Transportation Infrastructure Geotechnology (2025) 12:282
https://doi.org/10.1007/s40515-025-00740-3

RESEARCH

Interpretable Data-Driven Modeling of Bridge Pier Scour Using Gene Expression Programming and Physics-Informed Neural Networks

Laiba Gulaly[1] · Majid Khan[2] · Husna Usman[1] · Aaila Asghar[1]

Received: 19 August 2025 / Accepted: 29 October 2025

Abstract
Bridge pier scour was investigated using advanced machine learning approaches to enhance the prediction of normalized scour depth. Gene expression programming (GEP), three variants of physics-informed neural networks (PINNs), and a baseline artificial neural network were developed and evaluated using the USGS pier-scour database comprising 569 laboratory-scale data. GEP achieved the highest predictive accuracy, with a coefficient of determination R^2 of 0.883 and the lowest prediction errors, demonstrating strong generalization capability. Among the PINN variants, the GreyBox model achieved a balanced trade-off between physical consistency and predictive accuracy, while PINN-Lite reduced computational cost with moderate precision, and PINN-Melville highlighted the limitations of embedding simplified empirical relationships. Explainable artificial intelligence (XAI) techniques, including SHapley additive explanations, local interpretable model-agnostic explanations, individual conditional expectation plots, and partial dependence plots, were applied to interpret model predictions. Analysis identified velocity ratio and sediment ratio as dominant predictors, with coupled hydraulic-sediment interactions shaping scour responses. The results showed that changes in scour depth mainly depend on how strong the water flow is and how the sediment behaves. The integration of GEP, PINNs, and XAI provides a powerful framework combining predictive accuracy with physical interpretability, offering valuable insights for bridge design, safety assessment, and scour risk management.

Keywords Bridge pier scour · Gene expression programming · Physics-informed neural network · Explainable machine learning

Extended author information available on the last page of the article

Published online: 13 November 2025

1 Introduction

Bridge scour, defined as the removal of sediment from around bridge piers due to flowing water, is a critical phenomenon that directly impacts the safety and stability of bridges (Wang et al. 2025). Statistics highlight the severity of the problem, as more than 41% of recorded scour depths fall between 0.5 and 5 m, with maximum depths reaching up to 15 m. In the United States alone, approximately 60% of bridge collapses have been attributed to local scour, while more than 2,500 bridges were damaged by flood-related scour between 1980 and 1990 in the Northeast and Midwest regions (Briaud et al. 1999). Current estimates suggest there are over 80,000 scour-susceptible bridges, including nearly 21,000 classified as scour-critical. These figures underscore the urgent need for accurate prediction and mitigation strategies (Kim et al. 2024a, b).

The formation of scour around bridge piers is a dynamic process governed by interactions between the flow field, sediment transport, and structural geometry. The scour process typically develops in three stages: initiation, where the flow obstruction induces complex vortices such as downflow, horseshoe vortices, and wake vortices; progression, marked by intensified sediment erosion due to turbulent flow; and equilibrium, reached when erosive forces and resisting sediment forces balance. Various types of scour have been identified, including local scour caused by direct flow impact, contraction scour from channel constriction, general scour resulting from long-term bed changes, lateral scour along riverbanks, and ponding scour due to water-level differences across the pier. Failure statistics indicate that approximately 64% of bridge failures are caused by local scour, followed by 14% due to channel migration and 5% due to contraction scour. Among these, nearly 70% of failures manifest in lateral and vertical failure modes (Lin et al. 2014).

Accurate prediction of scour depth remains a critical challenge due to the highly unpredictable nature of the phenomenon, particularly under extreme hydrological events such as floods. Even minor inaccuracies in scour estimation can lead to catastrophic structural failures. Contributing factors include noisy and incomplete input data, nonuniform load distributions at pier-abutment interfaces, debris and vegetation accumulation, and site-specific variability in sediment and hydraulic conditions. These complexities necessitate robust predictive tools that can capture the nonlinear interactions influencing scour.

Traditionally, two main approaches have been employed for scour depth estimation. The first involves empirical models that express scour depth as a function of flow characteristics, pier geometry, and sediment properties (Omara et al. 2020). While widely used, these formulas are generally reliable only within the dataset on which they were developed and often produce inconsistent estimates under similar conditions. The second approach relies on three-dimensional computational fluid dynamics (CFD) models, which provide valuable insights into flow-sediment interactions and scour hole formation. However, these models are computationally demanding, require extensive calibration and validation, and are difficult to generalize for diverse field conditions (Ettema et al. 2011; Richardson and Panchang 1998).

Several engineering guidelines have also been developed for practical scour prediction. The federal highway administration's (FHWA's) hydraulic engineering circular-18 (HEC-18) method is the most widely adopted, offering comprehensive guidance across a range of hydraulic and sediment conditions, though it often provides conservative estimates based on uniform sediment assumptions. The Modified HEC-18 improves accuracy for site-specific conditions but requires extensive calibration, while the Old HEC-18 offers simplicity at the expense of reduced accuracy in complex scenarios (Schuring et al. 2017). Other region-specific methods include the florida department of transportation (FDOT) method, which accounts for tidal influences but is geographically limited, and the scour rate in cohesive soils (SRICOS) method (Kwak and Briaud 2002), which is particularly effective in cohesive soils and for time-dependent scour predictions, though its complexity and testing costs limit widespread use.

Machine learning (ML) has recently emerged as a powerful tool for improving predictive accuracy in complex engineering problems, including scour depth estimation around bridge piers. Traditional data-driven algorithms such as artificial neural networks (ANN), random forests, and support vector machines (SVM) have demonstrated notable success in capturing nonlinear relationships between hydraulic, geometric, and sediment variables (Wang et al. 2025; Parsaie et al. 2019; Qaderi et al. 2021; Kim et al. 2024a, b). For example, ANN trained on FHWA datasets containing hundreds of field and laboratory measurements have achieved superior performance compared to empirical formulas, particularly under diverse hydraulic conditions. Similarly, evolutionary approaches such as gene-expression programming (GEP) have provided accurate functional relationships for scour prediction (Azamathulla et al. 2010; Khan et al. 2012), offering interpretable alternatives to purely black-box models.

Recent studies have explored advanced modeling and mitigation strategies for scour around bridge piers and hydraulic structures. Molavi et al. (2025). compared several ML techniques, GEP, support vector regression, and ANN, to simulate temporal bed profile evolution around cylindrical piers. Their findings confirmed the high predictive capability of GEP; however, the study relied exclusively on controlled laboratory data, lacking validation under field-scale hydraulic variability and omitting interpretability analysis. Fathi et al. (2023) experimentally evaluated the combined use of riprap and submerged vanes as abutment scour countermeasures, demonstrating substantial scour reduction and riprap stability enhancement. Yet, such findings are configuration-specific and provide limited insight into underlying flow–sediment interactions that could generalize to diverse field conditions. Fuladipanah et al. (2023a, b) applied ML-based models for predicting scour depth downstream of flip bucket spillways and around bridge piers, achieving improved performance over empirical regressions; however, these models were treated as black boxes, offering minimal explanation of variable influence or model reasoning. Similarly, Raeisi et al. (2024) experimentally assessed sacrificial piles as countermeasures against local scour around underwater pipelines, successfully demonstrating physical mitigation mechanisms but without incorporating predictive modeling or generalizable frameworks. Collectively, these studies highlight the growing role of ML and hybrid approaches in scour prediction and control. Nonetheless, most efforts remain limited either by their reliance on laboratory data, the absence of physical constraints within the models, or a lack of interpretability, restricting their applicability in real-world bridge design and risk assessment.

Addressing these shortcomings, the present research develops explainable and physics-informed ML frameworks that couple predictive accuracy with physical consistency and transparency, thereby enhancing trust and usability for engineering practice (Ribeiro et al. 2016). It has moved toward physics-informed neural networks (PINNs), which embed governing physical equations into the training process. Unlike conventional black-box ML models, PINNs serve as grey-box models, enhancing both predictive power and physical consistency. In this study, a baseline ANN is first developed as a reference model. Building upon it, three PINN variants are introduced: a general grey-box PINN, a computationally simplified PINN-Lite, and a domain-specific PINN-Melville (2000) formulation that integrates established scour equations directly into the network architecture. These models aim to balance predictive accuracy, computational efficiency, and physical realism. In parallel, symbolic regression methods are explored, with GEP employed to derive explicit functional relationships. Together, these approaches bridge the gap between traditional empirical equations and modern ML models, providing both predictive strength and interpretability, thereby making the outcomes more transparent and accessible to practicing engineers.

To further enhance trust and usability, this research incorporates state-of-the-art explainable artificial intelligence (XAI) techniques. Tools such as SHAP (SHapley Additive exPlanations), LIME (local interpretable model-agnostic explanations), individual conditional expectation (ICE) plots, and partial dependence plots (PDP) are employed to interpret the predictions of ANN, PINN, and GEP models. These methods quantify feature importance, visualize variable interactions, and provide case-specific explanations, offering engineers a transparent understanding of how each model arrives at its predictions. In summary, this study integrates multiple modeling approaches: a baseline ANN, three grey-box PINN variants, GEP, coupled with advanced interpretability tools (LIME, SHAP, ICE, PDP). Together, these methods aim to develop accurate, physically consistent, and transparent models for predicting bridge pier scour depth. By addressing both accuracy and interpretability, the research contributes to safer bridge design, improved maintenance planning, and the advancement of trustworthy AI applications in hydraulic engineering.

2 Research Methodology

2.1 Overview of Predictive Machine Learning Algorithms

In this study, multiple ML algorithms are employed to develop predictive models for estimating bridge pier scour depth. The selected algorithms include GEP and PINN, with a baseline ANN serving as a reference model. These approaches are chosen to capture the nonlinear interactions among hydraulic, sediment, and structural parameters while also addressing the critical need for interpretability in engineering applications.

2.1.1 Gene-Expression Programming (GEP)

GEP is an evolutionary computation technique commonly used for symbolic regression (Ferreira 2006). It evolves mathematical expressions by representing them as linear chromosomes, which are later expressed as nonlinear parse trees. This flexible encoding enables GEP to capture complex nonlinear relationships while maintaining interpretability through explicit functional forms. In the context of this study, GEP is employed as a tool for exploring potential empirical-like relationships for pier scour prediction. Although no finalized equation has been derived at this stage, the methodology provides a systematic framework for discovering simplified and interpretable models that complement data-driven approaches.

2.1.2 Physics-Informed Neural Networks (PINN)

PINN is a grey-box machine learning framework that integrates physical laws with data-driven modeling. Unlike traditional black-box neural networks, PINNs embed governing differential equations, conservation principles, or empirical scour relationships directly into their loss function (Lawal et al. 2022). This allows the model to respect physical consistency while still learning complex nonlinear interactions from data. In this study, a baseline ANN model is first developed as a reference, and then three variants of PINNs are explored: a general grey-box PINN that incorporates fundamental flow-sediment interaction equations; a PINN-Lite model, which reduces computational cost by applying simplified physical constraints; and a PINN-Melville model, which embeds Melville's widely used empirical scour equations into the network structure. These different implementations enable a systematic assessment of PINNs in terms of predictive accuracy, computational efficiency, and their ability to maintain physical realism in bridge scour prediction.

2.2 Dataset Acquisition

The success of ML models in scour prediction depends heavily on the quality and representativeness of the dataset. For this research, dataset was obtained from the USGS pier-scour database (PSDb-2014), which was compiled through an extensive review of published studies conducted by the U.S. geological survey (Benedict and Caldwell 2014). This database integrates both laboratory and field measurements, covering diverse hydraulic and sediment conditions. Specifically, it contains 569 laboratory tests and 1,858 field observations, gathered from 23 U.S. states and six additional countries, making it one of the most comprehensive resources for scour research.

We used 569 laboratory data and, to ensure robust model development and fair evaluation, divided them into three subsets. Approximately 70% of the records were allocated for training to optimize network weights and hyperparameters. Another 15% of the data was used for validation, which helped monitor generalization capability and minimize overfitting. The remaining 15% served as the independent testing set, providing an unbiased measure of predictive accuracy once the models were finalized.

Scour depth at bridge piers is governed by a combination of hydraulic, sediment, and geometric factors (Shahriar et al. 2021). In general form, this relationship can be expressed as in **Eq. (1)**:

$$\mathrm{Y_s} = \mathrm{f}\,(\mathrm{b},\ \mathrm{V},\ \mathrm{V_c},\ \mathrm{y},\ \mathrm{D_{50}},\ \sigma_{\mathrm{g}}) \tag{1}$$

where y_s (ft) is the pier scour depth, b (ft) is the pier width normal to flow, V (ft/s) is the approach flow velocity, V_c (ft/s) is the sediment critical velocity, y (m) is the approach flow depth, D_{50} (mm) is the median grain size, and σ_g is the gradation coefficient.

For the machine learning models, the scour depth is normalized with respect to the pier width, and five dimensionless input variables were used, as described in the **Eq. (2)**:

$$\frac{y_s}{b} = f\left(\frac{y}{b}, \frac{V}{V_c}, \frac{b}{D_{50}}, \sigma\,\mathrm{g}\,, F_r\right) \tag{2}$$

where y_s/b is the scour depth normalized with respect to the pier width, flow depth to pier width ratio (y/b) is a factor related to pier geometry, flow velocity to sediment critical velocity (V/V_c), and F_r (F_r is the Froude number = $\frac{V}{\sqrt{g\times y_s}}$) are factors related to flow, and pier width normal to flow to sediment size (b/D_{50}), and sediment uniformity factor σ_g are factors related to sediment size.

2.3 Pre-Processing and Statistical Description of the Dataset

The raw dataset underwent several preprocessing steps to enhance data quality and ensure reliable model predictions. Outlier detection was performed using both the Z-score ($|z| > 3$) and interquartile range (IQR) methods (1.5×IQR rule) for comparison; however, the Z-score approach was adopted for final analysis, identifying 56 outliers out of 569 samples. After screening, extreme anomalies that could distort the learning process were removed, reducing the dataset size from 569 to 513 samples. The distributions of input variables (y/b, V/V_c, b/D_{50}, σ_g, F_r) and the output variable (y_s/b) are shown in the histograms (Figs. 1 and 2), which highlight skewness in several predictors, particularly b/D_{50} and σ_g, whereas y_s/b follows a relatively bell-shaped distribution. Next, missing value treatment was applied; 16 missing entries were detected in the σ_g column. Since σ_g exhibits a skewed distribution, median imputation with a value of 1.3 was used instead of mean imputation to ensure robustness against outliers. After imputation, no missing values remained. Finally, the cleaned dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets, enabling model optimization, generalization checks, and unbiased evaluation. These pre-processing steps, supported by statistical plots, ensured that the dataset was consistent, representative, and suitable for developing both machine learning and physics-informed neural network models.

To better understand the data distribution and variability, a statistical summary was generated for the main input and output parameters as described in Table 1. Descrip-

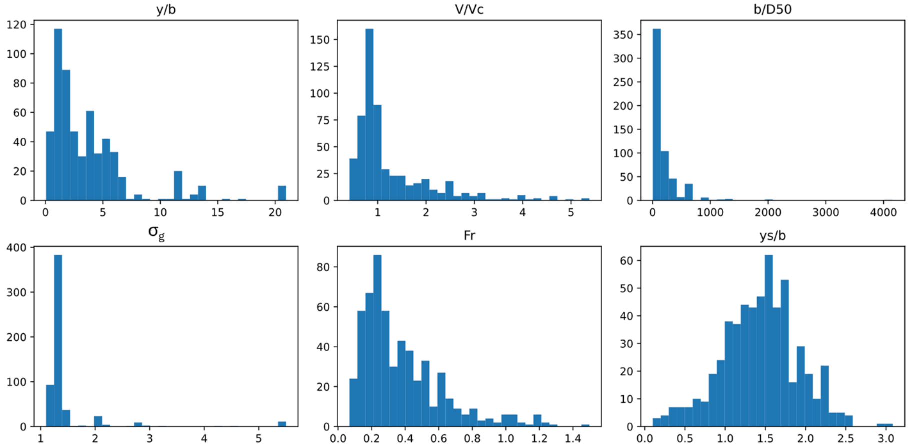


Fig. 1 Feature values distribution histograms

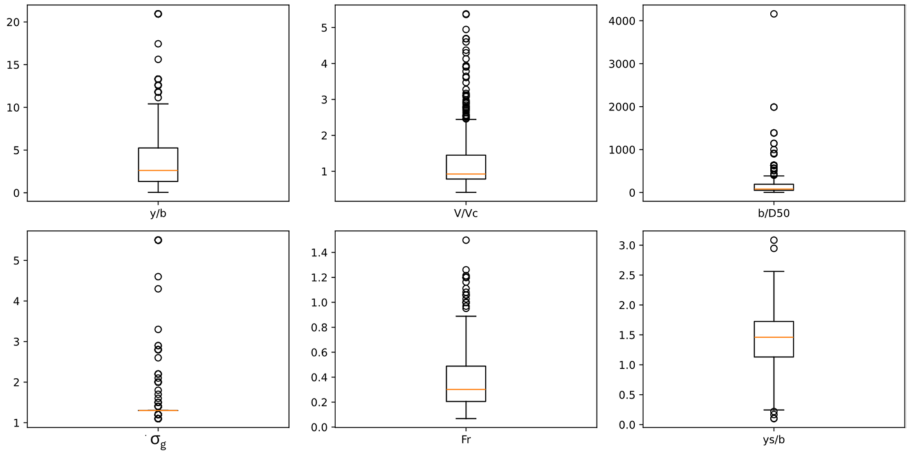


Fig. 2 Feature values distribution violin plots

tive statistics, including mean, standard deviation (SD), skewness, and kurtosis, were calculated, along with minimum and maximum values for each subset. For the training dataset ($n=399$), the mean values of y/b, V/V_c, b/D_{50}, F_r, and y_s/b were 3.97, 1.23, 170.07, 0.37, and 1.44, respectively. The skewness values indicated positive skew in most variables, particularly b/D_{50} (8.30), while ys/b showed near symmetry (–0.08). The validation dataset ($n=85$) exhibited a slightly lower mean y/b (3.38) and higher velocity ratio (1.30), with b/D_{50} averaging 204.65. The testing dataset ($n=85$) was broadly consistent with training, with a mean y_s/b of 1.40 and a similar spread in F_r (mean 0.39, SD 0.24). Overall, the statistical characteristics demonstrate that the three subsets are comparable, with some variation in higher-order statistics (skewness, kurtosis) due to natural variability in hydraulic conditions. This structured

Table 1 Statistical description of the dataset input and output parameters

Parameter	y/b	V/V_c	b/D_{50}	F_r	y_s/b
Training data					
Count	399	399	399	399	399
Mean	3.97	1.23	170.07	0.37	1.44
SD	3.87	0.81	283.73	0.24	0.47
Kurtosis	6.67	6.00	102.86	2.92	0.37
Skewness	2.38	2.26	8.30	1.55	−0.08
Minimum	0.05	0.42	3.67	0.07	0.10
Maximum	20.95	5.38	4159.14	1.50	3.08
Validation data					
Count	85	85	85	85	85
Mean	3.38	1.30	204.65	0.40	1.45
SD	3.24	0.91	299.05	0.29	0.49
Kurtosis	2.31	6.10	14.92	1.58	−0.34
Skewness	1.66	2.37	3.30	1.50	−0.07
Min	0.10	0.48	5.36	0.08	0.41
Max	13.30	5.36	1988.49	1.26	2.53
Testing data					
Count	85	85	85	85	85
Mean	3.77	1.32	210.70	0.39	1.40
SD	4.07	0.87	284.70	0.24	0.48
Kurtosis	6.71	1.94	6.77	0.68	0.53
Skewness	2.42	1.63	2.50	1.10	−0.24
Min	0.10	0.42	3.67	0.11	0.10
Max	20.95	4.30	1386.84	1.11	2.56

pre-processing ensures that the ML and PINN models are trained on representative and balanced data, improving the reliability of performance assessments.

Pearson's correlation analysis (Fig. 3) provided additional insights into the relationships among variables. As expected, V/V_c was strongly correlated with F_r (r=0.76), reflecting their joint dependence on hydraulic intensity. The target variable ys/b showed its strongest positive correlation with V/V_c (r=0.23), confirming the role of excess velocity in promoting scour, while negative correlations were observed with b/D_{50} ($r = -0.28$) and σg ($r = -0.19$), indicating that larger sediment sizes and greater non-uniformity act to suppress scour through armoring effects. The geometric parameter y/b showed negligible correlation with ys/b (r≈0), suggesting that flow depth relative to pier width plays a minor direct role in predicting normalized scour depth. Taken together, the statistical and correlation analyses emphasize the dual influence of hydraulic and sediment characteristics on scour processes, providing a solid foundation for subsequent predictive modeling.

2.4 Development and Hyperparameter Tuning of ML Models

The development and tuning of machine learning models followed a systematic pipeline combining environment setup, data preparation, model training, and rigorous hyperparameter optimization. All models were implemented in Python 3.10 within a Conda environment, executed in Jupyter Notebook. Widely used scientific libraries such as pandas, numpy, scipy, scikit-learn, TensorFlow, and Keras were employed

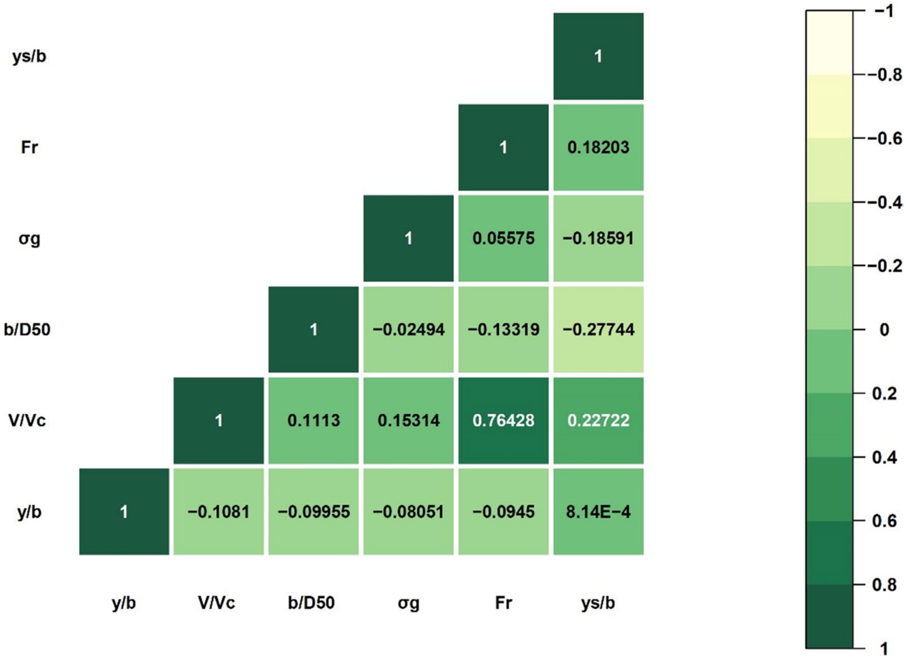


Fig. 3 Pearson's correlation heat plot for the dataset

to handle data preprocessing, model construction, and evaluation. The dataset was partitioned into training, validation, and testing subsets following a 70/15/15 split, ensuring both reliable model development and independent assessment of generalization. To further enhance robustness, a 5-fold cross-validation strategy was applied on the training portion, reducing the risk of overfitting and ensuring stability across different subsets of the data.

Hyperparameter optimization was conducted using a grid search strategy, systematically exploring parameter combinations to identify the configuration that minimized prediction error while maintaining numerical stability. For neural models, parameters such as the number of neurons, activation functions, learning rate, and optimizer settings were tuned. For GEP, evolutionary parameters including the number of genes, head size, operator set, and mutation rates were calibrated to balance exploration and convergence. The incorporation of physics-informed penalty terms in the PINN variants required additional tuning to ensure that the constraints improved physical realism without destabilizing convergence. Table 2 presents the final set of hyperparameters and configurations adopted for each model. The baseline ANN provided a benchmark for purely data-driven learning, while the PINN-Lite demonstrated the benefits of introducing minimal physics-based constraints. The PINN + Melville Prior blended empirical theory with data, and the Grey-Box PINN advanced toward symbolic discovery of engineering-style equations. Finally, GEP offered interpretable analytical expressions, bridging empirical regression and machine learning approaches.

Table 2 Hyperparameters and settings of the developed ML models

Model / Parameter	Optimal Value
GEP	
Genes	6
Head size	15
Chromosomes	500
Linking function	Division
Set of functions	+, −, ×, ÷, sqrt, Exp, x^2, $\sqrt[3]{}$, $\sqrt[4]{}$, Sin
Constant per gene	10
Upper bound	10
Lower bound	−10
Data type	Floating-point
Mutation rate	0.00138
Permutation rate	0.00546
Inversion rate	0.00546
Random cloning	0.00102
IS transportation	0.00546
RIS transportation	0.00546
Gene transportation	0.00277
Recombination rate	0.00277
RNC mutation	0.00206
Baseline ANN	
Hidden layers	(64, 64)
Activation function	ReLU
Output layer	1 (linear)
Optimizer	Adam
Max iterations (epochs)	750
Loss function	Mean Squared Error (MSE)
Batch size	Default (sklearn)
Random state	SEED
PINN-Lite	
Architecture	Same as ANN (64–64–1)
Optimizer	Adam
Training iterations	750
Loss function	MSE + physics penalty terms
Physics penalties	Clear-water threshold, upper bound cap, monotonicity, Shields-type proxy
Batch size	Default
Random state	SEED
PINN + Melville Prior	
Architecture	Same as ANN (64–64–1)
Optimizer	Adam
Training iterations	750
Loss function	MSE + physics penalties + Melville prior
Prior form	Simplified Melville/Coleman relation
Physics penalties	Same as PINN-Lite
Batch size	Default
Random state	SEED
Grey-Box PINN	
Base structure	ANN + symbolic power-law prior

Table 2 (continued)

Model / Parameter	Optimal Value
Learnable prior	Exponents and scale of power-law
Optimizer	Adam
Training iterations	750
Loss function	MSE + physics penalties + prior regularization
Physics penalties	Clear-water threshold, upper bound, monotonicity, Shields-type proxy
Random state	SEED

2.5 Model's Performance Assessment

Evaluating the predictive capability of scour depth models requires robust statistical measures that capture both error magnitude and model fit. In this study, three widely used metrics are adopted: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). The RMSE provides an estimate of the average size of prediction errors, with larger deviations weighted more heavily, making it sensitive to significant mismatches between observed and predicted scour values. The MAE, in contrast, reports the mean of the absolute differences, offering an intuitive and less sensitive measure of error distribution. The R^2 value reflects the proportion of variance in the observed data that can be explained by the model; values approaching unity indicate a close alignment between predictions and actual measurements.

The mathematical forms of these metrics are expressed as in **Eqs. (3)-(5)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (ai - mi)^2}{\sum_{i=1}^{n} (ai - \overline{a}i)^2} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{n} |ai - mi|}{n} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (ai = mi)^2}{n}} \tag{5}$$

where a.i. and mi denote the observed and predicted scour depths, respectively, $\overline{a}i$ represents the mean of the observed values, and n is the total number of data points. Together, these three metrics provide a balanced assessment of prediction accuracy, capturing both average performance and the extent of variability in errors.

2.6 Models' Interpretation.

Beyond numerical accuracy, understanding why a model makes certain predictions is essential for engineering applications. To enhance interpretability, this study integrates several explainable AI (XAI) techniques, including SHAP, LIME, ICE plots, and PDP. SHAP quantifies the relative contribution of each input variable, providing a global view of feature importance across the dataset, while LIME explains individual predictions by locally approximating the model around specific instances. ICE plots capture how predictions for a single observation vary as one input changes, thereby revealing nonlinear effects at the instance level. PDP, in contrast, offers a broader perspective by illustrating the average influence of one or more features on predicted scour depth across all data points. Together, these methods provide both global and local interpretability, transforming complex machine learning into transparent frameworks. This dual perspective not only improves trust in the predictions but also ensures alignment with physical reasoning, making the models more reliable for decision-making in bridge scour assessment.

3 Results and Discussion

3.1 Performance Analysis of ML Models

The results in Table 3 reveal clear trade-offs between predictive accuracy and physical consistency among the developed models. The baseline ANN achieved the best fit on training data ($R^2 = 0.817$, RMSE = 0.206), but its reduced accuracy on validation and testing ($R^2 \approx 0.60$) suggests mild overfitting and weaker generalization. By contrast, the PINN variants yielded slightly lower predictive accuracy ($R^2 = 0.49$–0.56 on testing) but ensured predictions adhered to hydraulic principles such as clear-water thresholds and monotonic scour behavior. This highlights their strength in producing physically credible outputs, even if at the expense of raw statistical performance.

The GEP model showed the most robust generalization, outperforming all other approaches with $R^2 = 0.883$ and the lowest errors (RMSE = 0.162, MAE = 0.118) on the testing dataset. Its ability to produce explicit symbolic equations makes it not only highly accurate but also interpretable, offering practical utility for engineers. In summary, while ANN and GEP deliver higher predictive accuracy, PINNs provide valuable physical realism, making them complementary approaches depending on whether interpretability, accuracy, or physical consistency is prioritized.

Table 3 Performance metrics of models' predictions

Performance metrics	Baseline_ANN	PINN_Lite	PINN_Melville	PINN_Grey-Box	GEP
Training Dataset					
R^2	0.817	0.748	0.694	0.718	0.703
MAE	0.153	0.450	0.409	0.456	0.176
RMSE	0.206	0.501	0.461	0.509	0.258
Validation Dataset					
R^2	0.582	0.583	0.528	0.550	0.804
MAE	0.210	0.444	0.394	0.430	0.150
RMSE	0.304	0.520	0.473	0.513	0.215
Testing Dataset					
R^2	0.602	0.543	0.490	0.556	0.883
MAE	0.203	0.477	0.450	0.489	0.118
RMSE	0.288	0.543	0.510	0.554	0.162

3.2 Regression and Error Envelope Analysis of ML Models

Figure 4 compares the experimental and predicted normalized scour depths for GEP, baseline ANN, and the three PINN variants. The 1:1 line and ± 10% bounds serve as benchmarks to evaluate prediction bias and dispersion. The GEP model exhibits near-symmetric scatter around the 1:1 line, with minimal systematic bias. Slight under-prediction appears at higher y_s/b (> 2.5), likely reflecting data scarcity in extreme conditions, but the overall distribution within ± 10% bounds indicates excellent generalization and balanced behavior. The baseline ANN tends to overpredict at lower scour depths (< 1.0) and underpredict at higher values (> 2.0), suggesting sensitivity saturation, a common issue when the model lacks physical constraints. Among

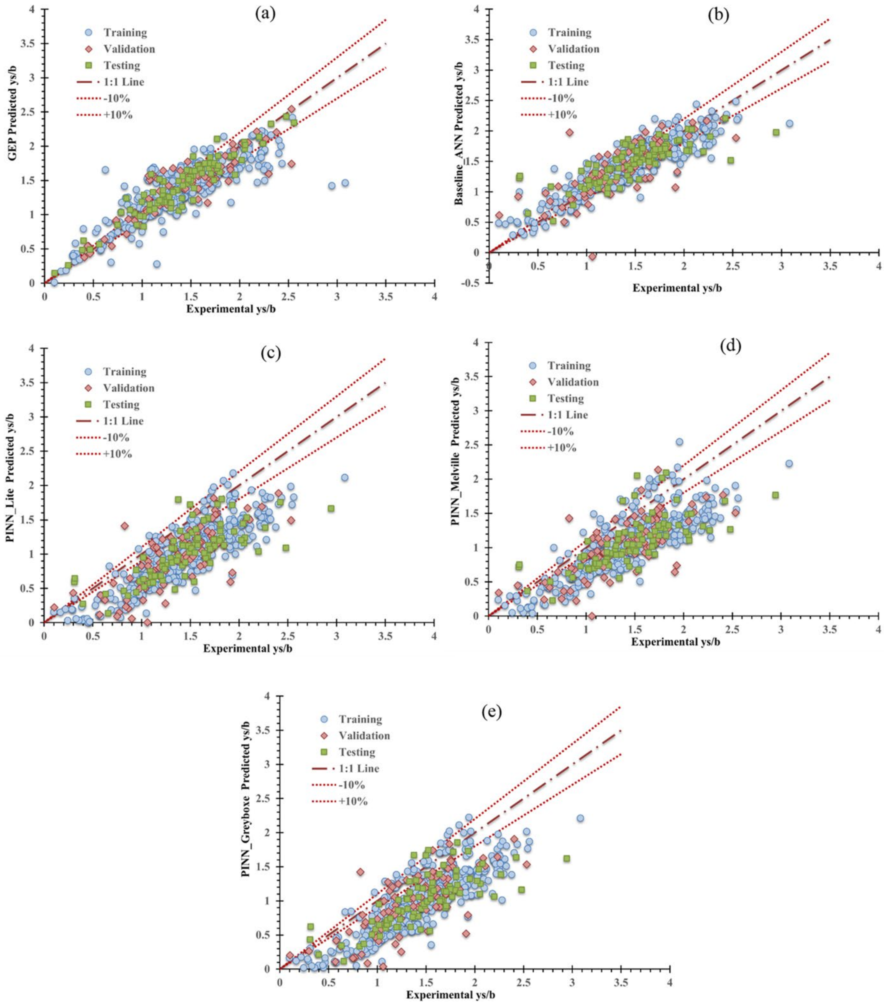


Fig. 4 Regression plot of models' predictions vs. experimental values of ys/b (**a**) GEP (**b**) Baseline_ANN (**c**) PINN_Lite (**d**) PINN_Mellville (**e**) PINN_Greybox

the physics-informed variants, PINN-Lite shows similar but slightly reduced dispersion, demonstrating improved consistency through partial physics regularization. PINN-Melville, however, displays a clearer underprediction trend across the range, indicating that embedding simplified empirical relationships constrained the model's flexibility, causing bias toward conservative estimates. The PINN-GreyBox performs most robustly among the PINNs, maintaining close alignment with the 1:1 line across all data splits, with minor underprediction only at very high scour depths. This balanced pattern highlights its ability to integrate data-driven learning with physical priors effectively. From an engineering perspective, mild underprediction implies potential risk in design applications (unsafe side), whereas overprediction leads to conservative but costlier designs. Therefore, the GEP and PINN-GreyBox models strike the most practical balance between predictive accuracy and safety, offering reliable scour estimates for design and risk evaluation.

3.3 Comprehensive Comparison of ML Models with Empirical Models

The comparison highlights the clear limitations of traditional empirical models and the advantages of ML approaches in predicting bridge pier scour depth. As shown in Table 4, empirical models such as HEC-18 (Arneson et al. 2013), Melville (2000), Wilson (1995), and Briaud (2015) are based on simplified assumptions of flow, sediment, and pier geometry. When evaluated on the present dataset, their predictive ability was weak, with R^2 values between 0.10 and 0.35 and high errors (e.g., RMSE = 1.06 for HEC-18 and 1.15 for Briaud). Even the better-performing Melville and Wilson models ($R^2 \approx 0.34$–0.35) showed limited generalization beyond laboratory conditions. By contrast, ML models delivered substantial improvements (Table 5). The GEP model achieved $R^2 = 0.745$ with RMSE = 0.164, while the baseline ANN reached $R^2 = 0.749$ with RMSE = 0.258. These values represent a marked reduction in error compared to empirical models, reflecting the ability of ML approaches to capture nonlinear interactions among multiple factors.

Moreover, interpretability, often a concern with ML, can be enhanced through XAI tools such as SHAP, LIME and PDP, which reveal how variables like velocity, sediment size, and pier width influence predictions. This balance of accuracy and interpretability underscores the superiority of ML over empirical formulas, making models such as GEP and ANN more reliable for practical bridge design and scour risk assessment.

3.4 Model's Interpretation

3.4.1 SHAP Interpretation

The SHAP analysis provides a global perspective on feature relevance and directional influence in predicting scour depth. The bar chart in Fig. 5 indicates that the sediment-related parameter b/D_{50} contributes substantially to the model's predictions, followed by V/V_c and F_r, while σ_g and y/b show comparatively lower contributions. This pattern suggests that the model attributes scour behavior primarily to the combined influence of hydraulic intensity and sediment characteristics. The SHAP

Table 4 Empirical models' equations used in comparison (Kim et al. 2024a, b)

Model	Reference	Equation	Parameters
HEC-18	(Arneson et al. 2013)	$\frac{y_s}{b} = 2K_1K_2K_3\left(\frac{y}{b}\right)^{0.35}{F_r}^{0.43}$	K_1 = pier nose shape factor K_2 = pier alignment factor K_3 = bed condition factor F_r = Froude number
Melville	(Melville 2000)	$y_s = K_{yb}K_IK_dK_sK_\theta\,K_G$	K_{yb} = pier depth size factor K_I = flow intensity factor K_d = sediment size factor K_s = pier nose shape factor K_θ = pier alignment factor K_G = channel geometry factor (= 1 for pier)
Wilson	(Wilson 1995)	$\frac{y_s}{b'} = 0.9\frac{y}{b'}$	b' = projected pier width
Briaud	(Briaud 2015)	$\frac{y_s}{b'} = 2.2K_{pw}K_{psh}K_{pa}K_{psp}\left(2.6F_{pier} - F_{c(pier)}\right)^{0.7}$	K_{pw} = water depth influence factor K_{psh} = pier shape influence factor $K_{p.a.}$ = aspect ratio influence factor K_{psp} = pier spacing influence factor F_{pier} = pier Froude number, $F_{c(pier)}$ = critical pier Froude number

Table 5 Predictive performance of empirical models in comparison with ML models

Model	Performance	
	R^2	RMSE
HEC-18	0.11	1.06
Melville	0.34	0.79
Wilson	0.35	0.73
Briaud	0.1	1.15
GEP (whole data)	0.745	0.164
ANN_Baseline (whole data)	0.749	0.258

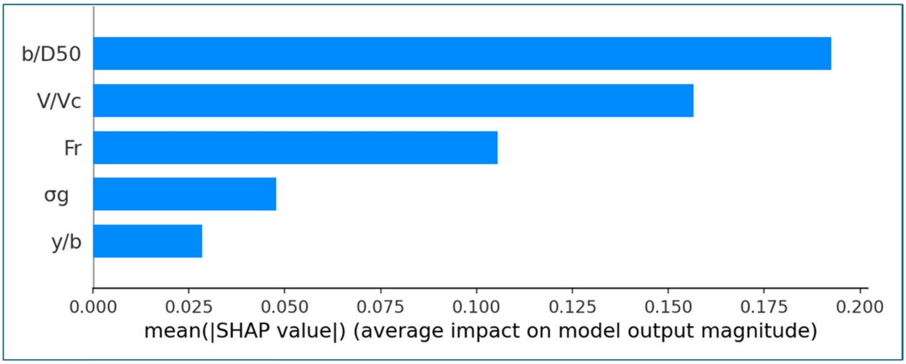


Fig. 5 SHAP global feature importance plot for testing dataset

beeswarm plot in Fig. 6 further illustrates the direction of influence for each input variable. Increasing values of V/V_c correspond to elevated predicted scour depths, consistent with the role of velocity ratio in driving sediment entrainment and bed degradation. In contrast, larger b/D_{50} values tend to reduce predicted scour, reflecting the stabilizing influence of coarser sediments. For F_r, higher values are generally associated with reduced scour predictions, which may indicate that, beyond certain flow regimes, additional turbulence exerts a limited incremental effect. The influence of σ_g suggests that higher gradation levels correspond to reduced scour, implying enhanced resistance provided by non-uniform sediments. Variations in y/b exhibit limited directional impact, suggesting a relatively stable role across the examined range. The SHAP interaction heatmap in Fig. 7 highlights pairwise dependencies among features that influence model outputs. A pronounced interaction between V/V_c and b/D_{50} indicates that the influence of flow intensity on predicted scour is modulated by sediment size relative to pier width. Specifically, elevated velocities lead to increased scour, but this response is dampened when coarser sediments dominate. Additionally, the interaction between F_r and V/Vc implies that turbulence may enhance velocity-driven scour under certain hydraulic conditions. Interactions involving σ_g and y/b appear less prominent but still indicate that gradation and flow depth contribute to secondary modulation of scour responses. Overall, the SHAP analysis reveals that model predictions reflect a dynamic interplay between hydraulic loading (V/V_c, F_r) and sediment scaling (b/D_{50}, σ_g), with geometric effects (y/b) contributing more subtly. These insights suggest that scour development is shaped

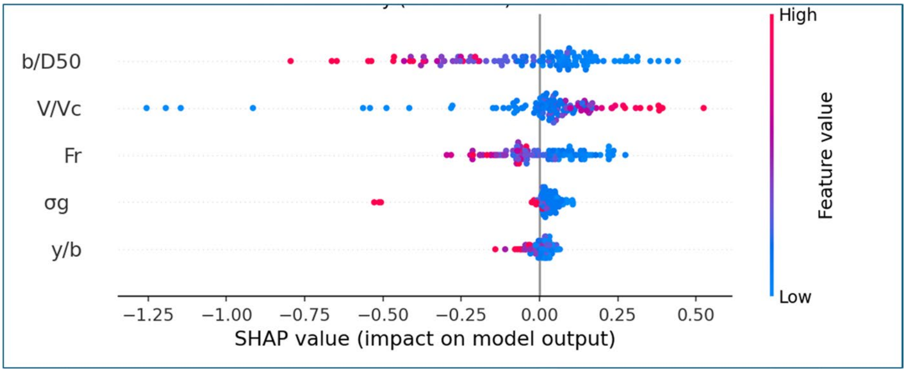


Fig. 6 SHAP summary beeswarm plot for the testing dataset

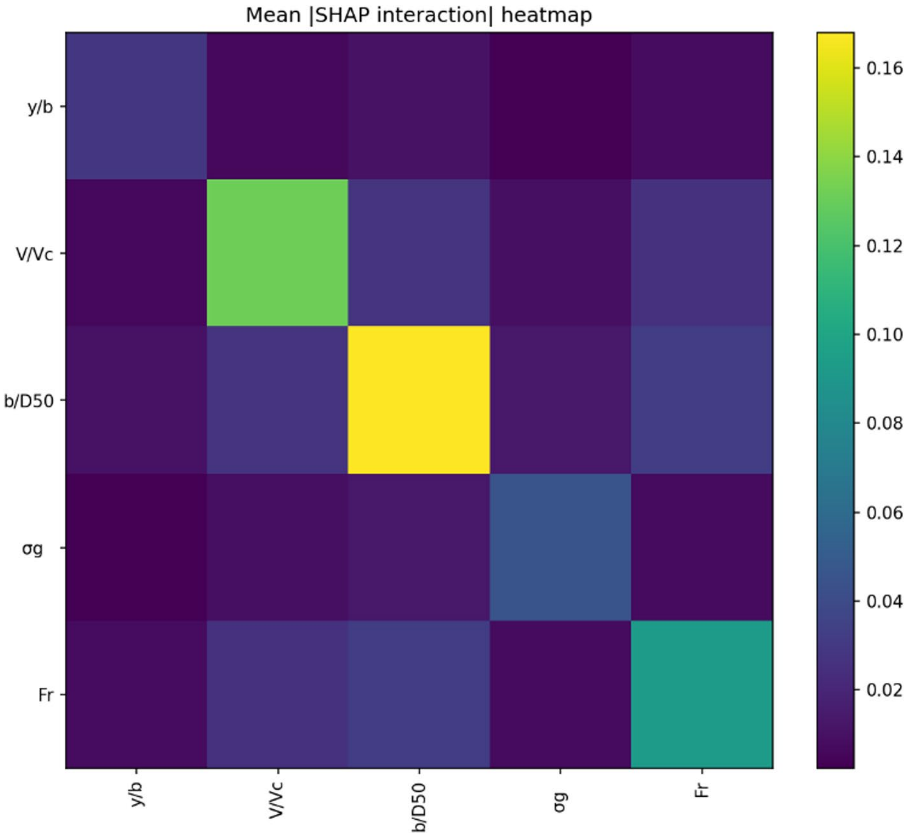


Fig. 7 SHAP interaction heatmap

by the joint action of flow intensity and sediment resistance rather than by individual parameters in isolation.

3.4.2 LIME Interpretation

The LIME analysis provides localized explanations of individual predictions, showing how different input conditions influence scour depth estimates. As some sample cases are summarized in Table 6, the velocity ratio (V/V_c) consistently emerges as the dominant factor, driving sharp increases in scour near threshold conditions and producing the highest predictions under extreme flows. Sediment effects moderated these responses: finer sediments amplified scour, while coarser sediments reduced it through armoring. Gradation (σ_g) further supported scour reduction by introducing stabilizing non-uniformity, though with a modest influence compared to V/V_c. The Froude number reinforced hydraulic intensity under low-flow conditions but became less influential at higher turbulence levels. In contrast, the geometric factor y/b had minimal effect across cases. Overall, LIME confirms that scour depth is locally governed by hydraulic drivers, moderated by sediment characteristics, with geometry playing only a secondary role. These localized insights complement the global explanations from SHAP and PDP.

3.4.3 ICE and PDP Interpretation

The PDP plots in Fig. 8 identify the velocity ratio (V/V_c) as the most influential factor, with scour depth rising sharply once the critical threshold is exceeded and then stabilizing at higher values. ICE plots in Fig. 9 confirm this global trend but reveal case-specific variability: some curves show abrupt initiation near $V/V_c \approx 1$, while

Table 6 Local feature contributions from LIME for selected cases

Case	Predicted y_s/b	Dominant feature	Contribution (weight)	Secondary features	Remarks
1	1.84	V/V_c (0.33)	Strong positive	b/D_{50} (0.11), F_r (0.11)	Threshold exceedance drives scour
2	1.31	V/V_c (0.35)	Strong positive	b/D_{50} (0.13), F_r (0.11)	Moderate velocity, sediment coarseness reduces scour
3	1.57	V/V_c (0.34)	Positive	F_r (0.12), b/D_{50} (0.11)	Interaction of flow intensity and sediment
4	1.77	V/V_c (0.22)	Moderate positive	F_r (0.10), b/D_{50} (0.09)	High velocity ratio but sediment armoring mitigates
5	1.58	V/V_c (0.34)	Positive	b/D_{50} (0.12), F_r (0.10)	Consistent with global trend

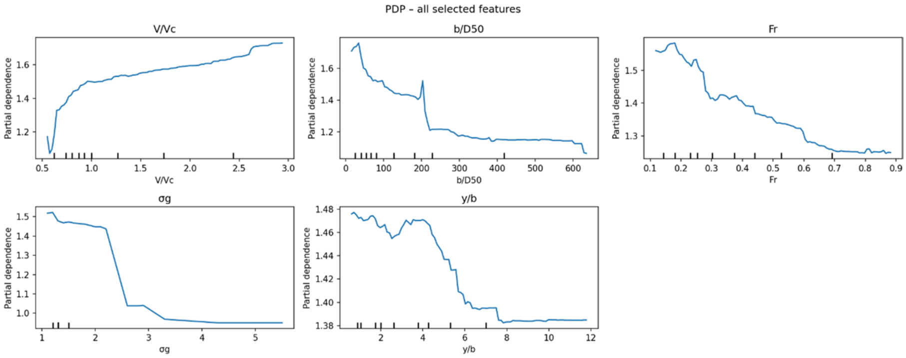


Fig. 8 PDP interpretation plots for all parameters

others display smoother increases, reflecting interactions with sediment and geometric conditions. For the relative pier width (b/D_{50}), both PDP and ICE demonstrate a consistent negative effect, with scour decreasing as sediment size increases relative to pier width. The trend levels off beyond $b/D_{50} \approx 400$, indicating the onset of sediment armoring, a finding consistent with earlier experimental studies. The Froude number (F_r) exhibits an opposite behavior, where scour depth decreases with increasing F_r, although ICE curves reveal variability in the rate of decline across different cases. The gradation coefficient (σ_g) further highlights the stabilizing role of sediment non-uniformity, with scour markedly reduced once $\sigma_g > 2$, after which responses flatten; ICE plots suggest that this effect is stronger under certain hydraulic settings. The relative flow depth (y/b) emerges as the weakest predictor, producing only minor decreases in scour depth, with largely flat ICE curves confirming its limited role.

Together, PDP captures the dominant global effects while ICE plots highlight localized variations, reinforcing that scour is primarily controlled by hydraulic intensity (V/V_c, F_r) and sediment properties (b/D_{50}, σ_g), whereas geometric scaling (y/b) plays only a secondary role.

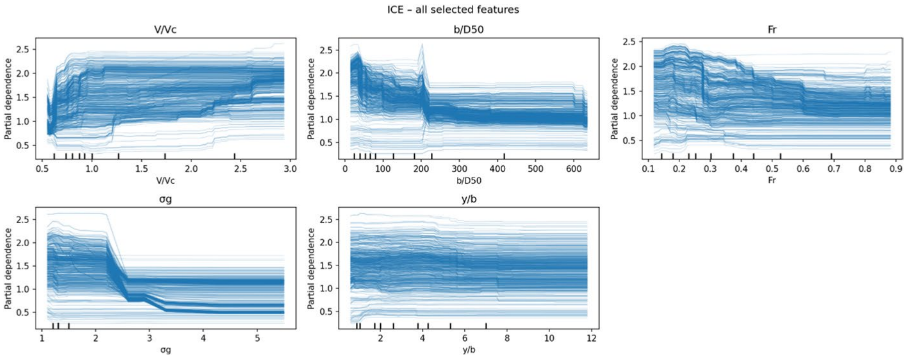


Fig. 9 ICE interpretation plots for all parameters

3.4.4 Additional Interpretation Plots

In addition to SHAP, PDP, ICE, and LIME, supplementary visualizations were employed to further validate interpretability findings. The Sobol sensitivity bar chart in Fig. 10 compares first-order (S1) and total-order (ST) indices. The ST indices rank features in decreasing order (b/D_{50}, V/V_c, σ_g, F_r, y/b), confirming the dominance of sediment scaling and hydraulic intensity. Notably, the S1 contributions for σ_g and F_r diminish almost to zero, indicating that their effects on scour predictions arise mainly through interactions with other variables rather than in isolation. In contrast, b/D_{50} and V/V_c show both direct (S1) and interaction-driven (ST) effects, while y/b retains only a minor influence in both cases.

The permutation importance plot in Fig. 11 quantifies the drop in model performance when each feature is randomly shuffled. Consistent with SHAP results, b/D_{50} and V/V_c emerge as the most critical predictors, while σ_g and F_r contribute moderate importance, and y/b remains the least influential. Unlike SHAP, which explains directionality, permutation importance provides a robustness check by directly measuring predictive reliance on each feature. The alignment between the two strengthens confidence in the feature hierarchy identified across interpretability methods.

The partial residual linear surrogate plots (Fig. 12) offer further validation by approximating the black-box model with a simple linear surrogate and then examining feature contributions. These plots display the residual-adjusted relationships between each input variable and predicted scour depth. A strong linear trend in the partial residuals indicates that the surrogate adequately captures the feature's effect, whereas curvature or scatter reveals non-linearities or interactions. For V/V_c, the partial residual plot shows a steep linear increase near the threshold (≈1.0), consistent with the dominant role of hydraulic intensity, though deviations at higher values indi-

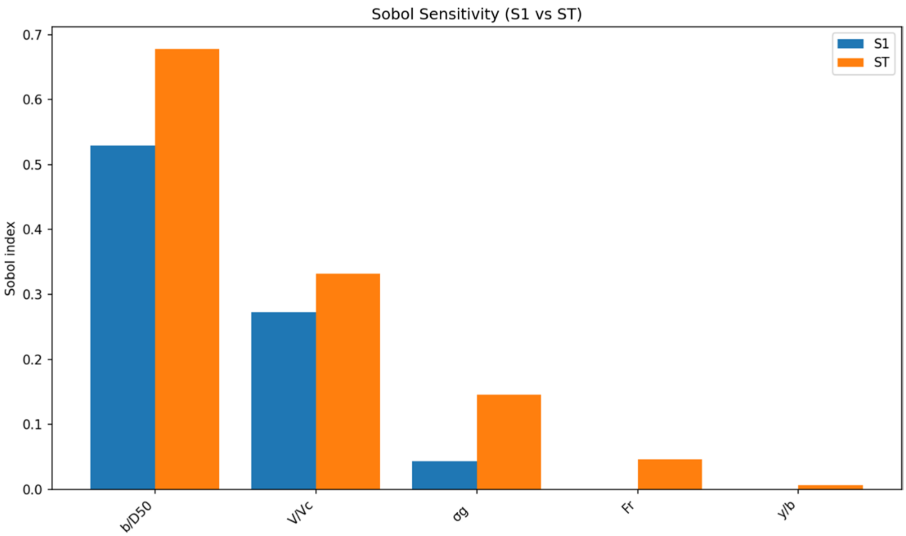


Fig. 10 Sobol sensitivity indexes for all input parameters

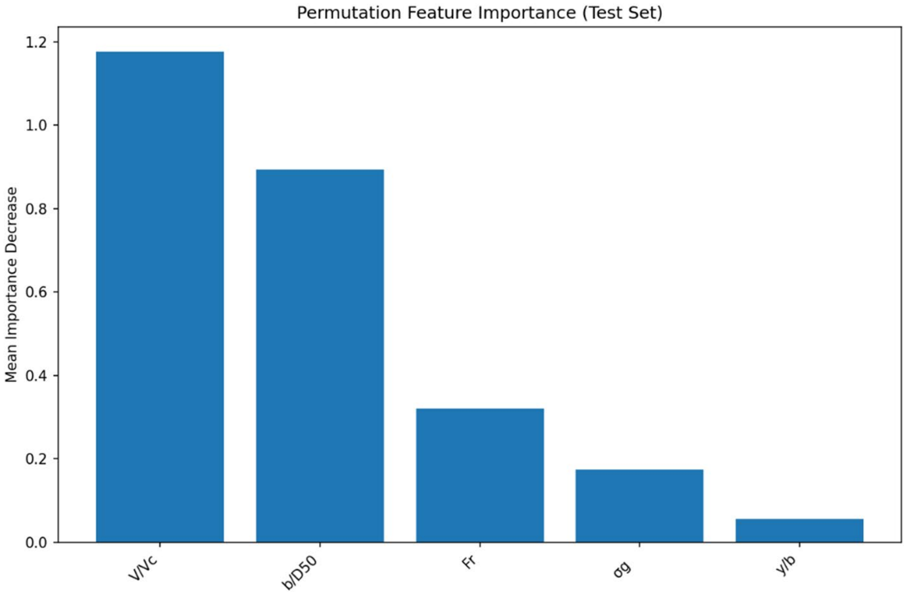


Fig. 11 Permutation feature importance plot

cate non-linear acceleration captured by the original model. For b/D_{50}, a clear negative linear slope is observed. σ_g exhibits a modest but stabilizing effect, with partial residuals curving downward once $\sigma_g > 2$, confirming the non-linear moderation of scour. F_r shows weaker and more scattered patterns, reinforcing that its influence is primarily interaction-driven rather than direct. Finally, y/b produces nearly flat residuals, further confirming its minimal role.

Finally, the Morris sensitivity scatter plot (Fig. 13) provides a complementary perspective by displaying the mean absolute effect (μ^*) against the standard deviation (σ) for each feature. Features with high μ^* indicate strong overall influence, while high σ values signal non-linear effects or strong interactions. The plot positions b/D_{50} and V/V_c in the upper-right quadrant, confirming them as dominant drivers with both strong effects and interaction tendencies. σ_g and F_r show moderate μ^* but elevated σ, reflecting their interaction-driven contributions. y/b again lies near the origin, confirming its negligible role.

Together, these supplementary plots complement XAI-based interpretations and reinforce the conclusion that scour depth is primarily governed by the interplay of hydraulic intensity and sediment scaling, with geometric effects playing a minor role.

4 Conclusions

This study demonstrated the potential of ML approaches, particularly GEP and PINN, to improve bridge pier scour prediction. Among all models, GEP achieved the highest predictive accuracy of $R^2 = 0.88$, RMSE = 0.16 on the testing dataset, showing excellent generalization and balanced predictions across the data range. The baseline ANN

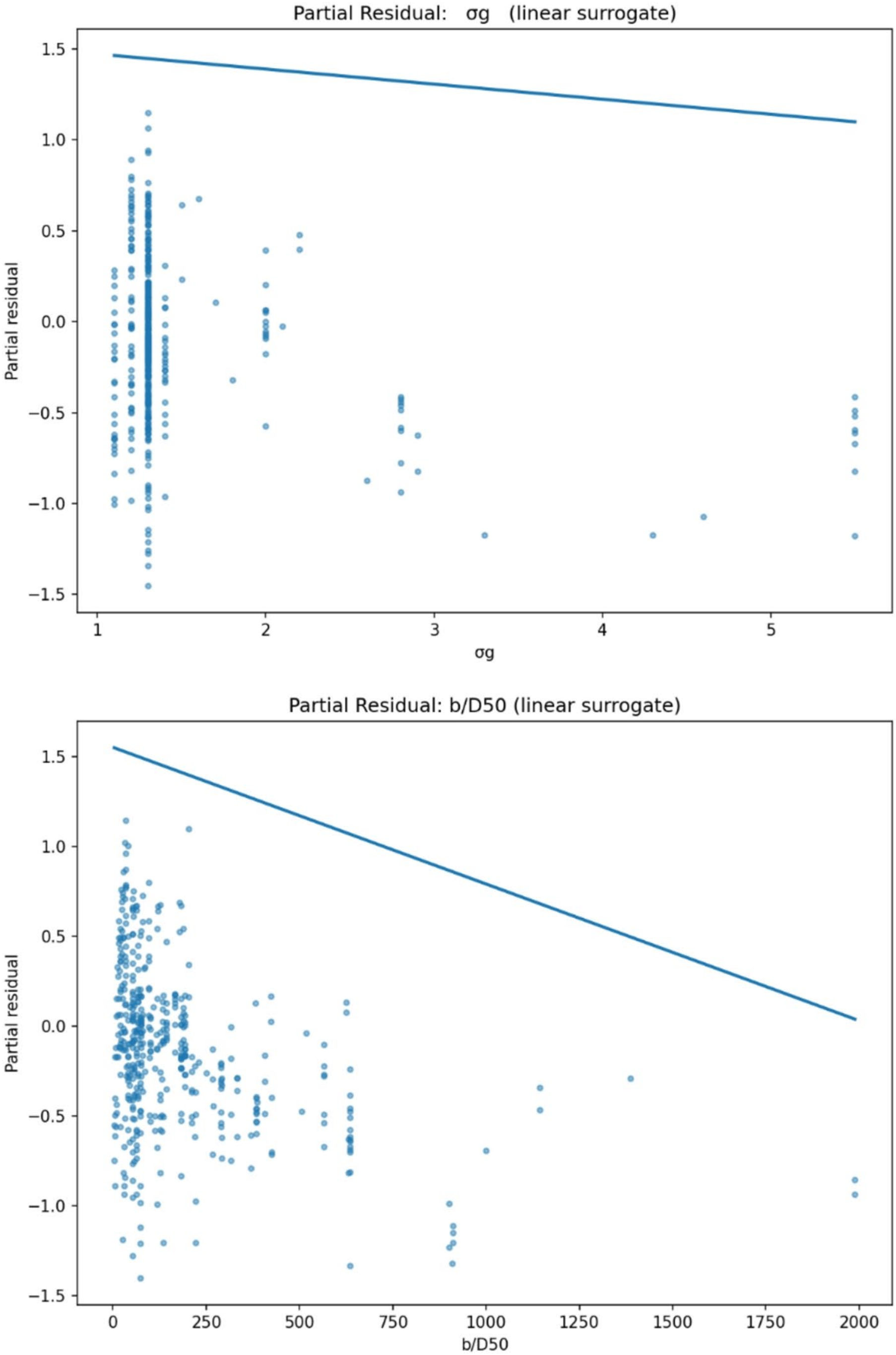
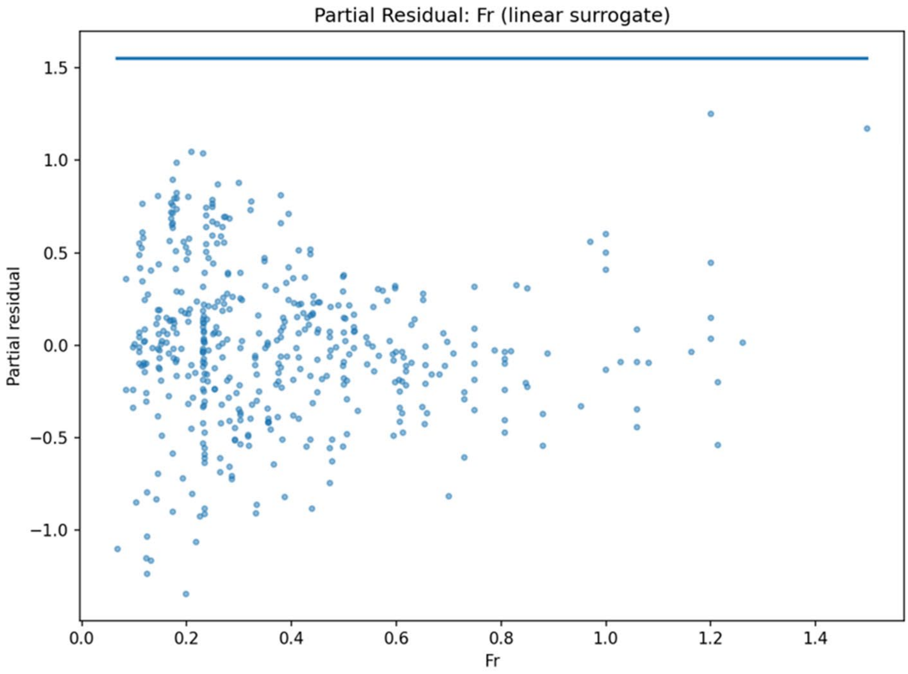


Fig. 12 Partial residual plots for all input features



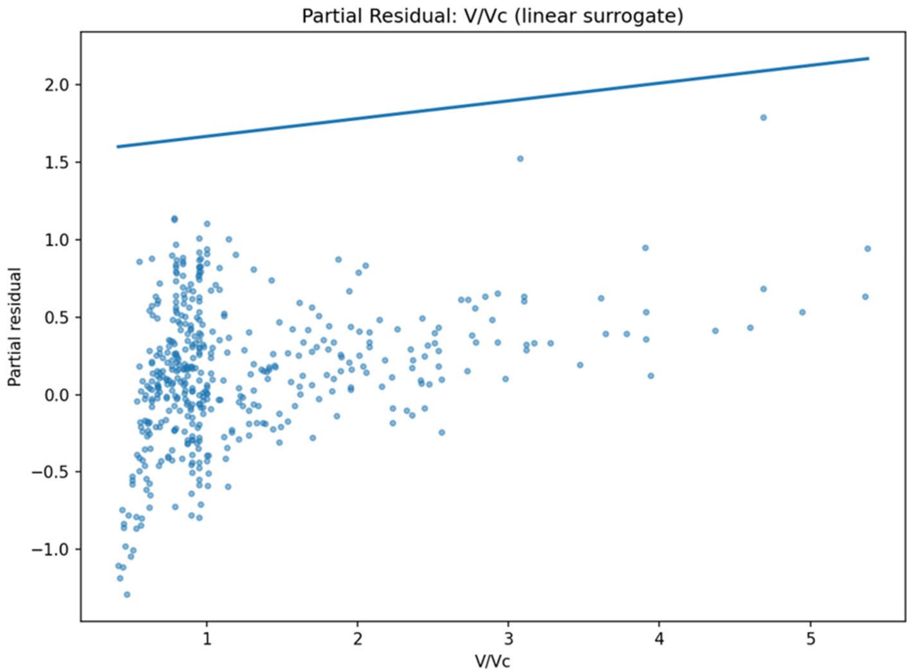


Fig. 12 (continued)

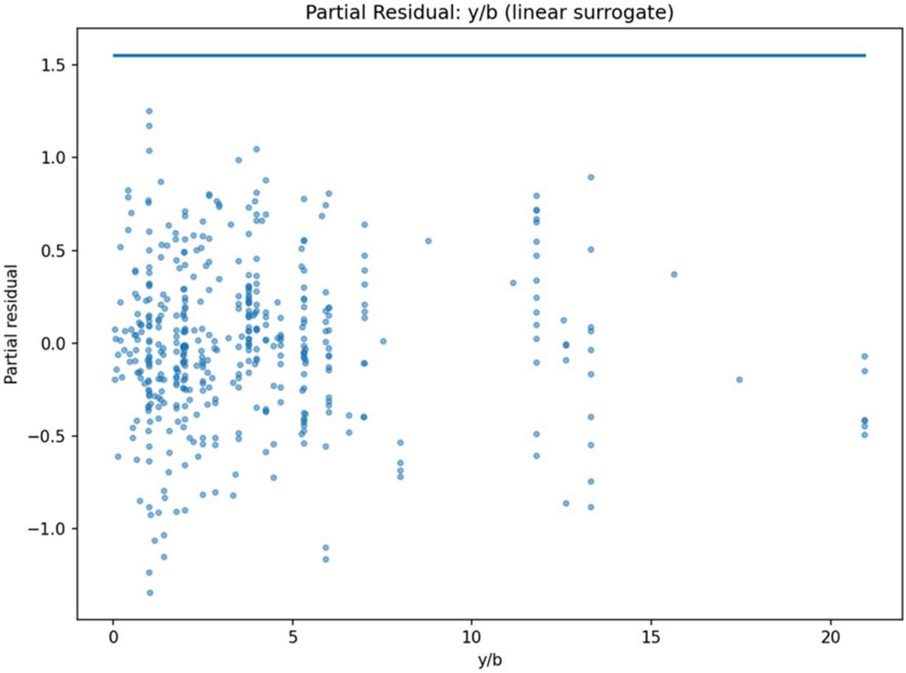


Fig. 12 (continued)

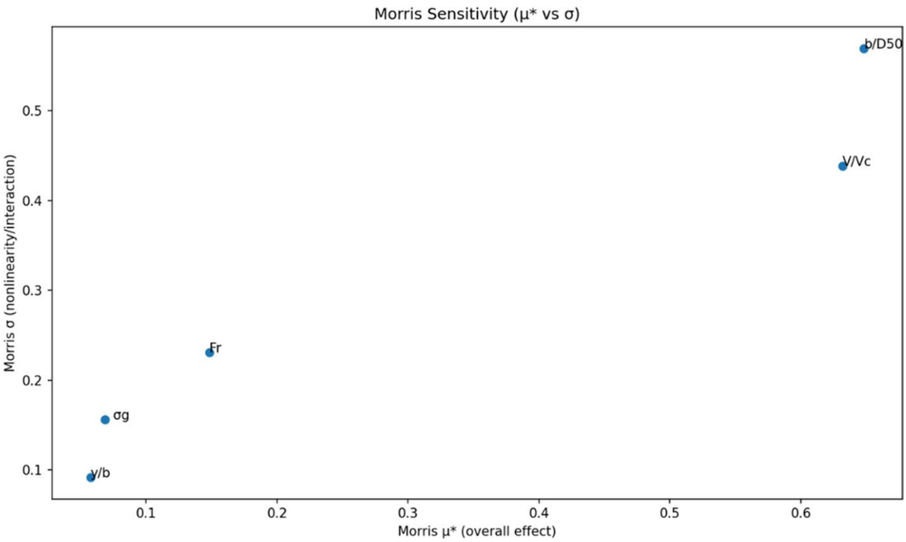


Fig. 13 Morris sensitivity scatter plot for all features

performed moderately ($R^2 = 0.60$), revealing mild overfitting, whereas the PINN variants produced slightly lower statistical accuracy ($R^2 = 0.49$–0.56) but ensured physically consistent and hydraulically credible predictions. Among them, PINN-GreyBox achieved the most balanced performance, PINN-Lite reduced computational cost while retaining partial physical regularization, and PINN-Melville tended toward conservative underprediction due to its reliance on empirical constraints. Equally significant is the role of explainable AI in enhancing model transparency. SHAP confirmed that scour is predominantly governed by the velocity ratio (V/V_c) and sediment scaling (b/D_{50}), with F_r and σ_g exerting secondary influences, and y/b playing a minor role. Interaction effects between V/V_c and b/D_{50} underscored the balance between hydraulic loading and sediment armoring. PDP and ICE plots captured global and local nonlinear trends, while LIME provided case-specific insights into how different input conditions shaped predictions. Limitations of this work include reliance on a laboratory-scale dataset, which may restrict the model's generalizability under prototype or field-scale conditions. The PINN models, while physically interpretable, exhibited reduced numerical accuracy, and GEP, despite its strong performance, may struggle with extrapolation beyond the trained parameter space. Future research should aim to integrate hybrid GEP-PINN architectures, incorporate larger, field-calibrated datasets, and explore physics-guided explainable AI frameworks to bridge accuracy, interpretability, and scalability. Overall, the synergy between GEP's predictive precision, PINNs' physical realism, and XAI's interpretability represents a promising direction for developing reliable, transparent, and practically applicable scour prediction tools in hydraulic engineering.

Author Contributions L.G: Conceptualisation, methodology design, and original draft preparation. M.K: Data collection, software setup, and validation. H.U: Writing, data analysis, and visualisation. A.A: Literature review, editing, and analysis.

Funding 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.

Data Availability Data will be made available upon request by the corresponding author.

Declarations

Ethical Approval This manuscript has not been published in whole or in part elsewhere and is not currently being considered for publication in another journal.

Informed Consent Not applicable.

Competing interests The authors declare no competing interests.

References

Arneson, L.A., Zevenbergen, L.W., Lagasse, P.F., CLopper, P.E.: Evaluating Scour at bridges, United States. Federal Highway Administration. Office of Bridge Technology; National Highway Institute (U.S.). United States. Federal Highway Administration (2013)

Azamathulla, H.M., Ghani, A.A., Zakaria, N.A., Guven, A.: Genetic programming to predict Bridge pier scour. J. Hydraul Eng. **136**, 165–169 (2010)

Benedict, S.T., Caldwell, A.W.: Prepared in Cooperation with the South Carolina department of transportation A Pier-Scour database: 2,427 field and Laboratory Measurements of Pier Scour Data Series 845 (2014)

Briaud, J.-L.: Scour depth at bridges: Method including soil Properties. I: Maximum scour depth prediction. J. Geotech. Geoenviron Eng. **141**, 4014104 (2015). https://doi.org/10.1061/(ASCE)GT.1943-5606.0001222

Briaud, J.-L., Ting, F.C.K., Chen, H.C., Gudavalli, R., Perugu, S.: Prediction of scour rate in cohesive soils at Bridge piers. J. Geotech. Geoenvironmental Eng. **125**, 237–246 (1999). https://doi.org/10.1061/(ASCE)1090-0241(1999)125:4(237)

Ettema, R., Constantinescu, G., Melville, B.: Evaluation of bridge scour research: Pier scour processes and predictions. National Cooperative Highway Research Program (NCHRP) Report 682, Transportation Research Board, Washington, D.C.; (2011)

Fathi, A., Zomorodian, S.M.A., Zolghadr, M., Chadee, A., Chiew, Y.-M., Kumar, B., et al.: Combination of riprap and submerged Vane as an abutment scour countermeasure. Fluids. **8**, 41 (2023)

Ferreira, C.: Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Springer, Berlin (2006)

Fuladipanah, M., Hazi, M.A., Kisi, O.: An in-depth comparative analysis of data-driven and classic regression models for scour depth prediction around cylindrical Bridge piers. Appl. Water Sci. **13**, 231 (2023a)

Fuladipanah, M., Azamathulla, H.M., Tota-Maharaj, K., Mandala, V., Chadee, A.: Precise forecasting of scour depth downstream of flip bucket spillway through data-driven models. Results Eng. **20**, 101604 (2023b)

Khan, M., Azamathulla, H.M., Tufail, M.: Gene-expression programming to predict pier scour depth using laboratory data. Journal of Hydroinformatics **14**, 628–645 (2012)

Kim, T., Shahriar, A.R., Lee, W.-D., Gabr, M.A.: Interpretable machine learning scheme for predicting Bridge pier scour depth. Comput. Geotech. **170**, 106302 (2024a)

Kim, T., Shahriar, A.R., Lee, W.D., Gabr, M.A.: Interpretable machine learning scheme for predicting Bridge pier scour depth. Comput. Geotech. **170**, 106302 (2024b). https://doi.org/10.1016/j.compgeo.2024.106302

Kwak, K.S., Briaud, J.L.: Case study: An analysis of pier scour using the SRICOS method. KSCE Journal of Civil Engineering **6**, 243–253 (2002)

Lawal, Z.K., Yassin, H., Lai, D.T.C., Che Idris, A.: Physics-informed neural network (PINN) evolution and beyond: A systematic literature review and bibliometric analysis. Big Data Cogn. Comput. **6**, 140 (2022)

Lin, C., Han, J., Bennett, C., Parsons, R.L.: Case history analysis of Bridge failures due to scour. Clim. Eff. Pavement Geotech. Infrastruct., pp. 204–216. (2014)

Melville, B.W.: Bridge Scour, vol. 112. Water Resources (2000)

Molavi, A., Kaleybar, F.A., Rathnayake, N., Rathnayake, U., Fuladipanah, M., Azamathulla, H.M.: Machine learning approaches for simulating Temporal changes in bed profiles around cylindrical Bridge pier: A comparative analysis. Hydrology. **12**, 238 (2025)

Omara, H., Abdeelaal, G.M., Nadaoka, K., Tawfik, A.: Developing empirical formulas for assessing the scour of vertical and inclined piers. Marine Georesources & Geotechnology **38**, 133–143 (2020)

Parsaie, A., Haghiabi, A.H., Moradinejad, A.: Prediction of scour depth below river pipeline using support vector machine. KSCE Journal of Civil Engineering **23**, 2503–2513 (2019)

Qaderi, K., Javadi, F., Madadi, M.R., Ahmadi, M.M.: A comparative study of solo and hybrid data driven models for predicting Bridge pier scour depth. Marine Georesources & Geotechnology **39**, 589–599 (2021)

Raeisi, F., Zomorodian, S.M.A., Zolghadr, M., Azamathulla, H.M.: Sacrificial piles as a countermeasure against local scour around underwater pipelines. Water Science and Engineering **17**, 187–196 (2024)

Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. data Min., pp. 1135–44. (2016)

Richardson, J.E., Panchang, V.G.: Three-dimensional simulation of scour-inducing flow at Bridge piers. J. Hydraul Eng. **124**, 530–540 (1998)

Schuring, J.R., Dresnack, R., Golub, E.: Design and Evaluation of Scour for Bridges Using HEC-18, U.S. Department of Transportation. Federal Highway Administration, Washington, D.C. (2017)

Shahriar, A.R., Ortiz, A.C., Montoya, B.M., Gabr, M.A.: Bridge pier scour: An overview of factors affecting the phenomenon and comparative evaluation of selected models. Transp. Geotech. **28**, 100549 (2021). https://doi.org/10.1016/j.trgeo.2021.100549

Van Wilson, K.: Scour at Selected Bridge Sites in Mississippi, vol. 94. US Department of the Interior, US Geological Survey (1995)

Wang, T., Reiffsteck, P., Chevalier, C., Chen, C.-W., Schmidt, F.: An interpretable model for Bridge scour risk assessment using explainable artificial intelligence and engineers' expertise. Structure and Infrastructure Engineering **21**, 643–655 (2025)

Authors and Affiliations

Laiba Gulaly[1] · Majid Khan[2] · Husna Usman[1] · Aaila Asghar[1]

✉ Laiba Gulaly
21pwciv5593@uetpeshawar.edu.pk

✉ Majid Khan
majkhan@siue.edu

Husna Usman
21pwciv5607@uetpeshawar.edu.pk

Aaila Asghar
21pwciv5614@uetpeshawar.edu.pk

[1] University of Engineering and Technology, Peshawar, Pakistan

[2] Department of Civil Engineering, Southern Illinois University Edwardsville, EdwardsvilleIL 62026, United States