Research on Engineering Structures & Materials
www.jresm.org
MIM Research Group

Research Article

Explainable machine learning for predicting cost and schedule overruns in irrigation projects: A safety-driven approach

Abbas Ali [*,a], Hatim Abd Al-Karim [b]

Department of Civil Engineering, College of Engineering, Al-Nahrain University, Baghdad, Iraq

Article Info

Article History:

Received 24 May 2026

Accepted 16 July 2026

Keywords:

Machine learning;
Irrigation projects;
Safety indicators;
Cost overrun;
Time delay;
SHAP

Abstract

Irrigation projects are essential for ensuring water security and sustainable agricultural development; however, they are highly susceptible to human, managerial, equipment, and environmental risks that contribute to cost overruns and schedule delays. Although machine learning (ML) has been widely applied to project performance prediction, its integration with validated safety indicators and explainable artificial intelligence for irrigation projects remains largely unexplored. Therefore, this study proposes an explainable machine learning framework for predicting cost overruns and schedule delays using safety indicators. A comprehensive dataset comprising 50 irrigation projects implemented in Iraq between 2022 and 2025 was developed. Twenty-three safety indicators were identified through a systematic literature review, refined through expert consultation, and validated using Partial Least Squares Structural Equation Modeling (PLS-SEM). Four ML algorithms, namely Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost), were developed and evaluated using 5-fold cross-validation. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The cross-validation results demonstrated that Random Forest achieved the most robust predictive performance for both cost overruns (R^2 = 0.912, RMSE = 1.685) and schedule delays (R^2 = 0.894, RMSE = 3.236). Furthermore, SHAP analysis identified human-related factors as the primary drivers of cost overruns and managerial factors as the dominant contributors to schedule delays. The proposed framework provides an interpretable decision-support tool for proactive safety management, early risk identification, and improved planning of irrigation projects.

1. Introduction

Irrigation projects are a vital component of infrastructure development, and water resource management, which plays a crucial role in food security and economic stability. These projects are complex, requiring multi-disciplinary coordination, significant use of heavy equipment, exposure to the environment, and considerable human interaction. Worldwide, irrigation projects often suffer significant delays and cost overruns, reducing their economic feasibility and use, and their development potential [1]. There is growing evidence that the lack of safety management is an indirect but very significant reason for these overruns and that operational risk and project performance, as measured by overruns, are closely connected [2]. Safety in irrigation projects has shifted from a reactive approach to accident prevention to being a proactive measure against project performance, based on the planning, coordination, communication, and control of its operations. Safety issues in irrigation projects can lead to work stoppages, equipment downtime,

*Corresponding author: st.abbas.ali@ced.nahrainuniv.edu.iq
[a]orcid.org/0009-0005-8520-6896; [b]orcid.org/0000-0002-2399-0735
DOI: http://dx.doi.org/10.17515/resm2026-1704ml0524rs

rework, loss of productivity and indirect losses that may be even more severe. The combination of these disruptions means that the customer experiences schedule slippage, leading to cost escalation and setting up a cause-and-effect relationship, such that the impact of safety incidents becomes indirect on the bottom line [3]. For this reason, time overruns are often precursors to cost overruns, and both serve as systemic indicators of inadequate safety and risk management systems. These risks are exacerbated in an irrigation project's operating context. Projects are often situated in remote regions, or sensitive to the environment, and exposed to weather variability, soil conditions, logistical constraints as well as equipment reliability, adding to the uncertainty. The interaction of human factors (i.e., lack of risk awareness, lack of communication, lack of safety training, etc.) and machine and environmental hazards heightens vulnerabilities to delays [4,5]. These multifactorial influences show that it is not enough to identify causes of delay; the interaction of these factors in the system within the project ecosystem should be understood and quantified. In the project management context, one of the key success factors is cost and schedule performance. Cost overruns are typically the amount by which the costs exceed the contract amount, while time overruns are the time that is gone over the planned time [6, 7]. Research has shown that the majority of projects in the global infrastructure sector have a planned budget to be exceeded by 30-40% due to several factors, including safety inefficiencies [2]. When safety measures are not met, it can interrupt the continuity of the work, lead to more corrective actions, and have cascading effects on coordinating stakeholders and on the overall project efficiency [8,9]. The predictive issues of cost and time overruns are met with recent developments in machine learning (ML) providing a powerful solution framework. In contrast to traditional statistical methods, ML algorithms can model complex, non-linear relationships in high dimensional data. This capability is especially relevant to irrigation projects, as there are safety indicators that have dynamic interactions with each other, such as training adequacy, maintenance practices, communication efficiency, procedural compliance, environmental exposure, and reporting near-miss [10]. The ML models can recognize hidden patterns in these variables, which can enable early identification of risk trajectories, that cannot be identified by traditional forecasting methods [11]. In recent years, several studies have been conducted to show the effectiveness of using ML algorithms to predict the outcome of an irrigation project. However, hybrid and ensemble models have been found to be more accurate than the single-method ones, and provide better early-warning capabilities for the schedule delays and cost escalation [12, 13]. Although machine learning has been successfully applied to predict cost and schedule performance in general construction projects, such use of ML for safety-driven performance prediction in irrigation projects is still under-explored because irrigation infrastructure is characterized by unique operational conditions, including extensive environmental exposure, remote locations, hydraulic structures, deep excavations, seasonal construction constraints, and water-related safety hazards. Consequently, there remains a critical need for an explainable machine learning framework specifically designed for irrigation projects using validated safety indicators. Irrigation infrastructure, with its components exposed to the environment, requiring equipment intensive operations, and distributed project locations, demands unique predictive frameworks, which incorporate safety management variables and performance forecasting. Existing research highlights the importance of data quality, feature engineering and regular model validation in improving the accuracy of the model, along with the choice of the algorithm [14] . Despite the important progress achieved in applying machine learning techniques for predicting projects performance, existing studies have mainly focused on general construction datasets and traditional project management factors. In addition, previous research rarely addressed irrigation projects, despite their unique operational characteristics, including environmental exposure, far locations, equipment concentrated activities, and complex stakeholder coordination. so, there remains a research gap in developing integrated predictive frameworks that combine safety indicators with advanced machine learning techniques for predicting cost and time overruns in irrigation projects. subsequently, the main aim of this study is to develop and evaluate machine learning predictive models for predicting cost and time overruns in irrigation projects using safety indicators. the study also aims to identify the most influential safety – related factors affecting irrigation projects through SHAP – based interpretability analysis, thereby supporting proactive decision – making and improving project planning and safety management practices.

2. Literature Review

2.1. Time Overrun Studies in Infrastructure Projects

Usually, time overrun is defined as exceeding the actual time limit or agreed delivery time of the project. In concept, it is the gap between time to completion and actual time to completion as shown in Eq (1), which is typically due to inefficiencies in the execution process in activities [15]. Delays are usually caused by various interruptions of a project's activities that occur at specific events, like weather variability, resource constraints, or design changes, and can be quantified in a measurable schedule overrun in a connected project system. Empirical evidence suggests that nearly 70% of projects exceed their planned schedules, with average delays ranging from 10% to 30% of the original project duration [16]. Previous studies have identified unexpected site conditions, financing delays, poor contract management, material delivery delays, and inadequate planning as major contributors to schedule overruns [17,18]. Similarly, contractor-related problems, construction errors, financial constraints, and communication deficiencies have also been recognized as significant causes of schedule slippage [19]. Collectively, these findings indicate that time overruns are systemic problems arising from planning deficiencies, operational uncertainty, and organizational inefficiencies, underscoring their persistent global challenge in construction project management.

$$Time\ overrun\ ratio\ = \frac{Actual\ schedule\ \ - Planned\ schedule}{Planned\ schedule} \tag{1}$$

2.2. Cost Overruns in Infrastructure Projects

Cost overruns are among the most persistent and extensively investigated challenges in project and infrastructure management because of their significant economic and organizational consequences [20]. In general, a cost overrun refers to the difference between the planned (or approved) project cost and the actual cost incurred upon project completion, as expressed in Eq. (2). The terms cost escalation and budget overrun are often used interchangeably to describe financial deviations that occur when initial planning assumptions fail to reflect actual project conditions (21-22). Empirical studies have found that project stakeholders are frequently unaware of the relationship between managerial decisions, operational actions, governance and final project costs [23]. Cost overruns are commonly attributed to deficiencies in project planning, technical uncertainties, risk-related changes, inadequate contingencies, and execution-stage disruptions. Previous studies have also identified material shortages, financing constraints, poor contract management, price fluctuations, procurement problems, design modifications, and organizational inefficiencies as major contributors to cost escalation [24], [25], [26]. Additionally, ineffective supervision, weak management practices, and financial constraints interact to adversely affect overall project cost performance [27]. The literature as a whole reinforces the fact that cost overruns are not just financial surprises, but the result of interrelated managerial, technical and environmental risks and risks and the need for a holistic, risk-informed cost management framework.

$$Cost\ overrun\ ratio\ = \frac{actual\ budget\ \ - Planned\ budget}{Planned\ budget} \tag{2}$$

Although much research has examined the reasons behind and implications of cost and time overruns in infrastructure projects, predicting these variations is difficult because of the interaction of management, engineering, environmental, and safety factors. The development of machine learning (ML) has provided advanced and, in many cases, more powerful predictive capabilities. More recently, explainable artificial intelligence and shapley additive explanations (SHAP) have made great improvements in the transparency and interpretability of predictive ML models. Table 1 summarizes previous studies that have applied machine learning techniques to predict project performance and highlights the remaining research gaps addressed in the present study.

As summarized in Table 1. previous studies have demonstrated the effectiveness of machine learning techniques in predicting project cost overruns and schedule delays across various construction sectors. Nevertheless, several important research gaps remain. First, the majority of existing studies focused on general construction projects, such as buildings, highways, marine, and residential projects, whereas irrigation projects possess unique operational characteristics, including extensive environmental exposure, hydraulic structures, remote locations, seasonal constraints, and complex excavation activities. Second, most studies emphasized prediction accuracy without incorporating validated safety-related indicators as the primary predictive variables. Third, although a limited number of studies employed explainable artificial intelligence through SHAP, model interpretability remains largely overlooked, thereby limiting the practical adoption of machine learning models in project management. Finally, no previous study has integrated validated safety indicators, explainable machine learning, and SHAP within a unified framework for simultaneously predicting cost overruns and schedule delays in irrigation projects. To address these limitations, this study proposes an explainable machine learning framework that integrates validated safety indicators with advanced predictive models and SHAP-based interpretation to predict both cost overruns and schedule delays in irrigation projects. The proposed framework combines safety indicators and machine learning for predictive modeling, and SHAP for model interpretability, thereby providing a transparent and practical decision-support tool tailored to the unique characteristics of irrigation projects.

Table 1. Comparative Review of Previous Machine Learning Studies for Cost and Schedule Overrun Prediction and Identified Research Gaps

Study	Project Type	ML Model	Explainability (SHAP)	Main Contribution	Research Gap
[28]	Construction projects	ML algorithms	✗	Compared ML algorithms for delay prediction	Focused only on delay prediction without explainability or safety indicators.
[29]	Construction projects	ANN, SVM	✗	Predicted construction cost overruns	Did not incorporate safety-related variables or model interpretability.
[30]	Construction projects	SVM	✗	Accurate cost overrun prediction	Used historical project variables only.
[31]	Construction projects	XGBoost	✓	Cost prediction with SHAP interpretation	No safety-driven framework; limited to general construction.
[32]	Marine construction	GRNN, SVM	✗	Time-delay prediction	Focused on delay prediction without explainability.
[33]	Highway projects	ANN	✗	Highway cost prediction	Infrastructure-specific but excluded safety indicators.
[34]	Construction projects	NGBoost, ANN, LR	✓	Cost prediction with SHAP	Focused on construction costs rather than integrated cost–schedule prediction.
[35]	Public buildings	RF, DT, XGBoost, KNN	✗	Time prediction	Did not analyze model interpretability or safety variables.
[36]	Public buildings	RF, KNN, NN	✗	Duration prediction	General construction context only.
[37]	Residential projects	RF, RANSAC	✗	Cost and schedule prediction	No explainable AI and no safety-oriented framework.
[38]	public projects	ML algorithms (CatBoost)	✗	Cost overrun prediction	Large dataset but excluded explainability and validated safety indicators.

3. Research Methodology

This research adopted a quantitative predictive research design to investigate the relationships between validated safety performance indicators and project performance, represented by cost overruns and time delays in irrigation projects. The methodological framework, illustrated in Fig. 1, integrates empirical field data, structured questionnaire-based measurements, and historical project records with advanced machine learning techniques. This integrated approach ensures methodological rigor, predictive reliability, and robust evaluation of the proposed predictive models.

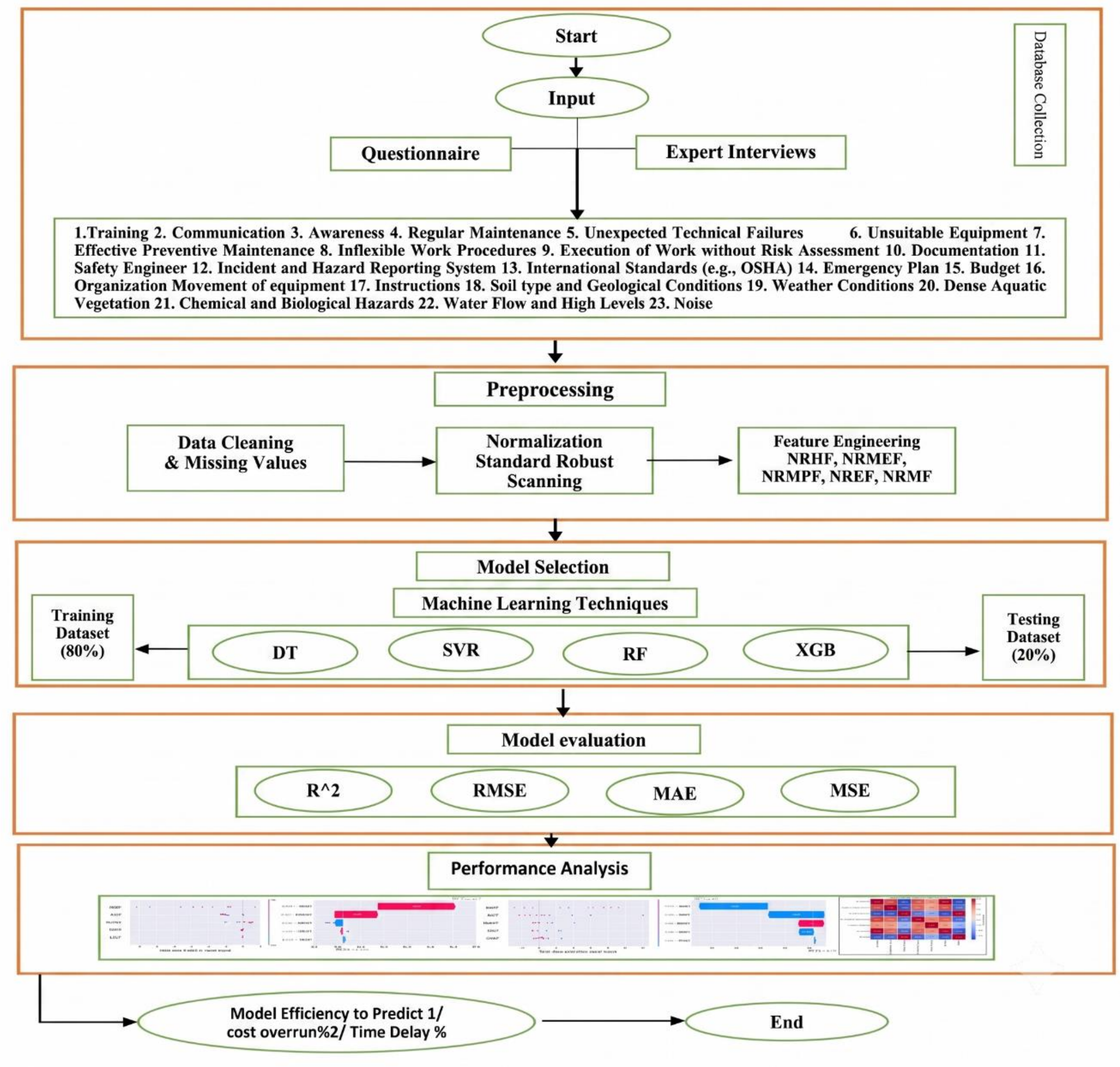

Fig. 1. Flowchart of the Research

3.1. Research Design

The overall research design is quantitative and analytical, designed to identify systematic and measurable relationships between safety performance indicators and project performance outcomes that are generalizable. In particular, the study follows the parameters of the effect that the safety variables have on the cost and schedule deviations encountered in the project. The framework integrates three principal components:

- Structured collection of archival data from completed irrigation projects
- Questionnaire-based measurement of safety performance indicators
- Machine learning–based predictive modeling and comparative evaluation

Python 3.10 was used as the main computation environment, while the use of Scikit-learn, Pandas, NumPy, Seaborn, Jupyter Notebook, and Matplotlib libraries was used for data preprocessing, machine learning model development, statistical analysis, and visualization. This methodological approach is integrated and increases the analytical strength, as it draws on empirical project data and sophisticated predictive modelling methods. It was decided to use the quantitative approach because it would allow for objective measurement, statistical reliability and predictive generalization across irrigation project settings.

3.2. Data Collection and Dataset Format Development

Historical data from 50 irrigation projects implemented in Iraq between 2022 and 2025 were collected and organized into a standardized electronic dataset. The dataset included planned and actual project costs and durations, along with 23 safety-related indicators identified through a comprehensive literature review, refined through expert consultation, and measured using a five-point Likert scale questionnaire. The indicators were validated using Partial Least Squares Structural Equation Modeling (PLS-SEM), including reliability and validity assessments (Previous chapter), before being used as input variables for the machine learning models. The dataset was organized in Microsoft Excel (xlsx) format to facilitate preprocessing and predictive modeling using Python.

3.3. Data Preprocessing and Preparation

One of the most important steps for guaranteeing the reliability of the analysis and stability of the model is the data preprocessing. The following sequence of procedures were followed:

3.3.1 Data Cleaning

Data cleaning was performed to ensure the quality, consistency, and reliability of the dataset prior to model development. The dataset was screened for missing values, outliers, and inconsistencies. Missing values were examined using the Pandas isnull () function, and no missing observations were identified; therefore, no imputation procedures were required. Potential outliers were assessed using the Interquartile Range (IQR) method. A small number of potential outliers were detected in the Human Factors (2 observations), Management Factors (1 observation), and Cost Overrun (5 observations) variables, whereas no outliers were identified in the remaining variables. These observations were verified against the original project records and confirmed to represent genuine project conditions rather than data entry or measurement errors. Therefore, all observations were retained for subsequent analysis to preserve the representativeness of the real-world irrigation project dataset.[39].

3.3.2 Scaling and Normalization

Since safety indicators were measured on different scales, they were scaled using the Min-Max normalization [40]. The change is given as Eq (3):

$$X' = \frac{X - min\ (X)}{max\ (X) - min\ (X)} \tag{3}$$

where, X'= normalized value: X= original value: min $(X)$and max (X)= minimum and maximum values in the dataset

This normalization procedure improves convergence efficiency and enhances model performance stability [41].

3.3.3 Train–Test Splitting

The dataset was split into training and testing sets to promote generalization of the model. Various splitting ratios (6:4, 7:3, 8:2, 9:1) were tested. The ratio of 8:2 (80% training, 20% testing) proved to be the best one and was used.

3.3.4 *Hyperparameter Tuning Using GridSearchCV*

Machine learning models are highly sensitive to hyperparameter selection, which directly influences prediction accuracy, model complexity, and generalization capability. Therefore, hyperparameter tuning was performed prior to model development to identify the optimal parameter combinations for each algorithm and ensure a fair comparison among the developed models. In this study, GridSearchCV was employed to systematically evaluate multiple hyperparameter combinations for the Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) models. The optimal hyperparameters were subsequently adopted to train the final predictive models. The optimal parameter settings are presented in Table 2.

Table 2. Optimal hyperparameters obtained using GridSearchCV

Model	Parameters	Interpretation
DT	max_depth=7	The decision tree was effectively pruned, reducing the risk of overfitting compared with an unpruned tree.
RF	n_estimators=200, max_features=sqrt	Increasing the number of trees to 200 and using the square root of the available features at each split improved model stability and enhanced its generalization capability.
SVR	C=100, gamma=0.1, epsilon=0.01	SVR required relatively strong hyperparameter settings to achieve optimal performance, indicating its high sensitivity to hyperparameter selection.
XGBoost	learning_rate=0.05, max_depth=3	The model selected a shallow tree depth and a low learning rate, which help reduce overfitting and improve generalization performance.

3.3.5 K-Fold Cross-Validation

To enhance the robustness and generalization capability of the developed machine learning models, K-Fold Cross-Validation was employed during the model evaluation process. Cross-validation is considered an effective technique for assessing predictive performance, particularly when the available dataset is relatively limited. In this study, the dataset was divided into K = 5 equal folds. During each iteration, four folds (80% of the data) were used for training, while the remaining fold (20%) was used for validation. This process was repeated five times, ensuring that each fold served once as a validation set. The final model performance was obtained by averaging the evaluation metrics across all folds, including the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The application of K-Fold Cross-Validation reduces the risk of overfitting, improves the reliability of performance estimates, and provides a more comprehensive assessment of model stability compared with a single train-test split. Therefore, it was adopted as an additional validation procedure for comparing the predictive performance of the SVR, DT, RF, and XGBoost models.

3.4. Data Analysis and Predictive Modeling

Historical data and safety metrics from various projects were combined to form an analytical data set and analyzed with the help of Python based machine learning techniques. The process of modeling involved training, validation and performance evaluation of the model to predict cost overrun and time delays. The proposed framework highlights the effectiveness of the indicators of safe water as predictive variables for taking proactive decisions in irrigation project management, which are based on data.

4. Algorithm Evaluation Techniques

For machine learning modeling, the dataset is often split into a training set and a test set for assessing the model's generalizability and mitigating the risk of overfitting [30,31]. The training set is employed to develop and calibrate the models, and the testing set is used to evaluate the models' predictive ability on data not included in the training set. Since this paper involves a regression problem, a number of statistics were used to measure the accuracy of the prediction such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The mean squared error measures the mean of the squares of the errors, while the mean absolute error measures the mean of the absolute errors. RMSE is an interpretable error measure, which is the same unit as the dependent variable. R^2 represents the percentage of the variance in the model that is explained.

5. Machine Learning Algorithms

The study transformed the cost overrun and time delay prediction problem into a regression task and tested four machine learning algorithms: SVR, DT, RF and XGBoost. The models were realized in Python to assure methodological diversity and a reliable comparative predictive analysis as follows:

5.1. Support Vector Regression (SVR)

Considering the capability of handling high dimensional and noisy data with good generalization performance and modelling nonlinear relationships, Support Vector Regression (SVR) was chosen. The model is especially appropriate for the current research as safety indicators have complex interaction with cost overruns and time delays [33,34]. Moreover, SVR can work well in the limited project data conditions by taking advantage of the kernel functions, especially Radial Basis Function (RBF) [40]. The predictive function is the ability to predict the consequences of a situation.

5.2. Decision Tree (DT)

The choice of Decision Tree (DT) was made because it was able to show nonlinear and hierarchical relationships between safety indicators and project performance [45]. The model is also applicable to the existing data set because it is capable of detecting important thresholds and interactions of safety with cost overruns and time delays [46]. DT also offers an intuitive analysis structure that can be useful for comprehending how the 23 safety performance indicators affect project outcomes. The Sum of Squared Errors (SSE) is used to determine splits in regression trees.

5.3. Random Forest (RF)

Random Forest (RF) was selected due to its strong predictive performance and robustness against overfitting in complex regression problems [47] . The model is suitable for the current dataset because it effectively captures nonlinear relationships and multidimensional interactions among safety indicators through ensemble learning [38, 39]. Furthermore, RF enhances prediction stability and generalization capability under limited and noisy project data conditions. For regression tasks, the final prediction is obtained by averaging individual tree predictions.

5.4. Extreme Gradient Boosting (XGBoost)

XGBoost prefer using because it is both fast and accurate, and is capable of modelling complex nonlinear relationships with high quality [50]. This model is especially relevant for the existing data set as it is able to capture subtle interactions between safety indicators. in addition to, it provides boosting and regularization methods which improve prediction performance and stop overfitting when the amount of project data is limited.

6. SHAP algorithm

A comprehensive framework for interpreting Multilevel Models (MLMs) was known as SHAP Analysis. It relies on Shapley Additive exPlanations, as a single framework for model interpretation. Shapely values are employed to showcase the relative influence and contribution of individual input variables to the final output variable [40, 41]. This concept shares similarities with parametric analysis, where all variables except one are held constant, and the impact of the variable in question on the target attribute is observed. In the SHAP methodology, the model generates predicted outcomes for each anticipated sample, and the SHAP value indicates the value assigned to each feature in the given sample [42, 43]in Eq (4).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \tag{4}$$

where $\phi_i$represents the Shapley value of feature i, $F$is the set of all input features, $S$is a subset of features excluding i, and $f(\cdot)$denotes the prediction function of the machine learning model. The SHAP value quantifies the contribution of each feature to the final prediction.

7. Results and Discussion

7.1. Comparison of Prediction Accuracy

To assess the predictive capability of the developed model, the performance of random forest (RF), support vector regression (SVR), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) algorithms (DT) algorithms was compared. Table 3 presents the evaluation metrics for the test and training dataset, specifically the coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and mean absolute error (MAE), which are the key performance metrics for a typical cost overrun and time delay evaluation.

Table 3. Comparison for both cost overruns and time delays in testing and training data

Algorithms	Cost Overrun							
	R^2		MSE		RMSE		MAE	
	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	0.996	-0.348	0.164	9.544	0.405	3.089	0.245	1.823
Random Forest	0.972	0.695	1.192	2.160	1.092	1.470	0.680	0.885
XGBoost	0.9600	0.779	0.001	1.565	0.028	1.251	0.018	0.962
SVR	0.977	0.707	0.966	2.076	0.983	1.441	0.499	1.081
Algorithms	Time Delay							
	R^2		MSE		RMSE		MAE	
	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	0.998	0.822	0.402	16.982	0.634	4.121	0.374	3.589
Random Forest	0.987	0.876	2.436	11.827	1.561	3.439	1.156	2.603
XGBoost	0.970	0.839	0.001	15.359	0.038	3.919	0.017	3.102
SVR	0.985	0.916	2.757	8.064	1.660	2.840	0.815	2.115

Among the evaluated models, the Decision Tree exhibited a clear overfitting problem in cost overrun prediction, achieving an excellent training performance (R^2 = 0.996) but a poor testing performance (R^2 = −0.348). This substantial performance gap indicates that the model memorized the training data rather than learning generalizable patterns, which is a common limitation of single decision trees when applied to relatively small datasets. In contrast, XGBoost achieved the highest testing accuracy for cost overrun prediction (R^2 = 0.779), demonstrating superior generalization capability. This performance can be attributed to its boosting strategy and built-in regularization mechanisms, which effectively reduce model complexity and improve predictive robustness. For time delay prediction, SVR achieved the highest testing performance (R^2 = 0.916), indicating its ability to capture the nonlinear relationships between safety indicators and schedule deviations under the adopted train–test split

7.2. Fold Cross-Validation Performance Evaluation

To comprehensively assess the predictive accuracy and generalization capability of the developed machine learning models, K-fold cross-validation procedure was employed.

Table 4. Five-Fold Cross-Validation results for cost overrun and time delay prediction models

Model	Cost overrun %				Time Delay%			
	Mean R^2	Mean RMSE	Mean MAE	Std R^2	Mean R^2	Mean RMSE	Mean MAE	Std R^2
SVR	-0.145	5.626	4.383	0.286	0.024	10.798	8.596	0.189
DT	0.893	1.758	0.955	0.109	0.869	3.504	1.815	0.106
RF	0.912	1.685	1.027	0.075	0.894	3.236	2.115	0.086
XGBoost	0.825	2.060	1.045	0.098	0.876	3.417	1.862	0.090

The predictive performance of Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) was evaluated using the average coefficient of determination (Mean R^2), root mean square error (Mean RMSE), mean absolute error (Mean MAE), and the standard deviation of R^2 (Std R^2) across the five folds. The results, summarized in Table 4, provide a robust comparison of model accuracy, prediction error, and stability for both cost overrun and time delay prediction.

7.3. Comparison Between Train–Test Split and 5-Fold Cross-Validation

To further evaluate the robustness of the developed models, the results obtained using the conventional 80:20 train–test split were compared with those from 5-fold cross-validation, as presented in Table 3. While the train–test split reflects model performance for a single random partition of the dataset, cross-validation provides a more reliable estimate of predictive performance by averaging the results across multiple independent folds, particularly for relatively small datasets. For cost overrun prediction, XGBoost achieved the highest performance under the 80:20 train–test split (R^2 = 0.779), whereas Random Forest outperformed all other models during 5-fold cross-validation, achieving the highest average R^2 (0.912) together with the lowest RMSE (1.685) and the smallest variability among folds (Std R^2 = 0.075). This improvement indicates that Random Forest provides superior generalization capability owing to its ensemble learning strategy, which effectively reduces model variance and minimizes overfitting. The Decision Tree model showed relatively consistent performance between the two evaluation approaches. Although it achieved high prediction accuracy under cross-validation (R^2 = 0.893), its performance remained slightly lower than Random Forest because a single decision tree is inherently more sensitive to variations in the training data, increasing the likelihood of overfitting when the available dataset is limited. For time delay prediction, Support Vector Regression (SVR) produced the highest prediction accuracy using the 80:20 train–test split (R^2 = 0.916). However, its average cross-validation performance declined substantially to R^2 = 0.024, indicating that the excellent result obtained under the single train–test split was highly dependent on the selected random partition and was not consistently reproducible. This behavior suggests limited generalization capability and highlights the sensitivity of SVR to small datasets and hyperparameter selection. In contrast, random forest and XGBoost maintained consistently high predictive performance under both evaluation approaches. For time delay prediction, Random Forest achieved the highest cross-validation accuracy (R^2 = 0.894), followed closely by XGBoost (R^2 = 0.876). The relatively small differences between their train–test and cross-validation performances indicate greater model stability and robustness compared with SVR. Generally, although XGBoost and SVR achieved the highest prediction accuracy under the single 80:20 train–test split for cost overruns and time delays, respectively, the 5-fold cross-validation results demonstrate that Random Forest is the most reliable and robust model for both prediction tasks. The consistency of its predictive accuracy, combined with the lowest prediction errors and variability across folds, indicates superior generalization performance and supports its practical application for predicting cost overruns and schedule delays in irrigation projects.

7.4. Correlation Matrix Analysis

The correlation matrix among the study variables is presented at Fig .2., which shows the relation between the factors related to safety and the performance measures of the projects such as cost overrun and time delay in irrigation projects. Pearson correlation coefficient was used to assess the relationships; the closer the values are to +1, the stronger the positive relationship will be. Based on the results, there were moderate to strong correlations between most of the variables indicating that safety-related factors within irrigation project environments are interconnected. The primary relationships are as follows:

- The correlation that was found to be the strongest was between the Machine and Equipment Factors and Time Delay, with the correlation coefficient being 0.81, which shows that equipment failures, maintenance and machinery inefficiency are major contributors to delays on projects.
- A similar coefficient was found between Human Factors and Methods and Procedures Factors (0.80), indicating that the efficiency of operational procedures and implementation practices is closely related to workforce performance.

- The Machine and Equipment Factors also had a high correlation with Cost Overrun (0.77) which means that issues related to equipment significantly contribute to project cost overrun.
- Moreover, Human Factors was correlated highly with Machine and Equipment Factors (0.73), showing the relationship between the capability of the human resource and the effectiveness of the operation of machinery and equipment in irrigation projects.
- Moderate correlations were shown between Management Factors and Time Delay (0.63), and also between Management Factors and Machine and Equipment Factors (0.67) indicating that managerial practices have significant influence on controlling the project performance.
- Environmental Factors showed comparatively moderate relationships with the other variables, with the highest correlation observed with Machine and Equipment Factors (0.67), indicating that environmental conditions may influence equipment efficiency and field operations.

Overall, the results indicate that human factors, machinery and equipment, and operational procedures are the most influential factors affecting time delays and cost overruns in irrigation projects.

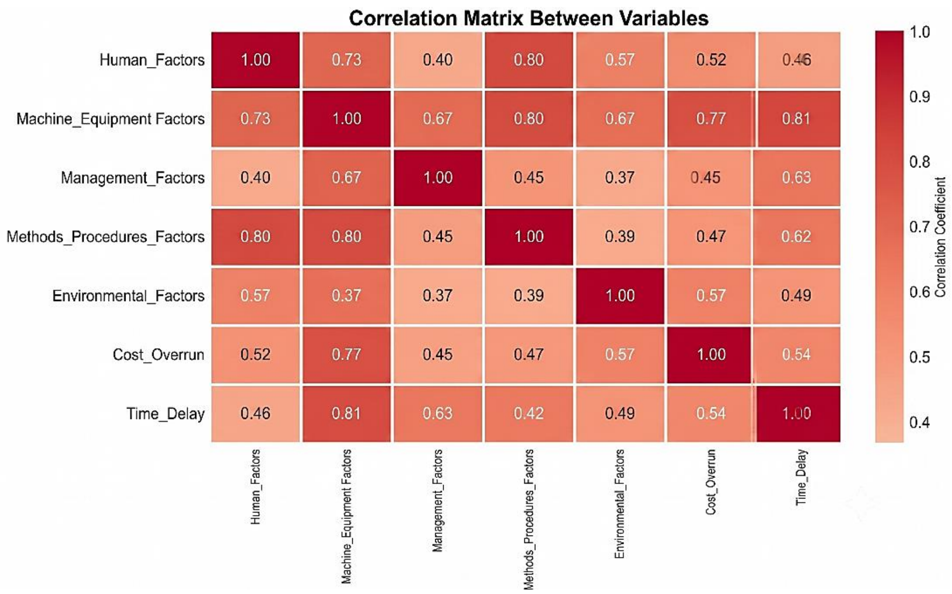

Fig. 2. Correlation matrix analysis

7.5. SHAP Values Analysis

7.5.1 SHAP Values Analysis for Cost Overrun

According to Fig .3 SHAP Summary (Bee Warm) results, NRHF (Normalized Human Factors) was identified as the most influential variable affecting cost overrun prediction. Higher values of human-related problems generally produced positive SHAP values, indicating an increase in predicted cost overruns, while improved workforce skills, training, and safety awareness contributed to reducing cost deviations. NRMF (Normalized Management Factors) also showed a strong positive influence, where weak planning, ineffective supervision, and poor coordination increased the likelihood of cost escalation. In contrast, NRMEF (Normalized Machine and Equipment Factors) demonstrated a moderate positive effect on cost overruns, indicating that equipment failures, insufficient maintenance, and machinery inefficiency contributed to increased project costs. NRMP (Normalized Methods and Procedures) and NREF (Normalized Environmental Factors) had comparatively smaller impacts with SHAP values clustered around zero, indicating that their influence on cost overrun prediction was comparatively limited.

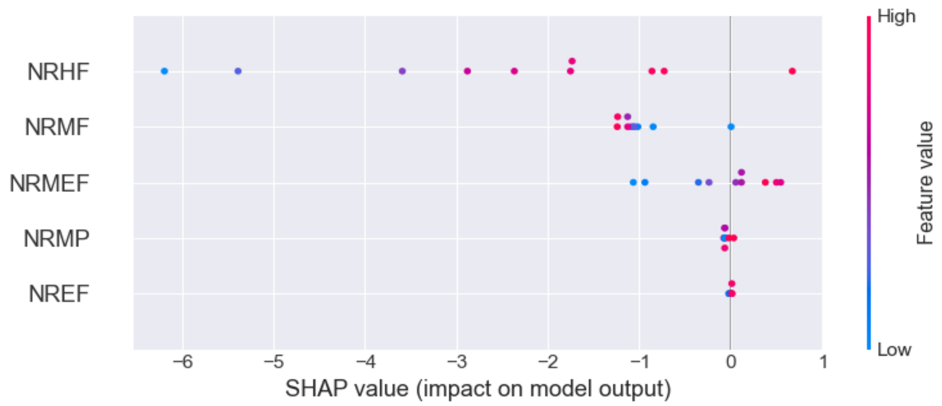

Fig. 3. SHAP summary, input variables on the prediction of cost overrun

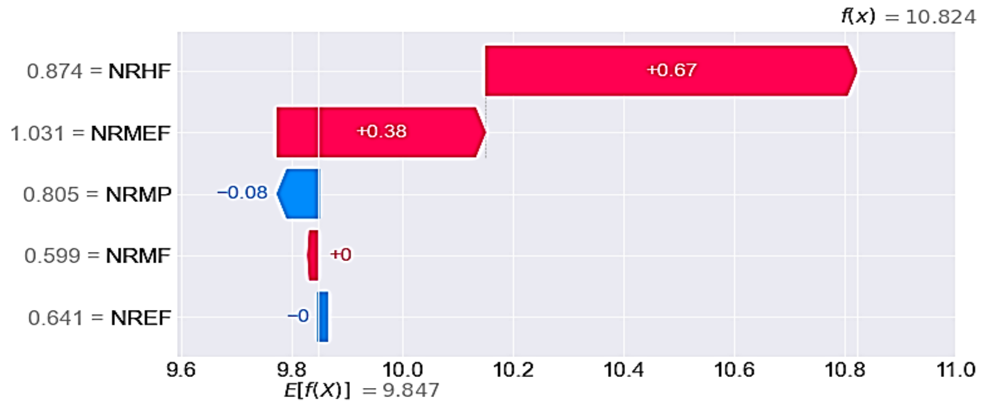

Fig. 4. SHAP waterfall plot, input variables on the prediction of cost overrun

The SHAP value for NRHF (Normalized Human Factors) is the highest of all the SHAP values, (+0.67), as seen in Fig.4, thus making NRHF the most positive factor for the predicted cost overrun. This positive impact indicates that deficiencies in workforce skills and training, in communication, and in safety compliance greatly increased the forecasted project cost. In the same way, NRMEF (Normalized Machine and Equipment Factors) had a positive impact (+0.38), meaning that machine failures, poor maintenance and machinery inefficiency contributed to increase cost overruns due to operating disruptions and extra costs. NRMP (Normalized Methods and Procedures), on the other hand, had a slight negative contribution (−0.08), which suggests that proper construction processes and organized procedures in the construction processes actually helped mitigate the anticipated cost overruns. NRMF (Normalized Management Factors) and NREF (Normalized Environmental Factors) had minimal impact near zero, indicating that in this particular observation management practices and environmental conditions did not significantly affect the cost overrun prediction. These findings can be interpreted in light of the operational characteristics of irrigation projects. Human-related factors exert the greatest influence on cost overruns because inadequate workforce skills, insufficient safety awareness, and poor communication frequently lead to construction errors, rework, material waste, and reduced productivity. In irrigation projects, where excavation works, concrete lining, canal structures, and hydraulic installations require high operational accuracy, even minor human errors can result in substantial additional labor, equipment, and material costs. Similarly, machinery and equipment deficiencies increase project expenditures through equipment downtime, maintenance costs, and delays in critical construction activities. In contrast, the relatively small contributions of management and environmental factors in this observation suggest that direct operational issues associated with workforce performance and equipment reliability had a more immediate impact on project costs than broader managerial or environmental conditions. These findings emphasize that improving workforce competency, strengthening safety training, and implementing preventive equipment maintenance can substantially reduce the likelihood of cost overruns in irrigation projects.

7.5.2 SHAP Values Analysis for Time Delay

The SHAP Summary Plot results Fig. 5 indicated that NRMF (Normalized Management Factors) had the greatest influence on the time delay prediction in irrigation projects. The positive SHAP values for higher values of management-related problems suggested that they generally led to an increase in predicted project delays, whereas the effective planning, supervision, coordination, and risk management activities reduced schedule deviations. The influence of NRHF (Normalized Human Factors) was also very positive, with low workforce skill, low productivity and poor communication making it more likely that there would be time delays. In addition, NRMEF (Normalized Machine and Equipment Factors) showed a moderate positive relationship with project delays, signifying that machinery failures, maintenance issues and machinery inefficiencies resulted in construction continuity disruption and heightened delay risks. NREF (Normalized Environmental Factors) demonstrated a relatively smaller and more variable influence: For projects, environmental factors might either accelerate or retard delays. Lastly, the effect of NRMP (Normalized Methods and Procedures) was the least, with SHAP values clustering around zero meaning that it has a comparatively minor role in time delay prediction.

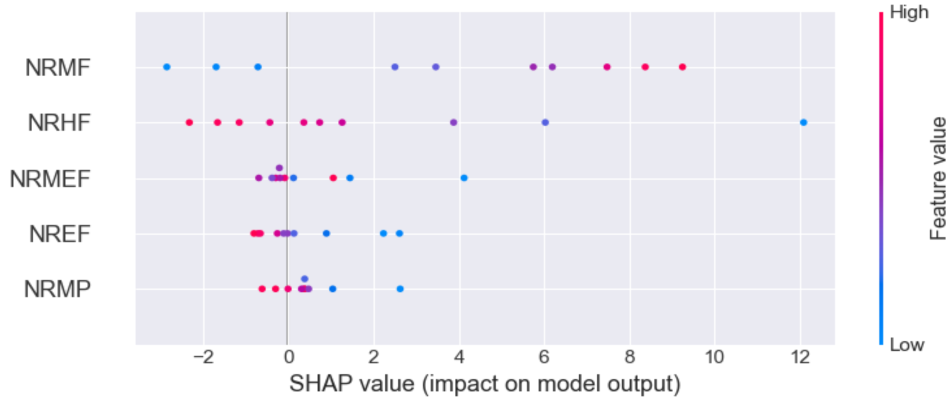

Fig. 5. SHAP summary, input variables on the prediction of time delay

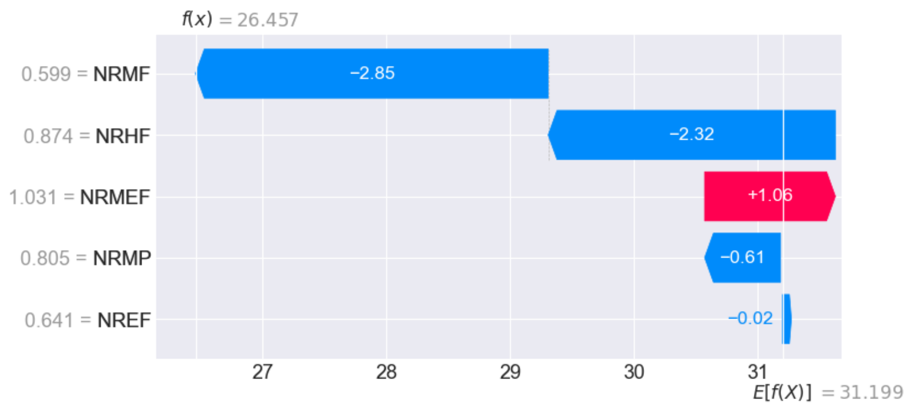

Fig. 6. SHAP waterfall plot**,** input variables on the prediction of duration delay

The SHAP Waterfall Plot Fig. 6 indicates that NRMF (Normalized Management Factors) had the strongest negative contribution to the predicted time delay, with a SHAP value of (−2.85). This negative effect indicates that the expected project delay was indeed substantially reduced as a result of effective project planning, supervision, coordination and risk management efforts. Likewise, the NRHF (Normalized Human Factors) had a negative contribution of (−2.32), meaning that the more skilled workers, effective communication, and more productive workforce, the better the project progressed and the lower the schedule deviations. NRMEF (Normalized Machine and

Equipment Factors) showed a positive contribution of (+1.06), suggesting that machinery failure, lack of equipment and maintenance issues contributed to the increased estimated delay for the project. In the meantime, the contribution of organized workflows and efficient operational procedures (NRMP) was moderate (−0.61), indicating that NRMP helped mitigate delays. Lastly, the influence of the environmental condition (NREF, Normalized Environmental Factors) was almost negligible (−0.02), suggesting that the environment was of little importance for the prediction of the time delay for this specific observation. These findings can be explained by the operational characteristics of irrigation projects. Effective management practices, including accurate planning, timely decision-making, efficient resource allocation, and continuous site supervision, play a fundamental role in maintaining project schedules., delays in procurement, equipment mobilization, or coordination among project stakeholders can interrupt sequential construction activities and substantially extend project duration. Similarly, competent and safety-aware workers contribute to reducing operational errors, rework, and productivity losses, thereby minimizing schedule deviations. In contrast, machinery breakdowns and inadequate equipment availability directly interrupt construction operations, particularly in excavation, earthworks, and canal construction activities, leading to schedule extensions. The relatively small contribution of environmental factors suggests that, for the analyzed projects, managerial and operational practices had a greater influence on project completion time than environmental conditions.

8. Features Importance Analysis (Extra Trees) for Cost Overrun and Time Delay

The "Features Importance" analysis presented a machine learning model, which shows the importance of each safety factor to the variation of the key performance indicators of irrigation project, which are the cost overrun and the time delay. The value of this analysis is that it can be used for each outcome separately to identify intervention and improvement priorities with greater accuracy and differentially, depending on the outcome desired.

8.1. Analysis of Factor Importance in Predicting Cost Overrun and Time Delay

The analysis of feature importance in Fig .7 for the prediction of cost overrun showed that the factors with the highest importance values were Human Factors (0.5838). This suggests that budget overruns in irrigation projects are mainly due to workforce skills, supervision, experience, and safety procedure compliance. Management Factors scored a (0.2787), indicating that planning, monitoring and decision-making is a critical factor in cost control of a project.

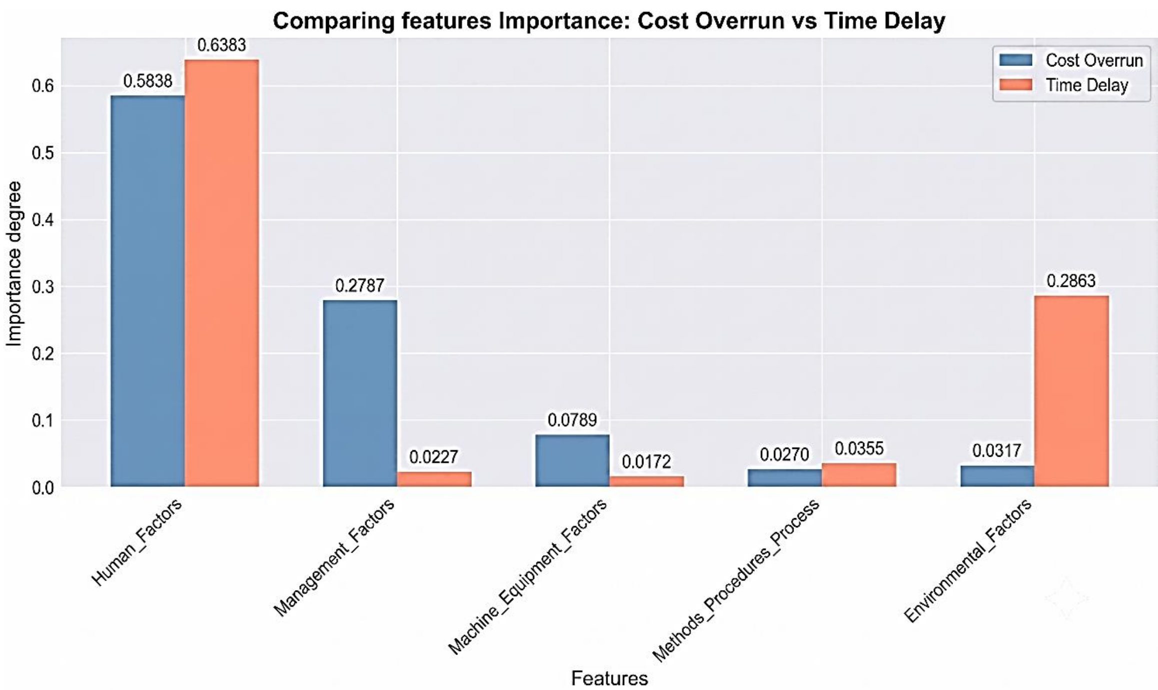

Fig. 7. Features importance analysis for cost overrun and time delay

The third place was taken by the Machine and Equipment Factors which had an importance value of (0.0789), which means that these factors have a moderate influence due to failures of equipment and maintenance problems. Environmental Factors and Methods and Procedures Factors had relatively small effects, with values of (0.0317) and (0.0270) respectively, indicating a relatively small contribution to the prediction of cost overrun. The importance analysis of the time delay prediction indicated that the most important variables were Management Factors, with an importance value of (0.2787), because planning, monitoring and decision-making are key factors in keeping projects on schedule. Human Factors was ranked second with a value of (0.0789), meaning that productivity and experience of the employees greatly influence the project's completion time. Machine and Equipment Factors were in third position with a value of (0.0317), which indicates that problems with machines and equipment can cause project delays. Environmental Factors and Methods and Procedures Factors also had relatively low values of importance values (0.0270 and 0.0207, respectively), suggesting a comparatively small contribution to time delay prediction.

8.2. Comparative and Integrated Analysis Between the Two Indicators

Comparing the two performance indicators shows that there are basic differences in the structure of influence and the hierarchy of priority: The differences between the most important factor are the most striking difference. There is an acknowledged difference in the top three responses between the two: Human Factors leads the list for cost overrun, and Management Factors leads the list for time delay. This means that financial difficulties are mainly attributed to human errors and time delay to administrative planning and follow-up. Other factors show a difference in ordering as well. Next, in both indicators, are the Machine and Equipment Factors, but they are more closely related to the costs than to delay. Environmental and Procedural factors combine in the lower ranks, albeit in a more noticeable way when looking at time delay as opposed to costs. As far as concentration versus distribution of impact is concerned, it is observed that cost overrun shows high concentration in two main factors (Human and Management) together explaining more than 86% of the impact. The effects on time delay, on the other hand, are more evenly divided among factors, so that a more general and more detailed approach to solving the problems of schedules is needed.

9. Limitations

Although the proposed framework demonstrated promising predictive performance, several limitations should be acknowledged. First, the study was developed using a dataset of 50 irrigation projects. While this represents a relatively small sample for machine learning applications, obtaining comprehensive datasets that simultaneously integrate cost, schedule, and safety information from real-world irrigation projects is inherently challenging due to the limited availability and implementation of such projects. Therefore, this study establishes an important empirical foundation for safety-driven predictive modeling in irrigation infrastructure. Second, the developed models were calibrated using data collected from irrigation projects in Iraq; consequently, their predictive performance may reflect the local managerial, environmental, and operational conditions. Nevertheless, the proposed methodological framework, which integrates validated safety indicators, machine learning-based prediction, and SHAP-based model interpretability, is not geographically constrained and can be readily adapted and validated using project data from other regions and infrastructure sectors.

10. Conclusion

This study introduced an explainable machine learning framework where predictive modeling is combined with safety indicators to forecast cost and time risks in irrigation projects. Previous studies, which mostly dealt with construction projects and focused on the accuracy of the predictors, were insufficient in this field. This framework integrates safety indicators and machine learning with SHAP explainability, thus offering actionable and time-sensitive recommendations for the management of irrigation projects. This research demonstrated that safety management indicators affect the performance of the projects. The safety indicators were validated using PLS-

SEM to confirm their reliability and readiness to be used as predictors, and only then were they introduced to the machine learning models. It was found that the biggest contributors to cost overruns were the factors that were of a human nature. On the other hand, the factors that had the highest contribution to delays in the project schedules were of a management nature. SHAP analysis advanced the transparency of the predictive models and explained the nature of the models by describing the contribution of safety indicators. Predictive performance was assessed using the standard 80:20 train-test split and 5-fold cross-validation. Good predictive performance was achieved in the train-test split. The most rigorous evaluation of model robustness and generalization for the small size of the data set was achieved by the 5-fold cross-validation. The results of the cross-validation showed that for the prediction of cost overruns and delays in project schedules, Random Forest was the most reliable, and for the prediction of cost overruns, XGBoost was the most accurate. In comparison, Support Vector Regression (SVR) showed a notable drop in performance during cross-validation. This suggests that SVR also has a poor ability to generalize, even when it provides a strong performance on a single train-test split. The developed models were originally developed for irrigation in Iraq, but the framework is flexible and can be modified and implemented in other areas and sectors.

Acknowledgement

The authors would like to convey their heartfelt gratitude to the employees and engineers of the Ministry of Water Resources of Iraq for their cooperation in collecting the required data.

References

[1] Fay M, Roze J. Beyond the Gap: How Countries Can Afford the Infrastructure They Need while Protecting the Planet. Washington (DC): World Bank; 2021.

[2] Unegbu HCO, Yawas DS, Dan-asabe B, Alabi AA. Advanced machine learning models for predicting project performance in complex construction environments. MSA Eng J. 2023;5339(Ml):1-14.

[3] Ismail MZ, Ramly ZM, Adul R. Systematic review of cost overrun research in the developed and developing countries. Int J Sustain Constr Eng Technol. 2021;12(1):196-211. https://doi.org/10.30880/ijscet.2021.12.01.020

[4] Brokbals S, Wapelhorst V, Ivan Č. Calculation of risk costs in construction projects: Empirical analysis of construction risks applying the Monte Carlo method. Constr Eng Des. 2019;2019:120-128. https://doi.org/10.1002/cend.201900014

[5] Viles E, Santilli A. Causes of delay in construction projects: a quantitative analysis. Eng Constr Archit Manag. 2020;27(4):917-935. https://doi.org/10.1108/ECAM-01-2019-0024

[6] Asiedu RO, Adaku E. Cost overruns of public sector construction projects: a developing country perspective. Int J Manage Proj Bus. 2019;12(4):1000-1022.

[7] Lee K, Kim K. Analyzing cost and schedule growths of road construction projects, considering project characteristics. Sustainability. 2021;13(24):13694. https://doi.org/10.3390/su132413694

[8] Hoseini E, Bosch-Rekveldt M, Hertogh M. Cost contingency and cost evolvement of construction projects in the preconstruction phase. J Constr Eng Manage. 2020;146(6):04020048. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001842

[9] Noaman AES, Aws S. Methodology of estimating construction projects contingency. IOP Conf Ser Earth Environ Sci. 2023;1129(1):012044. https://doi.org/10.1088/1755-1315/1129/1/012044

[10] Nath ND, Behzadan AH, Paal SG. Deep learning for site safety: Real-time detection of personal protective equipment. Autom Constr. 2020;112:103085. https://doi.org/10.1016/j.autcon.2020.103085

[11] Environmental Science. Improving safety through leveraging machine learning and safety-related data in the construction industry. IOP Conf Ser Earth Environ Sci. 2024;1389(1):012012. https://doi.org/10.1088/1755-1315/1389/1/012012

[12] Lathong K, Wisaeng K. An innovative hybrid machine learning techniques for predicting construction cost estimates. Int J Comput Civ Struct Eng. 2024;20(3):69-83.

[13] Peško I, Mučenski V, Šešlija M, Radujković M, Bibić D, Krklješ M. Estimation of costs and durations of construction of urban roads using ANN and SVM. Complexity. 2017;2017:2450370. https://doi.org/10.1155/2017/2450370

[14] Quo S. Big data adoption in the Chinese construction industry. Buildings. 2024;14(7):1891. https://doi.org/10.3390/buildings14071891

[15] Erlita A, Amin M, Purwoko B, Bintoro K. Risk management of time overrun in multiple phases of construction: Consultant perspective. Int J Res Rev. 2021;8(7):311-319. https://doi.org/10.52403/ijrr.20210743

[16] Zafar I, Wuni IY, Shen GQP, Ahmed S. A fuzzy synthetic evaluation analysis of time overrun risk factors in highway projects of terrorism-affected countries: the case of Pakistan. Int J Constr Manage. 2019;21(11):1-14.
[17] Kabadi HM. Review paper on impact of recourse constraints on the time overrun of construction projects. Int J Res Appl Sci Eng Technol. 2024;11(7):831-838.
[18] Aziz RF. Ranking of delay factors in construction projects after Egyptian revolution. Alexandria Eng J. 2013;52(3):387-406. https://doi.org/10.1016/j.aej.2013.03.002
[19] Hameed AA. Development of duration estimation model for irrigation and drainage projects using neural network. NeuroQuantology. 2022;20(6):3041-3053.
[20] Ahiaga-Dagbui DD, Smith SD. Neural networks for modelling the final target cost of water projects. In: Smith SD, editor. Procs 28th Annual ARCOM Conference; 2012 Sep 3-5; Edinburgh, UK. Association of Researchers in Construction Management; 2012. p. 307-316.
[21] Ahiaga-Dagbui DD. Dealing with construction cost overruns using data mining. Constr Manage Econ. 2014;32(7-8):37-41. https://doi.org/10.1080/01446193.2014.933854
[22] Sing C. Cost profiling of water infrastructure projects. J Infrastruct Syst. 2018;24(4):04018029. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000441
[23] Love PED, Smith J, Simpson I, Regan M. Understanding the landscape of overruns in transport infrastructure projects. Environ Plan B Plan Des. 2015;42(3):490-509. https://doi.org/10.1068/b130102p
[24] Risk P, Approaches M, Analysis C. Deterministic and probabilistic risk management approaches in construction projects: A systematic literature review and comparative analysis. Buildings. 2023;13(5):1312. https://doi.org/10.3390/buildings13051312
[25] Eliasson J. Cost overruns of infrastructure projects - distributions, causes and remedies. Transp Res Part A Policy Pract. 2025;198:104532. https://doi.org/10.1016/j.tra.2025.104532
[26] Molinari L, Haezendonck E, Mabillard V. Cost overruns of Belgian transport infrastructure projects: Analyzing variations over three land transport modes and two project phases. Transp Policy. 2023;134:167-179. https://doi.org/10.1016/j.tranpol.2023.02.017
[27] Belay SM, Tilahun S, Yehualaw M, Matos J, Sousa H, Workneh ET. Analysis of cost overrun and schedule delays of infrastructure projects in low income economies: Case studies in Ethiopia. Adv Civ Eng. 2021;2021:4991204. https://doi.org/10.1155/2021/4991204
[28] Egwim CN, Alaka H, Details C, Technologies BD. A comparative study on machine learning algorithms for predicting construction projects delay. Proc Int Conf Ind Eng Oper Manage. 2019;2019:1-10.
[29] Aung T, Liana SR, Htet A, Bhaumik A. Using machine learning to predict cost overruns in construction projects. J Technol Innov Energy. 2023;2(2):1-7. https://doi.org/10.56556/jtie.v2i2.511
[30] Coffie GH, Cudjoe SKF. Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. Int J Constr Manage. 2024;24(16):1742-1750. https://doi.org/10.1080/15623599.2023.2289754
[31] Coffie GH, Cudjoe SKF. Toward predictive modelling of construction cost overruns using support vector machine techniques. Cogent Eng. 2023;10(2):2269656. https://doi.org/10.1080/23311916.2023.2269656
[32] Nassar AH, Elbisy AM. A machine learning approach to predict time delays in marine construction projects. Eng Technol Appl Sci Res. 2024;14(5):16125-16134. https://doi.org/10.48084/etasr.8173
[33] Abd AM, Kareem YA, Zehawi RN. Prediction and estimation of highway construction cost using machine learning. Eng Technol Appl Sci Res. 2024;14(5):17222-17231. https://doi.org/10.48084/etasr.8285
[34] Das P, Kashem A, Hasan I, Islam M. A comparative study of machine learning models for construction costs prediction with natural gradient boosting algorithm and SHAP analysis. Asian J Civ Eng. 2024;25(4):3301-3316. https://doi.org/10.1007/s42107-023-00980-z
[35] Debero DW, Sinesilassiea EG. Machine learning model for construction time prediction: A case of selected public building projects in Hosanna, Ethiopia. J Eng. 2024;2024:5653690. https://doi.org/10.1155/2024/5653690
[36] Liben SM, Belachew DA, Elsaigh WA. Comparing advanced and traditional machine learning algorithms for construction duration prediction: a case study of Addis Ababa's public sector. Eng Res Express. 2024;6(4):045012. https://doi.org/10.1088/2631-8695/ad979f
[37] Restrepo Ramirez AF, Rua-Machado CA. Predicting delays and cost overruns in construction projects: A machine learning approach. Rev EIA. 2025;22(44):1861. https://doi.org/10.24050/reia.v22i44.1861
[38] Hamdan MM, Thneibat M, Hyari K. Predicting cost overrun in construction projects using machine learning algorithms: the case of Jordan. Eng Constr Archit Manag. 2026;33(6):4732-4763. https://doi.org/10.1108/ECAM-09-2024-1209
[39] Abdelalim AM, Salem M, Al-adwani M. An analysis of factors contributing to cost overruns in the global construction industry. Buildings. 2025;15(1):18. https://doi.org/10.3390/buildings15010018

[40] Tengtrairat PP. Efficient noisy sound-event mixture classification using adaptive-sparse complex-valued matrix factorization and O vs O SVM. Sensors. 2020;20(16):4368. https://doi.org/10.3390/s20164368
[41] Plevris V, Solorzano G, Bakas NP, El M, Ben A. Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: Proceedings of the 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022); 2022 Jun 5-9; Oslo, Norway. 2022. https://doi.org/10.23967/eccomas.2022.155
[42] Jos P, García E, S BMP, Paredes S. Modelling energy performance of residential dwellings by using the MARS technique, SVM-based approach, MLP neural network and M5 model tree. Energy Build. 2023;341:112000. https://doi.org/10.1016/j.apenergy.2023.121074
[43] Audemard G, Bellart S, Bounia L, Koriche F, Lagniez J, Marquis P. On the explanatory power of Boolean decision trees. Data Knowl Eng. 2022;142:102088. https://doi.org/10.1016/j.datak.2022.102088
[44] Blanquero R, Carrizosa E, Molero-Río C, Romero D. On optimal regression trees to detect critical intervals for multivariate functional data. Comput Oper Res. 2023;152:106152. https://doi.org/10.1016/j.cor.2023.106152
[45] Ho A, Czarnecki S. Random forest algorithm and support vector machine for nondestructive assessment of mass moisture content of brick walls in historic buildings. Autom Constr. 2023;149:104780. https://doi.org/10.1016/j.autcon.2023.104793
[46] Ali M, Prasad R, Xiang Y, Jamei M, Mundher Z. Ensemble robust local mean decomposition integrated with random forest for short-term significant wave height forecasting. Renew Energy. 2023;205:731-746. https://doi.org/10.1016/j.renene.2023.01.108
[47] Afandizadeh S, Amoei N, Mirzahossein H. Analyzing the effects of congestion on planning time index - Grey models vs. random forest regression. Int J Transp Sci Technol. 2023;12(2):578-593. https://doi.org/10.1016/j.ijtst.2022.05.008
[48] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA, USA. New York: ACM; 2016. p. 785-794. https://doi.org/10.1145/2939672.2939785
[49] Islam M, Das P, Rahman M, Naz F, Kashem A. Prediction of compressive strength of high-performance concrete using optimization machine learning approaches with SHAP analysis. J Build Pathol Rehabil. 2024;9(1):44. https://doi.org/10.1007/s41024-024-00445-z
[50] Nasir M, Iqbal M, Zahid M, Khan K, Nasir M, Jalal FE. Evaluating the relevance of eggshell and glass powder for cement-based materials using machine learning and SHapley Additive exPlanations (SHAP) analysis. Case Stud Constr Mater. 2023;19:e02278. https://doi.org/10.1016/j.cscm.2023.e02278
[51] Abdulalim A, Iqbal M, Zahid M, Khan K, Nasir M, Jalal FE. Prediction of rapid chloride penetration resistance of metakaolin based high strength concrete using light GBM and XGBoost models by incorporating SHAP analysis. Constr Build Mater. 2022;345:128296. https://doi.org/10.1016/j.conbuildmat.2022.128296
[52] Cakiroglu C, Demir S, Ozdemir MH, Aylak BL, Sariisik G, Abualigah L. Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating SHAP analysis. Expert Syst Appl. 2024;237:121464. https://doi.org/10.1016/j.eswa.2023.121464
[53] Das P, Kashem A. Hybrid machine learning approach to prediction of the compressive and flexural strengths of UHPC and parametric analysis with shapley additive explanations. Case Stud Constr Mater. 2024;20:e02723. https://doi.org/10.1016/j.cscm.2023.e02723