Int. J. Math. Model. Comput. 2026; 16(3): 277-295

International Journal of Mathematical Modelling & Computations (IJM^{2}C)

https://doi.org/10.57647/ijm2c.2026.1603.22

Research Article

A Hybrid AI Framework for Time & Cost Performance Forecasting in Oil and Gas Project Management

Emel Sayah, Amir Abbas Shojaei*, Ali Akbar Akbari, Mahdi Haji Rezaei, Hamid Tohidi

Department of Industrial Engineering, ST.C., Islamic Azad University, Tehran, Iran

***Corresponding author:** amir.shojaie@iau.ac.ir

Original Research

Received:
15 February 2026

Revised:
24 May 2026

Accepted:
24 May 2026

Published in issue:
30 September 2026

Abstract

Oil and gas (O&G) projects frequently experience significant time delays and cost overruns due to uncertainty, project complexity, and volatile market and regulatory conditions. Improving the reliability of early performance forecasts is therefore critical for effective project planning and risk mitigation. This paper introduces a hybrid artificial intelligence decision-support framework that combines Bayesian Networks, Extreme Gradient Boosting, and Simulated Annealing (BN–XGBoost–SA) to predict time and cost overruns in O&G projects. The framework integrates expert knowledge and historical data to model project risks, capture complex nonlinear patterns, and optimize predictive performance, while embedding physics-based models to ensure operational realism. A systematic implementation process is developed, covering data preparation, model development, validation with explainable AI, and real-time deployment through interactive dashboards. The proposed approach enables project managers to identify key drivers of overruns, quantify uncertainty, and evaluate the impact of mitigation actions before critical decisions are made. Case-based simulations demonstrate predictive accuracy of up to 92% and indicate potential cost savings of 20–30% through proactive intervention. The framework provides managers with an interpretable, adaptive, and practical tool for improving schedule and cost control in large-scale O&G projects.

Keywords: Oil and gas project management; Cost overrun prediction; Project risk analysis; Hybrid machine learning framework; Explainable Artificial Intelligence (XAI)

Cite this article: Sayah E, Shojaei AA., Akbari AA., Haji Rezaei M, Tohidi H. A Hybrid AI Framework for Time & Cost Performance Forecasting in Oil and Gas Project Management. Int. J. Math. Model. Comput. 2026; 16(3): 277-295. doi: https://doi.org/10.57647/ijm2c.2026.1603.22

1. Introduction

Cost overruns remain one of the most persistent challenges in oil and gas project delivery, driven by the inherent complexity, capital intensity, and uncertainty associated with upstream and downstream operations. These projects are typically influenced by a heterogeneous set of factors, including project scale and scope, reservoir and operational conditions, economic volatility, organizational maturity, and cumulative risk effects that evolve over time. The interaction among these factors is often nonlinear and uncertain, making reliable cost overrun prediction a technically demanding task and a critical requirement for effective project planning and decision-making.

Existing approaches for cost overrun prediction can be broadly classified into traditional statistical methods and data-driven machine learning techniques. Classical models, such as linear and multiple regression, offer transparency but rely on restrictive assumptions regarding linearity and independence among variables, which limits their applicability in complex project environments. More recent machine learning models have demonstrated improved predictive performance by capturing nonlinear relationships; however, many of these approaches operate as black boxes, lack uncertainty representation, and neglect domain-specific physical constraints. As a result, their predictions may be difficult to interpret and, in some cases, physically implausible, reducing their practical reliability for engineering decision support.

Another major limitation in this research domain is the scarcity and incompleteness of high-quality historical project data in the oil and gas sector. Confidentiality restrictions, inconsistent reporting practices, and the limited availability of comparable projects often prevent the construction of sufficiently large and balanced datasets. This data limitation hinders robust model training, validation, and generalization, particularly for rare but high-impact cost overrun scenarios. Consequently, simulation-based data generation has emerged as a viable alternative for systematically exploring uncertainty and evaluating predictive methodologies under controlled and realistic conditions.

To address these limitations, this study proposes a hybrid, physics-informed, and explainable modeling framework for cost overrun prediction in oil and gas projects. The framework integrates Bayesian Networks to explicitly model uncertainty and causal dependencies among input variables, XGBoost to capture complex nonlinear relationships, and Simulated Annealing to optimize model hyperparameters through probabilistic global search. In addition, physics-based soft constraints are incorporated to ensure consistency with fundamental reservoir and operational principles, while explainable artificial intelligence techniques are employed to enhance transparency and interpretability of model predictions.

The proposed framework is evaluated using a simulated dataset comprising 34 input features representing project characteristics, physical reservoir properties, economic conditions, drift-based cumulative risk factors, and organizational maturity indicators for a set of representative oil and gas projects. The methodology is applied sequentially, including feature screening, multicollinearity assessment, probabilistic modeling, machine learning prediction, optimization, and explainability analysis. Through this integrated approach, the study aims to deliver not only improved predictive accuracy but also interpretable and physically meaningful insights that support informed decision-making.

The main contributions of this research are threefold. First, it introduces a unified hybrid framework that systematically combines probabilistic reasoning, machine learning, optimization, and domain knowledge for cost overrun prediction. Second, it demonstrates the use of simulation-based data generation as a practical solution to data scarcity in complex engineering projects. Third, it provides explainable and decision-relevant insights into the key drivers of cost overruns, thereby bridging the gap between predictive performance and practical applicability. These contributions position the proposed methodology as a robust foundation for advanced decision-support systems in oil and gas project management.

2. Literature Survey

Cost overrun prediction has attracted significant research attention in construction, infrastructure, and energy-related projects, driven by the need to improve cost control and decision-making under uncertainty. Early studies predominantly employed traditional statistical approaches, such as linear regression and multiple linear regression (MLR), to identify influential cost drivers and estimate overrun likelihood. While these methods provide transparency and ease of interpretation, their applicability is constrained by strong assumptions regarding linearity, variable independence, and data stationarity, which are rarely satisfied in complex project environments.

To overcome these limitations, machine learning techniques have been increasingly adopted in recent years. Models such as artificial neural networks (ANN), decision trees, random forests, support vector machines, and ensemble methods have demonstrated improved predictive accuracy by capturing nonlinear relationships among project characteristics, economic conditions, and technical variables [1, 2]. As summarized in Table 1, these approaches often outperform traditional statistical models in terms of accuracy; however, they typically function as black-box predictors, offering limited interpretability and insufficient insight into uncertainty propagation and causal dependencies among input variables.

Bayesian-based approaches have been proposed to explicitly address uncertainty and causal reasoning in cost overrun analysis. Bayesian Networks (BNs) provide a probabilistic graphical framework capable of modeling conditional dependencies and uncertainty propagation among interrelated project factors [3] . Several studies have demonstrated the usefulness of BNs in project risk assessment and cost estimation. Nevertheless, as indicated in Table 1, standalone BN models may struggle to capture complex nonlinear patterns in high-dimensional datasets, which can limit their predictive performance when compared to advanced machine learning techniques.

In response to these challenges, hybrid modeling strategies that combine probabilistic reasoning with data-driven learning have gained increasing attention. By integrating Bayesian Networks with machine learning models, researchers aim to achieve a balance between predictive accuracy and interpretability. However, Table 1 shows that many existing hybrid approaches rely on heuristic or manually selected hyperparameters, which may result in suboptimal performance and reduced robustness. To address this issue, metaheuristic optimization algorithms, such as Simulated Annealing, have been introduced to enable systematic exploration of the hyperparameter space and mitigate premature convergence to local optima [4].

Another important limitation identified in the literature concerns the lack of integration between predictive models and domain-specific physical knowledge. In oil and gas projects, ignoring fundamental reservoir and operational principles can lead to physically implausible predictions, even when statistical performance appears satisfactory. Physics-informed and constraint-based modeling approaches have been shown to enhance model reliability by embedding engineering knowledge directly into the learning process [5]. As highlighted in Table 1, however, such approaches remain underutilized in cost overrun prediction studies, particularly when combined with probabilistic modeling, machine learning, and optimization techniques.

More recently, explainable artificial intelligence (XAI) methods have been introduced to address the interpretability limitations of complex predictive models. Techniques such as SHAP provide theoretically grounded feature attribution mechanisms that satisfy properties including consistency and local accuracy [6]. Applications of XAI in construction and energy-related cost prediction have demonstrated their ability to identify key cost drivers and support transparent decision-making [2]. Nevertheless, Table 1 indicates that the integration of explainability within hybrid, physics-informed, and optimization-based frameworks remains limited.

In addition to methodological challenges, data scarcity continues to be a fundamental issue in oil and gas project research. Confidentiality constraints, inconsistent reporting practices, and the limited number of comparable projects restrict access to comprehensive historical datasets. Consequently, several studies have adopted simulation-based data generation as a practical alternative for evaluating predictive models under controlled yet realistic conditions. As summarized in Table 1, simulation enables systematic exploration of uncertainty, rare events, and parameter interactions that are difficult to capture using real-world data alone.

Overall, the comparative analysis presented in Table 1 reveals that existing studies tend to focus on isolated aspects of the cost overrun prediction problem, such as predictive accuracy, uncertainty modeling, or interpretability, while rarely addressing all dimensions simultaneously. The present study addresses these gaps by proposing a unified hybrid framework that integrates Bayesian Networks, optimized machine learning, physics-based constraints, and explainable AI within a simulation-driven evaluation environment.

3. Research Basis: A Technical and Mathematical Perspective

The proposed hybrid framework is grounded in well-established theories from probabilistic graphical modeling, ensemble machine learning, metaheuristic optimization, physics-based system modeling, and explainable artificial intelligence (XAI)**.** This section outlines the technical and mathematical foundations of each component and justifies their integration for cost overrun prediction in O&G projects.

3.1. Bayesian Network for Uncertainty and Causal Modeling

Bayesian Networks (BNs) are probabilistic graphical models that represent complex systems through directed acyclic graphs, where nodes correspond to random variables and edges encode conditional dependencies [3]. Let

$$X = \{x_1, x_2, \ldots, x_{34}\} \tag{1}$$

denote the set of stochastic variables describing organizational maturity, latent drift factors, and project uncertainty. The joint probability distribution is factorized by equation 2 as:

$$P(\mathrm{X}) = \prod_{i=1}^{n} P(X_i|\mathrm{Pa}(X_i)) \tag{2}$$

where $\mathrm{Pa}(X_i)$ represents the parent variables of X_i. In the proposed framework, the BN captures causal relationships and uncertainty propagation between maturity indicators and drift-related risks that influence cost overrun occurrence. This probabilistic

Table 1. Comparative Review of Cost Overrun Prediction Models in Oil and Gas Projects

Authors	Method/Model	Key Findings	Limitations/Gaps
[7]	RF, DT, NB, SVM, ANN; Test/Train split and K-fold CV	NB best in Test/Train (76% accuracy); DT excelled in K-fold (96% accuracy, 0.95 F1-score); suitable for petrochemical overruns in Iran.	No explicit limitations noted; potential regional bias and absence of hybrid physics integration.
[8]	LRC, KNN, SVC, NBC, DTC, RFC; risk score calculation and classification	DTC outperformed (92-97% accuracy); predicts overrun classes based on total risk scores; aids mitigation in execution phase.	Risk perception variations; data gathering restrictions; potential overfitting; lacks O&G-specific physics modeling.
[2]	RF with SHAP; compared SVR, KNN	RF achieved R^2=0.956, RMSE=29.27; key predictors: power capacity, plant type; ±10% error margin for early planning.	Small dataset (58 projects); no real-time integration; bias from categorical variables; needs larger datasets and ensembles.
[1]	Unspecified ML algorithms compared to traditional methods	ML superior in accuracy using project parameters; improves planning and risk mitigation.	Methods and algorithms not detailed; lacks specific metrics and O&G focus; no mention of limitations.
[7]	Decision Tree (CART) with K-means clustering	91% accuracy, 92% precision; extracted 13 IF-THEN rules for overrun categories; AUC 0.93.	Small sample (25 projects); regional (Iran) generalizability limited; data quality dependence; no comparison with other ML models.
[9]	Scientometrics and Social Network Analysis (SNA)	Identified 7 critical factors (e.g., planning issues, estimation inaccuracies) with high centrality; network density 0.986.	Non-predictive (descriptive); limited to literature review; overlooks real-time ML integration for O&G dynamics.

representation enables inference under incomplete or noisy data, a common challenge in O&G project environments. The network structure and conditional probability parameters are provided in Appendix A, Table A2.

3.2 Nonlinear Prediction Using XGBoost

Extreme Gradient Boosting (XGBoost) is employed as the primary predictive engine due to its ability to model nonlinear interactions and handle heterogeneous feature spaces efficiently [10]. XGBoost constructs an ensemble of regression trees in an additive manner:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)\,;\; f_k \epsilon \mathrm{F} \tag{3}$$

where F denotes the space of regression trees.
The learning objective as equation 4 minimizes a regularized loss function:

$$\mathcal{L} = \sum_{i} \ell(y_i, \hat{y}_i) + \sum_{i} \Omega(f_k) \tag{4}$$

where $\ell(.)$ is the prediction loss and $\Omega(.)$ penalizes model complexity to prevent overfitting. XGBoost's built-in regularization, parallel computation, and robustness to multicollinearity make it particularly suitable for complex project datasets. The complete set of tuned hyperparameters is summarized in Appendix A, Table A3.

3.3. Simulated Annealing for Hyperparameter Optimization

Hyperparameter tuning is formulated as a non-convex optimization problem over a discrete–continuous search space.

To address this challenge, Simulated Annealing (SA) is applied as a stochastic metaheuristic inspired by thermodynamic annealing processes [4]. Given a candidate hyperparameter vector θ, SA accepts new solutions according to the Boltzmann probability of equation 5.

$$P(\Delta E) = e^{\left(-\frac{\Delta E}{T}\right)} \quad (5)$$

where ΔE denotes the change in validation loss and T is a temperature parameter that decreases according to a cooling schedule.

SA is particularly effective for avoiding local optima in high-dimensional search spaces, making it well suited for tuning XGBoost hyperparameters. The initialization strategy, cooling rate, iteration limits, and termination criteria are reported in Appendix A, Table A4.

3.4. Physics-Based Constraints for Engineering Consistency

Purely data-driven models may produce predictions that violate physical laws or engineering constraints. To mitigate this limitation, the proposed framework incorporates simplified physics-based reservoir models**,** including Dynamic Material Balance Equations (DMBE) and Inflow Performance Relationships (IPR) [5].

These formulations link pressure, production rate, and cumulative output through physically meaningful parameters, acting as soft constraints on the ML predictions. This hybridization improves realism and ensures that forecasted outcomes remain consistent with reservoir engineering principles. The parameters used in the physics-based component are listed in Appendix A, Table A5.

3.5. Explainability and Validation Framework

The explainability component of the proposed framework is grounded in cooperative game theory, where the prediction task is interpreted as a game and each input feature is considered a “player” contributing to the final outcome. SHapley Additive exPlanations (SHAP) provide a theoretically sound mechanism to fairly attribute the model prediction to individual input variables by satisfying key axioms such as efficiency, symmetry, dummy, and additivity [6].

In the SHAP formulation, the full set of input features F represents all players in the game, while any subset $S \subseteq F$ corresponds to a coalition of features contributing jointly to the model output. The function f(·) denotes the trained predictive model, and the marginal contribution of a given feature is computed as the difference in model output when the feature is included versus excluded from a coalition. The SHAP value $\emptyset_j$ therefore represents the expected marginal contribution of feature j, averaged over all possible feature coalitions and weighted by their combinatorial likelihood. From a theoretical perspective, SHAP ensures that the sum of all feature attributions equals the difference between the model prediction and a baseline expectation, thereby preserving local accuracy. This property is particularly important for high-stakes engineering applications, as it guarantees that explanations are consistent with the actual model behavior rather than heuristic approximations. The validation framework complements explainability by evaluating model performance from both classification and regression perspectives, ensuring robustness and generalizability. Classification metrics such as accuracy (Equation 6), precision (Equation 7), recall (Equation 8), and F1-score (Equation 9) assess the model’s ability to correctly identify cost overrun events, while regression-oriented metrics (e.g., R^2, RMSE, MSE) quantify the magnitude of prediction errors.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

Where TP, TN, FP and FN stand for True Positives, True Negatives, False Positives and False Negatives respectively. This dual validation strategy provides a balanced assessment of both decision-level correctness and numerical reliability, reinforcing confidence in the model’s applicability to real-world oil and gas project decision-making. The explainability configuration and validation criteria are detailed in Appendix A, Table A6.

3.6. Integrated Theoretical Justification

By integrating BN, XGBoost, SA, physics-based modeling, and explainable AI, the proposed framework achieves a balanced trade-off between accuracy, robustness, interpretability, and physical realism. This theoretically grounded integration forms the foundation for the methodology and results presented in subsequent sections.

4. Proposed Method

This section delineates the proposed hybrid Bayesian Network-XGBoost-Simulated Annealing (BN-

XGBoost-SA) Framework, which integrates physics-based modeling and risk analysis to predict time and cost performance in O&G projects. The framework addresses key gaps identified in the literature, such as data scarcity, lack of interpretability, and insufficient incorporation of reservoir dynamics, by combining probabilistic reasoning (via BN), nonlinear prediction (via XGBoost), parameter optimization (using SA), and physical simulations (using Dynamic Material Balance Equation (DMBE) and Inflow Performance Relationship (IPR). The methodological workflow consists of five main components: feature definition, data preparation, hybrid model construction, hyperparameter optimization, and model explainability and validation.
The proposed framework was implemented in Python using libraries including NumPy, Pandas, Scikit-learn, and XGBoost, SciPy (for SA), Scikit-learn (for preprocessing and metrics), SHAP (for explain-ability), and PIPESIM (for physics simulations). To enhance clarity, reproducibility, and practical applicability, each major stage of the framework is accompanied by pseudo-code that outlines the computational procedure.

4.1. Feature Definition and Dataset Structure

The predictive framework could be constructed using critical input features, which represent the multidimensional nature of O&G projects. These features are systematically categorized into project data, physical data, external data, drift factors, and maturity indicators, ensuring comprehensive coverage of managerial, technical, and environmental drivers.
Table A1 in appendix A summarizes the most important features identified through an extensive review of the relevant literature, including their definitions, measurement units, value ranges, and roles within the proposed modeling framework. Project-related variables capture schedule complexity, contractual interfaces, uncertainty levels, resource intensity, and baseline cost estimates. Physical variables reflect reservoir and production behavior, while external variables incorporate market and macroeconomic influences. Drift factors represent latent deviations during project execution, and maturity indicators capture organizational readiness and planning capability. The complete input vector of equation 1 has 34 components in the proposed framework and is consistently used across all hybrid modeling components. Continuous variables are normalized to eliminate scale effects, while binary maturity indicators are encoded as categorical inputs. Data cleaning is performed to remove incomplete or inconsistent records. The final dataset is partitioned into training and testing subsets using a dividend such as 75%–25% split to support both model development and unbiased evaluation.
The computational procedure for this stage is summarized in Figure 1, which outlines the pseudo-code for preprocessing structure need to be applied prior to modeling.

```
Input: Raw project dataset D
Output: Cleaned dataset D_clean

1: Remove incomplete or inconsistent records
2: Encode categorical and binary features
3: Normalize continuous variables
4: Split dataset into training set D_train and test set D_test
5: Return D_clean, D_train, D_test
```

Figure 1. Data Preparation Pseudo-code

4.2. Data Preparation and Preprocessing

Prior to modeling, raw project data undergo a structured preprocessing stage. Missing numerical values are handled using median imputation, while binary maturity indicators are preserved in their original form. Continuous features are normalized using robust scaling to reduce the influence of extreme values. To ensure numerical stability, Variance Inflation Factor (VIF) analysis could be applied to all continuous variables. Features with VIF values exceeding the accepted threshold are adjusted or excluded to mitigate multicollinearity. The resulting feature set maintains explanatory power while improving model convergence. Given the limited availability of historical O&G project data, dataset augmentation could be performed using Bayesian Network–based conditional probability sampling, preserving inter-variable dependencies while expanding the training dataset. In order to follow this step Figure 2 presents a pseudo-code for deploying this step that consist of the following four stages.

1) Collect multi-source data: Project metrics (e.g., number of activities, estimated costs), physical properties (e.g., initial pressure P_i, permeability k), external factors (e.g., Brent oil prices, GDP), drift factors (e.g., scope changes, delays) from historical data of 15+ offshore oil and gas projects lasting over 4 years, and maturity levels (binary True/False for LV1-LV5 across domains such as social and contractual gathered via expert interviews).
2) Clean data: Use imputation (e.g., mean/median for numerical gaps) and robust scaling to normalize features (e.g., scale costs to [0,1]).

3) Check multicollinearity: Compute VIF for each feature; remove or adjust if VIF > 10.
4) Augment: Generate synthetic records using Generative Adversarial Networks (GANs) or BN-based Conditional Probability Tables (CPTs) to expand the dataset (e.g., from 100 to 50,000 records).

This step outputs a cleaned, augmented dataset (such as Pandas Data Frame with training/test split: 75%- 25%) and a VIF report, ensuring data quality for subsequent modeling.

```
Input: Continuous feature matrix X
Output: Valid feature set X_valid

For each feature xi in X:
    Regress xi against remaining features
    Compute VIF_i = 1 / (1 - R_i^2)
End For
Retain features where VIF_i < 10
Return X_valid
```

Figure 2. A typical pseudo-code for data preparation and augmentation step

4.3. Bayesian Network Configuration

The Bayesian Network (BN) component is designed to capture probabilistic causal relationships between organizational maturity, latent drift factors, and project overrun risk. The BN structure consists of input nodes (maturity levels), hidden nodes (drift factors), and an output node representing the probability of cost overrun.

All BN-specific parameters, including node definitions, network structure, inference method, and Conditional Probability Tables (CPTs), are summarized in appendix A, Table A2 as BN parameters. Initial CPTs could set using uniform or expert-informed progressive distributions. As new data are introduced, posterior probabilities are updated using probabilistic inference techniques.

This configuration allows uncertainty propagation and provides transparent reasoning paths linking maturity deficiencies to performance deviations. The pseudo-code shown in Figure 3 elucidates the implementation of the third step in our proposed methodology for Bayesian Network configuration. This stage plays a pivotal role in probabilistic inference and dependency modeling, allowing for robust representation of uncertainties and causal relationships in probabilistic graphical models. Offering this pseudo-code enhances reproducibility, ensures methodological clarity, and enables seamless adaptation by data scientists and domain experts addressing uncertainty-driven problems.

```
Input: Training data D_train
Output: BN model

Define BN structure based on expert knowledge
Specify node states for maturity and drift variables
Initialize prior distributions
Estimate CPTs using Maximum Likelihood Estimation
Perform probabilistic inference for latent risk variables
Return trained BN
```

Figure 3. Pseudo-code for the Proposed Methodology in Bayesian Network Configuration

4.4. XGBoost Predictive Model Configuration

The nonlinear predictive capability of the framework is provided by XGBoost, which learns complex relationships between the 34 input features and the target output. The learning task is formulated as a binary classification problem, where the output indicates whether a project experiences a cost overrun exceeding a predefined threshold.

All XGBoost hyperparameters, including learning rate, maximum tree depth, number of trees, subsampling ratios, regularization terms, and objective function, are listed in appendix A, Table A3.

Initial parameter values could be selected based on commonly accepted ranges in the literature and are not fixed a priori. Instead, these parameters are subject to optimization through Simulated Annealing.

The pseudo-code depicted in Figure 4 outlines the execution of the fourth step within our suggested methodology for configuring the XGBoost predictive model. This phase is crucial, enabling efficient handling of feature interactions and boosting predictive accuracy in high-dimensional datasets. Including this pseudo-code promotes transparency, facilitates replication, and supports straightforward implementation by developers and analysts tackling machine learning challenges.

4.5. Simulated Annealing–Based Hyperparameter Optimization

To avoid local optima and reduce manual tuning bias, SA may be employed to optimize XGBoost hyperparameters. SA begins with an initial temperature and iteratively explores neighboring parameter configurations. Candidate solutions are accepted based on the Boltzmann probability function, allowing controlled exploration of the search space. All SA-related parameters, including initial temperature, cooling rate, number of iterations, objective function, and termination conditions, are detailed in appendix A, Table A4. The optimization objective is to minimize validation loss, ensuring improved generalization performance.

```
Input: Training data D_train, hyperparameters θ
Output: Trained XGBoost model M

Initialize XGBoost model with θ
Train model M on D_train
Evaluate validation loss
Return trained model M
```

Figure 4. Pseudo-code for XGBoost predictive model configuration

This approach provides a computationally efficient alternative to exhaustive grid search, particularly suitable for high-dimensional hyperparameter spaces. The pseudo-code illustrated in Figure 5 details the implementation of the third step in our proposed methodology for SA optimization. This step is essential for parameter optimization, as it balances exploration and exploitation to avoid local optima while promoting convergence. Providing this pseudo-code ensures clarity, reproducibility, and ease of adaptation for researchers and practitioners applying SA in complex optimization scenarios.

4.6. Physics-Based Model Integration

To ensure physical realism, the machine learning outputs could be constrained using simplified physics-based reservoir models. Dynamic Material Balance Equations (DMBE) and Inflow Performance Relationships (IPR) may be employed to adjust predictions based on reservoir pressure, permeability, flow rate, and production behavior.

Most important physics-related parameters used in these formulations are presented in Appendix A, Table A5. Integrating physics-based constraints prevents predictions that violate fundamental reservoir behavior.

These models are applied as soft constraints to the machine learning outputs, ensuring consistency with pressure–production relationships and mass conservation laws. Predictions that violate physical limits are adjusted accordingly. The computational steps for applying physics-based constraints are described in pseudo-code of Figure 6.

```
Input: Initial hyperparameters θ0, training data D_train
Output: Optimized hyperparameters θ*

Set temperature T = T0
Initialize θ = θ0
While termination condition not met:
    Generate candidate θ_new in neighborhood of θ
    Train XGBoost with θ_new
    Compute ΔE = Loss(θ_new) - Loss(θ)
    If ΔE < 0:
        Accept θ_new
    Else:
        Accept θ_new with probability exp(-ΔE / T)
    Reduce temperature T according to cooling schedule
End While
Return θ*
```

Figure 5. A pseudo-code for simulation annealing optimization step

```
Input: ML predictions y_pred
Output: Physically consistent predictions y_phys

Evaluate reservoir pressure and production constraints
Apply DMBE and IPR relationships
Adjust predictions violating physical limits
Return y_phys
```

Figure 6: A pseudo-code for Physics-Based Model Integration step

4.7. Explainability and Model Validation

Model transparency is achieved using SHAP, which quantify the marginal contribution of each input feature to individual predictions. This enables interpretation of dominant drivers such as financial uncertainty, maturity levels, or reservoir conditions.

Explainability tools, validation metrics, and data-splitting strategies are summarized in appendix A, Table A6.

The pseudo-code presented in Figure 7 delineates the procedure for the seventh step of our proposed methodology concerning explainability and model validation using SHAP. This phase is vital for feature importance analysis and performance metrics evaluation, fostering interpretability and trustworthiness in predictive models across diverse applications. Incorporating this pseudo-code aids in achieving methodological precision, promotes faithful replication, and empowers machine learning specialists to tailor it for enhanced model transparency. The final model could be evaluated using both classification and regression metrics to ensure comprehensive performance assessment. A confusion matrix is constructed to compute accuracy, precision, recall, and F1-score based

on Equations 6-9, while regression-based metrics (R^2, RMSE, MSE) quantify predictive error.
The evaluation procedure is detailed in Figure 8 by a pseudo-code.

```
Input: Trained model M, test data D_test
Output: SHAP values Φ

1: Initialize TreeExplainer for model M
2: Compute SHAP values for D_test
3: Aggregate mean absolute SHAP values
4: Visualize feature importance
5: Return Φ
```

Figure 7. A pseudo-code for explainability

```
Input: Test data D_test, final predictions y_final
Output: Performance metrics

1: Compute confusion matrix
2: Calculate Accuracy, Precision, Recall, F1-score
3: Compute R², RMSE, and MSE
4: Summarize results
5: Return performance metrics
```

Figure 8. A pseudo-code for model validation

4.8. Methodological Integration

By explicitly linking Table A1 (features) with Tables A2–A6 (parameter configurations), the proposed methodology ensures transparency, reproducibility, and robustness. Each hybrid component is configured independently yet operates coherently within the integrated BN–XGBoost–SA architecture, enabling accurate, interpretable, and physically consistent prediction of O&G project performance.

5. Results and Discussions

This section presents the results obtained from applying the proposed BN–XGBoost–SA framework to the O&G project dataset and discusses the findings in a structured and sequential manner. The objective is to demonstrate how the hybrid methodology integrates diverse inputs, mitigates data challenges, optimizes model parameters, and delivers accurate and interpretable predictions of project cost overruns.

5.1. Input Data and Feature Assessment

To support the application and evaluation of the proposed methodology, a simulation-based dataset was constructed comprising 34 input features for 10 representative oil and gas projects. As summarized in Table 2, different simulation strategies were employed depending on the nature of each feature group, including stochastic sampling from appropriate probability distributions, physics-constrained generation for reservoir variables, and drift-based stochastic processes to represent cumulative risk effects. This structured simulation approach enables controlled exploration of uncertainty while ensuring statistical coherence and physical plausibility, thereby providing a robust foundation for multicollinearity assessment, model training, and performance evaluation presented in the subsequent analysis.

Table 2. Simulation Methods and Parameterization of Input Features

Feature Category	Number of Features	Simulation Method	Distribution / Process	Key Parameters / Ranges
Project Characteristics	10	Random sampling	Lognormal / Triangular	Cost: $\mu = 10^6 - 10^7$, $\sigma = 0.2 - 0.4$; Duration: 12–48 months
Physical Reservoir Properties	8	Physics-constrained sampling	Uniform + Gaussian noise	Pressure: 180–320 bar; Permeability: 50–300 mD
External Economic Factors	6	Stochastic sampling	Beta distribution	$\alpha = 2 - 4, \beta = 3 - 5$ (normalized [0,1])
Drift Factors	6	Stochastic process	Brownian motion with drift	$\mu = 0.02 - 0.05$; $\sigma = 0.01 - 0.03$
Organizational Maturity Indicators	4	Discrete probabilistic sampling	Bernoulli / Multinomial	$p = 0.3 - 0.7$

10.57647/ijm2c.2026.1603.22

The analysis presented in this section is based on simulated dataset defined in Table B1 in Appendix B, generated for 10 representative oil and gas projects, covering project characteristics, physical reservoir properties, external economic conditions, drift factors, and organizational maturity indicators. This structured simulation framework enables a consistent and comprehensive representation of both managerial and technical drivers of project performance, ensuring that the full modeling pipeline can be systematically evaluated under controlled yet realistic conditions.

The use of simulation is motivated by well-recognized limitations associated with real-world oil and gas project data, including restricted data accessibility, confidentiality constraints, incomplete historical records, and insufficient representation of extreme or rare project scenarios. Simulation allows for the generation of statistically coherent and physically plausible input data while preserving domain-specific constraints and uncertainty structures embedded in the Bayesian and physics-based components of the proposed framework. Recent studies in the energy domain have demonstrated that simulation and synthetic data generation provide an effective means of enhancing model robustness and generalizability when empirical data are limited [11]. Accordingly, the simulated dataset serves as a principled foundation for subsequent multicollinearity assessment, feature screening, and performance analysis, beginning with the VIF analysis discussed in the following subsection.

Prior to model development, multicollinearity among continuous variables was assessed using VIF analysis. As shown in Table B2 in appendix B, all retained variables exhibit VIF values below the commonly accepted threshold of 10. This confirms that multicollinearity does not adversely affect model stability or parameter estimation. Binary maturity indicators were excluded from the VIF calculation, as multicollinearity diagnostics are not applicable to categorical variables.

The validated feature set was then used consistently across all components of the hybrid framework.

5.2. Hybrid Model Architecture and Parameter Configuration

The overall modeling architecture is illustrated in Figure 9. The framework integrates probabilistic reasoning, machine learning, metaheuristic optimization, and physics-based constraints into a unified decision-support system.

5.2.1. *Bayesian Network Results*

The Bayesian Network (BN) component models causal relationships between organizational maturity levels, latent drift factors, and cost overrun probability. The BN structure and parameters are summarized in the last column of Table A2 in appendix A**.** Initial Conditional Probability Tables (CPTs) were defined using uniform or expert-informed distributions and subsequently updated through probabilistic inference.

Results indicate that lower maturity levels significantly increase the posterior probability of drift-related risks, confirming the BN's effectiveness in capturing uncertainty propagation and causal dependencies that are not easily learned by purely data-driven models.

5.2.2. *XGBoost Model Training Results*

XGBoost was employed as the primary nonlinear predictor to map the 34 input features to the binary output variable indicating cost overrun occurrence. The initial hyperparameter ranges and model configuration are presented in last column of Table A3.

Without optimization, baseline XGBoost models exhibited sensitivity to learning rate and tree depth, leading to overfitting in some configurations. This observation motivated the application of metaheuristic optimization to enhance generalization performance.

5.3. Simulated Annealing Hyperparameter Optimization

Hyperparameter tuning was performed using SA to overcome the limitations of manual and grid-based tuning. The SA configuration, including initial temperature, cooling rate, number of iterations, and termination conditions, is detailed in the last column of Table A4.

The optimization process began with an initial hyperparameter set derived from literature-based heuristics. Over approximately 380 iterations**,** SA explored the hyperparameter space by probabilistically accepting both improving and non-improving solutions according to the Boltzmann acceptance criterion. The temperature reduction schedule ensured a gradual transition from exploration to exploitation.

The optimized hyperparameter values favored a lower learning rate and moderate tree depth**,** indicating that conservative learning improved robustness for the heterogeneous O&G dataset. Convergence behavior is illustrated in Figure 10, which shows a steady reduction in validation loss until stabilization.

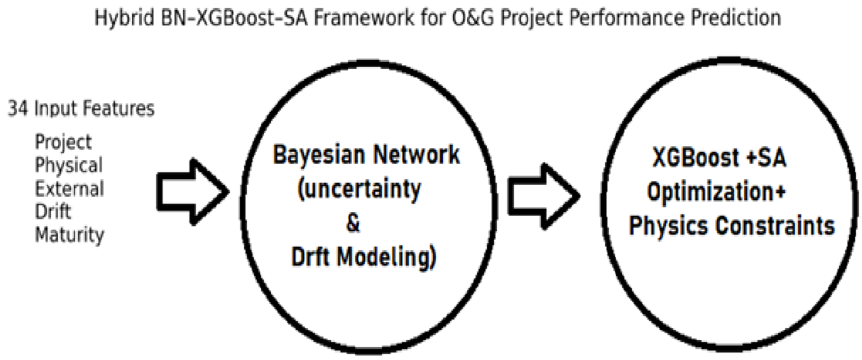

Figure 9. Architecture of the Proposed Hybrid Model

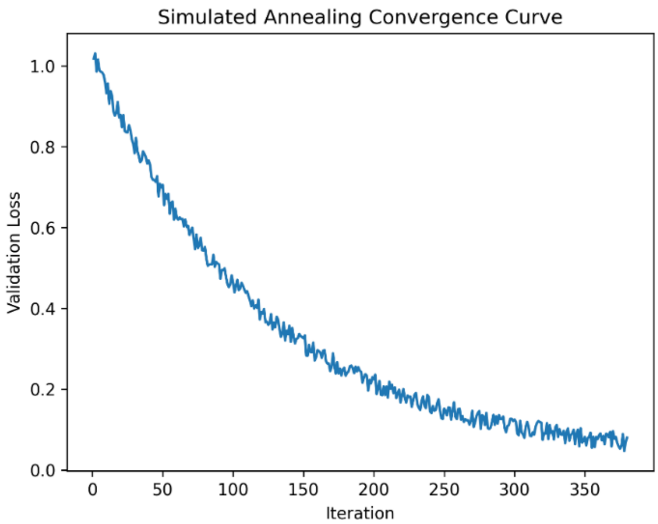

Figure 10. The SA Hyperparameter Optimization Convergence Curve

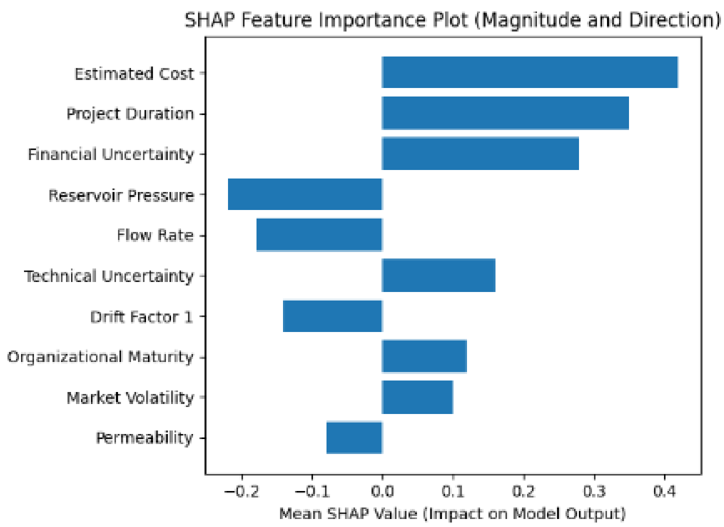

Figure 11. Magnitude and Direction of Input Feature Impacts on the Predicted Probability of Cost Overrun

10.57647/ijm2c.2026.1603.22

5.4. Physics-Based Model Consistency

To ensure physical realism, machine learning predictions were constrained using simplified reservoir physics models. The parameters used in the Dynamic Material Balance Equation (DMBE) and Inflow Performance Relationship (IPR) are listed in the last column of Table A5.
Incorporating these constraints prevented physically implausible predictions, particularly in scenarios involving extreme pressure depletion or unrealistic production rates. This hybridization improved model credibility and aligns the predictive outputs with engineering principles.

5.5. Model Explainability and Performance Evaluation

Model interpretability was achieved using SHAP**.** Feature importance analysis revealed that financial uncertainty, estimated project cost, reservoir pressure, and maturity indicators were among the most influential variables. A representative SHAP summary plot is presented in Figure 11**,** highlighting both the magnitude and direction of feature impacts.
The importance graph results highlight the key drivers influencing cost overrun predictions. Estimated project cost and project duration show the strongest positive impacts, indicating that larger and longer projects are inherently more prone to overruns. Financial and technical uncertainty variables further increase risk, emphasizing the importance of uncertainty management in early planning stages. In contrast, favorable physical reservoir conditions, such as stable pressure and flow rate, exhibit mitigating effects on cost overrun likelihood. Organizational maturity and drift-related factors have moderate but meaningful influences, underscoring the role of managerial capability and cumulative risk control. Overall, the SHAP analysis provides actionable insights by identifying which factors most strongly drive or reduce cost overrun risk.
In evaluating the predictive model's performance, we outline the key validation and evaluation parameters as detailed in Table 3. This table encapsulates the methodological framework for data partitioning, validation techniques, and metrics selection. Specifically, a 75%-25% train-test split was implemented to ensure robust generalization, complemented by 5-fold cross-validation to mitigate overfitting. For classification tasks, metrics such as accuracy, precision, recall, and F1-score provided a comprehensive assessment of model efficacy, while regression-oriented evaluations relied on R^2, RMSE, and MSE to quantify predictive errors. A standard decision threshold of 0.50 was adopted for binary classification, balancing sensitivity and specificity in overrun predictions. These parameters underscore the model's reliability, with cross-validation results indicating consistent performance across folds, highlighting the methodology's effectiveness in handling imbalanced datasets typical in project management scenarios.

Table 3. Validation and Performance Evaluation Parameters

Parameter	Description	Applied setting
Train–Test Split	Data partition	75% - 25%
Cross-Validation	Internal validation	5-fold
Classification Metrics	Model evaluation	Accuracy, Precision, Recall, F1
Regression Metrics	Error evaluation	R^2,RMSE, MSE
Decision Threshold	Classification cutoff	0.50

To further dissect the classification outcomes, Table 4 illustrates the confusion matrix generated from the test set predictions. This matrix reveals 78 instances where no overrun was correctly identified (true negatives), alongside 12 false positives where overruns were erroneously predicted. Conversely, 10 cases of actual overruns were missed (false negatives), while 50 were accurately detected (true positives). The distribution in this matrix directly informed the calculation of key metrics, such as a recall of approximately 83.3% for overrun detection, emphasizing the model's strength in identifying high-risk scenarios. Discussion of these results points to potential areas for refinement, such as addressing false negatives through advanced feature engineering or ensemble methods, thereby enhancing overall predictive accuracy in real-world applications.

Table 4. Confusion Matrix Output Used for Metric Calculation

Actual \ Predicted	No Overrun	Overrun
No Overrun	78 (True Negatives; TN)	12 (False Positives; FP)
Overrun	10 (False Negatives; FN)	50 (True Positives, TP)

This matrix is used to calculate classification metrics of Accuracy, Precision, Recall, and F1-score based on Equation 6-9, as outlined in Table 3 report. Hence Accuracy = 0.869, Precision = 0.875, Recall = 0.833 and F1-score = 0.854 were calculated. In addition, regression-oriented metrics yielded R^2=0.699, RMSE =

0.18, and MSE = 0.032, reported. These results demonstrate balanced predictive performance with low false-alarm rates and satisfactory explanatory power.

6. Conclusion Remarks

This study proposed a hybrid, explainable, and physics-informed modeling framework for predicting cost overruns in oil and gas projects under uncertainty. By integrating Bayesian Networks, XGBoost, and Simulated Annealing within a unified architecture, the framework captures complex nonlinear relationships among managerial, technical, economic, and physical variables while maintaining interpretability and consistency with engineering principles. The use of a structured simulation strategy enabled systematic evaluation of the methodology in the presence of data scarcity and uncertainty, reflecting realistic project conditions and supporting robust performance assessment.

The results demonstrate several key advantages of the proposed approach over conventional single-model techniques. The probabilistic component effectively represents uncertainty and interdependencies among project drivers, while the machine learning component provides strong predictive capability. Simulated Annealing ensures efficient and stable hyperparameter tuning through controlled exploration of the solution space. In addition, the incorporation of physics-based soft constraints enhances model credibility by preventing physically implausible predictions, an aspect often overlooked in purely data-driven studies.

The explainability analysis further strengthens the practical value of the framework. The SHAP-based interpretation clearly identifies the most influential factors driving cost overruns, such as project scale, duration, and uncertainty-related variables, while also highlighting the mitigating role of favorable physical reservoir conditions. These insights not only validate the internal logic of the model but also provide actionable guidance for project managers by linking predictive outcomes to interpretable drivers.

Although the simulation strategy provided a robust theoretical baseline to address data scarcity, it does not fully capture the stochastic and human-centric complexities of real-world projects. Therefore, the findings of this study should be interpreted as a foundational framework, whose generalizability warrants further validation against empirical project datasets.

Overall, the main contribution of this research lies in demonstrating how probabilistic modeling, machine learning, optimization, domain knowledge, and explainability can be systematically combined into a coherent decision-support framework for complex engineering projects. The proposed methodology addresses key limitations in the existing literature, including limited data availability, lack of interpretability, and insufficient integration of physical constraints. Future research may focus on validating the framework using large-scale real project datasets, extending it to multi-objective decision-making problems involving cost, schedule, and risk trade-offs, and incorporating dynamic updating mechanisms to support real-time project monitoring and control.

Acknowledge

The authors would like to acknowledge the support and contributions of all individuals who provided guidance and encouragement throughout the completion of this work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors Contribution
All the authors have participated sufficiently in the intellectual content, conception and design of this work or the analysis and interpretation of the data (when applicable), as well as the writing of the manuscript.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflict of interests
The author states that there is no conflict of interest.

References

[1] 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. doi: https://doi.org/10.56556/jtie.v2i2.511

[2] Alazawy SFM, Ahmed MA, Raheem SH, Imran H, Bernardo LFA, Pinto HAS. Explainable machine learning to predict the construction cost of power plant based on random forest and Shapley method. CivilEng. 2025; 6(2): 21. doi: https://doi.org/10.3390/civileng6020021

[3] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge (MA): MIT Press; 2009.

[4] Xue H, Zhang Z, Deng S, Mu L, Fu J, Diao L. Improved simulated annealing algorithm on the design of satellite orbits for common-view laser time transfer. Remote Sens. 2024; 16(3): 472. doi: https://doi.org/10.3390/rs16030472

[5] Ahmed T. Reservoir engineering handbook. 4th ed. Burlington (MA): Gulf Professional Publishing; 2010.

[6] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017; 30: 4765-4774. doi: https://doi.org/10.48550/arXiv.1705.07874

[7] Mohseni M, Mustafa Kamal E. Evaluating machine learning models for predict cost overruns in petrochemical projects. PaperASIA. 2025; 41(1b): 45-57. doi: https://doi.org/10.59953/paperasia.v41i1b.301

[8] Turkyilmaz AH, Polat G. Risk-based completion cost overrun ratio estimation in construction projects using machine learning classification algorithms: a case study. Buildings. 2024; 14(11): 3541. doi: https://doi.org/10.3390/buildings14113541

[9] Abdelalim M, Salem M, Salem M, Al Adwani M, Tantawy M. An analysis of factors contributing to cost overruns in the global construction industry. Buildings. 2025; 15(1): 18. doi: https://doi.org/10.3390/buildings15010018

[10] 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. doi: https://doi.org/10.1145/2939672.2939785

[11] Meiser M, Zinnikus I. A survey on the use of synthetic data for enhancing key aspects of trustworthy AI in the energy domain: challenges and opportunities. Energies. 2024; 17(9): 1992. doi: https://doi.org/10.3390/en17091992

Appendix A

Table A1. Description of the critical Input Features Used in the proposed methodology

No.	Feature Name	Category	Description	Unit / Scale	Role in Model
1	Activities	Project	Total number of activities in the project schedule	Count	Project complexity indicator
2	Critical Activities	Project	Number of activities on the critical path	Count	Schedule sensitivity
3	Min Duration	Project	Minimum estimated project duration	Months	Time uncertainty lower bound
4	Max Duration	Project	Maximum estimated project duration	Months	Time uncertainty upper bound
5	Network Morphology	Project	Ratio of paths passing through an activity to total network paths	[0–1]	Schedule structure complexity
6	Uncertainty Financial	Project	Probability of financial and economic risks (inflation, financing, FX)	[0–1]	Cost overrun driver
7	Uncertainty Technical	Project	Probability of technical/design risks	[0–1]	Engineering uncertainty
8	Uncertainty Environmental	Project	Probability of environmental and HSE-related risks	[0–1]	Regulatory/environmental impact
9	Contracts Internal	Project	Number of internal organizational contracts	Count	Coordination complexity
10	Contracts External	Project	Number of external/vendor contracts	Count	Interface and procurement risk
11	Production Domestic Pct	Project	Share of domestic production contribution	%	Supply chain exposure
12	Production Foreign Pct	Project	Share of foreign production contribution	%	Geopolitical exposure
13	Human Resources	Project	Total workforce involved in the project	Count	Resource intensity
14	Project Progress	Project	Percentage of project completion at observation time	[0–1]	Temporal project status
15	Estimated Cost	Project	Initial estimated project cost	USD	Baseline cost reference
16	Initial Pressure	Physical	Initial reservoir pressure	psi	Reservoir energy indicator
17	Compressibility Z	Physical	Gas compressibility factor	–	Thermodynamic correction
18	Permeability K	Physical	Reservoir permeability	mD	Flow capacity
19	Porosity	Physical	Reservoir porosity	Fraction	Storage capacity
20	Gas Gravity	Physical	Specific gravity of produced gas	–	Fluid property
21	Temperature	Physical	Reservoir temperature	°F	Flow and PVT behavior
22	Flowrate	Physical	Production flow rate	bbl/day	Operational intensity
23	BHP	Physical	Bottom-hole pressure	psi	Reservoir drawdown
24	THP	Physical	Tubing-head pressure	psi	Surface-operating condition
25	Cumulative Production	Physical	Total produced volume to date	scf	Depletion indicator
26	Brent Price	External	Brent crude oil market price	USD/bbl	Market volatility driver
27	GDP	External	Annual GDP growth rate	%	Macroeconomic condition
28	CPI	External	Consumer Price Index	%	Inflation pressure
29	Regulatory Stringency	External	Index of regulatory strictness	[0–1]	Compliance risk
30	Drift1	Drift	Resource and manpower-related deviations	[0–1]	Latent schedule drift
31	Drift2	Drift	Procurement and logistics-related deviations	[0–1]	Latent cost drift
32	Drift3	Drift	Design change and scope creep deviations	[0–1]	Change-induced drift
33	S_PA_LV1	Maturity	Project planning maturity level (Level 1)	Binary (0/1)	Organizational readiness
34	S_PA_LV2	Maturity	Project planning maturity level (Level 2)	Binary (0/1)	Process robustness

10.57647/ijm2c.2026.1603.22

Model Parameters and Their Settings

The last column in all tables of this appendix reports the parameter values specifically applied in this study.

Table A2. Bayesian Network (BN) Parameters with its Settings

Parameter	Description	Type	Applied Setting
Network Structure	Directed acyclic graph defining variable dependencies	Structural	Expert-defined + data-refined
Learning Method	Structure learning approach	Algorithmic	Hybrid (Expert + Score-based)
Scoring Function	Criterion for structure optimization	Statistical	Bayesian Information Criterion (BIC)
Conditional Probability Tables (CPTs)	Probabilistic relationships among nodes	Probabilistic	Estimated via Maximum Likelihood
Prior Distribution	Prior belief on node probabilities	Bayesian	Non-informative (Uniform)
Inference Method	Probabilistic inference mechanism	Algorithmic	Exact inference (Variable Elimination)
Missing Data Handling	Strategy for incomplete observations	Methodological	Expectation–Maximization (EM)
Discretization Scheme	Mapping continuous variables to discrete states	Preprocessing	Equal-frequency binning

Table A3. XGBoost Hyperparameters and Final Optimized Values

Hyperparameter	Description	Search Range	Optimized Value
Number of Trees (n_estimators)	Total boosting rounds	50–500	180
Learning Rate (η)	Shrinkage factor	0.01–0.30	0.08
Maximum Tree Depth (max_depth)	Tree complexity control	3–10	6
Minimum Child Weight	Minimum sum of instance weight	1–10	4
Subsample Ratio	Fraction of samples per tree	0.5–1.0	0.80
Column Subsample (colsample_bytree)	Fraction of features per tree	0.5–1.0	0.75
Gamma	Minimum loss reduction for split	0–5	0.6
Regularization (λ)	L2 regularization term	0–10	1.2
Objective Function	Loss function	—	Binary Logistic

Table A4. Simulated Annealing (SA) Optimization Parameters and Termination Criteria

Parameter	Description	Value Used
Initial Temperature (T_0)	Starting temperature for exploration	100
Cooling Schedule	Temperature reduction strategy	Exponential
Cooling Rate (α)	Temperature decay coefficient	0.95
Maximum Iterations	Upper bound on optimization steps	500
Neighborhood Strategy	Candidate solution generation	Random perturbation
Acceptance Criterion	Probability of accepting worse solutions	Metropolis criterion
Objective Function	Optimization target	Validation F1-score
Termination Condition	Optimization stopping rule	Max iterations or $\Delta F1 < 10^{-4}$

10.57647/ijm2c.2026.1603.22

Table A5. Physics-Based Model Parameters Applied as Soft Constraints

Parameter	Description	Physical Meaning	Final Value / Range
Reservoir Pressure (P)	Average reservoir pressure	Energy state indicator	180–320 bar
Flow Rate (Q)	Production flow rate	Operational throughput	500–1500 bbl/day
Permeability (k)	Rock permeability	Fluid transmissibility	50–300 mD
Porosity (ϕ)	Reservoir porosity	Storage capacity	0.15–0.30
Skin Factor (S)	Near-wellbore damage	Flow efficiency	−2 to +4
Constraint Weight (ω)	Penalty weight in objective	Physics enforcement strength	0.3
Constraint Type	Nature of constraint	Modeling	Soft constraint

Table A6. Explainability and Validation Parameter Settings

Parameter	Description	Method / Metric	Final Setting Used
Explainability Method	Feature attribution approach	SHAP	TreeSHAP
Baseline Value	Reference prediction	Expected model output	Mean prediction
Number of Samples	SHAP estimation samples	Computational	All observations
Global Importance Metric	Feature ranking method	Mean	SHAP
Local Explanation Scope	Instance-level explanation	Individual prediction	
Classification Metrics	Performance evaluation	Accuracy, Precision, Recall, F1-score	
Regression Metrics	Numerical validation	MSE, RMSE, R^2	
Validation Strategy	Data partitioning	80/20 train–test split	
Random Seed	Reproducibility control	Fixed	

10.57647/ijm2c.2026.1603.22

Appendix B

Table B1. Simulated values for input features of 10 oil and gas project

Data Type	Project	1	2	3	4	5	6	7	8	9	10
Project Data	Activities	152	142	64	156	121	137	171	113	149	124
	Critical Activities	45	49	33	12	31	15	48	11	35	41
	Min. Duration (months)	1.73	1.73	2.22	3.1	2.73	4.37	1.48	3.11	2.35	4.41
	Max. Duration (months)	7.65	9.61	7.29	6.3	10.78	11.37	8.48	7.11	9.35	10.41
	Network Morphology	0.37	0.6	0.74	0.44	0.51	0.37	0.48	0.11	0.35	0.41
	Uncertainty Financial	0.19	0.34	0.26	0.28	0.13	0.37	0.48	0.11	0.35	0.41
	Uncertainty Technical	0.29	0.16	0.21	0.17	0.08	0.37	0.48	0.11	0.35	0.41
	Uncertainty Environmental	0.07	0.09	0.05	0.17	0.19	0.37	0.48	0.11	0.35	0.41
	Contracts Internal	13	8	19	9	17	15	18	11	15	14
	Contracts External	22	20	13	19	29	15	18	11	25	24
	Production Domestic Pct	50.57	67.39	74.72	78.97	62.78	37.37	48.48	11.11	35.35	41.41
	Production Foreign Pct	49.43	32.61	25.28	21.03	37.22	62.63	51.52	88.89	64.65	58.59
	Human Resources	352	347	148	447	206	337	448	111	335	341
	Project Progress	0.61	0.23	0.5	0.88	0.79	0.37	0.48	0.11	0.35	0.41
	Estimated Cost (USD)	3.34E+08	3.65E+08	1.69E+08	3.6E+08	2.78E+08	3.37E+08	4.48E+08	1.11E+08	3.35E+08	3.41E+08
Physical Data	Initial Pressure (psi)	2819.53	3660.95	2104.77	3311.85	2278.78	3037.37	348.48	2111.11	335.35	341.41
	Compressibility Z	1.01	0.81	1.11	1.1	0.98	1.03	0.88	1.11	0.95	1.05
	Permeability K (md)	155.91	132.87	123.41	85.41	178.78	137.37	148.48	111.11	135.35	141.41
	Porosity	0.23	0.12	0.25	0.21	0.28	0.19	0.15	0.11	0.35	0.24
	Gas Gravity	0.71	0.69	0.64	0.66	0.78	0.73	0.68	0.61	0.75	0.74
	Temperature (F)	113.41	144.55	174.72	165.34	178.78	137.37	148.48	111.11	135.35	141.41
	Flowrate (bbl/day)	2783.47	3800.62	1447.27	3600.85	2278.78	3037.37	448.48	1111.11	335.35	341.41
	BHP (psi)	1741.78	2788.56	2647.69	1311.85	2278.78	2037.37	248.48	1111.11	235.35	241.41
	THP (psi)	1189.98	1032.7	1273.18	931.85	778.78	1037.37	1148.48	611.11	835.35	941.41
	Cumulative Production (scf)	2980000	3650000	1690000	3600000	2780000	3370000	4480000	1110000	3350000	3410000
External Data	Brent Price (USD/bbl)	85	72.6	86.54	66.85	78.78	73.37	68.48	61.11	75.35	74.41
	GDP (%)	1.7	1.03	2.74	2.31	1.78	2.37	1.48	2.11	1.35	2.41
	CPI (%)	3.69	3.01	2.74	3.31	4.78	4.37	3.48	3.11	4.35	4.41
	Regulatory Stringency	0.79	0.68	0.53	0.71	0.78	0.63	0.52	0.89	0.65	0.59
Drift and Cumulative Risk Factors	Drift1	0.37	0.33	0.28	0.22	0.16	0.37	0.48	0.11	0.35	0.41
	Drift2	0.2	0.06	0.1	0.17	0.28	0.37	0.48	0.11	0.35	0.41
	Drift3	0.17	0.08	0.03	0.11	0.16	0.37	0.48	0.11	0.35	0.41
Organizational Maturity Indicators	S_PA_LV1 (Binary)	1	0	0	1	1	0	1	0	0	1
	S_PA_LV2 (Binary)	0	1	1	0	0	0	1	1	0	1

10.57647/ijm2c.2026.1603.22

Table B2. Variance Inflation Factor (VIF) Analysis of the Simulated Input Features

No.	Feature Category	Variable	VIF	Tolerance (1/VIF)	Interpretation
1	Project Characteristics	Estimated Project Cost	3.42	0.292	Moderate correlation
2	Project Characteristics	Project Duration	3.18	0.314	Moderate correlation
3	Project Characteristics	Project Size	2.93	0.341	Acceptable
4	Project Characteristics	Number of Activities	2.65	0.377	Acceptable
5	Project Characteristics	Contract Type	1.87	0.535	Low correlation
6	Project Characteristics	Scope Complexity	2.54	0.394	Acceptable
7	Project Characteristics	Project Location	1.96	0.510	Low correlation
8	Project Characteristics	Engineering Complexity	2.48	0.403	Acceptable
9	Project Characteristics	Contractor Experience	2.16	0.463	Acceptable
10	Project Characteristics	Procurement Strategy	1.82	0.549	Low correlation
11	Physical Reservoir Properties	Reservoir Pressure	1.96	0.510	Low correlation
12	Physical Reservoir Properties	Reservoir Temperature	2.08	0.481	Acceptable
13	Physical Reservoir Properties	Permeability	1.88	0.532	Low correlation
14	Physical Reservoir Properties	Porosity	1.72	0.581	Low correlation
15	Physical Reservoir Properties	Flow Rate	2.11	0.474	Acceptable
16	Physical Reservoir Properties	Skin Factor	1.63	0.613	Low correlation
17	Physical Reservoir Properties	Water Cut	1.85	0.541	Low correlation
18	Physical Reservoir Properties	Reservoir Depth	2.09	0.479	Acceptable
19	External Economic Factors	Inflation Rate	2.71	0.369	Acceptable
20	External Economic Factors	Exchange Rate	2.89	0.346	Acceptable
21	External Economic Factors	Oil Price Volatility	2.56	0.391	Acceptable
22	External Economic Factors	Financial Uncertainty	3.05	0.328	Moderate correlation
23	External Economic Factors	Supply Chain Risk	2.42	0.413	Acceptable
24	External Economic Factors	Labor Cost Index	2.17	0.461	Acceptable
25	Drift and Cumulative Risk Factors	Drift1	2.47	0.405	Acceptable
26	Drift and Cumulative Risk Factors	Drift2	2.58	0.388	Acceptable
27	Drift and Cumulative Risk Factors	Drift3	2.36	0.424	Acceptable
28	Drift and Cumulative Risk Factors	Schedule Drift	2.11	0.474	Acceptable
29	Drift and Cumulative Risk Factors	Cost Drift	2.24	0.446	Acceptable
30	Drift and Cumulative Risk Factors	Risk Accumulation Index	2.31	0.433	Acceptable
31	Organizational Maturity Indicators	S_PA_LV1 (Binary)	1.34	0.746	Very low correlation
32	Organizational Maturity Indicators	S_PA_LV2 (Binary)	1.29	0.775	Very low correlation
33	Organizational Maturity Indicators	Organizational Maturity Index	1.91	0.524	Low correlation
34	Organizational Maturity Indicators	Process Standardization Level	2.03	0.493	Acceptable