Simulation-Driven Machine Learning Framework for Semi-Dynamic Project-Delay Prediction

Siyona Behera, Sahana Ramachandran, Varsha Giridharan

***Abstract*—Schedule overruns remain a persistent issue across large-scale projects, highlighting the challenge of modeling uncertainty and interdependent risks in complex project networks. Existing analytical methods such as CPM/PERT scheduling, fuzzy earned-value analysis, and conventional Monte Carlo simulation remain largely static and treat risks as independent, limiting their ability to capture cascading effects. This study introduces a simulation-driven machine-learning framework that combines CPM/PERT-based Monte Carlo simulation with supervised learning to enable semi-dynamic delay-risk prediction. The approach captures risk interactions and uncertainty distributions empirically, rather than through predefined fuzzy or grey equations. Three classifiers—Logistic Regression, Naïve Bayes, and Random Forest—are trained on 1050 synthetically generated project networks, and evaluated on a held-out test set to predict delayed completion. The predictive accuracy is then compared against a baseline analytical model from prior work, demonstrating the benefit of simulation-based feature engineering for early-stage schedule-delay forecasting.**

***Index Terms*—Project scheduling, Monte Carlo simulation, Bayesian networks, machine learning, delay prediction, risk interaction, uncertainty modeling, project performance forecasting, earned value management, CPM/PERT.**

I. Introduction

PROJECTS are inherently dynamic and uncertain systems that often lead to risks emerging as execution progresses. These risks may propagate through interdependent activities, amplifying disruptions and causing significant schedule and cost overruns. The inability to forecast such delays results in substantial losses, particularly in construction, software, and engineering projects. Existing analytical methods either struggle to capture the cascading nature of risks or perform poorly when project data are dynamic. Even recent machine-learning approaches often rely on static or subjective inputs, which limits their predictive accuracy. This study addresses these limitations by training machine-learning models on simulation-generated uncertainty features rather than static inputs, enabling early-stage delay prediction.

Conventional scheduling and risk-assessment techniques such as the Critical Path Method (CPM), Program Evaluation and Review Technique (PERT), Earned Value Management (EVM), and Monte Carlo (MC) simulation form the foundation of project-performance analysis. These frameworks quantify task dependencies, estimate expected durations under uncertainty, and evaluate progress through cost and schedule indices. However, they are largely static: each forecast is computed from fixed input parameters and must be recalculated manually as new information becomes available. As a result, such methods fail to capture evolving project conditions or interdependent risks that cascade through the network. Later extensions such as fuzzy EVM, grey-system estimation, and hybrid analytical models introduced probabilistic representations of uncertainty, yet they remain snapshot-based and non-adaptive.

Consequently, recent research has attempted to overcome the limitations of static risk-analysis by explicitly modelling risk interdependencies and uncertainty. For example, the PN–RIN model integrates a Project Network (PN) and a Risk Interaction Network (RIN) to capture how risks propagate through causal links and affect activity durations, and applies simulation-based control metrics to sensitive activities. In another stream, fuzzy earned-value analysis extends classical EVM by representing schedule and cost indices as fuzzy sets, thereby accounting for ambiguous estimates of performance metrics. Further, SEM-BN models combine structural equation modelling with Bayesian networks to forecast project outcomes under uncertainty; however, despite their sophistication these approaches rely on predefined causal structures or static data snapshots. Collectively, these studies demonstrate that while risk interaction and uncertainty modelling have become more mature, forecasting methods remain constrained by implicit or static assumptions. Hence, there remains a need for a framework that learns risk propagation patterns directly from data and adapts early-stage project signals into predictive models.

To address these remaining limitations, this paper proposes a simulation-driven machine-learning framework for early-stage project-delay prediction. The framework integrates CPM/PERT-based Monte Carlo simulation with supervised learning to generate and analyze synthetic project data that capture structural, uncertainty, and early-performance characteristics. Instead of relying on predefined fuzzy or causal equations, the proposed approach learns risk-propagation patterns empirically from simulation outcomes. Three classifiers—Logistic Regression, Naïve Bayes, and Random Forest—are trained on 1050 synthetically generated project networks and evaluated on a held-out test set. The inclusion of early Earned Value Management (EVM) metrics such as the Schedule Performance Index (SPI) and Cost Performance Index (CPI) at 20% progress enables semi-dynamic forecasting that reflects evolving project conditions. Comparative experiments against an analytical baseline from prior work assess the predictive improvement introduced by simulation-based feature engineering.

The proposed framework offers a scalable and reproducible approach for combining simulation outputs with data-driven learning, providing project managers with a quantitative tool to anticipate schedule delays before they escalate. Although demonstrated here on synthetic datasets, the methodology can be extended to real-world project environments and evolved toward fully dynamic, real-time forecasting systems. The remainder of this paper is organized as follows: Section II reviews the literature; Section III describes the proposed methodology; Section IV presents experimental results and discussion; and Section V concludes the paper with key findings and directions for future research.

II. Literature Review

Project-delay prediction has been approached from multiple analytical and probabilistic perspectives. Past studies have attempted to capture uncertainty, interdependence, and causal relationships through different modelling methods. This section reviews three representative works that serve foundational to the present study and identifies the gaps that motivate the proposed framework.

Song and Vanhoucke proposed a two-layer Project Network–Risk Interaction Network (PN–RIN) model to analyse how project-level risks propagate through activity dependencies. The project network represents task precedence, while the risk network encodes interdependencies among risk factors such as resource unavailability, design errors, or procurement delays. The integration of both layers enables simulation of cascading failures across the network. The authors introduced a Risk Criticality Index (RCI) and a Risk Significance Index (RSI) to quantify each risk's contribution to overall delay. Their experiments demonstrated that the PN–RIN model captures indirect and compounding effects that traditional CPM or PERT overlook. Although PN–RIN successfully formalises risk interdependence, it remains an analytical framework that does not generalise across projects. Its parameters and causal links must be defined manually, and the model does not learn from empirical or simulated data. Hence, while it advances the representation of risk interactions, it remains static and descriptive rather than predictive.

Fan et al. advanced project-performance evaluation under uncertainty by integrating fuzzy set and grey system theories within an Earned Value Management (EVM) framework. Their Grey-Fuzzy EVM (GF-EVM) model expresses project progress using interval grey triangular fuzzy numbers (IGTFNs) that simultaneously capture linguistic vagueness and confidence intervals. By quantifying expert-based assessments of cost and schedule performance, the approach provides a more realistic representation of uncertainty than classical EVM. The authors validated the model on a three-phase industrial project and compared its performance with Z-number EVM and traditional fuzzy EVM, reporting improved cost- and schedule-forecast accuracy. However, the model's uncertainty representation remains subjective and snapshot-based. Both fuzzy and grey parameters are pre-defined by experts and do not evolve with incoming progress data. Consequently, while GF-EVM enhances interpretability, it lacks the adaptivity required for dynamic or data-driven forecasting. The present study builds on this limitation by generating uncertainty features through Monte Carlo simulation rather than expert definition, allowing supervised-learning models to empirically learn uncertainty patterns from data.

More recent studies by Unsal-Altuncan and Vanhoucke combine Structural Equation Modelling (SEM) with Bayesian Networks (BN) to capture both causal relationships and probabilistic dependencies among project variables. The SEM component quantifies latent factors such as management efficiency or communication quality, while the BN component infers conditional probabilities of delay or cost overrun. The hybrid SEM–BN model was trained on synthetically generated projects using RanGen2 (a random project generator) and compared with traditional Monte Carlo forecasts and a few well known ML models. Results indicated improved predictive accuracy for time and cost performance relative to purely analytical baselines. Despite this improvement, the SEM–BN framework still relies on predefined causal structures and simulated static datasets. The model parameters are calibrated once and do not adapt as project progress data become available. Its so-called "dynamic" simulation merely reproduces progress snapshots through a second Monte Carlo run rather than learning from evolving data. Thus, while it provides a stronger probabilistic foundation than earlier analytical models, it remains pseudo-dynamic and limited in generalisability.

Collectively, the PN–RIN, fuzzy EVM, and SEM–BN approaches advance the understanding of project-risk interaction and uncertainty quantification. Yet all three share two fundamental limitations: (1) their reliance on predefined structures or expert judgment, and (2) their static treatment of project data. None of the models adaptively learn from empirical or simulated progress information.

The present study addresses this gap by employing Monte Carlo simulation as a data generator rather than merely a baseline forecaster and by training machine-learning classifiers on early-stage performance features, including SPI and CPI. This simulation-driven learning paradigm enables semi-dynamic delay prediction and establishes a bridge between analytical project-risk models and modern data-driven forecasting methods.

III. Methodology

A. Novelty

This study introduces a simulation-driven data-generation pipeline that differs fundamentally from static analytical methods and expert-defined fuzzy or grey methods. The novelty lies in extracting empirical uncertainty behaviour from project networks, which shifts the core idea from representing uncertainty to learning it. This study uses artificial project data from RanGen2 RG30 (Vanhoucke et al., 2008) to ensure compatibility with established benchmarks. However, instead of using the project data as is, PERT triplets are created around the durations to introduce uncertainty. The data is then subjected to CPM and 200-run Monte Carlo simulations from which distributional features—such as mean and variance of activity durations, critical-path sensitivity, probability of late completion, and indicators of network-level uncertainty—are extracted. Hence, this study produces simulation-derived uncertainty features that implicitly encode risk interactions without requiring expert rules, linguistic uncertainty scales, or predefined causal links. Additionally, early EVM signals including Schedule Performance Index (SPI) and Cost Performance Index (CPI) are computed at 20% progress, enabling forecasting that partially reflects evolving project reality. This introduces semi-dynamic prediction, unlike snapshot-based methods in prior work.

Machine-learning models then learn patterns embedded in the simulation data. This allows risk interactions to emerge naturally through Monte Carlo sampling with no need for explicit RIN or SEM-BN structures. This also helps avoid reliance on linguistic fuzzy sets or expert-defined grey intervals. To contextualize the benefit of simulation-derived features, the predictive performance of the machine-learning models is compared against a simple analytical baseline, where delay classification is derived directly from deterministic CPM estimates or the Monte Carlo late-finish probability.

Collectively, these contributions establish a data-driven, generalizable, and causal-structure-free framework for project-delay prediction that bridges classical project-risk analysis with modern machine-learning methodologies.