variance Inflation Factor (VIF) Results.

Feature	VIF (Base)	VIF (XGBoost-SF)	Interpretation
GWPZ Gradient	N/A	72,139.30	Extreme Multicollinearity (with Slope)
Slope	1.13	71,660.94	Extreme Multicollinearity (with GWPZ Gradient)
Elevation	1.83	8,081.75	Extreme Multicollinearity (with GWPZ Mean)
GWPZ Mean	N/A	8,075.25	Extreme Multicollinearity (with Elevation)
GWPZ Std	N/A	8.85	High Correlation
Runoff	5.25	5.27	High Correlation
Rainfall	4.56	4.72	Moderate Correlation
NDVI	1.88	1.90	Low Correlation

Mean WL	1.67	1.68	Low Correlation
Drainage Density	1.54	1.57	Low Correlation
Lineament Density	1.16	1.20	Low Correlation