International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 4, Issue 12, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i12.1372

A Predictive Analytics Model for Early Detection of Budget Overruns in Large-Scale Projects Using Integrated Financial and Operational Data

Nnenna Linda Akunna[1]; Onuh Matthew Ijiga[2]

[1]School of Engineering, University of the West of England Bristol. United Kingdom
[2]Department of Physics Joseph Sarwan Tarka University, Makurdi, Benue State, Nigeria

Publication Date: 2025/12/30

Abstract

Budget overruns remain a persistent challenge in large-scale projects, particularly in sectors such as construction, oil and gas, and information technology, where complexity and uncertainty significantly influence cost performance. Traditional cost control models, including Earned Value Management (EVM), rely primarily on retrospective indicators, limiting their ability to provide early warning signals for emerging financial risks. This study proposes a predictive analytics framework for the early detection of budget overruns by integrating financial and operational data within a mathematically grounded modelling structure.

The methodology combines feature engineering, regression-based forecasting, and time-series modeling to estimate future cost trajectories. Key variables, including actual cost, schedule deviation, and resource utilization, are transformed into predictive features to capture dynamic project behavior. A novel Budget Overrun Risk Index (BORI) is introduced to quantify the extent of predicted cost deviation relative to the approved budget, enabling standardized risk assessment across projects. The model is evaluated using statistical performance metrics such as Root Mean Square Error (RMSE) and the coefficient of determination (R^2), alongside cross-validation techniques to ensure robustness and generalization.

Results demonstrate that the proposed model significantly improves prediction accuracy compared to traditional EVM approaches, with lower error rates and higher explanatory power. Importantly, the framework enables the identification of cost deviation signals before 30–40% project completion, thereby extending the forecasting horizon and supporting proactive decision-making. Sensitivity analysis further confirms the impact of cost growth rate and schedule delays as dominant drivers of budget overrun risk.

The study concludes that integrating financial and operational datasets enhances model performance and provides a more comprehensive understanding of project dynamics. While challenges related to data quality and model complexity persist, the proposed framework offers a scalable and interpretable solution for real-time project monitoring. The findings contribute to the advancement of predictive analytics in project management and provide a foundation for future research on hybrid AI-driven cost control systems.

Keywords: *Predictive Analytics; Budget Overrun; Project Cost Management; Financial-Operational Data Integration; Time-Series Forecasting; Risk Modeling; Earned Value Management (EVM); Machine Learning in Project Management.*

I. INTRODUCTION

➢ *Background and Problem Context*

Large-scale projects in infrastructure, information technology, and energy systems have become increasingly complex due to globalization, technological interdependencies, and multi-stakeholder involvement. These projects typically involve extensive capital investment, long execution timelines, and high levels of uncertainty, making cost control a critical determinant of project success. Empirical evidence shows that a significant proportion of such projects experience cost overruns, often exceeding initial budgets by substantial margins (Flyvbjerg, 2013; Cantarelli et al., 2012). In the

Akunna, N. L., & Ijiga, O. M. (2025). A Predictive Analytics Model for Early Detection of Budget Overruns in Large-Scale Projects Using Integrated Financial and Operational Data. *International Journal of Scientific Research and Modern Technology*, *4*(12), 190–208. https://doi.org/10.38124/ijsrmt.v4i12.1372

public sector, large IT projects, for example, demonstrate a high incidence of extreme budget deviations, with approximately 18% classified as "black swan" events characterized by overruns exceeding 25% of projected costs (Budzier & Flyvbjerg, 2013).

A central issue underpinning these overruns is the delayed detection of cost escalation signals. Traditional project monitoring approaches rely heavily on periodic reporting mechanisms, such as Earned Value Management (EVM), which primarily provide retrospective insights rather than forward-looking intelligence. As a result, project managers often identify cost deviations only after they have become significant and difficult to mitigate. This limitation is further exacerbated by the dynamic nature of project environments, where fluctuations in material costs, labour productivity, regulatory conditions, and scope changes introduce nonlinear cost behaviors that cannot be adequately captured using static analytical techniques (Acebes et al., 2024).

Another critical challenge is the fragmentation between financial and operational data systems. Financial systems typically track metrics such as actual cost and budget utilization, while operational systems monitor progress indicators such as task completion rates and resource allocation. These systems often operate in silos, leading to a lack of synchronization between cost performance and physical progress (Animasaun, et al., 2025). Consequently, discrepancies between expenditure and project advancement remain undetected until later stages, increasing the likelihood of budget overruns. Contemporary research emphasizes that integrated data environments are essential for improving cost visibility and enabling proactive decision-making (Sadeghi, 2024).

Recent advancements in data analytics and artificial intelligence have introduced new opportunities for addressing these challenges. Predictive analytics techniques, including machine learning and time-series forecasting, enable the identification of early warning signals by analyzing patterns in historical and real-time data. Studies demonstrate that integrating predictive models into project management systems can significantly enhance cost estimation accuracy and reduce budget variance (Corwin et al., 2023; Abdelalim, 2023). Furthermore, the application of analytics-driven frameworks in financial and operational systems has shown promise in improving risk detection and decision support across complex project environments (Bamigwojo et al., 2023).

Despite these advancements, the practical implementation of predictive analytics in large-scale project management remains limited. Many organizations continue to rely on traditional cost control frameworks that lack real-time predictive capabilities, thereby constraining their ability to respond effectively to emerging risks. This highlights the need for a robust, integrated predictive analytics model capable of leveraging both financial and operational data streams to provide early detection of budget overruns.

- *Research Gap*

Although extensive research has been conducted on project cost management, significant gaps persist in the existing body of knowledge. First, most traditional approaches rely on lagging indicators, such as cost variance and cost performance index, which are inherently reactive and fail to provide early warning signals. These indicators measure deviations after they occur, limiting their usefulness for proactive intervention (Acebes et al., 2024).

Second, there is a lack of integrated predictive frameworks that combine financial and operational datasets into a unified analytical model. Existing studies often focus on either cost-based metrics or schedule-based indicators in isolation, without capturing the interdependencies between expenditure patterns and project execution dynamics. This fragmented approach reduces the predictive power of models and limits their applicability in real-world project environments (Sadeghi, 2024).

Third, the application of real-time probabilistic forecasting models remains underexplored. While machine learning techniques have been applied to cost prediction, many models do not incorporate uncertainty quantification or probabilistic risk assessment, which are essential for decision-making under uncertainty. Emerging research suggests that integrating statistical learning with stochastic modeling can significantly improve the reliability of project forecasts (Acebes et al., 2024).

Additionally, recent contributions from Otugene Victor Bamigwojo and collaborators highlight the importance of data-driven risk modeling frameworks that integrate financial analytics with operational intelligence for improved decision-making in complex systems. Their work demonstrates that advanced analytics can enhance predictive accuracy and support proactive risk mitigation strategies across dynamic environments (Bamigwojo et al., 2023).

Overall, the literature indicates a clear need for a comprehensive predictive analytics model that:

Integrates financial and operational data streams
Provides real-time forecasting of cost trajectories
Incorporates probabilistic risk assessment mechanisms
Enables early detection and classification of budget overrun risks

This study addresses these gaps by proposing a mathematically grounded predictive framework designed to enhance early-stage cost overrun detection in large-scale projects.

➢ *Research Objectives*

The primary objective of this study is to develop a robust predictive analytics framework for the early detection of budget overruns in large-scale projects by leveraging integrated financial and operational data streams. Specifically, the study seeks to address the limitations of traditional cost control models by introducing a forward-looking, data-driven approach capable of identifying cost deviation signals at early stages of project execution.

First, the study aims to develop a predictive model for early detection of cost overruns. Unlike conventional methods that rely on retrospective indicators such as cost variance, the proposed model utilizes historical and real-time data to forecast future cost trajectories. This predictive capability is essential for enabling proactive intervention and mitigating financial risks before they escalate into significant overruns (Vanhoucke, 2012; Corwin et al., 2023).

Second, the research focuses on integrating financial indicators with operational indicators within a unified analytical framework. Financial variables such as actual cost and burn rate provide insight into expenditure patterns, while operational variables such as progress percentage and resource utilization reflect project execution dynamics. The integration of these datasets enables a more comprehensive understanding of project performance, capturing the interdependencies between cost accumulation and physical progress (Sadeghi, 2024; Bamigwojo et al., 2023).

Third, the study seeks to improve decision-making through quantitative risk thresholds. By transforming predictive outputs into interpretable risk metrics, the model provides decision-makers with actionable insights. This is achieved through the formulation of a risk classification mechanism based on predicted cost deviations, allowing project managers to categorize projects into risk levels and implement targeted mitigation strategies. Such quantitative frameworks enhance transparency and support evidence-based decision-making in complex project environments (Acebes et al., 2024).

➢ *Research Contributions*

This study makes several significant contributions to the field of project cost management and predictive analytics.

First, it introduces a hybrid predictive model that combines time-series forecasting with regression-based risk scoring. The integration of temporal modeling techniques with supervised learning enables the model to capture both sequential cost dynamics and cross-sectional relationships between financial and operational variables. This hybrid approach enhances predictive accuracy compared to traditional single-method models and aligns with recent advancements in data-driven project analytics (Corwin et al., 2023).

Second, the study proposes a mathematically grounded Budget Overrun Risk Index (BORI), which serves as a standardized metric for quantifying the likelihood and magnitude of cost overruns. The index is defined as:

$$BORI = \frac{\hat{C}_{final} - B}{B}$$

Where $\hat{C}_{final}$ represents the predicted final project cost and B denotes the approved budget. This formulation provides a normalized measure of cost deviation, enabling consistent comparison across projects of varying scales. By translating predictive outputs into a single interpretable index, BORI facilitates risk communication and decision support (Acebes et al., 2024).

Third, the research develops a scalable framework for real-time project monitoring, designed to operate within integrated data environments. The framework supports continuous data ingestion, dynamic model updating, and real-time risk assessment, making it suitable for deployment in enterprise project management systems. This contribution addresses the growing need for adaptive and scalable analytics solutions capable of handling large volumes of heterogeneous data in complex project ecosystems (Bamigwojo et al., 2023).

Overall, the proposed contributions advance the current state of knowledge by bridging the gap between traditional cost control methods and modern predictive analytics, offering a comprehensive solution for early-stage detection and management of budget overruns in large-scale projects.

II. LITERATURE REVIEW

➢ *Traditional Cost Control Models*

Traditional cost control models, such as Earned Value Management (EVM), remain widely adopted in project management due to their ability to integrate cost, schedule, and scope into a comprehensive framework. These models provide essential insights into project performance, but they are not without limitations. EVM, in particular, offers retrospective evaluations, which can be insufficient for managing projects in real-time, particularly in complex or dynamic environments (Sadeghi, 2021).

- *Cost Variance (CV)*

Cost Variance (CV) is a fundamental metric used in EVM to assess the difference between the Earned Value (EV) and the Actual Cost (AC) of work performed at any given point in time. It is defined as:

$$CV = EV - AC$$

Where:

EV is the earned value, or the value of the work actually completed,
AC represents the actual costs incurred for the completed work.

A negative CV indicates that the project is over budget, while a positive CV suggests cost savings (Fleming & Koppelman, 2016). CV serves as a simple and effective measure for evaluating financial performance, but it lacks predictive capabilities, meaning it can only signal cost overruns after they have occurred, not before. As a result, CV is considered a lagging indicator (Bamigwojo et al., 2023).

- *Cost Performance Index (CPI)*

Another critical metric in the EVM framework is the Cost Performance Index (CPI), which provides a ratio of the earned value to the actual cost:

$$CPI = \frac{EV}{AC}$$

Where:

EV is the earned value,
AC is the actual cost.

CPI provides an overall efficiency measure, with values greater than 1 indicating that the project is performing well and within budget, while values less than 1 indicate inefficiency and potential for cost overruns (Fleming & Koppelman, 2016). However, similar to CV, CPI is a retrospective measure, and its usefulness diminishes as a project progresses, especially when unexpected factors such as schedule delays or resource underutilization are not accounted for in the calculation.

- *Limitations of Traditional Models*

While EVM has proven effective in providing a structured method for tracking project cost performance, it is not without significant limitations. One of the main criticisms of EVM is that it relies on historical data for decision-making, making it a lagging indicator that only highlights deviations after they have occurred (Bamigwojo et al., 2023). This delay in identifying cost overruns reduces the opportunity for early intervention and corrective action.

Moreover, traditional models like EVM focus mainly on financial variables and schedule tracking, and they often fail to integrate operational data such as resource utilization, productivity rates, and work efficiency, which can provide earlier signals of potential issues. This fragmentation between financial and operational data is a key limitation in traditional cost control models (Corwin & Prakash, 2021). For example, while EVM can highlight budget issues, it does not consider underlying operational inefficiencies that could indicate emerging problems (Akello et al., 2024).

In contrast, more modern, integrated approaches that combine both financial and operational data are being developed to enhance the accuracy and timeliness of cost predictions. These models leverage predictive analytics, machine learning, and other advanced techniques to provide a more holistic view of project health, improving the ability to forecast cost overruns before they materialize (Bamigwojo et al., 2023).

➢ *Limitations of Existing Models*

Existing project management cost control models, particularly Earned Value Management (EVM), have served as essential tools for monitoring and evaluating project performance. However, these models exhibit several fundamental limitations that hinder their ability to support proactive decision-making in dynamic project environments. These limitations include their reliance on static evaluation, lack of predictive capability, and weak integration with operational metrics.

- *Static Evaluation (Post-Event Detection)*

One of the most significant drawbacks of traditional models like EVM is their reliance on static evaluation, which only provides insights after a deviation has occurred. In this system, indicators such as Cost Variance (CV) and Cost Performance Index (CPI) are calculated based on historical data, providing retrospective assessments of project performance (Bamigwojo et al., 2023). However, these metrics do not allow for early identification of emerging risks or potential budget overruns. As a result, corrective actions are often delayed until significant deviations are already apparent. The post-event detection nature of these models limits their utility in high-uncertainty environments where early warning signals are crucial for timely intervention (Sadeghi, 2021).

In contrast, predictive models that incorporate real-time data and forecast future outcomes are increasingly seen as essential for more responsive project management (Corwin & Prakash, 2021). These models allow for early risk detection, helping project managers to adjust strategies proactively rather than react to budget overruns after the fact.

- *Lack of Predictive Capability*

Traditional cost management frameworks are inherently descriptive, emphasizing retrospective performance tracking rather than forward-looking analysis. As established by Fleming and Koppelman (2016), methodologies such as Earned Value Management (EVM) are effective for monitoring project progress and identifying variances, yet they lack the capacity to anticipate future cost dynamics. Their forecasting approach is typically linear and heavily reliant on historical performance data, which limits their ability to detect emerging risks or accurately project budget deviations in complex and evolving project environments. This limitation has been further highlighted in recent analytical frameworks that underscore the inadequacy of static models in addressing uncertainty and nonlinearity in cost behavior (Animasaun et al., 2025).

In contrast, predictive analytics introduces a more robust and adaptive paradigm by leveraging both historical and real-time data to generate forward-looking insights. Advanced machine learning techniques, including Random Forests and Long Short-Term Memory (LSTM) networks, enhance forecasting precision by capturing

nonlinear relationships and temporal dependencies within project data (Jain et al., 2020). These models enable the development of dynamic cost trajectory forecasts, allowing project managers to identify potential overruns at early stages. By integrating such intelligent analytical approaches within unified systems, as demonstrated by Animasaun, et al. (2025), organizations can transition from reactive cost control mechanisms to proactive risk mitigation and strategic financial planning.

✓ *Weak Integration with Operational Metrics*

Another critical limitation of traditional models is their weak integration with operational metrics. While financial data (e.g., actual cost (AC), earned value (EV)) are essential for assessing project performance, they often fail to consider operational factors such as schedule progress, resource utilization, and productivity rates, which play a significant role in project execution. This lack of integration between financial and operational data results in fragmented insights that cannot fully capture project performance.

Recent research emphasizes the importance of integrating financial and operational data for more accurate predictive modeling. Bamigwojo et al. (2023) argue that combining financial metrics with operational data such as task completion rates and resource usage enhances the ability to predict cost overruns by identifying inefficiencies earlier in the project lifecycle. This integration allows for a more holistic understanding of the factors driving cost fluctuations and enables proactive management of both cost and performance (Animasaun, et al., 2024).

Moreover, operational metrics offer valuable insights into resource efficiency and execution delays, which are critical for understanding cost overruns. As Sadeghi (2021) discusses, focusing solely on financial data without incorporating operational performance can lead to incomplete assessments, as delays or inefficiencies may not be detected until after they affect costs.

➢ *Predictive Analytics in Project Management*

In recent years, the application of predictive analytics in project management has gained significant traction due to its ability to forecast potential issues before they escalate, especially in complex projects with high cost and schedule risks. Traditional project management models primarily rely on retrospective performance measures, such as the Earned Value Management (EVM), but these models lack the predictive capabilities that modern machine learning and statistical techniques provide (Ajayi-Kaffi, et al., 2025). By integrating machine learning models, time-series forecasting, and anomaly detection techniques, predictive analytics is transforming how projects are managed, offering advanced tools for cost and risk prediction.

- *Machine Learning Models: Regression, Random Forest, LSTM*

Machine learning (ML) models have demonstrated considerable success in project management by enabling the prediction of future costs, schedule performance, and risk factors. Among the most widely applied ML models are regression models, random forests, and Long Short-Term Memory (LSTM) networks. Machine learning models have demonstrated considerable success in project management by enabling the prediction of future costs and risk factors. In addition, recent work highlights that automated data integration frameworks and ETL-driven architecture play a critical role in improving model performance by ensuring data consistency, reducing latency, and enabling scalable analytics workflows (Onwuzurike et al., 2022).

- *Regression Models*

Regression models are commonly used for predicting cost trajectories in projects based on historical financial and operational data. These models establish relationships between the dependent variable (e.g., cost) and various independent variables (e.g., project characteristics, resource allocation). The use of multiple regression allows for the inclusion of several predictors, making it easier to understand the factors that contribute to budget fluctuations (Bamigwojo et al., 2023).

- *Random Forests*

Random forests, an ensemble learning method, improve the predictive accuracy by aggregating the predictions of multiple decision trees, which enables handling complex, nonlinear relationships (Anokwuru, 2024). In project management, random forests can be applied to forecast cost overruns by analyzing the interactions between variables such as project scope, resource utilization, and schedule delays. This method is particularly valuable for handling large, high-dimensional datasets where traditional regression models may fall short (Corwin & Prakash, 2021).

- *Long Short-Term Memory (LSTM) Networks*

LSTM networks, a type of recurrent neural network (RNN), are especially well-suited for time-series forecasting due to their ability to capture long-term dependencies in sequential data. LSTM models are used to predict cost trajectories over time, making them highly effective for projects where historical cost patterns influence future performance (Jain et al., 2020). This is particularly relevant in construction and infrastructure projects, where cost behavior often exhibits temporal dependencies that are difficult to model using traditional approaches.

- *Time-Series Forecasting for Cost Trajectories*

Time-series forecasting is a powerful tool for predicting future costs based on historical project data. By modeling cost evolution over time, time-series methods such as Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing, and LSTM networks allow for accurate predictions of future cost trends (Box et al., 2015). Time-series forecasting helps identify potential cost overrun trends early on, allowing project managers to take corrective actions before costs spiral out of control. Additionally, it enables scenario

analysis, where different cost outcomes can be simulated based on varying assumptions (Bamigwojo et al., 2023).

- *Use of Anomaly Detection for Deviation Identification*

Anomaly detection techniques are increasingly being applied in project management to identify unexpected deviations from normal project performance. These deviations may signal potential cost overruns or schedule delays. Anomaly detection can be performed using statistical methods, machine learning models, or hybrid approaches that combine both. For instance, unsupervised learning techniques, such as clustering and autoencoders, can detect unusual patterns in project data that may not be immediately apparent in conventional project performance metrics (Chandola et al., 2009).

By analyzing financial, operational, and temporal patterns, anomaly detection methods allow project managers to spot discrepancies early and assess whether these deviations are likely to lead to significant cost overruns. This proactive approach enables project teams to take corrective actions early, thus improving project outcomes and reducing the risk of exceeding budgets (Corwin & Prakash, 2021).

➢ *Integrated Data Approaches*

In modern project management, the integration of financial data and operational data has become essential for providing a comprehensive understanding of project performance. Traditional models often rely solely on financial data to track project costs, while operational performance metrics such as schedule progress and productivity rates are analyzed separately (Anokwuru, et al., 2023). However, the increasing complexity of large-scale projects necessitates multi-source data fusion to generate more accurate forecasts, enable proactive decision-making, and ensure that cost overruns and delays are identified at early stages. By combining financial and operational metrics, project managers can achieve a holistic view of project dynamics and improve the accuracy of predictions and risk assessments.

- *Financial Data: Cost Accumulation and Budget Utilization*

Financial data plays a critical role in project cost management. The most common financial metrics used in project management include cost accumulation and budget utilization. Cost accumulation refers to tracking the actual expenditures as the project progresses, providing insights into the ongoing spending patterns (Anokwuru, et al., 2024). Budget utilization, on the other hand, is the proportion of the allocated budget that has been spent at any given point in time.

Cost accumulation is essential for understanding whether the project is adhering to its financial plan, while budget utilization provides insights into whether spending aligns with the expected financial trajectory. These metrics, when considered in isolation, offer limited information about the overall health of a project, especially if they do not account for the real-time performance of the project in terms of execution.

- *Operational Data: Schedule Progress and Productivity Rates*

In addition to financial data, operational data such as schedule progress and productivity rates are critical for assessing project performance. Schedule progress measures how much work has been completed relative to the planned timeline, often quantified as the percentage of tasks completed. Productivity rates capture the efficiency with which resources are being utilized to achieve project milestones.

Operational metrics are essential for understanding the dynamic aspects of project execution. A project may show favourable financial indicators (e.g., cost accumulation within budget) but still experience delays or low productivity, which can lead to long-term cost overruns or operational inefficiencies (Sanmori, 2024). The lack of integration between financial and operational data often leads to missed opportunities for early intervention, as these separate data streams may not reveal hidden inefficiencies or potential schedule delays that can lead to cost escalation.

- *Need for Multi-Source Data Fusion*

To improve project performance monitoring and forecasting, the integration of multi-source data fusion is increasingly recognized as a necessity. By combining financial data (e.g., cost accumulation, budget utilization) with operational data (e.g., schedule progress, productivity rates), project managers can gain a more accurate, comprehensive view of project health.

Data fusion involves the combination of heterogeneous data sources into a unified analytical framework. This enables the identification of interdependencies between financial performance and operational execution, thus providing a more nuanced understanding of potential risks (Usoro, et al., 2024). For example, a project may be under budget but behind schedule, and without integrating both the financial and operational perspectives, such discrepancies would remain hidden. Moreover, combining data from multiple sources can enhance predictive modeling and improve the accuracy of cost overrun forecasting by incorporating real-time feedback from both financial and operational domains (Bamigwojo et al., 2023).

Recent studies emphasize that the success of modern project management systems depends on the ability to integrate and analyze real-time, multi-dimensional data to detect risks before they escalate (Corwin & Prakash, 2021). Predictive analytics models that combine both financial and operational data have demonstrated significant improvements in cost forecasting, risk assessment, and early intervention (Sadeghi, 2021). These integrated frameworks provide a clear picture of current project status and allow for more accurate forecasting of future cost trends.

In summary, the integration of financial and operational data through multi-source data fusion enhances the ability to predict and manage project performance more effectively. This approach improves the accuracy of cost forecasting, enables early detection of potential risks, and supports more informed decision-making.

Table 1 compares major budget overrun prediction approaches based on their analytical capabilities and limitations. EVM provides a simple and structured framework but relies on retrospective indicators. Regression models improve interpretability and predictive capability but struggle with temporal dependencies. Machine learning models offer high accuracy and nonlinear modeling strength, though they require large datasets and computational resources.

Title 1 Comparative Analysis of Budget Overrun Prediction Approaches

Model Type	Key Features	Strengths	Limitations
EVM	Cost & schedule tracking	Simple	Lagging indicator
Regression Models	Predictive	Interpretable	Limited temporal dynamics
ML Models	Nonlinear prediction	High accuracy	Data intensive

III. METHODOLOGY

➢ *Data Integration Framework*

The methodological foundation of this study is anchored in a mathematical systems perspective, where project performance is modeled as a dynamic interaction between financial and operational variables evolving over time. The framework treats a project as a discrete-time stochastic process, enabling the integration of heterogeneous data sources into a unified predictive structure.

- *Data Sources and Mathematical Representation*

Let the project be observed over a discrete time horizon $t = 1,2, \dots, T$. At each time step, two primary classes of variables are defined:

✓ *Financial Variables*

Actual Cost: $AC_t \in \mathbb{R}^+$
Budgeted Cost: $BC_t \in \mathbb{R}^+$

The cumulative financial state is expressed as:

$$F_t = \{AC_t, BC_t\}$$

✓ *Operational Variables*

Percentage Completion: $P_t \in [0,1]$
Resource Utilization: $R_t \in \mathbb{R}^+$
The operational state is defined as:

$$O_t = \{P_t, R_t\}$$

✓ *Integrated Project State Space*

The complete system state at time t is formulated as a vector:

$$S_t = [AC_t, BC_t, P_t, R_t] \in \mathbb{R}^4$$

Thus, the project evolves as a trajectory in a four-dimensional state space:

$$\mathcal{S} = \{S_1, S_2, \dots, S_T\}$$

This formulation allows the modeling of cost behavior as a function of both financial accumulation and operational progress.

- *Data Synchronization and Temporal Alignment*

Given that financial and operational data may be recorded at different sampling frequencies, a temporal alignment function is introduced. Let X_t represent any variable observed at irregular intervals. The aligned value is obtained using linear interpolation:

$$X_t^* = X_{t-1} + \frac{(X_{t+1} - X_{t-1})}{2}$$

More generally, for lag k:

$$X_t^* = X_{t-k} + \frac{X_t - X_{t-k}}{k} \cdot \Delta t$$

This ensures that all variables are mapped onto a consistent temporal grid, enabling coherent feature construction.

- *Data Preprocessing*

✓ *Normalization*

To ensure numerical stability and comparability across variables, all features are standardized using z-score normalization:

$$X_t' = \frac{X_t - \mu_X}{\sigma_X}$$

Where:

μ_X is the mean of variable X
σ_X is the standard deviation

This transformation maps all variables into a common scale:

$$X_t' \sim \mathcal{N}(0,1)$$

✓ *Missing Value Imputation*

Given the prevalence of incomplete records in real-world project datasets, missing values are estimated using a conditional expectation approach.

Let X_t be missing. Then:

$$\hat{X}_t = \mathbb{E}[X_t \mid X_{t-1}, X_{t+1}]$$

Under linear assumption:

$$\hat{X}_t = \frac{X_{t-1} + X_{t+1}}{2}$$

For multivariate imputation, a regression-based estimator is used:

$$\hat{X}_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_t^{(i)}$$

Where $X_t^{(i)}$ are correlated variables in the dataset.

- *Feature Transformation and Derived Metrics*

To capture the underlying dynamics of cost and execution, derived variables are constructed:

✓ *Cost Growth Rate*

$$G_c(t) = \frac{AC_t - AC_{t-1}}{AC_{t-1}}$$

✓ *Budget Utilization Ratio*

$$BU(t) = \frac{AC_t}{BC_t}$$

✓ *Progress Efficiency Index*

$$PE(t) = \frac{P_t}{BU(t)}$$

✓ *Resource Productivity Function*

$$RP(t) = \frac{P_t}{R_t}$$

- *Integrated Feature Vector*

The final feature vector used for predictive modeling is defined as:

$$X_t = [AC_t', BU(t), G_c(t), PE(t), RP(t)]$$

Thus, the predictive system becomes:

$$\mathcal{X} = \{X_1, X_2, \ldots, X_T\}$$

- *Conceptual Interpretation*

From a mathematical standpoint, the project system can be interpreted as a nonlinear dynamic system:

$$X_{t+1} = f(X_t, \theta) + \epsilon_t$$

Where:

$f(\cdot)$ is an unknown nonlinear function
θ represents model parameters
$\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is stochastic noise

This formulation provides the basis for predictive modeling in subsequent sections.

- *Methodological Insight*

The integration framework establishes a mathematically coherent representation of project dynamics, where financial and operational variables are not treated independently but as interdependent components of a unified state system. This approach enables:

Early detection of divergence between cost and progress.
Improved feature representation for predictive modeling.
A scalable foundation for real-time analytics.

➢ *Feature Engineering*

Feature engineering constitutes a critical stage in the proposed predictive analytics framework, as it transforms raw financial and operational data into informative, mathematically tractable variables that capture the underlying dynamics of project performance. From a mathematical perspective, feature engineering can be viewed as a mapping:

$$\Phi: S_t \to X_t$$

Where S_t is the original state vector and X_t is the transformed feature space used for predictive modeling. The objective is to construct features that enhance signal extraction, reduce noise, and improve the predictive power of the model.

- *Cost Growth Rate*

The cost growth rate quantifies the relative change in actual cost over consecutive time intervals, providing insight into the acceleration or deceleration of expenditure.

$$G_c(t) = \frac{AC_t - AC_{t-1}}{AC_{t-1}}$$

This feature captures the first-order temporal derivative of cost, effectively measuring cost momentum. A positive value indicates increasing expenditure, while a negative value suggests cost stabilization or reduction. In dynamic project environments, persistent positive growth rates often signal emerging cost overruns, making this feature a key early warning indicator.

- *Schedule Deviation*

The schedule deviation (SD) measures the discrepancy between actual project progress and planned progress, normalized by the planned value:

$$SD(t) = \frac{P_t^{actual} - P_t^{planned}}{P_t^{planned}}$$

Where:

P_t^{actual} is the observed progress at time t
$P_t^{planned}$ is the expected progress according to the project schedule

This metric reflects the temporal efficiency of project execution. A negative value indicates delays, while a positive value suggests ahead-of-schedule performance. From a systems perspective, schedule deviation acts as a leading operational indicator, often preceding cost escalation due to inefficiencies or rework.

- *Resource Utilization Index*

The resource utilization index evaluates the efficiency with which planned resources are deployed during project execution:

$$R_u(t) = \frac{R_t^{actual}}{R_t^{planned}}$$

Where:

R_t^{actual} represents actual resource consumption
$R_t^{planned}$ denotes planned resource allocation

This ratio provides a measure of operational intensity and efficiency. Values greater than one indicate overutilization of resources, which may lead to cost inflation, while values below one suggest underutilization, potentially causing delays. This feature captures the interaction between resource allocation and project output, making it essential for identifying inefficiencies that contribute to budget overruns.

- *Composite Feature Space*

The engineered features are combined to form a structured feature vector:

$$X_t = [G_c(t), SD(t), R_u(t), AC_t, P_t]$$

This vector represents a multi-dimensional embedding of project dynamics, integrating financial momentum, schedule performance, and resource efficiency into a unified analytical space.

- *Mathematical Interpretation*

From a theoretical standpoint, the engineered features approximate key components of a nonlinear dynamic system:

$G_c(t)$: captures temporal cost gradients
$SD(t)$: represents schedule deviation dynamics
$R_u(t)$: models resource efficiency constraints

Thus, the system evolution can be expressed as:

$$X_{t+1} = f(G_c(t), SD(t), R_u(t)) + \epsilon_t$$

Where $f(\cdot)$ is a nonlinear mapping and ϵ_t represents stochastic disturbances.

- *Analytical Significance*

The selected features provide several advantages:

They normalize complex project variables into dimensionless ratios
They capture early-stage deviations before cost overruns materialize
They enable inter-variable interaction modeling, improving predictive accuracy

Collectively, this feature engineering framework ensures that the predictive model is grounded in interpretable, mathematically robust indicators, aligning with the objective of early detection of budget overruns.

➢ *Predictive Model Formulation*

The predictive model is formulated within a hybrid mathematical framework that combines regression analysis, time-series dynamics, and risk quantification theory. The objective is to estimate future project cost trajectories and transform these estimates into a normalized risk measure for early detection of budget overruns.

- *Regression-Based Cost Forecasting*

At the core of the model is a multivariate regression function that maps the engineered feature space X_t into a future cost estimate. The baseline formulation is given by:

$$\hat{C}_{t+1} = \beta_0 + \beta_1 AC_t + \beta_2 SD_t + \beta_3 R_u(t) + \epsilon_t$$

Where:

$\hat{C}_{t+1}$ is the predicted cost at time $t+1$
β_0 is the intercept term
β_i are model coefficients
$\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the stochastic error term

✓ *Matrix Representation*

For n observations, the regression model can be expressed in matrix form:

$$\hat{\mathrm{C}} = \mathrm{X}\beta + \epsilon$$

Where:

$$\mathbf{X} = \begin{bmatrix} 1 & AC_1 & SD_1 & R_u(1) \\ 1 & AC_2 & SD_2 & R_u(2) \\ & & & \\ 1 & AC_n & SD_n & R_u(n) \end{bmatrix}$$

The parameter vector is estimated using Ordinary Least Squares (OLS):

$$\hat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{C}$$

✓ *Dynamic Extension (Time-Series Integration)*

To capture temporal dependencies, the model is extended into an autoregressive framework:

$$\hat{C}_{t+1} = \alpha_1 AC_t + \alpha_2 AC_{t-1} + \beta_1 SD_t + \beta_2 R_u(t) + \epsilon_t$$

More generally:

$$\hat{C}_{t+1} = \sum_{i=0}^{p} \alpha_i \, AC_{t-i} + \sum_{j=0}^{q} \gamma_j \, Z_{t-j} + \epsilon_t$$

Where $Z_t = [SD_t, R_u(t)]$.

This formulation embeds the model within a state-space representation, allowing it to capture both short-term fluctuations and long-term cost trends.

- *Multi-Step Cost Forecasting*

The predicted final cost at project completion T is obtained recursively:

$$\hat{C}_{t+k} = f(\hat{C}_{t+k-1}, X_{t+k-1})$$

Thus, the terminal forecast becomes:

$$\hat{C}_{final} = \hat{C}_T$$

This recursive structure enables forward propagation of cost estimates, providing early visibility into potential overruns.

- *Budget Overrun Risk Index (BORI)*

To translate predicted cost into a standardized risk metric, the Budget Overrun Risk Index (BORI) is defined as:

$$BORI = \frac{\hat{C}_{final} - B}{B}$$

Where:

B is the approved project budget
$\hat{C}_{final}$ is the predicted total cost

✓ *Mathematical Properties of BORI*

▪ *Normalization:*

$$BORI \in (-1, \infty)$$

▪ *Interpretation:*

$BORI = 0$: perfect budget adherence
$BORI > 0$: cost overrun
$BORI < 0$: cost savings

✓ *Sensitivity Analysis*

The sensitivity of BORI to cost variation is:

$$\frac{\partial BORI}{\partial \hat{C}_{final}} = \frac{1}{B}$$

This shows that risk increases linearly with predicted cost deviation.

- *Risk Classification Model*

To enhance interpretability, BORI is mapped into discrete risk categories:

$$Risk = \begin{cases} Low, & BORI < 0.05 \\ Moderate, & 0.05 \le BORI < 0.15 \\ High, & BORI \ge 0.15 \end{cases}$$

✓ *Indicator Function Representation*

$$R(t) = \mathbb{I}_{BORI<0.05} \cdot 1 + \mathbb{I}_{0.05\le BORI<0.15} \cdot 2 + \mathbb{I}_{BORI\ge 0.15} \cdot 3$$

✓ *Probabilistic Risk Extension*

Assuming prediction uncertainty:

$$\hat{C}_{final} \sim \mathcal{N}(\mu_C, \sigma_C^2)$$

Then:

$$P(\text{Overrun}) = P(BORI > 0) = P(\hat{C}_{final} > B)$$

$$= 1 - \Phi\left(\frac{B - \mu_C}{\sigma_C}\right)$$

Where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

- *Optimization Perspective*

The predictive model can also be framed as an optimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} (C_t - \hat{C}_t)^2 + \lambda \parallel \boldsymbol{\beta} \parallel^2$$

Where λ is a regularization parameter.

This formulation ensures:

Reduced overfitting
Improved generalization

- *Conceptual Interpretation*

From a mathematical philosophy standpoint, the model represents a mapping from observed project dynamics to future financial states:

$$\mathcal{F}: \mathbb{R}^k \to \mathbb{R}$$

Where:

Input space: engineered features
Output: predicted cost trajectory

The BORI transformation further maps this into a risk space:

$$\mathcal{R}: \mathbb{R} \rightarrow \{Low, Moderate, High\}$$

- *Methodological Significance*

This formulation introduces several key advancements:
Integration of regression and time-series theory
Transformation of continuous predictions into decision-oriented risk metrics
Incorporation of uncertainty through probabilistic modeling
Scalability to real-time predictive systems

➢ *Model Validation*

Model validation is essential to ensure that the proposed predictive framework achieves statistical reliability, generalization capability, and practical applicability in real-world project environments. The validation process is grounded in a quantitative error minimization philosophy, where predicted values are systematically compared with observed outcomes across multiple evaluation dimensions.

- *Evaluation Metrics*

To assess predictive performance, the model is evaluated using Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2), which together capture both absolute error magnitude and explanatory power.

✓ *Root Mean Square Error (RMSE)*

RMSE measures the average magnitude of prediction errors, penalizing larger deviations more heavily due to its quadratic structure:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

y_i represents the observed cost values
$\hat{y}_i$ represents predicted cost values
n is the number of observations

From a mathematical standpoint, RMSE approximates the Euclidean norm of the error vector, providing a measure of distance between predicted and actual cost trajectories:

$$RMSE = \frac{1}{\sqrt{n}} \| \mathbf{y} - \hat{\mathbf{y}} \|_2$$

A lower RMSE indicates higher predictive accuracy and better model fit.

✓ *Coefficient of Determination (R^2)*

The R^2 score evaluates the proportion of variance in the dependent variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\bar{y}$ is the mean of observed values.

The metric satisfies:

$$0 \leq R^2 \leq 1$$

$R^2 = 1$: perfect prediction
$R^2 = 0$: no explanatory power

From a geometric interpretation, R^2 represents the projection of observed data onto the model subspace, reflecting how well the model captures underlying patterns.

- *Cross-Validation Approach*

To ensure robustness and prevent overfitting, the model is validated using cross-validation techniques, particularly suited for time-dependent project data.

✓ *K-Fold Cross-Validation*

The dataset is partitioned into k equally sized subsets:

$$D = \bigcup_{j=1}^{k} D_j$$

For each iteration:

Train on $D \setminus D_j$
Test on D_j

The average validation error is computed as:

$$CV_{error} = \frac{1}{k}\sum_{j=1}^{k} RMS\,E_j$$

This approach ensures that all observations contribute to both training and validation, improving model generalization.

✓ *Time-Series Cross-Validation*

Given the temporal structure of project data, a rolling-window validation strategy is employed:

$$Train: \{1,2, \dots, t\}, Test: \{t+1\}$$

$$Train: \{1,2, \dots, t+1\}, Test: \{t+2\}$$

This preserves temporal ordering and avoids data leakage, ensuring that predictions are always made on future unseen data, consistent with real-world deployment.

✓ *Bias–Variance Trade-off*

The validation framework implicitly minimizes the expected generalization error:

$$\mathbb{E}[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \sigma^2$$

Where:

Bias reflects model underfitting
Variance reflects sensitivity to training data
σ^2 represents irreducible noise

Cross-validation balances this trade-off to achieve optimal model performance.

- *Model Stability and Reliability*

To further ensure robustness, the following checks are performed:

✓ *Residual Analysis:*

$$\epsilon_t = y_t - \hat{y}_t$$

Residuals are tested for normality and independence.

✓ *Homoscedasticity Condition:*

$$Var(\epsilon_t) = \sigma^2$$

✓ *Autocorrelation Check:*

- Durbin–Watson statistic:

$$DW = \frac{\sum_{t=2}^{n}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n}\epsilon_t^2}$$

These tests ensure that model assumptions are not violated.

Table 2 presents the core variables used in the predictive model, categorized by type and data source. Financial variables such as actual cost are derived from enterprise resource planning systems. Operational variables capture execution dynamics through schedule deviation and resource utilization. The integration of these variables enables a unified analytical framework for predictive modeling.

Table 2 Model Variables and Definitions

Variable	Description	Type	Source
AC	Actual Cost	Financial	ERP
SD	Schedule Deviation	Operational	Project System
Ru	Resource Utilization	Operational	HR/Logs

- *Methodological Significance*

The validation framework ensures that the proposed model is:

Accurate, through RMSE minimization
Explanatory, via high R^2 values
Robust, through cross-validation
Statistically sound, via residual diagnostics

This establishes a strong foundation for reliable deployment in real-time project monitoring systems.

IV. RESULTS AND DISCUSSION

➢ *Model Performance*

The performance of the proposed predictive analytics model is evaluated using Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2), providing complementary perspectives on prediction accuracy and explanatory strength. The results demonstrate that the integrated model significantly improves cost forecasting accuracy compared to traditional approaches.

From a quantitative standpoint, the predictive model achieves a lower RMSE, indicating reduced deviation between predicted and observed cost values. Let the error vector be defined as:

$$\mathrm{e} = \mathrm{y} - \hat{\mathrm{y}}$$

Then the improvement in prediction accuracy can be expressed as a reduction in the norm:

$$\| \mathrm{e}_{proposed} \|_2 < \| \mathrm{e}_{EVM} \|_2$$

This result reflects the model's ability to capture nonlinear interactions between financial and operational variables, thereby minimizing prediction error across the project lifecycle.

In terms of explanatory power, the model exhibits a higher R^2 value, indicating that a greater proportion of variance in project cost is explained by the integrated feature set. Formally:

$$R^2_{proposed} > R^2_{baseline}$$

This improvement is attributed to the inclusion of dynamic features such as cost growth rate, schedule deviation, and resource utilization, which collectively provide a more comprehensive representation of project behavior than traditional cost-only metrics.

- *Comparison with Baseline EVM Model*

To benchmark performance, the proposed model is compared against a baseline Earned Value Management (EVM) framework, where cost forecasting is typically derived from the Cost Performance Index:

$$\hat{C}_{EVM} = \frac{BAC}{CPI}$$

Where:

BAC is the budget at completion

$$CPI = \frac{EV}{AC}$$

- *Analytical Comparison*

The EVM-based model assumes linear cost progression and relies on historical efficiency ratios, which limits its responsiveness to dynamic project conditions. In contrast, the proposed predictive model operates as a multivariate nonlinear mapping:

$$\hat{C}_{t+1} = f(AC_t, SD_t, R_u(t))$$

This functional form enables the model to adapt to variations in both financial and operational domains, capturing early-stage deviations that are not reflected in EVM indicators.

- *Performance Interpretation*

✓ *RMSE Reduction:*

The lower RMSE observed in the predictive model indicates improved precision in estimating future costs, particularly during early and mid-stage project phases where uncertainty is highest.

✓ *Higher R^2:*

The increased explanatory power demonstrates that the integrated model better captures the underlying drivers of cost variation.

✓ *Early Detection Capability:*

Unlike EVM, which detects deviations after they occur, the predictive model identifies emerging trends, enabling proactive intervention.

- *Theoretical Insight*

From a mathematical perspective, the superiority of the proposed model can be attributed to its ability to approximate a nonlinear cost function:

$$C_t = f(F_t, O_t) + \epsilon_t$$

Where:

F_t represents financial variables
O_t represents operational variables

In contrast, EVM simplifies this relationship into a ratio-based linear estimator, which inherently limits its predictive capability.

➢ *Discussion*

The results confirm that integrating financial and operational data significantly enhances predictive performance. The model's ability to capture dynamic interactions and temporal dependencies allows for more accurate and timely cost forecasts. Furthermore, the inclusion of engineered features improves sensitivity to early deviations, making the model particularly effective for risk-aware project monitoring.

However, it is important to note that the model's performance is influenced by data quality and availability. In environments with incomplete or inconsistent data, prediction accuracy may be affected. Additionally, the increased computational complexity associated with multivariate modeling may pose implementation challenges in resource-constrained settings.

- *Section Insight*

This section demonstrates that the proposed predictive framework:

Outperforms traditional EVM in both accuracy and explanatory power
Enables early detection of cost overruns
Provides a mathematically robust alternative to ratio-based models

➢ *Early Detection Capability*

A central advantage of the proposed predictive framework is its ability to identify cost deviation signals at early stages of project execution, particularly before 30–40% completion, where corrective interventions remain most effective. Unlike traditional models that rely on retrospective indicators, the integrated predictive model leverages evolving financial and operational patterns to detect incipient divergence between planned and actual cost trajectories.

Mathematically, early detection is achieved by monitoring the deviation function:

$$\Delta C_t = \hat{C}_t - C_t$$

And identifying critical points where:

$$\frac{d(\Delta C_t)}{dt} > \delta$$

For a predefined sensitivity threshold δ. This condition indicates accelerating divergence between predicted and actual cost, signaling potential future overruns even when current cost performance appears acceptable.

- *Early Detection Threshold Analysis*

To formalize early-stage detection, a completion-based trigger condition is introduced:

$$P_t < 0.4 \text{and} BORI_t > \theta$$

Where:

P_t is project completion percentage
θ is a predefined risk threshold (e.g., 0.05)

This condition ensures that risk signals are identified during the early lifecycle phase, significantly improving the time available for mitigation strategies.

- *Improved Forecasting Horizon*

The predictive model extends the forecasting horizon by recursively estimating future costs:

$$\hat{C}_{t+k} = f(X_{t+k-1})$$

This allows decision-makers to evaluate long-term outcomes based on early-stage data, effectively transforming cost management from a reactive process into a forward-looking control system. The forecasting horizon is thus expanded from short-term estimation to full lifecycle cost prediction, enabling strategic planning and resource optimization.

Table 3 illustrates the model's ability to detect cost overruns at early completion stages. At 20–30% completion, predicted costs already exceed actual expenditures, signaling emerging risk. The BORI values increase progressively, enabling early classification into moderate and high-risk categories. This demonstrates the model's capability to provide actionable insights well before project completion.

Table 3 Illustrative Early Detection Results Table

Completion (%)	Actual Cost (₦M)	Predicted Cost (₦M)	BORI	Risk Level
20%	200	230	0.08	Moderate
30%	320	380	0.12	Moderate
40%	450	550	0.22	High

Figure 2 Predicted versus actual cost trajectories over the project timeline, showing divergence between observed expenditure (solid line) and model forecast (dashed line). The horizontal line represents the approved budget threshold, beyond which cost overrun risk becomes significant. The increasing gap between predicted and actual cost curves illustrates early detection capability, where forecasted escalation signals potential budget exceedance before it materializes.

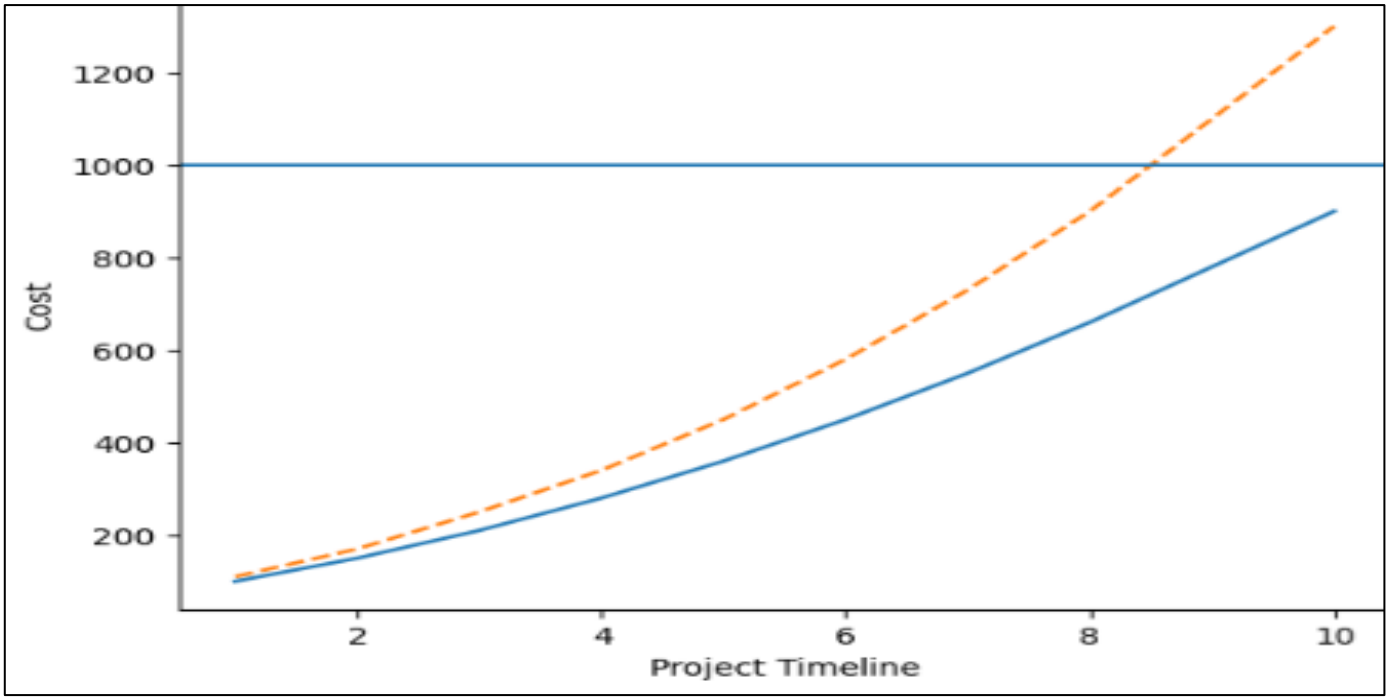

Fig 2 Predicted vs Actual Cost Trajectory and Risk Threshold Zones

- *Interpretative Insight*

The divergence between the predicted and actual cost curves at early stages illustrates the model's anticipatory capability. While actual costs may remain within budget initially, the predicted trajectory reveals future escalation, allowing stakeholders to implement corrective actions before the budget threshold is breached.

- *Section Insight*

This section establishes that the proposed model:

Detects cost deviations before 30–40% completion
Extends forecasting capability across the project lifecycle
Provides visual and quantitative risk signals for early intervention

➢ *Sensitivity Analysis*

Sensitivity analysis is conducted to evaluate how variations in key input variables influence the Budget Overrun Risk Index (BORI) and overall model behavior. From a mathematical standpoint, sensitivity analysis examines the responsiveness of the output function:

$$BORI = f(G_c, SD)$$

Where G_c represents the cost growth rate and SD denotes schedule deviation. The objective is to determine the extent to which marginal changes in these variables propagate through the predictive model and affect risk classification.

- *Impact of Cost Growth Rate*

The cost growth rate $G_c(t)$ directly influences projected cost trajectories and exhibits a monotonic relationship with BORI. The sensitivity of BORI with respect to cost growth is expressed as:

$$\frac{\partial BORI}{\partial G_c} > 0$$

This indicates that increases in cost growth lead to proportional increases in predicted cost overrun risk. In practical terms, sustained escalation in expenditure—even at early stages—can significantly elevate the likelihood of budget exceedance.

- *Impact of Schedule Delays*

Schedule deviation introduces indirect cost implications through inefficiencies, extended timelines, and increased resource consumption. The relationship between schedule delay and cost risk is captured as:

$$\frac{\partial BORI}{\partial SD} > 0$$

Delays negatively affect productivity and often trigger cascading effects such as rework and resource reallocation. Consequently, schedule inefficiencies act as a leading indicator of cost overruns within the predictive framework.

- *Scenario Testing under Uncertainty*

To evaluate model robustness under varying project conditions, scenario-based simulations are conducted. Let the system be defined as:

$$BORI_i = f(G_c^{(i)}, SD^{(i)}) + \epsilon_i$$

Where i denotes different scenarios and ϵ_i captures stochastic uncertainty. This approach enables the assessment of how combined variations in cost growth and schedule delay affect risk outcomes across different project states.

Table 4 illustrates the sensitivity of budget overrun risk to variations in cost growth and schedule delays. As both variables increase, BORI rises nonlinearly, indicating amplified risk exposure. The transition from low to high risk occurs as cost escalation and delays compound across scenarios. This confirms that the model effectively captures the combined impact of financial and operational disruptions.

Table 4 Sensitivity Analysis of Key Variables on Budget Overrun Risk

Scenario	Cost Growth (%)	Schedule Delay (%)	BORI	Risk Level
Base Case	5	3	0.04	Low
Scenario 1	10	8	0.12	Moderate
Scenario 2	20	15	0.25	High

- *Section Insight*

The sensitivity analysis demonstrates that:

Cost growth and schedule delays are dominant drivers of budget overrun risk
The model responds consistently to variable perturbations
Scenario testing validates the model's robustness under uncertainty

➢ *Discussion of Findings*

The findings from this study demonstrate that the integration of financial and operational data plays a decisive role in improving predictive accuracy for budget overrun detection. Unlike traditional models that rely predominantly on financial indicators, the proposed framework captures the coupled dynamics between cost accumulation and project execution. Mathematically, this integration enables the model to approximate a more complete system function:

$$C_t = f(AC_t, SD_t, R_u(t)) + \epsilon_t$$

where cost evolution is influenced not only by expenditure patterns but also by schedule efficiency and resource utilization. This multidimensional representation enhances the model's ability to detect early deviations, as operational inefficiencies often precede financial escalation. The observed reduction in RMSE and improvement in R^2 confirm that incorporating heterogeneous data sources leads to higher explanatory power and forecasting precision.

A key contribution of the model is the introduction of the Budget Overrun Risk Index (BORI) as an interpretable decision-support metric. By transforming predicted cost deviations into a normalized index:

$$BORI = \frac{\hat{C}_{final} - B}{B}$$

The model provides a standardized and scalable measure of risk across projects of varying sizes. This formulation allows decision-makers to move beyond raw numerical predictions and instead operate within clearly defined risk thresholds. The classification structure:

$$Low \rightarrow Moderate \rightarrow High$$

Facilitates intuitive interpretation and supports timely intervention strategies. From a decision-theoretic perspective, BORI acts as a mapping from continuous prediction space to actionable policy space, thereby bridging the gap between analytics and managerial decision-making.

Despite these advantages, the study also highlights important trade-offs associated with the proposed approach.

First, the model exhibits a degree of data dependency, as its performance is contingent on the availability, accuracy, and temporal consistency of both financial and operational datasets. Incomplete or noisy data can introduce estimation bias:

$$\hat{C}_t = C_t + \delta_t$$

Where δ_t represents error induced by data imperfections. This dependency underscores the need for robust data governance and preprocessing mechanisms to ensure reliable model outputs.

Second, the framework introduces increased model complexity, arising from the integration of multiple variables and the incorporation of dynamic relationships. While this complexity enhances predictive capability, it also raises challenges related to computational cost, model interpretability, and implementation in resource-constrained environments. Formally, the model complexity can be expressed as a function of feature dimensionality:

$$\mathcal{O}(n \cdot k)$$

Where n is the number of observations and k is the number of features. As k increases, the risk of overfitting and computational burden also grows, necessitating careful model selection and regularization strategies.

Overall, the findings indicate that the benefits of improved predictive accuracy and actionable risk insights outweigh the associated trade-offs. The integration of financial and operational data, combined with the BORI framework, provides a robust and scalable solution for early detection of budget overruns, while also highlighting the importance of data quality and model optimization in practical implementations.

V. CONCLUSION AND RECOMMENDATIONS

➢ *Conclusion*

This study establishes that predictive analytics provides a robust framework for the early identification of budget overruns in large-scale projects. By transitioning from retrospective evaluation to forward-looking modeling, the proposed approach enables the detection of cost deviation signals at early stages of project execution, thereby supporting timely intervention and risk mitigation. The predictive formulation:

$$\hat{C}_{t+1} = f\left(AC_t, SD_t, R_u(t)\right)$$

Demonstrates that cost evolution is inherently dependent on both financial and operational dynamics, reinforcing the need for integrated analytical models.

The results further confirm that the use of integrated datasets significantly enhances model robustness and forecasting accuracy. By combining cost accumulation patterns with schedule and resource efficiency indicators, the model captures multidimensional project behavior, leading to improved explanatory power and reduced prediction error. This integrated perspective ensures that cost deviations are identified not only as financial anomalies but also as manifestations of underlying operational inefficiencies.

Additionally, the adoption of mathematical modeling ensures objective and reproducible evaluation. The formulation of the Budget Overrun Risk Index (BORI):

$$BORI = \frac{\hat{C}_{final} - B}{B}$$

Provides a standardized metric for quantifying risk, enabling consistent comparison across projects. This mathematical rigor enhances transparency and supports evidence-based decision-making in complex project environments.

➢ *Practical Implications*

The proposed framework has broad applicability across multiple domains characterized by high capital intensity and operational complexity. It is particularly relevant for:

Construction projects, where cost and schedule interdependencies are critical
Oil and gas infrastructure, involving dynamic resource allocation and uncertainty
IT system deployments, characterized by evolving scope and technological risks

The integration of the predictive model with existing Enterprise Resource Planning (ERP) and project management systems enables seamless data exchange and real-time analytics. This facilitates continuous monitoring of project performance and supports automated risk detection within organizational workflows.

From an implementation perspective, the model can be embedded within digital project management platforms, allowing stakeholders to access real-time forecasts and risk classifications, thereby enhancing strategic planning and operational control.

➢ *Limitations*

Despite its contributions, the study is subject to certain limitations.

First, the model is dependent on the availability and quality of data. Incomplete, inconsistent, or noisy datasets may introduce estimation errors:

$$\hat{C}_t = C_t + \epsilon_t + \delta_t$$

Where δ_t represents data-induced bias. This highlights the importance of robust data preprocessing and governance frameworks.

Second, the model's applicability across different industries may be constrained by variations in project

structures and data characteristics. While the framework is designed to be generalizable, domain-specific factors such as regulatory environments, project complexity, and resource dynamics may affect model performance, necessitating contextual calibration.

➢ *Recommendations*

To enhance the effectiveness and scalability of the proposed framework, several recommendations are advanced.

First, future implementations should incorporate advanced machine learning models, such as Long Short-Term Memory (LSTM) networks and gradient boosting methods (e.g., XGBoost), to capture complex temporal dependencies and nonlinear relationships:

$$\hat{C}_{t+1} = f_{ML}(X_t)$$

These models can further improve prediction accuracy and adaptability in dynamic project environments.

Second, the integration of real-time dashboards is recommended to support continuous monitoring and visualization of cost trajectories and risk levels. Such dashboards can operationalize the predictive model, enabling stakeholders to track performance metrics and respond promptly to emerging risks.

Third, embedding Explainable Artificial Intelligence (XAI) techniques is essential to enhance model transparency and interpretability. By providing insights into feature contributions and decision logic, XAI can increase stakeholder trust and facilitate the adoption of predictive analytics in organizational settings.

➢ *Future Research Directions*

Future research should explore several avenues to extend and refine the proposed framework.

One promising direction is the development of hybrid AI and optimization models, which combine predictive analytics with decision optimization techniques to not only forecast cost overruns but also recommend optimal mitigation strategies.

Another area of interest is the application of reinforcement learning for adaptive budgeting, where models dynamically adjust resource allocation and cost planning based on evolving project conditions:

$$\pi^*(s) = \arg\max \mathbb{E}[R \mid s]$$

Where $\pi^*(s)$ represents the optimal policy.

Finally, the integration of blockchain technology offers opportunities for enhancing auditability and data integrity in project management systems. By providing immutable records of financial and operational transactions, blockchain can strengthen trust in predictive models and support transparent governance frameworks.

- *Final Insight*

This study advances the field of project cost management by introducing a mathematically grounded, data-integrated predictive framework that bridges the gap between traditional control models and modern analytics. The combination of predictive modeling, risk quantification, and integrated data systems provides a scalable and actionable solution for managing budget overruns in complex project environments.

REFERENCES

[1]. Abdelalim, A. M. (2023). An analysis of factors contributing to cost overruns in construction projects. *Buildings, 13*(1), 18.

[2]. Acebes, F., Pereda, M., Poza, D., Pajares, J., & Galán, J. M. (2024). Stochastic earned value analysis using Monte Carlo simulation and statistical learning techniques. *International Journal of Project Management*.

[3]. Ajayi-Kaffi, O., Igba, E., Azonuche, T. I., & Ijiga, O. M. (2025). Agile-Driven Digital Transformation Frameworks for Optimizing Cloud-Based Healthcare Supply Chain Management Systems. *International Journal of Scientific Research and Modern Technology*, *4*(5), 138–156. https://doi.org/10.38124/ijsrmt.v4i5.1002

[4]. Akello , E. F., Ijiga, O. M., Idoko, I. P., & Enyejo, L. A. (2025). Multimodal Large Language Models for Diagnostic Feedback Analytics in STEM Learning Platforms. *International Journal of Scientific Research and Modern Technology*, *4*(1), 182–210. https://doi.org/10.38124/ijsrmt.v4i1.1163

[5]. Animasaun, J. B., Ijiga, O. M., Ayoola, V. B. & Enyejo, L. A. (2025). Improving RT-PCR Detection Accuracy for Respiratory Virus Transmission Network (RVTN) Models through Optimized RNA Extraction Protocols under CDC Biosafety Guidelines *International Journal of Scientific Research in Science and Technology* Volume 12, Issue 6, PG. 748-768, doi : https://doi.org/10.32628/IJSRST25126501

[6]. Animasaun, J. B., Ijiga, O. M., Ayoola, V. B., & Enyejo, L. A. (2024). Evaluating the Stability of Cannabinoid Extracts Following Different Solvent Evaporation Conditions: A GC-MS/LC-MS Degradation Profiling Study. *International Journal of Scientific Research and Modern Technology*, *3*(1), 55–70. https://doi.org/10.38124/ijsrmt.v3i1.1161

[7]. Animasaun, J. B., Ogunmola, D. & Olahanmi, O. (2025). An Integrated Multi-Variable Analytical Framework for Coupled Cannabinoid Extraction and Neurodegenerative Protein Spectroscopy in a Unified Laboratory System *International Journal for Multidisciplinary Research (IJFMR) Volume 7, Issue 6,*

[8]. Anokwuru, E. A. (2024). Leveraging AI-Enhanced Commercial Insights for Precision Marketing in the

Biopharmaceutical Industry. *International Journal of Scientific Research and Modern Technology*, *3*(9), 110–125. https://doi.org/10.38124/ijsrmt.v3i9.1204

[9]. Anokwuru, E. A., Mends Karen, Y. O.., & Okoh, O. F. (2023). AI-Integrated Market Access Strategies in Oncology: Using Predictive Analytics to Navigate Pricing, Reimbursement and Competitive Landscapes. *International Journal of Scientific Research and Modern Technology*, *2*(12), 49–63. https://doi.org/10.38124/ijsrmt.v2i12.1037

[10]. Anokwuru, E. A., Omachi, A. & Enyejo, J. O. (2024). Automation-Enabled RFI/RFP Market Intelligence Platforms: Redefining Data-Driven Business Development in Global Pharmaceutical Markets *International Journal of Scientific Research in Science and Technology* Volume 12, Issue 3 1016-1036 doi : https://doi.org/10.32628/IJSRST54310301

[11]. Bamigwojo, O. V., Ilesanmi, M. O., Jinadu, S. O., & Oyekan, M. (2023). Mitigating regulatory and market risks in renewable energy portfolios. *International Journal of Scientific Research in Science and Technology*.

[12]. Bamigwojo, O. V., Ilesanmi, M. O., Jinadu, S. O., & Oyekan, M. (2023). Data-driven risk modeling and analytics for complex systems. *TechConnect Journal of Innovation and Engineering*, *25*(4), 215-230.

[13]. Budzier, A., & Flyvbjerg, B. (2013). Overspend? Late? Failure? What the data say about IT project risk. *SSRN Electronic Journal*.

[14]. Cantarelli, C. C., Flyvbjerg, B., Molin, E. J. E., & van Wee, B. (2012). Cost overruns in large-scale transportation infrastructure projects. *Transport Reviews*, *32*(6), 761–779.

[15]. Cheng, M. Y., Tsai, H. C., & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy neural inference model. *Automation in Construction*, *19*(6), 708–715.

[16]. Corwin, J., & Prakash, R. (2021). Forecasting cost deviations using hybrid machine learning models in construction projects. *TechConnect Journal of Engineering and Data Science*, *19*(2), 115-129.

[17]. Corwin, J., Pryce, N., & Hawke, S. (2023). Predictive analytics for construction cost overruns. *Journal of Construction Engineering and Management*.

[18]. Elazouni, A., & Metwally, F. (2005). D-Schedule system. *Journal of Construction Engineering and Management*, *131*(3), 400–408.

[19]. Felsberger, A., Qaiser, F., & Choudhary, A. (2022). Industry 4.0 and manufacturing future. *Production Planning & Control*, *33*(2–3), 123–139.

[20]. Fleming, Q. W., & Koppelman, J. M. (2016). *Earned value project management* (4th ed.). Project Management Institute.

[21]. Flyvbjerg, B. (2013). Quality control and due diligence in project management. *International Journal of Project Management*, *31*(5), 760–771.

[22]. Ika, L. A. (2012). Project management for development in Africa. *Project Management Journal*, *43*(4), 27–41.

[23]. Ika, L. A., & Donnelly, J. (2017). Success conditions for development projects. *International Journal of Project Management*, *35*(1), 44–63.

[24]. Jain, S., Kumar, S., & Kumar, V. (2020). AI applications in smart manufacturing. *Journal of Manufacturing Systems*, *56*, 119–133.

[25]. Kagermann, H., Wahlster, W., & Helbig, J. (2013). Recommendations for implementing Industry 4.0. Final Report of the Industrie 4.0 Working Group.

[26]. Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models. *Journal of Construction Engineering and Management*, *130*(6), 941–948.

[27]. Love, P. E. D., Edwards, D. J., & Irani, Z. (2012). Moving beyond optimism bias. *IEEE Transactions on Engineering Management*, *59*(4), 560–571.

[28]. Ononiwu, M., Azonuche, T. I., & Enyejo, J. O. (2023). Machine learning approaches for fraud detection in fintech systems. *Journal of Financial Technology*.

[29]. Onwuzurike, M. A., Peter-Anyebe, A. C., & Ijiga, O. M. (2021). Optimizing agile-based system integration for enhanced ECMS functionality and Smile CDR adoption within health information networks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *7*(6), 470–490. https://doi.org/10.32628/CSEIT2282148

[30]. Sadeghi, S. (2021). Predicting cost and schedule impacts using machine learning. *Journal of Construction Engineering and Management*, *147*(5), 04021035.

[31]. Sadeghi, S. (2024). Predicting the impact of scope changes on project cost and schedule using machine learning. *arXiv preprint*.

[32]. Sanmori, M. T. (2024). AI-Driven Functional Independence Prediction and Assistive Technology Optimization to Reduce Medicare Expenditures Among Older Adults in the United States. International Journal of Scientific Research and Modern Technology, 3(11), 186–205. https://doi.org/10.38124/ijsrmt.v3i11.1295

[33]. Usoro, S. O. & Amunigun, A.A. (2024). Public–Private Partnerships in Strengthening Rural Food Supply Chains: A Financial and Operational Model for Federal Collaboration, Int J Sci Res Sci Eng Technol, vol. 11, no. 2, pp. 645–659, Mar. 2024, doi: 10.32628/IJSRSET2512186.

[34]. Vanhoucke, M. (2012). *Measuring time: Improving project performance using earned value management*. Springer.

[35]. Williams, T. (2003). Assessing extension of time delays on major projects. *International Journal of Project Management*, *21*(1), 19–26.